Improving Alignment and Controllability in GANs and Diffusion Models

Inauguraldissertation

zur Erlangung des akademischen Grades

eines Doktors der Naturwissenschaften

der Universität Mannheim

vorgelegt von

Yumeng Li

aus Tianjin, China

Mannheim, 2025

Dekan: Prof. Dr. Claus Hertling, Universität Mannheim
Referent: Prof. Dr.-Ing Margret Keuper, Universität Mannheim
Korreferent: Prof. Dr.-Ing Seong Joon Oh, Universität Tübingen
Korreferent: Dr. Dan Zhang, Bosch Research
Tag der mündlichen Prüfung: 21. Februar 2025

Abstract

Visual generative modeling is a transformative area that aims to synthesize diverse, realistic-looking visual content, e.g., images and videos. These models are widely applied in various domains, ranging from creative art design and the visual effects industry to data augmentation for downstream computer vision tasks. Over the past decade, this field has made tremendous progress, with significant advancements evolving from Generative Adversarial Networks (GANs) to diffusion models. Despite achieving higher fidelity and improved training stability, it remains challenging to control the synthesis process and generate content precisely as desired. To this end, this thesis presents several new techniques aimed at improving alignment and controllability in GANs and diffusion models across various tasks, such as GAN inversion, layout-to-image, text-to-image, and text-to-video generation.

Firstly, we propose a novel GAN inversion encoder that can faithfully reconstruct complex scene-centric datasets and disentangle content and style information from the given image. Based on this, we built an exemplar-based style synthesis pipeline, which can assist various downstream tasks, such as improving domain generalization in semantic segmentation. Secondly, we developed new training strategies for layout-to-image diffusion models, which can significantly improve the alignment with the layout condition while maintaining data diversity. Thirdly, we strengthened pretrained text-to-image diffusion models via inference time optimization. Our work enables the model to better follow the complex text prompt and synthesize multiple desired objects with their corresponding attributes. Lastly, we introduce the concept of Generative Temporal Nursing, where we intervene in the generative process on the fly during inference to improve control over the temporal dynamics in generated videos. This approach empowers pretrained text-to-video diffusion model to synthesize longer dynamic videos.

In summary, this thesis pushes the boundaries of visual generative modeling by enhancing control over the generation process and strengthening alignment with the provided conditional information. These enhancements make the models more effective in a wide range of real-world applications.

Zusammenfassung

Die visuelle generative Modellierung ist ein transformativer Bereich, der darauf abzielt, vielfältige, realistisch aussehende visuelle Inhalte wie Bilder und Videos zu synthetisieren. Diese Modelle finden in verschiedenen Bereichen Anwendung, von Gestaltung kreativer Kunst und visueller Effekte bis hin zur Datenaugmentation für diverse Computer-Vision-Aufgaben. Dieses Feld hat in den letzten zehn Jahren enorme Fortschritte gemacht, insbesondere durch die Entwicklungen von Generative Adversarial Networks (GANs) und Diffusionsmodellen. Trotz der Erreichung höherer Wiedergabetreue und verbesserter Trainingsstabilität bleibt es jedoch eine Herausforderung, den Syntheseprozess zu kontrollieren und Inhalte präzise nach den gewünschten Vorgaben zu generieren. Zu diesem Zweck stellt diese Dissertation mehrere neue Techniken vor, die darauf abzielen, die Ausrichtung und Steuerbarkeit von GANs und Diffusionsmodellen in verschiedenen Aufgabenbereichen, wie GAN-Inversion, Layout-zu-Bild, Text-zu-Bild und Text-zu-Video Synthese, zu verbessern.

Erstens stellen wir einen neuartigen GAN-Inversions-Encoder vor, der komplexe, szenen-zentrierte Datensätze original-getreu rekonstruieren und Inhalte von Stilinformationen des Eingabebildes trennen kann. Darauf aufbauend haben wir eine beispielbasierte Stilsynthese-Pipeline entwickelt, die vielfältige nachgelagerte Aufgaben unterstützen kann, unter anderem die Verbesserung der Domänengeneralisierung in der semantischen Segmentierung. Zweitens haben wir neue Trainingsstrategien für Layout-zu-Bild-Diffusionsmodelle entwickelt, die die Ausrichtung an der Layout-Vorgabe signifikant verbessern, während die Diversität der generierten Daten erhalten bleibt. Drittens haben wir vortrainierte Text-zu-Bild-Diffusionsmodelle durch Optimierung zur Inferenzzeit optimiert. Unsere Arbeit ermöglicht es dem Modell, komplexen Textvorgaben besser zu folgen und mehrere gewünschte Objekte mit ihren jeweiligen Attributen zu synthetisieren. Zuletzt führen wir das Konzept des Generativen Temporalen Nursing ein, bei dem wir während der Inferenz dynamisch in den Generierungsprozess eingreifen, um die Kontrolle über die zeitlichen Abläufe zu verbessern. Dieser Ansatz befähigt das

vortrainierte Text-zu-Video-Diffusionsmodell, längere dynamische Videos zu synthetisieren.

Zusammenfassend erweitert diese Dissertation die Grenzen der visuellen generativen Modellierung, indem sie den Generierungsprozess besser kontrolliert und die Übereinstimmung mit den bereitgestellten Eingabebedingungen stärker unterstützt. Diese Verbesserungen machen die Modelle effektiver für eine Vielzahl von realen Anwendungen.

Acknowledgements

I would like to express my sincere gratitude to Prof. Dr. Margret Keuper, Dr. Anna Khoreva and Dr. Dan Zhang for giving me the opportunity to pursue an industrial PhD at Bosch in collaboration with the University of Mannheim. I am profoundly grateful to Prof. Margret Keuper for her insightful guidance, invaluable scientific advice, and immense emotional support. She consistently demonstrated faith in our projects, especially when the initial results were not promising.

I also want to thank Dr. Anna Khoreva and Dr. Dan Zhang for their dedication and continuous support throughout my PhD journey. I am immensely grateful for their engagement in our scientific discussions, their incredible support in paper writing, and their genuine advice on my career development. Under their guidance, I have become not only a better researcher but also a better person.

I sincerely appreciate the company of my fellow PhD students: Edgar Schönfeld, Nadine Behrmann, Vadim Sushko, Zhakshylyk Nurlanov, Massimo Bini, Nikita Kister, Haiwen Huang, Jovita Lukasik, Julia Ge, Steffen Jung, Shashank Agnihotri, Katharina Prasse, Tejaswini Medi and Mishal Fatima. Their discussions and support have been invaluable and have made this journey amazingly joyful and memorable.

Special thanks go to the people at Bosch that I had the pleasure of working with: Jiayi Wang, Kevin Laube, William Beluch, Shin-I Cheng, Jan Metzen, Kaspar Sakmann, Finnie Nicole, Julio Borges, Alexander Kugele. I have gained countless insights from them, both scientifically and professionally.

Finally, I want to thank my family and friends for being an immense source of encouragement. Their unconditional love and support have been incredibly crucial in helping me navigate the ups and downs of my PhD journey. Heartfelt thanks to my mum for standing by me all the time and always having the unwavering faith in me.

Contents

1 Introduction

Imagine if machines could dream — conjuring visions as vivid and varied as those of a painter's brush strokes on canvas. This is the realm of generative models, the creative heart of artificial intelligence, where data serves not just to inform but to inspire. Unlike discriminative models, which excel at parsing the world, recognizing, localizing and categorizing elements within diverse datasets, generative models are akin to artists, endowed with the ability to create. These models, usually neural networks, are trained to synthesize novel data, such as images and videos, following the distribution of a given training dataset. Equipped with generative capabilities, they have a wide spectrum of applications across various industries. In art design and visual effects, they facilitate the rapid prototyping of creative visual

content, allowing artists and designers to create visually stunning environments and characters that would be challenging or time-consuming to craft manually. In fields requiring vast amounts of training data, such as autonomous driving and robotic learning, generative models can enrich real datasets with diverse synthetic data, especially for scenarios that are either rare or difficult to capture in real-world data collection, such as adverse weather conditions.

In recent years, the field of generative models has witnessed rapid development, driven by significant advancements from Variational Autoencoders (VAEs) (Kingma and Welling, 2014) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to the recent emergence of diffusion models (DMs) (Ho et al., 2020, Song et al., 2020). As one of the early generative models, VAE builds upon traditional Autoencoder (AE), consisting of an encoder and a decoder. VAE compresses the data into a compact latent space, where the encoded representation is formulated as a normal distribution. This probabilistic formulation enables sampling from the posterior distribution, allowing the decoder to generate new data. Despite being conceptually simple, the outputs produced by VAEs are often observed to be blurry (Burgess et al., 2018). Further, the introduction of GAN has marked a significant leap forward in enhancing the visual quality of the generated samples. GAN consists of a generator and a discriminator that engage in an adversarial training process. The generator aims to create realistic samples, while the discriminator works to differentiate between real and generated samples. This dynamic rivalry enables GANs to progressively produce content that is both more convincing and lifelike. However, this adversarial setup can also lead to training instabilities or mode collapse, where the generator tends to produce samples with limited diversity. The recently emerged diffusion model operates by iterative denoising from pure noise to generate clean data. The model is trained to reverse the forward process, which adds noise to real data gradually. DM exhibits the intriguing properties of stable training and the capability to produce diverse outputs.

Despite significant advancements in generative models that have enhanced visual quality and diversity, controlling the generation process and ensuring alignment with potential input conditions, such as text and layout, remains a substantial challenge. Unconditional random sampling will naturally lead to random results without any specific guidelines or constraints. To gain more precise control over

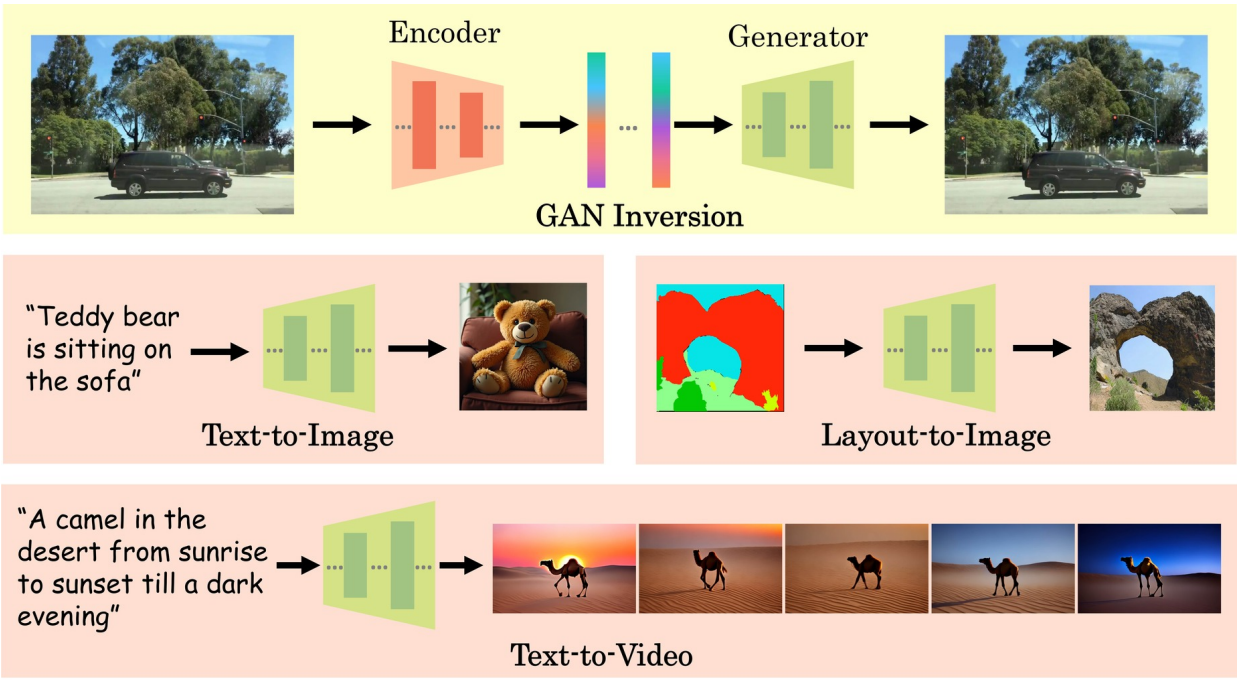


Figure 1.1: An overview of the topics addressed in this thesis. We focus on four main tasks: GAN inversion, layout-to-image, text-to-image, and text-to-video synthesis. For the latter three tasks, we implement innovative techniques using more recent diffusion models. For the illustration of the text-to-video task, the images shown are subsampled from the video sequence.

the generation process, multiple sources of auxiliary information have been utilized, e.g., an existing image (Abdal et al., 2019, Collins et al., 2020, Richardson et al., 2021), semantic label map (Park et al., 2019, Tan et al., 2021a, Sushko et al., 2022, Xue et al., 2023) or textual prompt (Rombach et al., 2022, Ramesh et al., 2022, Nichol et al., 2022). For instance, an existing image can serve as a reference for semantic content or as a style exemplar, which can be used for style transfer. Alternatively, a label map specifying the pixel-wise semantic class can also guide the layout content. Textual descriptions allow for more accessible and creative content creation, as users can simply describe in natural language what they would like to see in the outputs.

In this thesis, we aim to enable a more controllable generation process and enhance alignment with the given conditional information, particularly for GANs and diffusion models. As outlined in Fig. 1.1, this thesis concentrates on four tasks: GAN inversion, layout-to-image, text-to-image, and text-to-video generation. For

the latter three tasks, we introduce innovations using more recent diffusion models. GAN inversion aims to map a given image back to the latent space of a pretrained GAN generator, allowing the resulting latent codes to be used for further image editing and manipulation. In Chapter 3, we propose a novel GAN inversion encoder - Masked Noise Encoder (MNE), which can project the image into style and content latents respectively. Meanwhile, the MNE is capable of handling complex scene-centric data, as opposed to the object-centric focus seen in previous works. In Chapter 4, we propose two training strategies for layout-to-image diffusion models, to achieve better alignment with the label map condition. We integrate adversarial supervision into the diffusion model training pipeline so that the label map condition can be explicitly leveraged. Considering the iterative nature of DMs during inference, we introduce a multistep unrolling strategy that provides supervision over a time window rather than at a single timestep. In Chapter 5, to mitigate the issues of object negligence and attribute binding in text-to-image (T2I) synthesis, we intervene in the generation process of a pretrained T2I model with improved optimization objectives on the cross-attention maps, where cross-attention is a crucial link between the text prompt and the diffusion model. In Chapter 6, we introduce the concept of Generative Temporal Nursing (GTN), where we aim to alter the generative process on the fly during inference to improve control over the temporal dynamics and enable the generation of longer dynamic videos using a pretrained text-to-video diffusion model.

In the remainder of this chapter, we will discuss the research challenges and our contributions in Section 1.1. We then provide a detailed outline of the thesis structure in Section 1.2. Finally, we list the publications and open-source software that contributed to this thesis in Section 1.3 and Section 1.4, respectively.

1.1 Contribution Overview

This thesis focuses on improving alignment and controllability in GANs and Diffusion Models across four concrete tasks: GAN inversion, layout-to-image, text-to-image, and text-to-video generation. In the following sections, we will delve into the challenges and our contributions for each subtask.

1.1.1 Exemplar-Based Synthesis with Content-Style Disentanglement

GAN inversion aims to encode a given image into the latent codes of a pre-trained GAN generator, which facilitates image editing and manipulation. These latent codes are derived by reconstructing the image from the latent space. Previous studies (Richardson et al., 2021, Yao et al., 2022, Roich et al., 2022, Alaluf et al., 2022, Dinh et al., 2022, Šubrtová et al., 2022) have predominantly focused on simple, single-object-centric datasets such as FFHQ (Karras et al., 2019), CelebA-HQ (Karras et al., 2018), and LSUN (Yu et al., 2015). However, when applied to complex scene-centric datasets like Cityscapes (Cordts et al., 2016) and BDD100K (Yu et al., 2020), these methods often result in significant reconstruction errors, leading to unsatisfactory visual outcomes.

To address these limitations, in Chapter 3 we propose a novel GAN inversion encoder, dubbed as Masked Noise Encoder (MNE), which enables high quality reconstruction of complex scenes. We discovered that using latent vectors without spatial dimension alone are not sufficient for the faithful reconstruction of scene-centric datasets. Therefore, we design the encoder to map the image not only to latent vectors but also to intermediate noise maps with spatial dimensions. By doing so, the reconstruction quality is significantly improved. Meanwhile, MNE learns to encode the content and style information into the noise map and latent code, respectively. Favorably, MNE is equipped with strong plug-n-play ability, i.e., readily usable on novel domains without retraining or fine-tuning needed.

With MNE, we build an exemplar-based style synthesis pipeline. Given two exemplar images, we extract the noise map from the content reference and latent codes from the style reference. These elements are then combined to synthesize a novel sample. Our pipeline ensures that the resulting image reflects the semantic content of the content exemplar and the style of the style exemplar, effectively combining their properties into a cohesive visual output.

We further explored the application of our style synthesis pipeline in real-world scenarios, such as semantic segmentation. Specifically, we utilized our pipeline to generate stylized synthetic data, enabling the reuse of original labels from the content exemplar thanks to the precise control over the semantic content. We demonstrated that our pipeline can significantly enhance the domain generalization per-

formance of various segmenters without requiring additional annotation efforts. In addition to boosting performance during training, a style-augmented validation set based on known labeled data can serve as a good proxy test set, where we observe that there is a strong correlation between the performance on the synthetic test set and the real test set. This can facilitate effective model selection for deployment, thus offering great practical value.

1.1.2 Improved Layout-to-Image Diffusion Models Via Adversarial Supervision

The task of layout-to-image synthesis (L2I), also referred to as semantic image synthesis (SIS), involves generating realistic and diverse images from provided semantic label maps, which specify per-pixel semantic class labels. This task requires the generative model to effectively interpret and transform these maps into visually coherent images that adhere closely to the depicted layouts. Early studies on this task were conducted using GANs (Wang et al., 2018b, Park et al., 2019, Wang et al., 2021c, Tan et al., 2021b, Sushko et al., 2022), and have been recently extended to diffusion models (Wang et al., 2022, Xue et al., 2023, Zhang and Agrawala, 2023). GAN-based approaches often suffer from the mode collapse issue, where the generator produces similar-looking images despite sampling from different noise inputs, resulting in limited data diversity.

On the other hand, large-scale pretrained diffusion models, which have been trained on extensive datasets, are equipped with the capability of synthesizing more diverse data. Recent efforts have been devoted to adopting such pretrained diffusion models for the L2I task. FreestyleNet (Xue et al., 2023) proposed to finetune the diffusion model on the paired layout-image data and rectify the attention within the model based on the layout condition. However, fully fine-tuning the model tends to lead to overfitting on the training data. Built upon the text-to-image diffusion model Stable Diffusion (Rombach et al., 2022), FreestyleNet's overfitting is also evident in the diminished text controllability, which reflects the loss of powerful pretrained knowledge. Regardless of how the user varies the text prompt, the output remains largely insensitive to prompt changes, exhibiting little visual differentiation. Another line of work such as ControlNet (Zhang and Agrawala, 2023) has become

a more attractive option, which freezes the pretrained T2I model and introduces an additional adapter to accommodate the layout information. By doing so, the rich prior knowledge can be preserved to the maximum extent. However, a common observation is that the model outputs are often not closely aligned with the provided layout conditions. We attribute this to the suboptimal training pipeline, wherein the traditional diffusion model training loss, i.e., L_2 reconstruction loss, is employed without explicit supervision on the layout.

To achieve the goal of synthesizing diverse samples that are well aligned with the layout condition, in Chapter 4 we propose two novel training strategies for L2I diffusion models: adversarial supervision integration and multistep unrolling. Specifically, we employ a discriminator based on a semantic segmentation model, leveraging the layout condition explicitly to provide per-pixel feedback to the diffusion model generator on how closely the denoised images adhere to the input layout. Instead of employing supervision at a randomly sampled single timestep, as used in prior works, we additionally unroll multiple steps over a specified time window to mimic the inference time sampling, which provides a more comprehensive learning signal. Enabled by both training techniques, our approach can effectively ensure consistent layout alignment, while maintaining the text controllability inherent in the large-scale pretrained diffusion model.

1.1.3 Improved Generative Semantic Nursing for Text-to-Image Synthesis

Text-to-image synthesis is a rapidly advancing area in the field of generative models, where the goal is to create images from textual descriptions. This task involves generating visual content that accurately reflects the semantics and details conveyed by the text prompt. This technology has far-reaching applications in creative industries, digital art, and advertising, where T2I generation can significantly reduce manual effort and enhance creativity. Large-scale generative models such as GLIDE (Nichol et al., 2022), Stable Diffusion (Rombach et al., 2022), DALL·E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), eDiff-I (Balaji et al., 2022), have recently achieved significant progress and demonstrated exceptional capacity to generate stunning photorealistic images. However, it remains challeng-

ing to synthesize images that fully comply with the given prompt input (Marcus et al., 2022, Feng et al., 2023, Wang et al., 2023e, Chefer et al., 2023). There are two challenging semantic issues in text-to-image synthesis, i.e., "missing objects" and "attribute binding". "Missing objects" refers to the phenomenon that not all objects mentioned in the input text faithfully appear in the generated image. "Attribute binding" represents the critical compositionality problem, where the attribute information, e.g., color or texture, is not properly aligned with the corresponding object or is incorrectly attached to another object.

To mitigate these issues, Attend & Excite (A&E) (Chefer et al., 2023) has introduced the concept of Generative Semantic Nursing (GSN). The core idea is to update latent codes on the fly, enhancing the incorporation of semantic information from the given text into pretrained synthesis models. To enforce the object occurrence, A&E defined a loss objective that aims to maximize the maximum attention value for each object token. Although showing promising results on simple composition, e.g., "a cat and a dog", we observed unsatisfactory outcomes when the test prompts become more complex. We attribute this to the suboptimal loss objective, which considers only the single maximum value and fails to take spatial distribution into account.

In Chapter 5, we propose a novel objective function for GSN, termed as Divide & Bind, which exhibits outstanding capability in generating images that fully adhere to the prompt. We maximize the total variation of the attention map to encourage multiple, spatially distinct attention excitations. By distributing the attention spatially for each token, we facilitate the generation of all objects mentioned in the prompt, even in the presence of high token competition. Intuitively, this corresponds to dividing the attention map into several distinct regions. Furthermore, to mitigate the attribute binding issue, we propose a Jensen-Shannon divergence (JSD)-based binding loss. This loss explicitly aligns the distribution between the excitation of each object and its attributes. By combining both terms for optimization, our approach is able to improve the pretrained T2I model, enabling it to effectively generate multiple instances with correct attribute binding from complex textual descriptions.

1.1.4 Generative Temporal Nursing for Longer Dynamic Video Synthesis

Taking one step further from text-to-image synthesis, text-to-video (T2V) synthesis aims to generate dynamic video sequences from textual descriptions. Unlike static images, videos consist of multiple frames that must be temporally coherent and accurately reflect the input text. This task is inherently more complex, requiring the model not only to understand and generate visual content but also to manage the temporal evolution of scenes and actions as described in the text. Remarkable progress in T2V has been made by industry (OpenAI, 2024, Kaishou, 2024, AI, 2024, Runway, 2024), however, their models are not publicly accessible. Despite falling behind their industrial counterparts, open-sourced T2V diffusion models (Blattmann et al., 2023, Guo et al., 2024, Wang et al., 2023c, Chen et al., 2023, Wang et al., 2023b, Chen et al., 2024a) have still demonstrated promising results. We focused on investigating available open-sourced T2V models and identified two common issues: limited visual changes within the video, and poor ability to generate longer videos with coherent temporal dynamics. In particular, the scenes generated by these models tend to show a high degree of frame-to-frame similarity, often resembling a static image with slight changes rather than a video with dynamic and evolving content as the text prompt specified. Additionally, these models typically fail to extend beyond generating videos longer than the trained 16 frames per inference pass. Although recent studies (Qiu et al., 2024, Wang et al., 2023a) strive to produce longer videos using a sliding window approach, these methods incur significant overhead from multiple inference runs and confront the additional challenge of maintaining temporal coherence throughout these passes.

To mitigate the aforementioned issues, in Chapter 6, we propose the concept of *"Generative Temporal Nursing"* (GTN), which aims to enhance the temporal dynamics of (long) video synthesis in real-time during inference, without the need for re-training T2V models, and is designed to operate in a single pass to avoid excessive computational overhead. As a form of GTN, we propose VSTAR, consisting of **V**ideo **S**ynopsis Prompting (VSP) and **T**emporal **A**ttention **R**egularization (TAR). To disseminate the single input prompt and provide better guidance across frames, Video Synopsis Prompting leverages the capabilities of large language models (LLMs), e.g., ChatGPT (OpenAI, 2022), to decompose the single input prompt

that describes a dynamic transition into multiple stages of visual development. Furthermore, based on our systematic analysis of the architectural design of T2V models, we identified the temporal attention units as the critical component driving the dynamic aspects of video synthesis. Through detailed ablation studies comparing the temporal attention of real and synthesized videos, we found that the temporal attention of real videos exhibits a band-matrix-like structure, indicating a strong temporal correlation among adjacent frames and a diminishing correlation as the distance between frames increases. Intriguingly, the attention maps of the synthesized videos are less structured, especially for longer ones, which may account for their weaker temporal dynamics. Motivated by these findings, we introduce a straightforward yet effective Temporal Attention Regularization strategy to enhance the dynamics of generated videos. When combining both strategies, our VSTAR can effectively synthesize long dynamic videos that adhere to the input prompt, which describes a visual evolution.

1.2 Thesis Outline

In this section, we briefly describe the content of each subsequent chapter and list the corresponding publications.

Chapter 2: Related Work. In this chapter, we outline the foundational work upon which our contributions are based, along with other relevant research works. Specifically, we discuss three primary areas central to our study: generative modeling, GAN inversion, and conditional visual synthesis.

Chapter 3: Exemplar-Based Synthesis with Content-Style Disentanglement. In this chapter, we propose a novel GAN inversion encoder that can disentangle content and style information from a given image. Building upon this, we have established an exemplar-based style synthesis pipeline. We demonstrated that synthetic data generated by our pipeline not only improves the domain generalization of semantic segmenters during training but also assists in model validation and selection for deployment. This showcases the practical utility of our proposed approach.

The work presented in this chapter is published as the WACV 2023 paper titled

"Intra-Source Style Augmentation for Improved Domain Generalization" (Li et al., 2023b). An extension of the conference version has been accepted at IJCV 2024 (Li et al., 2024c). Part of the work was also presented as an extended abstract in an oral format at the "2nd Workshop and Challenge on Vision Datasets Understanding" at CVPR 2023.

Chapter 4: Improved Layout-to-Image Diffusion Models Via Adversarial Supervision. This chapter presents two training strategies for layout-to-image diffusion models, enhancing their alignment with the layout condition while simultaneously producing diverse data via text control. We first introduced a semantic segmenter-based discriminator to provide per-pixel feedback to the diffusion model generator, which explicitly leverages the input layout condition in the supervision process. Further, we propose a novel multistep unrolling strategy that encourages consistent adherence to the layout condition over a time window, rather than at just a single timestep. By employing our strategies, we improved several L2I diffusion models with different adapter designs, showcasing the model-agnostic effectiveness of our proposal.

The work presented in this chapter is based on the ICLR 2024 paper "Adversarial Supervision Makes Layout-to-Image Diffusion Models Thrive" (Li et al., 2024a).

Chapter 5: Improved Generative Semantic Nursing for Text-to-Image Synthesis. In this chapter, we focus on resolving two semantic alignment issues in pretrained T2I model, i.e., missing objects and attribute binding issue. We propose two novel inference-time objectives to update the latent codes during the generation process, ensuring that the semantic information in the given text prompt is better reflected in the final output. Our approach can effectively synthesize multiple objects with their corresponding attributes as mentioned in the text prompt.

The work presented in this chapter is published in the BMVC 2023 Oral paper "Divide & Bind Your Attention for Improved Generative Semantic Nursing" (Li et al., 2023a).

Chapter 6: Generative Temporal Nursing for Longer Dynamic Video Synthesis. In this chapter, we investigate the open-sourced T2V model, aiming to generate longer dynamic videos that involve a visual evolution. We introduce the concept of "Generative Temporal Nursing" (GTN), where we intervene in the video

generation process on-the-fly to enhance the temporal dynamics without any need for fine-tuning. As a form of GTN, we propose VSTAR, which comprises two components: Video Synopsis Prompting (VSP) and Temporal Attention Regularization (TAR). VSTAR enables the pretrained T2V model to generate longer dynamic videos beyond the typical short clips trained on, e.g., 16 frames.

The work in this chapter is based on the paper "VSTAR: Generative Temporal Nursing for Longer Dynamic Video Synthesis" (Li et al., 2025). Part of the work was presented as an extended abstract at the "AI for Content Creation Workshop" at CVPR 2024, and the full version is published at ICLR 2025. Yumeng Li is the lead author. William Beluch contributed to paper writing.

1.3 Publications

The content of this thesis is based on the following publications:

- Intra-Source Style Augmentation for Improved Domain Generalization
  Yumeng Li, Dan Zhang, Margret Keuper, Anna Khoreva
  Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023

- Divide & Bind Your Attention for Improved Generative Semantic Nursing
  Yumeng Li, Margret Keuper, Dan Zhang, Anna Khoreva
  Proceedings of the British Machine Vision Conference, BMVC 2023 (Oral)

- Intra- & Extra-Source Exemplar-Based Style Synthesis for Improved Domain Generalization
  Yumeng Li, Dan Zhang, Margret Keuper, Anna Khoreva
  International Journal of Computer Vision, IJCV 2024

- Adversarial Supervision Makes Layout-to-Image Diffusion Models Thrive
  Yumeng Li, Margret Keuper, Dan Zhang, Anna Khoreva
  Proceedings of the International Conference on Learning Representations, ICLR 2024

- VSTAR: Generative Temporal Nursing for Longer Dynamic Video Synthesis
  Yumeng Li, William Beluch, Margret Keuper, Dan Zhang, Anna Khoreva

Proceedings of the International Conference on Learning Representations, ICLR 2025

Additional publications not being part of this thesis:

- Anomaly-Aware Semantic Segmentation via Style-Aligned OoD Augmentation
  Dan Zhang, Kaspar Sakmann, William Beluch, Robin Hutmacher, **Yumeng Li**
  Proceedings of the IEEE/CVF International Conference on Computer Vision, Workshop on Robustness and Reliability of Autonomous Vehicles in the Open-world, ICCV 2023

- Domain-Aware Fine-Tuning of Foundation Models
  Ugur Ali Kaplan, Margret Keuper, Anna Khoreva, Dan Zhang, **Yumeng Li**
  International Conference on Machine Learning, Workshop on Foundation Models in the Wild, ICML 2024

- Label-Free Neural Semantic Image Synthesis
  Jiayi Wang, Kevin Alexander Laube, **Yumeng Li**, Jan Hendrik Metzen, Shin-I Cheng, Julio Borges, Anna Khoreva
  Proceedings of the European Conference on Computer Vision, ECCV 2024

1.4 Open-Source Software

As part of the work accomplished in this thesis, several software tools have been developed and released as open source. These tools are intended to facilitate further research and development in the community. Below is a list of open-sourced software published during this research:

- Accomplished for the work in Chapter 3: `https://github.com/boschresearch/ISSA`
- Accomplished for the work in Chapter 4: `https://github.com/boschresearch/Divide-and-Bind`
- Accomplished for the work in Chapter 5: `https://github.com/boschresearch/ALDM`

- Accomplished for the work in Chapter 6: https://github.com/boschresearch/VSTAR

2 Related Work

In this chapter, we aim to provide a broad overview of the related literature and background information, serving as preparation for the subsequent chapters of this thesis. We introduce generative modeling, with a particular focus on Generative Adversarial Networks (GANs) and diffusion models in Section 2.1. In Section 2.2, we provide a comprehensive overview of GAN inversion, which lays the foundation for Chapter 3. Furthermore, we delve into conditional visual synthesis in Section 2.3, covering layout-to-image, text-to-image, and text-to-video. These topics are relevant to Chapters 4 to 6 respectively.

2.1 Generative Modeling

Generative modeling has emerged as a pivotal area of research within the field of artificial intelligence, focusing on the development of models capable of generating new data samples that are indistinguishable from real data. These models, which include Variational Autoencoders (VAEs) (Kingma and Welling, 2014, Higgins et al., 2017, Van Den Oord et al., 2017), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014, Karras et al., 2018, 2019), autoregressive models (Van Den Oord et al., 2016, Esser et al., 2021, Chang et al., 2022) and diffusion models (DMs) (Ho et al., 2020, Song et al., 2020, Rombach et al., 2022), have demonstrated remarkable success across various domains, such as image synthesis, natural language processing, and audio generation. The fundamental objective of generative modeling is to learn the underlying distribution of a given dataset and to sample from this distribution to create novel instances that retain the intrinsic properties of the original data distribution.

Early work in generative modeling primarily centered around VAEs, which leverage probabilistic graphical models to encode data into a lower-dimensional latent space and subsequently decode it back to its original form with minimal loss of information. This probabilistic framework allows VAEs to generate diverse and coherent samples, though they often struggle with producing high-fidelity outputs compared to more recent approaches. The advent of GANs marked a significant breakthrough in the field, introducing a game-theoretic framework where a generator network and a discriminator network are trained simultaneously. This adversarial training process has led to the creation of highly realistic outputs, pushing the boundaries of what generative models can achieve. However, GANs are notoriously difficult to train and can suffer from issues such as mode collapse and instability. More recently, diffusion models have emerged as a powerful class of generative models. Diffusion models work by modeling the process of adding noise to data and then learning to reverse this process, effectively generating samples that closely resemble the original data distribution. DMs have shown impressive results in generating high-quality images and have been praised for their stability and diversity of outputs compared to GANs. The stable training process of DMs has enabled large-scale training and the generation of high-fidelity outputs on ex-

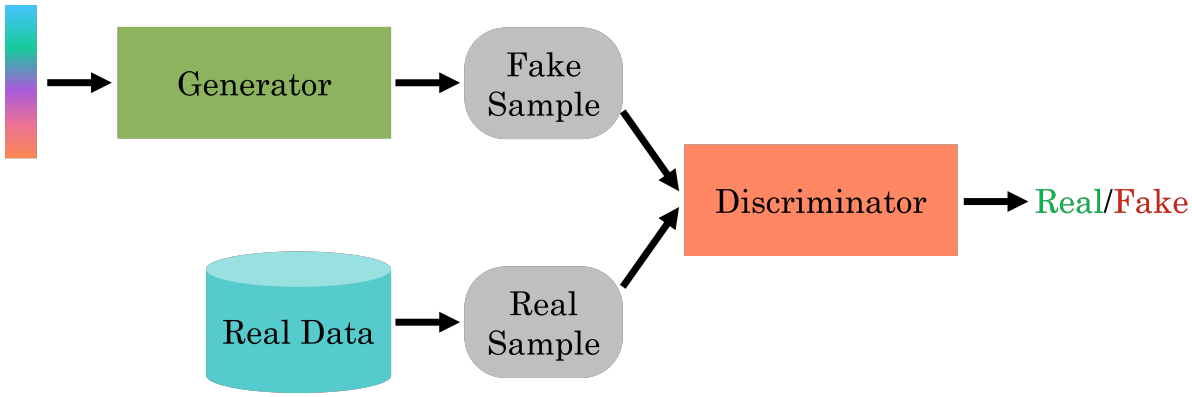


Figure 2.1: Illustration of Generative Adversarial Network (GAN).

tensive datasets. In the following, we will provide a more detailed examination of GANs in Section 2.1.1 and DMs in Section 2.1.2, which both have been studied in this thesis.

2.1.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) were introduced in 2014 by Goodfellow et al. (2014) and have revolutionized the field by introducing an adversarial training framework. GANs consist of two neural networks: the generator and the discriminator, which are trained simultaneously through a process of competition, as illustrated in Fig. 2.1. The generator network aims to create data samples that are indistinguishable from real data, fooling the discriminator. It takes a random noise vector as input and transforms it into a data sample, such as an image. The discriminator network, on the other hand, evaluates the authenticity of the samples it receives, distinguishing between real data from the training set and fake data produced by the generator. Therefore, the training of GANs can be seen as a two-player adversarial game. Formally, the training objective can be formulated as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(\boldsymbol{z})))], \tag{2.1}$$

where G and D represent the generator and discriminator, respectively. p_{data} and p_z denote the distribution of real data and input noise. $D(x)$ indicates the probability that x came from the real data distribution. In practice, the generator and discriminator are updated iteratively, typically in an alternating fashion. The respective

training objective can be derived from Eq. (2.1):

$$\min_D L_D = -\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))], \tag{2.2}$$

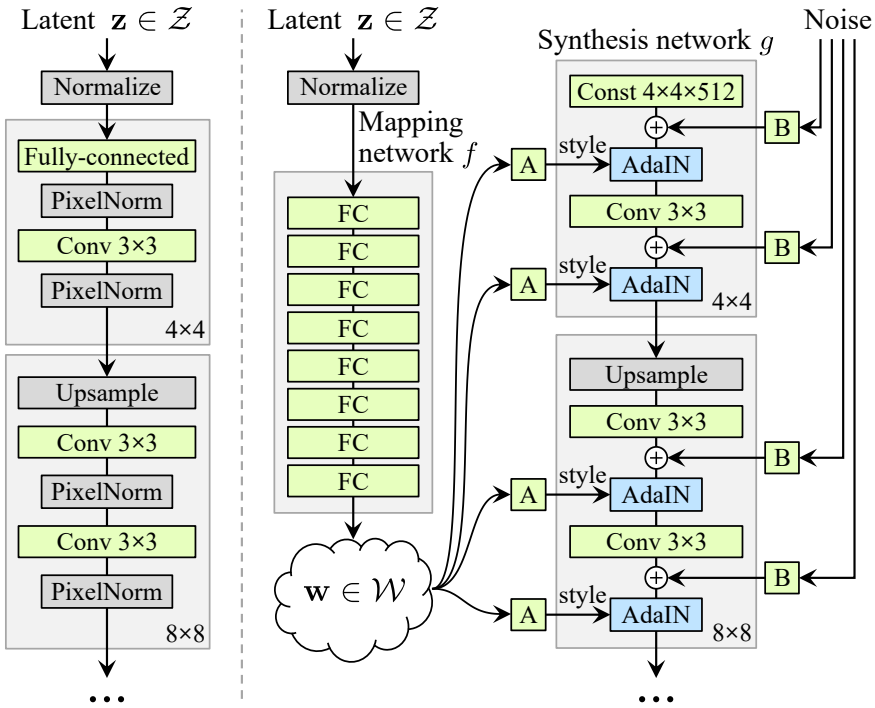
$$\min_G L_G = \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \tag{2.3}$$

However, it is observed that Eq. (2.3) may not provide sufficient gradient for G to learn well, if the discriminator can reject generated samples with high confidence as they are clearly different from the training data. To overcome this saturation problem, the non-saturating GAN loss is commonly used:

$$\min_G L_G = -\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log D(G(\boldsymbol{z}))]. \tag{2.4}$$

In theory, the solution of the two-player game is a Nash equilibrium, where neither the generator G nor the discriminator D can improve. Nevertheless, achieving Nash equilibrium in practice can be challenging due to the dynamic and often unstable nature of GAN training. Some works attempt to stabilize the training by modifying the loss function (Nowozin et al., 2016, Arjovsky et al., 2017, Mao et al., 2017).

Many efforts have also been devoted to the architectural design of GANs (Karras et al., 2018, 2019, Brock et al., 2019). The StyleGAN series (Karras et al., 2018, 2019, 2020a,b, 2021) has made notable contributions to this evolution. Introduced by Karras et al. (2019), StyleGAN brought a novel approach to the generator architecture by incorporating a style-based latent space, as illustrated in Fig. 2.2. Unlike the traditional generator that only feeds the latent code through the input layer, StyleGAN maps the input noise to the style latent, which is injected at various scales through adaptive instance normalization (AdaIN) to control different aspects of the generated outputs. This newly introduced style space enables better disentanglement of the latent space, offering greater opportunities for manipulation and editing tasks, such as via GAN inversion (Abdal et al., 2019, Tov et al., 2021, Patashnik et al., 2021), which will be described in more detail in Section 2.2. Following the success of StyleGAN, several iterations have been developed to enhance its capabilities and address its limitations. To mitigate the water droplet-like artifacts, StyleGAN2 (Karras et al., 2020b) replaced the instance normalization with weight demodulation, which is based on statistical assumptions about the signal in-



(a) Traditional (b) Style-based generator

Figure 2.2: Evolution from the traditional GAN generator to the style-based generator (Karras et al., 2019). The illustration is taken from Karras et al. (2019).

stead of actual contents of the feature maps. Additionally, StyleGAN2 revisited the progressive growing scheme and explored skip connections and residual connection design to produce high-quality images. StyleGAN2-ADA (Karras et al., 2020a) proposed an adaptive discriminator augmentation strategy that can effectively stabilize training when the training data is limited. Further, StyleGAN3 (Karras et al., 2021) redesigned all signal-processing aspects of the StyleGAN2 generator, which can reduce the texture sticking artifacts caused by aliasing, and make the morphing transition more natural.

2.1.2 Diffusion Models

Diffusion models (Ho et al., 2020, Song et al., 2020, Nichol and Dhariwal, 2021) have emerged as a powerful class of generative models capable of synthesizing high-quality results. Remarkably, they have demonstrated advantages in terms of training stability and output diversity compared to GANs. Diffusion models con-

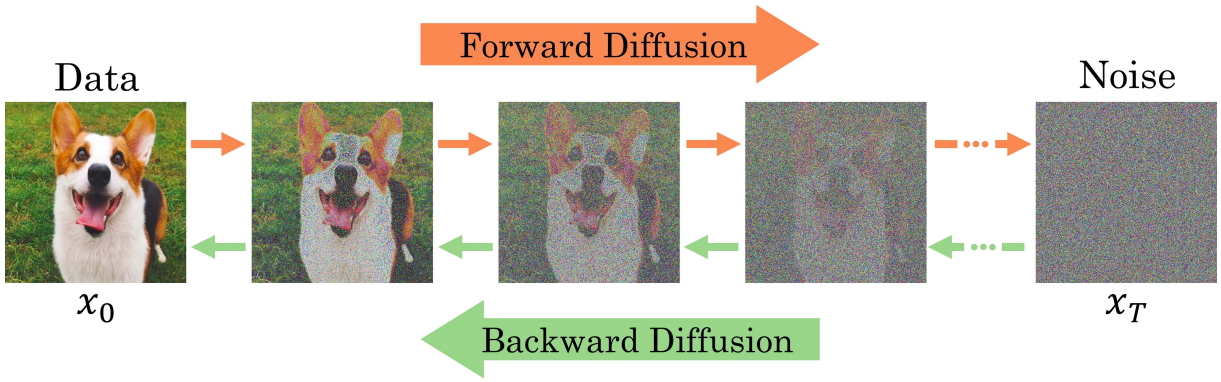


Figure 2.3: Illustration of Diffusion Model (DM).

sist of forward and backward diffusion processes, as illustrated in Fig. 2.3. In the forward process, the clean data is gradually turned into noise. A denoising model is then trained to conduct the backward diffusion process to recover the structure of the data, yielding a generative model. Formally, given a data point sampled from the real data distribution $x_0 \sim q(x)$, in the forward diffusion process, one can add Gaussian noise to the sample progressively in T steps, yielding a sequence of noisy samples $x_1, ..., x_T$:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{2.5}$$

where $\{\beta_t \in (0,1)\}_{t=1}^T$ represents the noise variance schedule. Let $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{s=0}^t \alpha_s$, one can sample x_t at any arbitrary timestep t in a closed form using reparameterization trick:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{ where } \epsilon \sim N(0, I) \tag{2.6}$$

$$\sim N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I). \tag{2.7}$$

As $T \to \infty$, the latent x_T is approximately an isotropic Gaussian distribution. Therefore, we can obtain new data points by sampling x_T from $N(0, I)$ if the reverse distribution $q(x_{t-1}|x_t)$ is learned.

Since $q(x_{t-1}|x_t)$ is intractable, in practice we learn a model p_θ to approximate

these conditional probabilities in order to run the backward diffusion process:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{2.8}$$

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{2.9}$$

The model can be trained to maximize the Evidence Lower Bound (ELBO):

$$\log p(x) \geqslant \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1)\right] \tag{2.10}$$

$$- D_{KL}\left(q(x_T|x_0)||p(x_T)\right) \tag{2.11}$$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[D_{KL}\left(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\right)\right]. \tag{2.12}$$

Empirically, Ho et al. (2020) observed that simplifying the training objectives leads to improved performance. The simplified loss term is summarized as follows:

$$L^{simple} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{2.13}$$

Typically, ϵ_θ is implemented as a UNet (Ronneberger et al., 2015) to predict the added noise at each timestep. More recently, transformer-based denoising models (Peebles and Xie, 2023) have showcased advantages in scalability and large-scale training.

While diffusion models operating in the pixel space have shown impressive results in generating high-quality images, they can be computationally intensive due to the high dimensionality of the data space. To address this challenge, Latent Diffusion Models (LDMs) (Rombach et al., 2022), also known as Stable Diffusion (SD), have been introduced, which operate in a lower-dimensional latent space rather than directly in the pixel space, as illustrated in Fig. 2.4. LDMs leverage an autoencoding model consisting of an encoder and a decoder to first map the high-dimensional data into a more compact latent space. More precisely, given an image $x \in \mathbb{R}^{H\times W\times 3}$ in RGB space, the encoder $\mathcal{E}$ projects x into a compressed latent representation $z = \mathcal{E}(x)$. Then z is mapped back to the image space by the decoder $\mathcal{D}$.

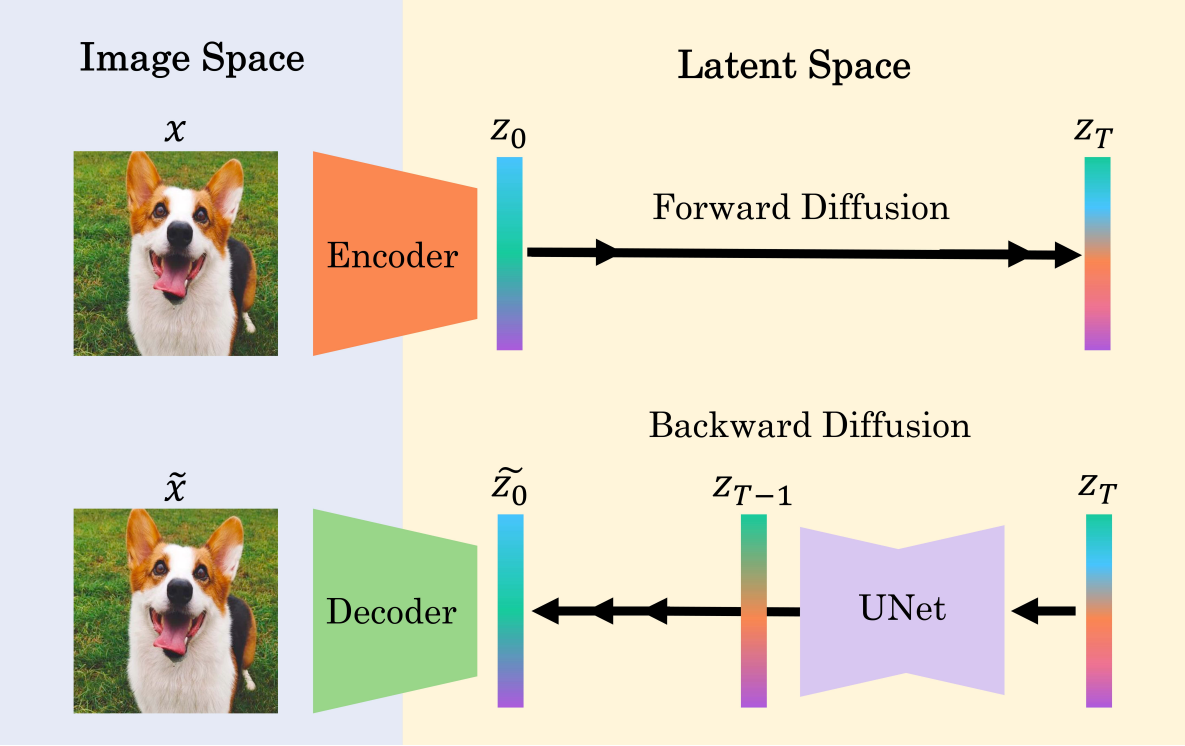


Figure 2.4: Illustration of Latent Diffusion Model (LDM).

The autoencoder is trained to reconstruct the data:

$$\tilde{x} = D(z) = D(\mathcal{E}(x)) \approx x. \tag{2.14}$$

Further, a diffusion model can be trained in the compact latent space of this trained autoencoder. The training objective described in Eq. (2.13) now reads:

$$L_{LDM} = \mathbb{E}_{z_0 \sim \mathcal{E}(x), t, \epsilon} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2 \right], \tag{2.15}$$

where $z \in \mathbb{R}^{h \times w \times c}$. Notably, the encoder downsamples the spatial dimension by a factor of $f = H/h = W/w$, which can greatly improve the computational efficiency. Thanks to the training efficiency and stability, LDMs also enables large-scale training with various conditional information (Rombach et al., 2022, Podell et al., 2024, Zhang and Agrawala, 2023), e.g., text, semantic label maps, which will be detailed in Section 2.3.

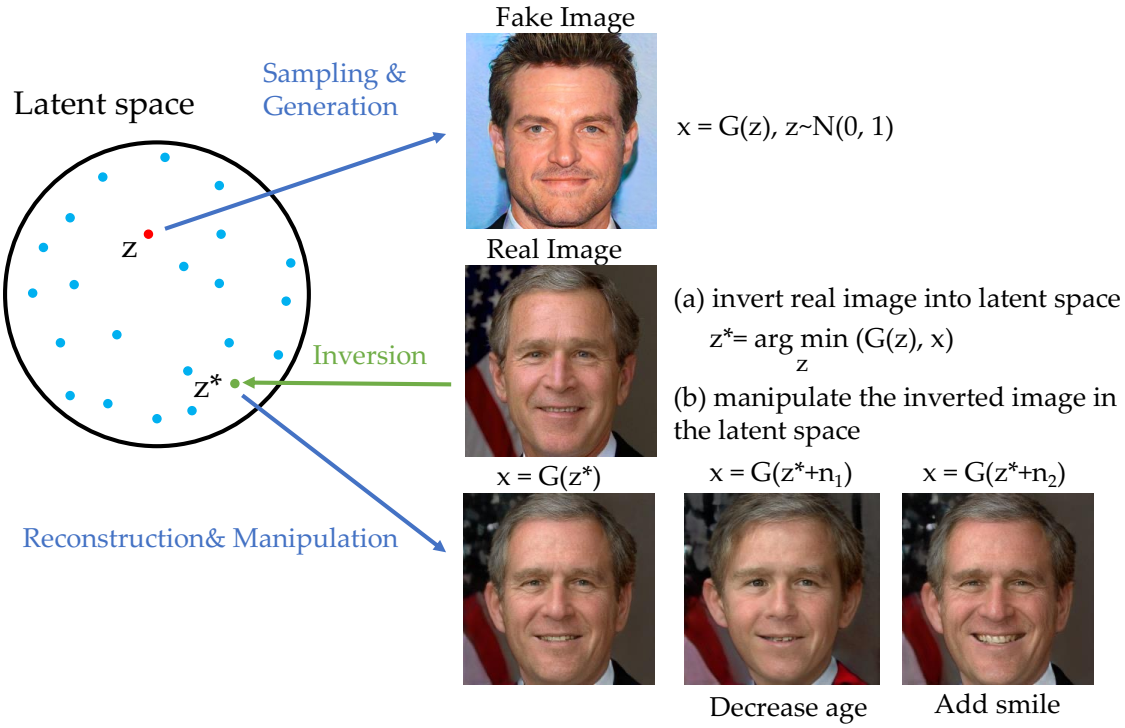


Figure 2.5: Illustration of GAN Inversion taken from Xia et al. (2022). GAN inversion maps a given real image x to the latent space and obtains the latent code z^*. The reconstructed image x^* is then obtained by $x^* = G(z^*)$. By varying the latent code $\mathbf{z}^*$ in different interpretable directions, we can edit the corresponding attribute of the real image.

2.2 GAN Inversion

GANs have demonstrated remarkable capability in generating high-quality samples that are often indistinguishable from real ones. Meanwhile, the latent space of GANs, such as StyleGAN (Karras et al., 2019), encodes rich semantic information, allowing for the manipulation and editing of outputs by varying the latent variables (Jahanian et al., 2020, Härkönen et al., 2020, Shen et al., 2020). While it is straightforward to obtain latent variables for synthesized samples, deriving latent codes for real images is less clear. To address this challenge, GAN inversion aims to map real images back into the latent space of a pre-trained GAN. Ideally, the inverted latent codes should be able to faithfully reconstruct the given image while maintaining editability, i.e., facilitating downstream manipulation and editing. However, there is a tradeoff between fidelity and editability (Tov et al., 2021, Roich et al., 2022, Moon and Park, 2022, Song et al., 2022). High-quality reconstruc-

tion may lead to latent codes that reside far away from the pretrained generator manifold, compromising the potential for further manipulation. This is due to the mismatch between the real data distribution and the synthesized one. To address this challenge, many methods have been proposed, which can be broadly categorized into three types: optimization-based (Creswell and Bharath, 2019, Abdal et al., 2019, 2020, Gu et al., 2020, Collins et al., 2020, Kang et al., 2021), encoder-based (Richardson et al., 2021, Bartz et al., 2021, Tov et al., 2021, Wei et al., 2022, Yao et al., 2022), and hybrid approaches (Zhu et al., 2020a, Chai et al., 2021, Roich et al., 2022, Dinh et al., 2022, Alaluf et al., 2022). In the remainder of this section, we will delve into each category in more detail.

2.2.1 Optimization-based GAN Inversion

Optimization-based GAN inversion methods (Creswell and Bharath, 2019, Abdal et al., 2019, 2020, Gu et al., 2020, Collins et al., 2020, Kang et al., 2021) directly optimize the latent vector on a per-image basis. Formally, given an image x and a pretrained GAN generator G, one strives to find an optimal latent code z^* by minimizing a distance metric ℓ via gradient descent:

$$z^* = \arg\min_z \ell(G(z), x), \; z \in \mathcal{Z} \sim N(0, I). \tag{2.16}$$

Typically, $\ell(\cdot)$ is a metric that measures the reconstruction quality and visual fidelity, such as ℓ_1, ℓ_2, perceptual loss (Johnson et al., 2016) and LPIPS (Zhang et al., 2018b). For instance, Abdal et al. (2019) employed a weighted combination of the perceptual loss and the pixel-wise ℓ_2 loss.

For StyleGAN inversion, it has been observed that the projected style space $\mathcal{W}$ and the extended style space $\mathcal{W}^+$ have better expressiveness and disentanglement property, rendering them more favorable choices for the GAN inversion task (Abdal et al., 2020, Cherepkov et al., 2021, Abdal et al., 2021). The extended style space $\mathcal{W}^+$ differs from the native $\mathcal{W}$ space in that it allows for separate latent codes for each layer of the generator network, rather than a single shared latent code, providing more flexibility. However, since the optimization problem is highly non-convex, it requires a good initialization to mitigate the local minima issue. Abdal et al. (2019) proposed using the average latent codes $\bar{w}$ for initialization. Some works (Zhu

et al., 2016, Roich et al., 2022, Alaluf et al., 2022) proposed hybrid approaches to provide a better starting point, which will be introduced in Section 2.2.3. Nevertheless, optimization-based methods generally have worse editability. It is also worth mentioning that the optimization is performed during inference and thus leads to significant computational overhead.

2.2.2 Encoder-based GAN Inversion

Being more computationally efficient, encoder-based methods (Richardson et al., 2021, Bartz et al., 2021, Tov et al., 2021, Wei et al., 2022, Yao et al., 2022) involve training of an encoder network E to map real images to latent codes in a single forward pass. Once trained, the encoder can be applied to different given images instead of being optimized on a per sample basis. Formally, the process can be formulated as:

$$\theta_E^* = \arg\min_{\theta_E} \ell(G(E(\theta_E, x)), x), \tag{2.17}$$

where x represents the training data, and the encoder E is parameterized with θ_E. The representative encoder-based work pSp encoder (Richardson et al., 2021) employed a feature pyramid (Lin et al., 2017) to extract three levels of feature maps, which are mapped to various latent codes in the extended style space $\mathcal{W}^+$of StyleGAN. The e4e encoder (Tov et al., 2021) further incorporated regularization losses to minimize the variance between the different style codes and enforce proximity to the original latent style space $\mathcal{W}$. In doing so, the e4e encoder achieved enhanced editability, while trading off some detail preservation. To improve the reconstruction quality, the Feature-Style encoder (Yao et al., 2022) further replaces the lower latent code prediction with a feature map prediction.

Despite much progress, most works only showcase applications on single object-centric datasets, such as CelebA-HQ (Karras et al., 2018), FFHQ (Karras et al., 2019), LSUN (Yu et al., 2015). They still fail on more complex scenes, thus restricting their application in practice. In Chapter 3, we propose a novel GAN inversion encoder, termed the Masked Noise Encoder. This encoder can reconstruct complex scene-centric data, such as driving scenes, and is capable of disentangling content and style. This enables style mixing augmentation, which is beneficial for various

downstream applications.

2.2.3 Hybrid GAN Inversion

Hybrid GAN inversion methods (Zhu et al., 2020a, Chai et al., 2021, Roich et al., 2022, Dinh et al., 2022, Alaluf et al., 2022, Song et al., 2022) seek to achieve a balance between reconstruction quality and computational efficiency, typically involving both encoder and optimization, or breaking the inversion process into two stages. Zhu et al. (2020a), Chai et al. (2021) first used an encoder to provide an initial estimate of the latent vector, which is then refined through optimization. PTI (Roich et al., 2022) similarly obtained a pivotal latent code through an encoder, then tuned the generator to further reduce distortion, especially for out-of-domain samples. Instead of direct fine-tuning, HyperStyle (Alaluf et al., 2022) and HyperInverter (Dinh et al., 2022) incorporate a hypernetwork to predict weight offsets for the generator. Designed for out-of-domain GAN inversion, Song et al. (2022) first invert the image and perform user-desired editing to obtain a coarse result. Further, a mask of interest is derived from the difference between unedited and edited results. After obtaining the mask, one can composite the edited image with the original input to resolve the out-of-domain issue. A deghosting network is employed to remove ghosting artifacts.

Recent works have explored the usage of additional inputs such as labeled regions of interest (Moon and Park, 2022) and segmentation masks (Šubrtová et al., 2022). Our method introduced in Chapter 3 only requires RGB images and a frozen generator. Importantly, it offers plug-n-play ability on unseen web-crawled images, broadening its applicability to downstream tasks.

2.3 Conditional Visual Synthesis

Instead of generating from randomly sampled noise, conditional visual synthesis aims to produce visual results, such as images and videos, that adhere to user-provided conditions. These conditions can include class labels (Mirza and Osindero, 2014, Dhariwal and Nichol, 2021, Sauer et al., 2022), spatial layouts (Wang et al., 2018b, Park et al., 2019, Sushko et al., 2022, Xue et al., 2023), and text prompts (Rom-

bach et al., 2022, Ramesh et al., 2022, Kang et al., 2023, Esser et al., 2024). The ability to guide the generative process with such specific inputs has opened new avenues in creative and practical applications, making conditional synthesis a vital area of research. Among these, three topics are particularly relevant to this thesis: layout-to-image, text-to-image and text-to-video.

Despite the rapid advancements in generative models, accurately controlling the synthesis process to produce the desired outcomes remains a significant challenge. Ensuring that the generated visuals not only meet the specified conditions but also maintain high fidelity and coherence is an active research topic. In Chapters 4 to 6, we propose several novel techniques to improve the alignment and controllability of the conditional generation process. These methods aim to enhance how generative models interpret and integrate user conditions, leading to more precise and high-quality outputs.

In what follows, we review and discuss the related works in the key areas of layout-to-image, text-to-image, and text-to-video synthesis. Finally, in Section 2.3.4, we discuss the evaluation metrics for conditional visual synthesis and datasets used.

2.3.1 Layout-to-Image Synthesis

The task of layout-to-image synthesis (L2I), also known as semantic image synthesis (SIS), is to generate realistic and diverse images given the semantic label maps, which previously has been studied based on Generative Adversarial Networks (GANs) (Wang et al., 2018b, Park et al., 2019, Wang et al., 2021c, Tan et al., 2021b, Sushko et al., 2022), and recently extended to diffusion models (Wang et al., 2022, Xue et al., 2023, Zhang and Agrawala, 2023).

GAN-Based Layout-to-Image Synthesis. The investigation can be mainly split into two directions: improving the conditional insertion in the generator (Park et al., 2019, Wang et al., 2021c, Tan et al., 2021b), and improving the discriminator's ability to provide more effective conditional supervision (Ntavelis et al., 2020, Sushko et al., 2022). For the generator design, early works such as Pix2Pix (Isola et al., 2017) and Pix2PixHD (Wang et al., 2018a) employed a UNet (Ronneberger et al., 2015), which takes the label map as input and produces an image as output. However, the semantic information cannot be well-preserved when the network goes deeper, i.e., at the

later stage of image generation. To mitigate this, SPADE (Park et al., 2019) proposed a spatially adaptive denormalization layer that directly modulates the generator's hidden representation at various scales, conditioned on the label map. Further, OASIS (Sushko et al., 2022) inherited the generator design of SPADE and introduced a pixel-wise semantic segmentation network as the discriminator. By doing so, the generator receives richer feedback instead of image-level real/fake classification. Notably, OASIS considerably improves the layout faithfulness compared to prior works. However, despite good layout alignment, samples of the above GAN-based L2I models often lack diversity and present high similarity between different generated images, especially when there is limited pixel-wise labeled training data. With the increasing prevalence of diffusion models, particularly the large-scale pretrained text-to-image diffusion models (Nichol et al., 2022, Ramesh et al., 2022, Balaji et al., 2022, Rombach et al., 2022), more attention has been devoted to leveraging pretrained knowledge for the L2I task and using diffusion models. Our work ALDM introduced in Chapter 4 falls into this field of study.

Diffusion Model Based Layout-to-Image Synthesis. Thanks to the training stability of diffusion models, it is easier to train them on large-scale data compared to GANs. Taking advantage of such powerful pretrained knowledge, most diffusion model-based L2I methods are built upon pretrained DMs, rather than being trained from scratch. PITI (Wang et al., 2022) learns a conditional encoder to match the latent representation of GLIDE (Nichol et al., 2022) in the first stage and finetune jointly in the second stage, which unfortunately leads to the loss of text editability of GLIDE. Training diffusion models in the pixel space is extremely computationally expensive as well. With the emergence of latent diffusion models, i.e., Stable Diffusion (SD) (Rombach et al., 2022), recent works (Xue et al., 2023, Zhang and Agrawala, 2023, Mou et al., 2024) made initial attempts to insert layout conditioning into SD. FreestyleNet (Xue et al., 2023) proposed to rectify the cross-attention maps in SD based on the label maps, while it also requires fine-tuning the whole SD, which largely compromises the text controllability, as shown in Fig. 4.1. On the other hand, OFT partially updates SD, T2I-Adapter (Mou et al., 2024) and ControlNet (Zhang and Agrawala, 2023) keep SD frozen, combined with an additional adapter to accommodate the layout conditioning. Despite preserving the intriguing editability via text, they do not fully comply with the label map (see Fig. 4.1). We

attribute this to the suboptimal diffusion model training objective, where the conditional layout information is only implicitly used without direct supervision. In light of this, in Chapter 4 we propose to incorporate the adversarial supervision to explicitly encourage alignment of images with the layout conditioning, and a multistep unrolling strategy during training to enhance conditional coherency across sampling steps.

Prior works (Xiao et al., 2022, Wang et al., 2023d) have also made links between GANs and diffusion models for unconditional generation. Nevertheless, they primarily build upon GAN backbones, and the diffusion process is considered as an aid to smoothen the data distribution (Xiao et al., 2022), and stabilize the GAN training (Wang et al., 2023d), as GANs are known to suffer from training instability and mode collapse. By contrast, our ALDM aims at improving L2I diffusion models, where the discriminator supervision serves as a valuable learning signal for layout alignment.

2.3.2 Text-to-Image Synthesis

Text-to-image (T2I) synthesis aims to create images based on input textual descriptions. This task requires the model to have a deep understanding of the prompt semantics and the ability to translate these semantics into visually coherent and contextually accurate images. Text-conditional GAN (Reed et al., 2016) is the first attempt for this task. It is a natural extension to the class conditional cGAN(Mirza and Osindero, 2014), where the class labels are replaced by the text embeddings. To improve the image quality, StackGAN(Zhang et al., 2017) decomposed this challenge into more manageable sub-problems. The first stage produces a low-resolution image with the primitive shape and colors of the object. The second stage further refines the output from the first stage and produces high-resolution images with more details. AttnGAN (Xu et al., 2018) was the first to develop an attention mechanism that enables better prompt understanding for GANs. Taking advantage of pretrained large foundation models such as CLIP in the generator training objective and DINO (Caron et al., 2021) in the discriminator, StyleGAN-T (Sauer et al., 2023) achieves impressive results for GAN-based T2I synthesis.

With the rapid emergence of diffusion models (Ho et al., 2020, Song et al., 2020, Nichol and Dhariwal, 2021), recent large-scale text-to-image diffusion models such

as GLIDE (Nichol et al., 2022), Stable Diffusion (Rombach et al., 2022), eDiff-I (Balaji et al., 2022), DALL·E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and SD3 (Esser et al., 2024), have showcased astonishing progress. Compared to GAN-based methods, the training stability of DMs enables large-scale training, which significantly improves the synthesis quality. One line of works directly generates images in the pixel space and is often combined with a super-resolution model to upsample the results to a higher resolution. Among them, GLIDE (Nichol et al., 2022) jointly trains the text encoder with the diffusion model, while Imagen (Saharia et al., 2022) and eDiff-I (Balaji et al., 2022) adopt a pretrained and frozen large language model, e.g., CLIP (Radford et al., 2021), T5 (Raffel et al., 2020), as the text encoder. The text embeddings generated by these language models are further utilized by the diffusion model via the cross-attention mechanism. The latter choice has become the de facto option, as it significantly reduces the computational burden during training. Another line of works involves first compressing the image to a low-dimensional latent space and then training the model in this latent space. The representative framework Stable Diffusion (SD) (Rombach et al., 2022) trained an autoencoder model in an adversarial manner following VQGAN (Esser et al., 2021) in the first stage. Subsequently, Stable Diffusion trains a denoising UNet (Ronneberger et al., 2015) in the latent space of the pretrained autoencoder model. SDXL (Podell et al., 2024) improves SD for high-resolution image synthesis by scaling up the architecture, e.g., employing a 3× larger UNet backbone. More recent work SD3 (Esser et al., 2024) builds upon the transformer-based DiT (Peebles and Xie, 2023) instead of UNet, achieving SoTA results.

Despite synthesizing high-quality images, it remains challenging to produce results that properly comply with the given text prompt. A few recent works (Feng et al., 2023, Chefer et al., 2023) aim at improving the semantic guidance purely based on the text prompt without model fine-tuning. StructureDiffusion (Feng et al., 2023) used language parsers for hierarchical structure extraction, to ease the composition during generation. Attend & Excite (Chefer et al., 2023) optimizes cross-attention maps during inference time by maximizing the maximum attention value of each object token to encourage object presence. However, we observed that Attend & Excite struggles with more complex prompts. In Chapter 5, we propose an improved generative semantic nursing approach Divide & Bind that in contrast fosters the

stimulation of multiple excitations. This improved optimization aids in holding the position amidst competition from other tokens. Additionally, we incorporate a novel binding loss that explicitly aligns the object with its corresponding attribute, yielding a more accurate binding effect.

2.3.3 Text-to-Video Synthesis

Text-to-video synthesis is a rapidly evolving field that extends the capabilities of text-to-image generation to dynamic visual content. The task involves generating video sequences that are coherent, temporally consistent, and accurately reflect the textual descriptions provided. This is inherently more challenging than text-to-image synthesis due to the need to maintain consistency across multiple frames while ensuring that the generated content aligns with the textual input.

Recent text-to-video diffusion models (Blattmann et al., 2023, Wang et al., 2023c, Chen et al., 2023, Wang et al., 2023b, Guo et al., 2024, Chen et al., 2024a) are commonly built upon large-scale pretrained T2I model, e.g., Stable Diffusion (Rombach et al., 2022). Such methods generally introduce a temporal dimension to the T2I model and incorporate temporal transformer for temporal modeling and fine-tune on a video dataset, however differ in their design choice of the temporal units and fine-tuning process. ModelScope (Wang et al., 2023b) and VideoCrafter (Chen et al., 2023) similarly inserting the temporal attention after spatial units within the UNet. LaVie (Wang et al., 2023c) and AnimateDiff (Guo et al., 2024) additionally employed Rotary Positional Encoding (Touvron et al., 2023) and Sinusoidal Positional Encoding based on the frame indices, respectively. More recently, VideoCrafter2 (Chen et al., 2024a) adopted the architecture of its predecessor, and advance the fine-tuning process by enriching existing video datasets with high-quality image data, achieving state-of-the-art T2V performance. More remarkable progress in T2V has been made by industry (OpenAI, 2024, Kaishou, 2024, AI, 2024, Runway, 2024), however, their models are not publicly accessible. Therefore, we focus on investigating open-sourced T2V diffusion models.

Due to the memory constraints, T2V models are typically trained on short video clips, i.e., 16 frames. Therefore, it is especially difficult to generate long videos beyond the trained duration. Some works (Wang et al., 2023a, Qiu et al., 2024) have specifically focused on long video generation, which typically require multiple in-

ference passes. FreeNoise (Qiu et al., 2024) proposed noise rescheduling combined with local window-based attention fusion. Gen-L-Video (Wang et al., 2023a) casts the problem as fusing multiple short video clips with temporal overlapping. However, they necessitate several passes for generation, significantly raising the inference overhead. Different from these methods, we propose VSTAR in Chapter 6 that targets long video generation with a pretrained T2V model in one *single* pass. VSTAR is more computationally efficient without running inference multiple times and bypasses the challenge of maintaining the coherency of different runs.

2.3.4 Evaluation Metric and Datasets

Evaluation of synthesis quality is crucial for advancing the development of generative models and for effectively comparing their progress in the field. As generative models continue to evolve, their outputs are judged not only by their ability to produce visually realistic results, but also by their capability to accurately and consistently reflect the provided conditions. Several commonly used evaluation metrics have been developed, each focusing on different aspects of the generated outputs.

Visual fidelity is a fundamental dimension often evaluated to determine how well a model performs in generating faithful content. One of the most widely used metrics for this is the Fréchet Inception Distance (FID) (Heusel et al., 2017). FID measures the distance between the distributions of feature representations, extracted from real images and synthesized ones using a pretrained Inception V3 network (Szegedy et al., 2016). A smaller FID value thus indicates higher visual fidelity of the results. Formally, FID can be defined as:

$$FID = \left\|\mu_{gen} - \mu_{real}\right\|^2 + \mathrm{Tr}\left(\Sigma_{gen} + \Sigma_{real} - 2(\Sigma_{gen}\Sigma_{real})^{1/2}\right), \tag{2.18}$$

where $\mu_{real}, \Sigma_{real}$ and μ_{gen}, Σ_{gen} are feature-wise mean and covariance matrix of the real and generated images, respectively.

In addition to perceived quality, diversity is another important dimension to consider. While FID has been widely adopted for evaluating perceptual quality, it fails to capture the lack of diversity, as a model replicating the real training set without much variation can achieve a very low FID score (Sajjadi et al., 2018). To complement FID, Precision and Recall metrics (Sajjadi et al., 2018) are often used

to provide a more comprehensive evaluation of generative models with respect to fidelity and diversity.

For conditional visual synthesis, alignment with the input conditions is an essential aspect as well. Given the layout condition, alignment evaluation is typically performed by employing a pretrained segmentation model and computing the Mean Intersection over Union (mIoU) between the predictions on synthesized images and the input condition. For text input conditions, evaluating alignment involves assessing how accurately the generated content reflects the details and semantics described in the text, with the aid of vision-language models. CLIPScore has been a popular metric to compute the embedding similarity between the text prompt and synthesized results, which leverages a pretrained multimodal CLIP model (Radford et al., 2021). The CLIP (Contrastive Language-Image Pre-training) model is trained on a large-scale image-text paired dataset, aligning visual and textual representations in a unified manner. However, some studies (Hu et al., 2023b, Lu et al., 2024) found that CLIPScore is suboptimal for complex prompts and more fine-grained evaluation, especially concerning specific objects and attributes. More recent evaluation metrics such as TIFA score (Hu et al., 2023b) utilize the power of Large Language Models (LLMs) to design questions of interest and then employ Visual Question Answering (VQA) models to verify if a certain aspect is properly synthesized in the output.

Additionally, for video synthesis, evaluation along the temporal dimension is a crucial aspect. Temporal consistency is typically measured between consecutive frames using an image-language model such as CLIP. Due to the current lack of powerful video-language foundation models, the textual alignment of T2V models is performed on each individual frame, similar to the evaluation process used in T2I tasks. Furthermore, user studies are often conducted to gather human feedback based on various criteria. The insights gained from these studies are particularly important in fields such as entertainment and advertising, where user satisfaction is critical.

In addition to evaluation metrics, datasets play a critical role in training and assessing the performance of generative models. Object-centric datasets such as FFHQ (Karras et al., 2019) and CelebA-HQ (Karras et al., 2018) are frequently used for GANs. These datasets contain high-resolution images of faces, offering diverse

attributes, e.g., age, ethnicity. Yet, they are relatively simple, as each image contains only a single face, and the facial structures often share a high similarity. In this thesis, particularly in Chapters 3 and 4, we focused on scene-centric datasets, such as Cityscapes (Cordts et al., 2016), ACDC (Sakaridis et al., 2021), and ADE20K (Zhou et al., 2017). Cityscapes is an urban driving dataset collected from 50 different cities, primarily in Germany. It contains 2975 finely annotated training images and 500 images for validation. Each image is annotated with pixel-level semantic segmentation, providing detailed labels for 19 semantic classes, including cars, pedestrians, buildings, roads, and other elements commonly found in urban environments. While Cityscapes focuses on urban scenes under normal weather conditions, the ACDC (Adverse Conditions Dataset with Correspondences) (Sakaridis et al., 2021) expands this by incorporating four adverse weather conditions: rain, fog, snow, and nighttime, which largely enhances the dataset diversity. ACDC is mostly captured in Zürich, containing 4006 images with high-quality per-pixel semantic class annotation. Unlike Cityscapes and ACDC, which focus on urban driving environments, ADE20K (Zhou et al., 2017) offers a more diverse range of scenes, including indoor and outdoor environments. It contains over 20000 images with pixel-level annotations across 150 semantic classes, covering a broad spectrum of objects such as buildings, furniture, animals, and vehicles. Compared to object-centric datasets, these scene-centric datasets pose a greater challenge for generative models, as they require an understanding not only of individual objects but also of their spatial relationships. Our proposed methods have demonstrated significant advantages in these complex settings, improving both visual fidelity and alignment with the conditional information. For text-to-image and text-to-video synthesis introduced in Chapters 5 and 6, our approaches are training-free and do not require any training data. Following prior works (Chefer et al., 2023, Yuan et al., 2024), we use a set of predefined text prompts to evaluate generation quality, alignment with the textual input, and other factors. More detailed evaluation protocols can be founded in the respective chapters.

3 Exemplar-Based Synthesis with Content-Style Disentanglement

In Chapter 3, we propose a novel masked noise encoder for StyleGAN2 inversion. Based on this, we have developed an exemplar-based style synthesis pipeline

that can significantly improve domain generalization in semantic segmentation. The model learns to faithfully reconstruct the image, preserving its semantic layout through noise prediction. Random masking of the estimated noise enables the style mixing capability of our model, i.e. it allows to alter the global appearance without affecting the semantic layout of an image. Using the proposed masked noise encoder to randomize style and content combinations in the training set, i.e., intra-source style augmentation (ISSA) effectively increases the diversity of training data and reduces spurious correlation. As a result, we achieve up to 12.4% mIoU improvements on driving-scene semantic segmentation under different types of data shifts, i.e., changing geographic locations, adverse weather conditions, and day to night. ISSA is model-agnostic and straightforwardly applicable with CNNs and Transformers. It is also complementary to other domain generalization techniques, e.g., it improves the recent state-of-the-art solution RobustNet by 3% mIoU in Cityscapes to Dark Zürich. In addition, we demonstrate the strong plug-n-play ability of the proposed style synthesis pipeline, which is readily usable for extra-source exemplars e.g., web-crawled images, without any retraining or fine-tuning. Moreover, we study a new use case to indicate neural network's generalization capability by building a stylized proxy validation set. This application has a significant practical sense for selecting models to be deployed in the open-world environment. Our code is available at `https://github.com/boschresearch/ISSA`. The content of this chapter corresponds to the WACV 2023 paper "Intra-Source Style Augmentation for Improved Domain Generalization" (Li et al., 2023b) and its extended version published at IJCV 2024 (Li et al., 2024c).

3.1 Introduction

The varying environment in real life with potentially diverse illumination and adverse weather conditions makes challenging the deployment of deep learning models in an open-world (Sakaridis et al., 2021, Zhang et al., 2021a). Therefore, improving the generalization capability of neural networks is crucial for safety-critical applications such as autonomous driving (see for example Fig. 3.1). While generally the target domains can be inaccessible or unpredictable at training time, it is important to train a generalizable model, based on the known (source) domain,

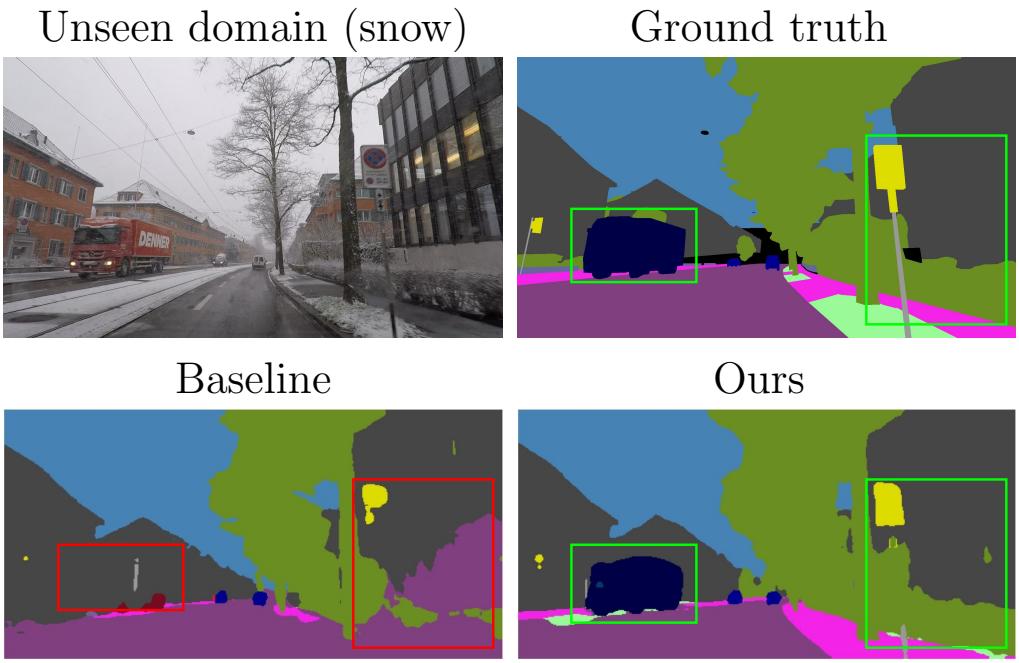


Figure 3.1: Semantic segmentation results of HRNet (Wang et al., 2021b) on unseen domain (snow), trained on Cityscapes (Cordts et al., 2016) and tested on ACDC (Sakaridis et al., 2021). The model trained with our ISSA can successfully segment the truck, while the baseline model fails completely.

which may offer only a limited or biased view of the real world (Burton et al., 2017, Shafaei et al., 2018).

Diversity of the training data is considered to play an important role for domain generalization, including natural distribution shifts (Taori et al., 2020). Many existing works assume that multiple source domains are accessible during training (Li et al., 2018a, Balaji et al., 2018, Li et al., 2018b, 2020, Jin et al., 2020, Zhou et al., 2020, Hu et al., 2020). For instance, Li et al. (2018a) applied meta-learning to better generalize to unseen domains, where source domains are divided into meta-source and meta-target domains to simulate domain shift; Hu et al. (2020) propose multi-domain discriminant analysis to learn a domain-invariant feature transformation. However, for pixel-level prediction tasks such as semantic segmentation, collecting diverse training data involves a tedious and costly annotation process (Caesar et al., 2018). Therefore, improving and predicting generalization from a *single source domain* is exceptionally compelling, particularly for semantic segmentation.

One pragmatic way to improve data diversity is by applying data augmenta-

tion. It has been widely adopted in solving different tasks, such as image classification (Zhang et al., 2018a, Hendrycks et al., 2019, Verma et al., 2019, Hong et al., 2021, Zhou et al., 2021), GAN training with limited data (Karras et al., 2020a, Jiang et al., 2021), or pose estimation (Peng et al., 2018, Bin et al., 2020, Wang et al., 2021a). One line of data augmentation techniques focuses on increasing the content diversity in the training set, such as geometric transformation (e.g., cropping or flipping), CutOut (DeVries and Taylor, 2017), and CutMix (Yun et al., 2019). However, CutOut and CutMix are ineffective on natural domain shifts, as reported in (Taori et al., 2020). Style augmentation, on the other hand, only modifies the style - the non-semantic appearance such as texture and color of the image (Gatys et al., 2016) - while preserving the semantic content. By diversifying the style and content combinations, style augmentation can reduce overfitting to the style-content correlation in the training set, improving robustness against domain shifts. Hendrycks corruptions (Hendrycks and Dietterich, 2018) provide a wide range of synthetic styles, including weather conditions. However, they are not always realistic looking, thus being still far from resembling natural data shifts. In this work, we propose an exemplar-based style synthesis pipeline for semantic segmentation, aiming to improve the style diversity in the training and validation set without extra labeling effort.

Our exemplar-based style synthesis technique is based on the inversion of StyleGAN2 (Karras et al., 2020b), which is the state-of-the-art unconditional Generative Adversarial Network (GAN) and thus ensures high quality and realism of synthetic samples. GAN inversion allows encoding a given image to latent variables, and thus facilitates faithful reconstruction with style mixing capability. To realize the synthesis pipeline, we learn to separate semantic content from style information based on a single source domain. This allows to alter the style of an image while leaving the content unchanged. In particular, we focus on intra-source style augmentation (ISSA). Namely, our exemplar-based style synthesis makes use of training samples from the source domain, extracting their styles and contents followed by randomly mixing them up. In doing so, we can increase the data diversity and alleviate the spurious correlation in the given training data.

The faithful reconstruction of images with complex structures such as driving scenes is non-trivial. Prior methods (Richardson et al., 2021, Yao et al., 2022,

Roich et al., 2022, Alaluf et al., 2022, Dinh et al., 2022, Šubrtová et al., 2022) are mainly tested on simple single-object-centric datasets, e.g., FFHQ (Karras et al., 2019), CelebA-HQ (Karras et al., 2018), or LSUN (Yu et al., 2015). As shown in (Abdal et al., 2020), extending the native latent space of StyleGAN2 with a stochastic noise space can lead to improved inversion quality. However, all style *and* content information will be embedded in the noise map, leaving the latent codes inactive in this setting. Therefore, to enable the precise reconstruction of complex driving scenes as well as style mixing, we propose a masked noise encoder for StyleGAN2. The proposed random masking regularization on the noise map encourages the generator to rely on the latent prediction for reconstruction. Thus, it allows to effectively separate content and style information and facilitates realistic style mixing, as shown in Fig. 3.2.

We further discover an excellent plug-n-play ability of the proposed style synthesis pipeline, i.e., it can be directly applied to unseen domains without requiring the re-training of the encoder or generator. For instance, in Fig. 3.11, we employ our pipeline directly on web-crawled images, where the model is only trained on Cityscapes. This appealing property opens up the opportunity to go beyond intra-source exemplar-based style mixing, and grants us more flexibility to harness extra-source data for style synthesis. Thus, we also experiment with extra-source style argumentation (ESSA) to further improve the generalization performance.

Besides data augmentation, we explore the usage of the proposed pipeline for assessing neural networks' generalization capability in Section 3.3.5. By transferring styles from unannotated data samples of the target domain to existing labelled data, we can build a style-augmented proxy set for validation without introducing extra-labelling effort. We observe that performance on this proxy set has a strong correlation with the real test performance on unseen target data, which could be used in practice to select more suitable models for deployment.

In summary, we make the following contributions:

- We propose a masked noise encoder for GAN inversion, which enables high quality reconstruction and style mixing of complex scene-centric datasets.
- We exploit GAN inversion for intra-source data augmentation, which can improve generalization under natural distribution shifts on semantic segmentation.

- Extensive experiments demonstrate that our proposed augmentation method ISSA consistently promotes domain generalization performance on driving-scene semantic segmentation across different network architectures, achieving up to 12.4% mIoU improvement, even with limited diversity in the source data and without access to the target domain.

- We discover the plug-n-play ability of our masked noise encoder, and showcase its potential of direct application on extra-source data such as web-crawled images.

- We further explore the usage of the proposed pipeline for assessing models' generalization performance on unseen data. By building a style-augmented proxy validation set on known labelled data, we observe that there is a strong correlation between the performance on the proxy validation set and the real test set, which offers useful insights for model selection without introducing any extra annotation effort.

3.2 Method

We introduce our exemplar-based style synthesis pipeline in Section 3.2.1, which relies on GAN inversion that can offer faithful reconstruction and style mixing of images. To enable better style-content disentanglement, we propose a masked noise encoder for GAN inversion in Section 3.2.2. Its detailed training loss is described in Section 3.2.3.

3.2.1 Exemplar-Based Style Synthesis Pipeline

The lack of data diversity and the existence of spurious correlation in the training set often lead to poor domain generalization. To mitigate them, the proposed style synthesis pipeline aims at 1) extracting styles from given exemplars, and 2) augmenting the training samples in the source domain with the new styles, while preserving their semantic content. For data augmentation, it employs GAN inversion to randomize the style-content combinations. In doing so, it diversifies the

Figure 3.2: Qualitative results (best view in color and zoom in) of StyleGAN2 inversion methods on Cityscapes, i.e., pSp (Richardson et al., 2021), pSp†, Feature-Style encoder (Yao et al., 2022) and our masked noise encoder. Note, pSp† is an improved version of pSp (Richardson et al., 2021) introduced by us, training pSp with an additional discriminator and incorporate synthesized images for better initialization. pSp† can reconstruct the rough layout of the scene but still struggles to preserve details. The Feature-Style encoder shows a better reconstruction quality, yet it cannot faithfully reconstruct small objects (e.g. pedestrian), and some objects (e.g. the vehicle, bicycle) are rather blurry. Our masked noise encoder has highest image fidelity, preserving finer details in the inverted image.

source dataset and reduces spurious style-content correlations. Because the content of images is preserved and only the style is changed, the ground truth label maps can be re-used for training and validation, without requiring any further annotation effort.

Our style synthesis pipeline is built on top of an encoder-based GAN inversion, given its fast inference. GANs, such as StyleGANs (Karras et al., 2019, 2020b,a), have shown the capability of encoding rich semantic and style information in intermediate features and latent spaces. For encoder-based GAN inversion, an encoder is

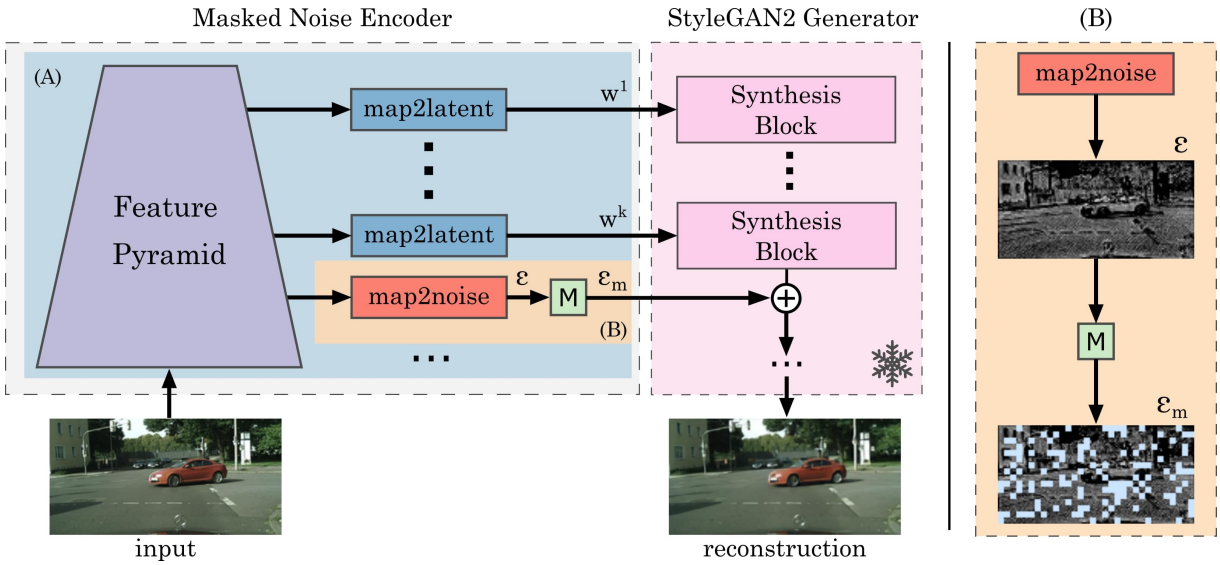


Figure 3.3: Method overview. Our encoder is built on top of the pSp encoder (Richardson et al., 2021), shown in the blue area (A). It maps the input image to the extended latent space $\mathcal{W}^+$ of the pre-trained StyleGAN2 generator. To promote the reconstruction quality on complex scene-centric dataset, e.g., Cityscapes, our encoder additionally predicts the noise map at an intermediate scale, illustrated in the orange area (B). M stands for random noise masking, regularization for the encoder training. Without it, the noise map overtakes the latent codes in encoding the image style, so that the latter cannot make any perceivable changes on the reconstructed image, thus making style mixing impossible.

trained to invert an input image back into the latent space of a pre-trained GAN generator. The encoder is desired to separately encode the style and content information of the input image. With such an encoder, it can synthesize new training samples with new style-content combinations. In particular, we are interested in intra-source style augmentation (ISSA), where the encoder should take the content and style codes from different training samples within the source domain and feed them to the pre-trained generator. If this encoder-based GAN inversion can also handle unseen data, we will further make use the styles of exemplars outside the source domain, such as web-crawled images, enabling extra-source style augmentation (ESSA). In both cases, since only the styles of the training samples in the source domain are modified, the newly synthesized training samples already have their ground truth label maps in place.

StyleGAN2 can synthesize natural looking images resembling complex scene-centric datasets such as Cityscapes (Cordts et al., 2016) and BDD100K (Yu et al.,

2020). However, existing GAN inversion encoders cannot provide the desired fidelity and style mixing capability to enable ISSA and ESSA for an improved domain generalization of semantic segmentation. Loss of fine details or inauthentic reconstruction of small-scale objects would even harm the model's generalization ability. Therefore, we propose a novel encoder design to invert StyleGAN2, termed *masked noise encoder* (see Fig. 3.3).

3.2.2 Masked Noise Encoder

We build our encoder upon the pSp encoder (Richardson et al., 2021). It employs a feature pyramid (Lin et al., 2017) to extract multi-scale features from a given image, see Fig. 3.3-(A). We improve over pSp by identifying in which latent space to embed the input image for the high-quality reconstruction of the images with complex street scenes. Further, we propose a novel training scheme to enable the style-content disentanglement of the encoder, thus improving its style mixing capability.

Extended Latent Space. The StyleGAN2 generator takes the latent code $w \in \mathcal{W}$ generated by an MLP network and randomly sampled additive Gaussian noise maps $\{\epsilon\}$ as inputs for image synthesis. As pointed out in Abdal et al. (2019), it is suboptimal to embed a real image into the original latent space $\mathcal{W}$ of StyleGAN2, due to the gap between the real and synthetic data distributions. A common practice is to map the input image into the extended latent space $\mathcal{W}^+$. The multi-scale features of the pSp feature pyramid are respectively mapped to the latent codes $\{w^k\}$ at the corresponding scales of the StyleGAN2 generator, i.e., map2latent in Fig. 3.3-(A).

Additive Noise Map. The latent codes $\{w^k\}$ from the extended latent space $\mathcal{W}^+$ alone are not expressive enough to reconstruct images with diverse semantic layouts such as Cityscapes (Cordts et al., 2016) as shown in Fig. 3.2-(pSp†). The latent codes of StyleGAN2 are one-dimensional vectors that modulate the feature vectors at different spatial positions identically. Therefore, they cannot precisely encode the semantic layout information, which is spatially varying. To address this issue, our encoder additionally predicts the additive noise map ε of the StyleGAN2 at an intermediate scale, i.e., map2noise in Fig. 3.3-(B). The noise map ε has spatial dimensions, making it inherently capable of encoding more information. It is particularly

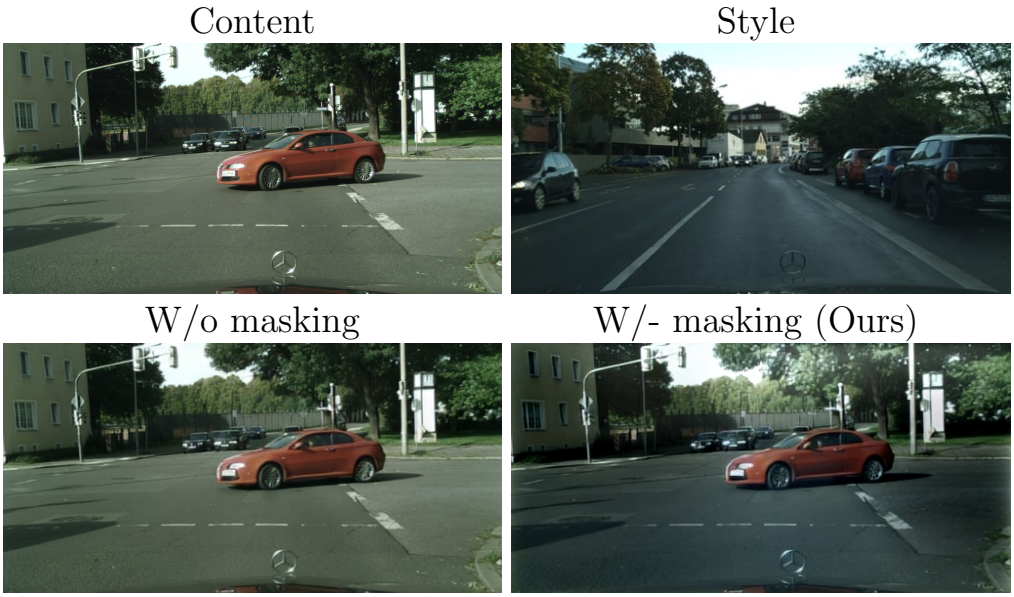


Figure 3.4: Style mixing effect enabled by random noise masking (best view in color). Despite the good reconstruction quality, the encoder trained without masking cannot change the style of the given Content image. In contrast, the encoder trained with masking can modify it using the style from the given Style image.

advantageous when dealing with content information that varies spatially, as the noise map can more readily accommodate such information. As evidenced by the visualization presented in Fig. 3.5, the noise map is adept at capturing the semantic content of the scene.

Random Noise Masking. While offering high-quality reconstruction, the additive noise map can be too expressive so that it encodes nearly all perceivable details of the input image. This results in a poor style-content disentanglement and can damage the style mixing capability of the encoder (see Fig. 3.4). To avoid this undesired effect, we propose to regularize the noise prediction of the encoder by random masking of the noise map. Note that the random masking as a regularization technique has also been successfully used in reconstruction-based self-supervised learning (Xie et al., 2022, He et al., 2022). In particular, we spatially divide the noise map into non-overlapping $P \times P$ patches, see $\boxed{\text{M}}$ in Fig. 3.3-(B). Based on a predefined ratio ρ, a subset of patches is randomly selected and replaced by patches of unit Gaussian random variables $\epsilon \sim N(0, 1)$ of the same size. $N(0, 1)$ is the prior distribution of the noise map at training the StyleGAN2 generator. We call this

Figure 3.5: Noise map visualization of our masked noise encoder. The noise map encodes the semantic content of the image.

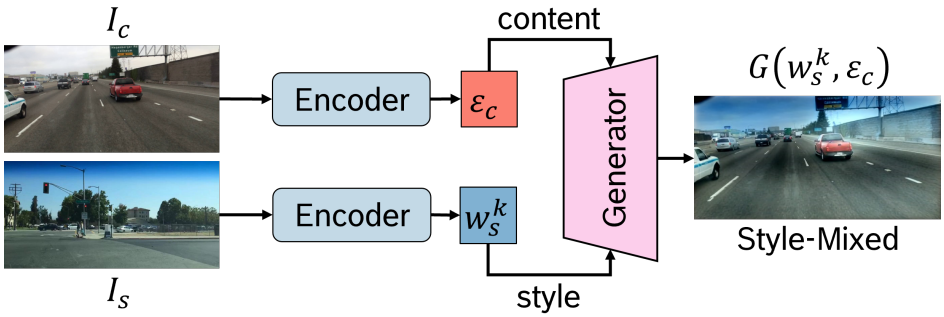


Figure 3.6: Style mixing process. The generator G takes the latent codes $\{w_s^k\}$ of I_s and the noise map ε_c of I_c, and produce the stylized image, i.e., $G(w_s^k, \varepsilon_c)$.

encoder *masked noise encoder* as it is trained with random masking to predict the noise map.

The proposed random masking reduces the encoding capacity of the noise map, hence encouraging the encoder to jointly exploit the latent codes $\{w^k\}$ for reconstruction. Fig. 3.7 visualizes the style mixing effect. The encoder takes the noise map ε_c and latent codes $\{w_s^k\}$ from the content image and style image, respectively. Then, they are fed into StyleGAN2 to synthesize a new image, i.e., $G(w_s^k, \varepsilon_c)$, as illustrated in Fig. 3.6. If the encoder is not trained with random masking, the new image does not have any perceptible difference with the content image. This means the latent codes $\{w^k\}$ encode negligible information of the image. In contrast, when being trained with masking, the encoder creates a novel image that takes the content and style from two different images. This observation confirms the enabling

Figure 3.7: Visual examples of style mixing on BDD100K (best view in color) enabled by our masked noise encoder. By combining the latent codes $\{w_s^k\}$ of I_s and the noise map ε_c of I_c, the synthesized images $G(w_s^k, \varepsilon_c)$ preserve the content of I_c with a new style resembling I_s.

role of masking for content and style disentanglement, and thus the improved style mixing capability. The noise map no longer encodes all perceptible information of the image, including style and content. In effect, the latent codes $\{w^k\}$ play a more active role in controlling the style. In Fig. 3.5, we further visualize the noise map of the masked noise encoder and observe that it captures well the semantic content of the scene.

Additionally, we discover that our masked noise encoder is equipped with strong plug-n-play ability, i.e., readily usable on novel domains without retraining or fine-

tuning. As shown in Fig. 3.11, the masked noise encoder together with the generator which is trained on Cityscapes not only reconstruct unseen domain data (e.g., north polar bear), but also remain the style mixing capability (e.g., turning bright day into a sunset scene). This generalization capability allows us to further exploit extra-source data for style synthesis, i.e., ESSA. Except that the styles are extracted from external exemplars, the style synthesis process of ESSA is identical to ISSA.

3.2.3 Encoder Training Loss

Mathematically, the proposed StyleGAN2 inversion with the masked noised encoder E^M can be formulated as

$$\begin{aligned} \{w^1, \ldots, w^K, \varepsilon\} &= E^M(x); \\ x^* &= G \circ E^M(x) = G(w^1, \ldots, w^K, \varepsilon). \end{aligned} \tag{3.1}$$

The masked noise encoder E^M maps the given image x onto the latent codes $\{w^k\}$ and the noise map ε. The StyleGAN2 generator G takes both $\{w^k\}$ and ε as the input and generates x^*. Ideally, x^* should be identical to x, i.e., a perfect reconstruction.

When training the masked noise encoder E^M to reconstruct x, the original noise map ε is masked before being fed into the pre-trained G

$$\varepsilon_M = (1 - M_{noise}) \odot \varepsilon + M_{noise} \odot \epsilon, \tag{3.2}$$

$$\tilde{x} = G(w^1, \ldots, w^K, \varepsilon_M), \tag{3.3}$$

where M_{noise} is the random binary mask, $\odot$ indicates the Hadamard product, and $\tilde{x}$ denotes the reconstructed image with the masked noise ε_M. The training loss for the encoder is given as

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda_1 \mathcal{L}_{lpips} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{reg}, \tag{3.4}$$

where $\{\lambda_i\}$ are weighting factors. The first three terms are the pixel-wise MSE loss, learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018b) loss and

adversarial loss (Goodfellow et al., 2014),

$$\mathcal{L}_{mse} = \left\|(1 - M_{img}) \odot (x - \tilde{x})\right\|_2, \tag{3.5}$$

$$\mathcal{L}_{lpips} = \left\|(1 - M_{feat}) \odot (\mathrm{VGG}(x) - \mathrm{VGG}(\tilde{x}))\right\|_2, \tag{3.6}$$

$$\mathcal{L}_{adv} = -\log D(G(E^M(x))). \tag{3.7}$$

which are the common reconstruction losses for encoder training (Zhu et al., 2020a, Richardson et al., 2021). Note that masking removes the information of the given image x at certain spatial positions, the reconstruction requirement on these positions should then be relaxed. M_{img} and M_{feat} are obtained by up- and downsampling the noise mask M_{noise} to the image size and the feature size of the VGG-based feature extractor. The adversarial loss is obtained by formulating the encoder training as an adversarial game with a discriminator D that is trained to distinguish between reconstructed and real images.

The last regularization term is defined as

$$\mathcal{L}_{reg} = \|\varepsilon\|_1 + \left\|E^M_w(G(w_{gt}, \epsilon)) - w_{gt}\right\|_2. \tag{3.8}$$

The L1 norm helps to induce sparse noise prediction. It is complementary to random masking, reducing the capacity of the noise map. The second term is obtained by using the ground truth latent codes w_{gt} of synthesized images $G(w_{gt}, \epsilon)$ to train the latent code prediction $E^M_w(\cdot)$ (Yao et al., 2022). It guides the encoder to stay close to the original latent space of the generator, speeding up the convergence.

3.3 Experiments

We start from the experiment setup in Section 3.3.1. Then, Section 3.3.2 and Section 3.3.3 respectively report our experiments on the masked noise encoder for StyleGAN2 inversion and ISSA for improved domain generalization of semantic segmentation.

3.3.1 Experimental Setup

Datasets. We conduct extensive experiments on four diverse driving scene datasets, which are Cityscapes (CS) (Cordts et al., 2016), BDD100K (BDD) (Yu et al., 2020), ACDC (Sakaridis et al., 2021), and Dark Zürich (DarkZ) (Sakaridis et al., 2019). Cityscapes is collected from different cities primarily in Germany, under good or medium weather conditions during daytime. BDD100K is a driving-scene dataset collected in the US, representing a geographic location shift from Cityscapes. Besides, it also includes more diverse scenes (e.g., city streets, residential areas, and highways) and different weather conditions captured at different times of the day. Both ACDC and Dark Zürich are collected in Switzerland. ACDC contains four adverse weather conditions (rain, fog, snow, night) and Dark Zürich contains night scenes. The default setting is to use Cityscapes as the source training data, whereas the validation sets of the other datasets represent unseen target domains with different types of natural shifts, i.e., used only for testing. Additionally, we also study the challenging day-to-night generalization scenario, where BDD100K-Daytime is used as the source set, ACDC-Night and Dark Zürich are treated as unseen domains. In both cases, we consider a *single source domain* for training.

Training details. We experiment with two image resolutions: 128×256 and 256×512. The StyleGAN2 (Karras et al., 2020a) model is first trained to *unconditionally* synthesize images and then fixed during the encoder training. To invert the pretrained StyleGAN2 generator, the masked noise encoder predicts both latent codes in the extended $\mathcal{W}^+$ space and the additive noise map. In accordance with the StyleGAN2 generator, $\mathcal{W}^+$ space consists of 14 and 16 latent code vectors for the input resolution 128×256 and 256×512, respectively. The additive noise map is always at the intermediate feature space with one fourth of the input resolution. We use the same encoder architecture, optimizer, and learning rate scheduling as pSp (Richardson et al., 2021). Our encoder is trained with the loss function defined in Eq. (3.4) with $\lambda_1 = 10$ and $\lambda_2 = \lambda_3 = 0.1$. For our random noise masking, we use a patch size P of 4 with a masking ratio $\rho = 25\%$. A detailed ablation study on the masking and noise map of the encoder can be found in Section 3.3.2.

We use the trained masked noise encoder to perform ISSA as described in Section 3.2.1. We experiment with several architectures for semantic segmentation, i.e.,

Figure 3.8: Influence of the noise map resolution on style-mixing ability. Using higher resolution noise map, e.g., $H \times W$, leads to poor style-mixing ability. While too low resolution, e.g., $\frac{H}{16} \times \frac{W}{16}$, cannot reconstruct the scene faithfully.

HRNet (Wang et al., 2021b), SegFormer (Xie et al., 2021), and DeepLab v2/v3+ (Chen et al., 2018a,b). The baseline segmentation models are trained with their default configurations and using the standard augmentation, i.e., random scaling and horizontal flipping.

3.3.2 Evaluation of Masked Noise Encoder

Reconstruction quality. Table 3.1 shows that our masked noise encoder considerably outperforms two strong StyleGAN2 inversion baselines pSp (Richardson et al., 2021) and Feature-Style encoder (Yao et al., 2022) in all three evaluation metrics. The achieved low values of MSE, LPIPS (Zhang et al., 2018b) and FID (Heusel et al., 2017) indicate its high-quality reconstruction. Both the masked noise encoder and the Feature-Style encoder adopt the adversarial loss $\mathcal{L}_{adv}$ and regularization using synthesized images with ground truth latent codes w_{gt}. Therefore, we also add them to train pSp and note this version as pSp†. While pSp† improves over pSp in MSE and FID, it still underperforms compared to the others. This confirms that inverting into the extended latent space $\mathcal{W}^{+}$ only allows limited reconstruction quality on Cityscapes. The Feature-Style encoder (Yao et al., 2022) replaces the prediction of the low level latent codes with feature prediction, which results in better reconstruction without severely harming style editability. However, its reconstruction on Cityscapes is still not satisfying and underperforms to our masked noise encoder. As noted in Yao et al. (2022), the feature size of the Feature-Style encoder is restricted. Using a larger feature map to improve reconstruction quality can only be done as a replacement of more latent code predictions. Consequently, it largely reduces the expressiveness of the latent embedding and leads to extremely poor editability, being no longer suitable for downstream applications, e.g., style mixing

Method	MSE ↓	LPIPS ↓	FID ↓
pSp (Richardson et al., 2021)	0.078	0.348	130.62
pSp† (Richardson et al., 2021)	0.049	0.339	14.60
Feature-Style (Yao et al., 2022)	0.025	0.220	7.14
Ours	**0.011**	**0.124**	**3.94**

Table 3.1: Reconstruction quality on Cityscapes at the resolution 128×256. MSE, LPIPS (Zhang et al., 2018b) and FID (Heusel et al., 2017) respectively measure the pixel-wise reconstruction difference, perceptual difference, and distribution difference between the real and reconstructed images. The proposed masked noise encoder (Ours) consistently outperforms pSp, pSp† and the feature-style encoder. Note, pSp† is introduced by us, by training pSp with an additional discriminator and incorporating synthesized images for better initialization.

data augmentation.

The visual comparison across pSp†, the Feature-Style encoder and our masked noise encoder is shown in Fig. 3.2 and is aligned with the quantitative results in Table 3.1. pSp† has overall poor reconstruction quality. The Feature-Style encoder cannot faithfully reconstruct small objects and restore fine details. In comparison, our masked noise encoder offers high-quality reconstruction, preserving the semantic layout and fine details of each class. Having a high-quality reconstruction is an important requirement for using the encoder for data augmentation. Unfortunately, neither pSp† nor the Feature-Style encoder achieve satisfactory reconstruction quality. For instance, they both fail at capturing the red traffic light in Fig. 3.2. Using such images for data augmentation can confuse the semantic segmentation model, leading to performance degradation.

Ablation on the masking effect. In Fig. 3.4 and Fig. 3.7, we visually observe that random masking offers a stronger perceivable style mixing effect compared to the model trained without masking. Next, we test the effect of masking on improving the domain generalization for the semantic segmentation task. In particular, we employ the encoder that is trained with and without masking to perform ISSA. In Table 3.2, while slightly degrading the source domain performance of the baseline model on Cityscapes, ISSA improves the domain generalization performance on BDD100K, ACDC and Dark Zürich. As ISSA with masked noise encoder is more

Method	Cityscapes	ACDC	BDD	Dark Zürich
Baseline	**70.47**	41.48	45.66	15.25
ISSA w/o masking	69.68	44.63	46.45	17.36
ISSA w/- masking	69.48	**47.43**	**47.87**	**26.10**

Table 3.2: The effect of random noise masking on improving domain generalization via ISSA. We report the mean Intersection over Union (mIoU) of HRNet (Wang et al., 2021b) trained on Cityscapes at the resolution 256×512. BDD100K (BDD), ACDC, and Dark Zürich (DarkZ) represent different domain shifts from Cityscapes.

Patch size	Ratio	MSE ↓	LPIPS ↓	FID ↓
2	25%	0.005	0.090	1.50
	50%	0.008	0.127	2.02
4	25%	**0.004**	**0.089**	**1.41**
	50%	0.009	0.129	2.01

Table 3.3: Ablation on the mask patch size and masking ratio. The influence of patch size on the reconstruction is minor, while masking ratio is more important, i.e., higher masking ratio has negative impact.

effective at diversifying the training set and reducing the style-content correlation, it achieves more pronounced gains in Table 3.2, e.g., more than 10% improvement in mIoU from Cityscapes to Dark Zürich.

Ablation on masking hyperparameters. We conduct an ablation study on the mask patch size P and masking ratio ρ, shown in Table 3.3. We observe that the mask patch size is a relatively insensitive hyperparameter, while higher masking ratio results in noticeable degradation on the reconstruction quality. Empirically, the patch size $P = 4$ with a masking ratio $\rho = 25\%$ achieves the best reconstruction performance. Therefore, we use the encoder trained with this parameter combination for our data augmentation ISSA.

Ablation on the noise map resolution. We investigate the effect of noise map size and experimentally observed that the reconstruction quality benefits the most from using the noise map at the intermediate feature space with one fourth of the input resolution. As shown in Table 3.4, using 32×64 noise, i.e., one fourth of the

Noise scale	MSE ↓	LPIPS ↓	FID ↓
$4 \times 8 \sim 8 \times 16$	0.041	0.317	14.90
32×64	0.008	0.101	2.30

Table 3.4: Effect of noise map resolution on reconstruction quality. Experiments are done on Cityscapes, 128×256 resolution.

image resolution, achieves better reconstruction quality than using lower resolution noise maps. Higher resolution noise map, e.g., full image resolution, in contrast, can be too expressive and encode nearly all perceivable details. This results in worse style mixing capability, as shown in Fig. 3.8. Therefore, we employ the intermediate noise map at one fourth of the input resolution in all of our experiments.

3.3.3 ISSA for Domain Generalization

Comparison with data augmentation methods. Table 3.5 reports the mIoU scores of Cityscapes to ACDC domain generalization using two semantic segmentation models, i.e., HRNet (Wang et al., 2021b) and SegFormer (Xie et al., 2021). Qualitative visualization is illustrated in Fig. 3.9. ISSA is compared with three representative data augmentations methods, namely, CutMix (Yun et al., 2019), Hendrycks's weather and digital corruptions (Hendrycks and Dietterich, 2018), and StyleMix (Hong et al., 2021). Remarkably, our ISSA is the top performing method, consistently improving mIoU in both models and across all four different scenarios of ACDC, i.e., rain, fog, snow and night. Compared to HRNet, SegFormer is more robust against the considered domain shifts.

In contrast to the others, CutMix mixes up the content rather than the style. It improves the in-distribution performance on Cityscapes, but this gain does not extend to domain generalization. Hendrycks's weather corruptions can be seen as the synthetic version of Cityscapes under the rain, fog, and snow weather conditions. While already mimicking ACDC at training, it can still degrade ACDC-Snow by more than 5.8% in mIoU using HRNet. Among the four Hendrycks' corruption types (i.e., noise, blur, digital and weather), Hendrycks-Digital, consisting of contrast, elastics transformation, pixelation and JPEG, is the best-performing one, but

Method	HRNet (Wang et al., 2021b) CS	Rain	Fog	Snow	Night	Avg.	SegFormer (Xie et al., 2021) CS	Rain	Fog	Snow	Night	Avg.
Baseline	70.47	44.15	58.68	44.20	18.90	41.48	67.90	50.22	60.52	48.86	28.56	47.04
ColorTransform	69.90	49.35	65.14	52.63	26.56	48.42	68.50	51.58	66.45	52.87	30.33	50.31
CutMix (Yun et al., 2019)	**72.68**	42.48	58.63	44.50	17.07	40.67	**69.23**	49.53	61.58	47.42	27.77	46.57
Hendrycks-Weather	69.25	**50.78**	60.82	38.34	22.82	43.19	67.41	54.02	64.74	49.57	28.50	49.21
Hendrycks-Digital	69.13	50.13	65.71	49.22	24.81	47.47	67.57	55.53	66.46	49.92	30.33	50.56
FDA (Yang and Soatto, 2020)	70.43	49.68	65.19	50.65	26.41	47.98	67.92	51.28	67.03	51.30	28.28	49.47
StyleMix (Hong et al., 2021)	57.40	40.59	49.11	39.14	19.34	37.04	65.30	53.54	63.86	49.98	28.93	49.08
ISSA (Ours)	70.30	50.62	**66.09**	**53.30**	**30.18**	**50.05**	67.52	**55.91**	**67.46**	**53.19**	**33.23**	**52.45**
Oracle	70.29	65.67	75.22	72.34	50.39	65.90	68.24	63.67	74.10	67.97	48.79	63.56

Table 3.5: Comparison of data augmentation for improving domain generalization, i.e., from Cityscapes (train) to ACDC (unseen). The mean Intersection over Union (mIoU) is reported on Cityscapes (CS), four individual scenarios of ACDC (Rain, Fog, Snow and Night) and the whole ACDC (Avg.). ColorTransform consists of various color transformations such as altering the contrast, brightness, saturation; luma flip and hue rotation. Hendrycks-Weather (Hendrycks and Dietterich, 2018) simulates weather conditions in a synthetic manner for data augmentation, and Hendrycks-Digital is composed of contrast, elastics transformation, pixelation and JPEG corruption. Oracle indicates the supervised training on both Cityscapes and ACDC, serving as an upper bound on ACDC for the other methods. Note, it is not supposed to be an upper bound on Cityscapes. Underline denotes worse results than the baseline on ACDC. ISSA performs the best and consistently improves the mIoU in all four scenarios of ACDC using both HRNet and SegFormer.

still underperforms ISSA. StyleMix (Hong et al., 2021) also seeks to mix up styles. However, it does not work well for scene-centric datasets, such as Cityscapes. Its poor synthetic image quality (see Fig. 3.10) leads to the performance drop over the HRNet baseline in many cases, e.g., on Cityscapes to ACDC-Fog from 58.68% to 49.11% mIoU.

More evaluation on the generalization performance from Cityscapes to BDD100K and Dark Zürich is provided in Table 3.6, where the observation is consistent with Table 3.5 explained above. In addition to weather changes, we further compare different data augmentation methods under the more challenging day-to-night setting in Table 3.7. ISSA present consistent advantages over competing methods, which again justifies the effectiveness of ISSA on improving generalization performance.

Comparison with domain generalization techniques. We further compare ISSA with two advanced feature space style mixing methods designed to improve

Method	HRNet (Wang et al., 2021b)				SegFormer (Xie et al., 2021)			
	CS	ACDC	BDD	DarkZ	CS	ACDC	BDD	DarkZ
Baseline	70.47	41.48	45.66	15.50	67.90	47.04	49.35	24.20
ColorTransform	69.90	48.42	50.22	24.13	68.50	50.31	51.09	25.04
CutMix (Yun et al., 2019)	**72.68**	40.67	45.57	15.34	**69.23**	46.57	48.93	22.98
Hendrycks-Weather	69.25	43.19	44.53	18.71	67.41	49.21	49.84	23.44
Hendrycks-Digital	69.13	47.47	47.60	22.32	67.57	50.56	51.11	25.11
FDA (Yang and Soatto, 2020)	70.43	47.98	48.74	22.46	67.92	49.47	50.47	22.45
StyleMix (Hong et al., 2021)	57.40	37.04	39.30	15.85	65.30	49.08	50.49	23.50
ISSA (Ours)	70.30	**50.05**	**50.29**	**27.24**	67.52	**52.45**	**51.92**	**27.39**

Table 3.6: Comparison of data augmentation for improving domain generalization, i.e., from Cityscapes (train) to ACDC, BDD100K and Dark Zürich (unseen). ISSA consistently outperforms the other data augmentation techniques across different datasets and network architectures, which is consistent with the Table 3.5.

domain generalization performance: MixStyle (Zhou et al., 2021) and DSU (Li et al., 2022c). Both extract the style information at certain normalization layers of CNNs. MixStyle (Zhou et al., 2021) mixes up styles by linearly interpolating the feature statistics, i.e., mean and variance, of different images, while DSU (Li et al., 2022c) models the feature statistics as a distribution and randomly draws samples from it.

We adopt the experimental setting of DSU with default hyperparameters, using DeepLab v2 (Chen et al., 2018a) segmentation network with ResNet101 backbone. Table 3.8 shows that ISSA outperforms both MixStyle and DSU by a large margin. We also observe that there is a slight performance drop on the source domain (i.e., CS) when applying DSU and MixStyle. As they operate at the feature-level, there is no guarantee that the semantic content stays unchanged after the random perturbation of feature statistics. Thus, the changes in feature statistics might negatively affect the performance, as also indicated in Li et al. (2022c). Note that, in contrast, ISSA operates on the image space. Combining ISSA with MixStyle and DSU leads to a strong boost in performance of these methods.

Being model-agnostic, ISSA can be combined with other networks designed specifically for the domain generalization of semantic segmentation. To showcase its complementary nature, we add ISSA on top of two state-of-the-art domain generalization methods for semantic segmentation, RobustNet (Choi et al., 2021) and

Method	BDD100K	ACDC-Night	Dark Zürich
Baseline	52.97	23.52	23.63
CutMix	**54.03**	24.37	23.99
Weather	52.10	23.79	24.21
Digital	52.10	24.17	23.24
StyleMix	46.33	19.13	19.27
ISSA(Ours)	53.37	**25.93**	**26.55**

Table 3.7: Comparison of data augmentation techniques for improving domain generalization using HRNet (Wang et al., 2021b), i.e., from BDD100K-Daytime to ACDC-Night and Dark Zürich. BDD100K-Daytime is a subset of BDD100K, which contains 2526 images in daytime under various weather conditions, but not in dawn/nighttime. Here, we evaluate the domain generalization with respect to day to night.

Method	Cityscapes	ACDC	BDD100K	Dark Zürich
Baseline (Chen et al., 2018a)	61.73	30.86	34.30	11.62
MixStyle (Zhou et al., 2021)	59.01	36.97	36.27	9.38
DSU (Li et al., 2022c)	59.59	38.31	35.53	12.29
ISSA (Ours)	**62.20**	**43.21**	**42.60**	**21.56**
MixStyle + ISSA	60.17	41.81	42.17	20.56
DSU + ISSA	60.20	43.31	42.24	24.63

Table 3.8: Comparison with feature-level augmentation methods on domain generalization performance of Cityscapes as the source. Following DSU (Li et al., 2022c), we conduct experiments using DeepLab v2 (Chen et al., 2018a) as the baseline for fair comparison.

SHADE (Zhao et al., 2022). RobustNet proposed a novel instance whitening loss to selectively remove domain-specific style information. SHADE on the other hand aims to learn style-invariant representation and preserve knowledge from the pre-trained backbone. Although color transformation has already been used for augmentation in both methods and SHADE additionally employs feature-level style augmentation, ISSA can introduce more natural style shifts, thus is able to bring further improvements. Table 3.9 verifies the effectiveness of ISSA, which brings extra gains for RobustNet and SHADE. For RobustNet, the performance of the challenging day to night scenario, i.e., Cityscapes to Dark Zürich is boosted from 20.11%

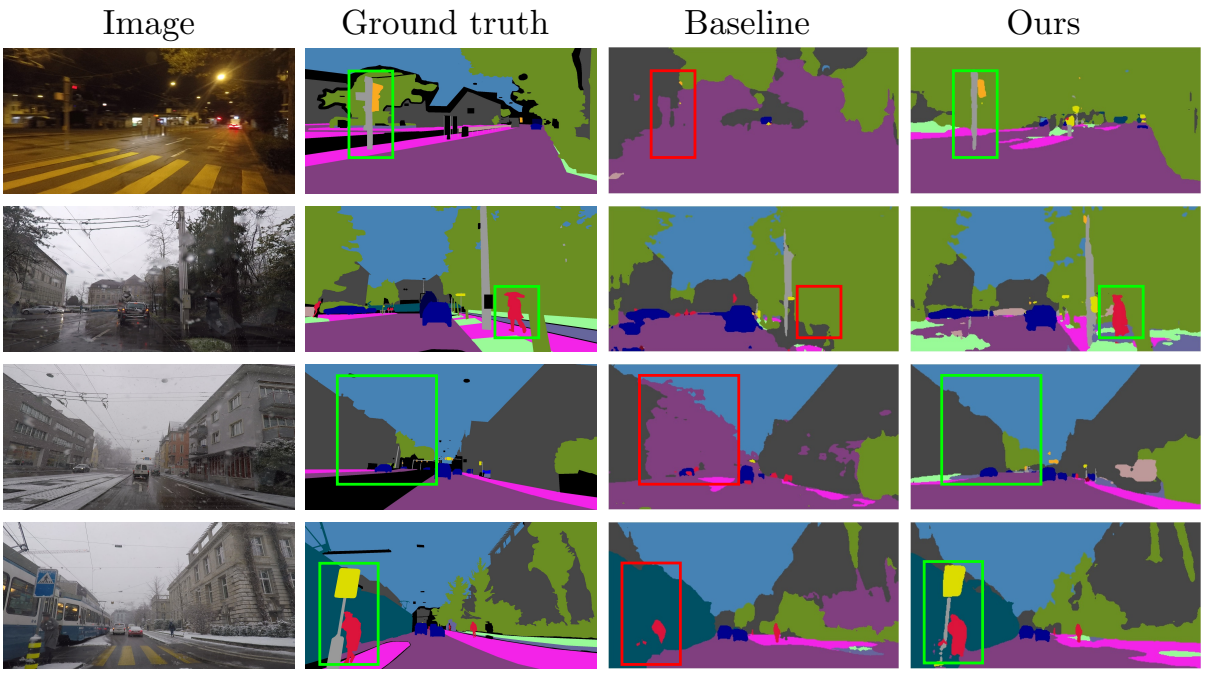


Figure 3.9: Semantic segmentation results of Cityscapes to ACDC generalization using HRNet. The HRNet is trained on Cityscapes only. The segmenter trained with ISSA provides more reasonable prediction under adverse weather conditions.

Figure 3.10: Comparison of StyleMix (Hong et al., 2021) and ISSA. StyleMix has rather low fidelity, while ISSA can preserve more details.

Method	Cityscapes	ACDC	BDD100K	Dark Zürich
Baseline (Chen et al., 2018b)	69.01	44.23	43.27	16.03
RobustNet (Choi et al., 2021)	**69.47**	47.25	46.94	20.11
+ ISSA	69.45	**47.55**	**48.44**	**23.09**
SHADE (Zhao et al., 2022)	**64.24**	47.30	46.44	25.37
+ ISSA	63.79	**47.64**	**47.76**	**25.58**

Table 3.9: Combination of ISSA and RobustNet (Choi et al., 2021). We adopt the experimental setting of RobustNet and use DeepLab v3+ (Chen et al., 2018b) as the baseline. Our ISSA is complementary to RobustNet and further improves its generalization performance.

to 23.09% in mIoU.

Comparison with unsupervised domain adaptation methods. We compare our method with multiple unsupervised domain adaptation (UDA) techniques, which not only have access to the source domain, but also use extra unlabeled samples of the target domain. The quantitative comparison of Cityscapes to ACDC adaptation/generalization is shown in Table 3.10. Our method has presented competitive performance, even without using images from the target domain.

3.3.4 Plug-n-Play Ability of the Exemplar-Based Style Synthesis Pipeline

In Section 3.3.3, we have focused on ISSA for improved domain generalization. Next, we investigate the plug-n-play ability of our exemplar-based style pipeline, which enables ESSA. Specifically, the generator and masked noise encoder which are trained on one dataset can be directly used for mixing styles from other datasets, thus avoiding retraining or fine-tuning the models. This ability is valuable in two perspectives: 1) harnessing external data for improved domain generalization via ESSA; and 2) saving computationally complexity. Compared to other data augmentation techniques, e.g., CutMix (Yun et al., 2019), Hendrycks corruption (Hendrycks and Dietterich, 2018), our style synthesis requires training GAN and an encoder, which could take considerable computational resources. Therefore, it is of practical interest if the trained models can be readily useable for novel domains.

Method	Network	Use Target	mIoU
Baseline		—	30.9
BDL (Li et al., 2019)	DeepLabv2	✓	32.7
CRST (Zou et al., 2019)		✓	32.8
AdaptSegNet (Tsai et al., 2018)		✓	33.4
SIM (Wang et al., 2020)		✓	34.6
MRNet (Zheng and Yang, 2021)		✓	36.1
ADVENT (Tsai et al., 2019)		✓	37.7
CLAN (Luo et al., 2019)		✓	39.0
FDA (Yang and Soatto, 2020)		✓	45.7
ISSA(Ours)		✗	43.2
DAFormer (Hoyer et al., 2022)	DAFormer	✓	55.4
ISSA(Ours)	SegFormer	✗	52.5

Table 3.10: Comparison with UDA methods on Cityscapes to ACDC generalization. Remarkably, our domain generalization method (without access to the target domain, neither images nor labels), is on-par or better than unsupervised domain adaptation (UDA) methods, which requires knowledge of the target domain during training. Results of UDA methods are from (Sakaridis et al., 2021).

ISSA using arbitrary encoders. Favorably, thanks to the plug-n-play ability of the synthesis pipeline, we observe that ISSA can still be effective even when encoder and generator are trained on a different dataset of a similar task, and re-training is not required. Note that here the source is with respect to the segmenter training for domain generalization, not the encoder training. As shown in Table 3.11, when training the segmenter on Cityscapes using ISSA, we can directly use generator and encoder trained on BDD100K without fine-tuning. Even though the models have not seen any samples of Cityscapes, they can still reconstruct and augment styles within Cityscapes, and the effectiveness of ISSA is not compromised. This implies that, once the generator and encoder are trained on one dataset, they are also straightforwardly applicable for augmenting novel datasets.

Extra-source exemplar based style synthesis. Furthermore, we exploit the usage of extra-source data as the style exemplar. Visual examples in Fig. 3.11 showcase the plug-n-play style-mixing ability of our encoder on web-crawled images,

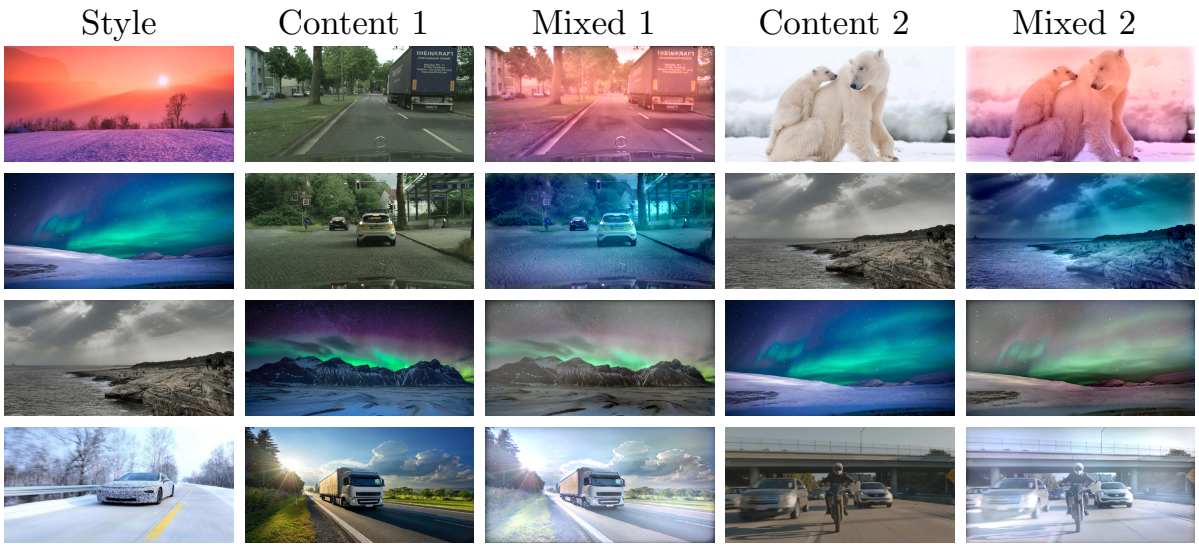


Figure 3.11: Extra-source exemplar based style synthesis using web-crawled images, where the generator and encoder are only trained on Cityscapes. Except for the Content 1 image of the first 2 rows, all the others are web-crawled images.

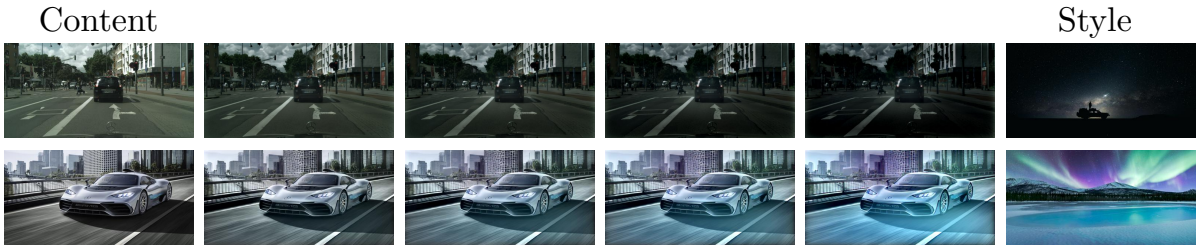


Figure 3.12: Visualization of interpolation in the style latent space. As illustrated, we can control the style mixing strength and achieve a smooth transition on both trained Cityscapes and unseen web-crawled images.

where the model is only trained on Cityscapes. It can be observed that the style of unseen images can still be successfully transferred to the content images, which grants us the opportunity to further utilize images on the web to enhance the effectiveness of style augmentation beyond intra-source styles. Also, we illustrate the interpolation capability in the style latent space on both trained Cityscapes and unseen web-crawled image. This property enables more control on the style mixing strength.

To further explore the usage of images on the web, we take Landscape Pictures[1]

[1] https://www.kaggle.com/datasets/arnaud58/landscape-pictures?resource=

Figure 3.13: Visual examples of stylized data by transferring style from one unannotated ACDC sample (target domain) to Cityscapes (source domain). Best view in color.

Method	Cityscapes	Rain	Fog	Snow	Night	Avg.
Baseline	**70.5**	44.2	58.7	44.2	18.9	41.5
ISSA: CS-G-E	70.3	50.6	66.1	**53.3**	30.2	50.1
ISSA: BDD-G-E	70.3	**52.2**	**66.3**	52.2	**31.0**	**50.4**

Table 3.11: Comparison on Cityscapes to ACDC generalization using ISSA with generator and encoder trained on Cityscapes (CS-G-E) and BDD100K (BDD-G-E), respectively. Despite never seeing Cityscapes samples, ISSA with BDD-G-E is still highly effective.

dataset as the extra-source exemplars for style augmentation. Table 3.12 justifies that by exploiting additional image styles, ESSA can further improve the generalization performance of ISSA on unseen target domains.

3.3.5 Stylized Proxy Validation Set Synthesis

Beyond the usage of data augmentation for network training, we further explore if our exemplar-based style synthesis pipeline can be used to assess the generalization capability of semantic segmentation models for both source and target domain

download

Method	Cityscapes	ACDC	BDD100K	Dark Zürich
Baseline	70.47	41.48	45.66	15.50
ISSA: CS-G-E	70.30	50.05	50.29	27.24
ESSA: CS-G-E	69.85	**50.87**	**51.42**	**29.06**

Table 3.12: Utilizing Landscape Pictures as extra-source exemplars for style augmentation, where the generator and encoder are only trained on Cityscapes (CS-G-E). ESSA can further improve the generalization performance from Cityscapes to other unseen datasets.

without extra data annotation effort. Prior work (Zhang et al., 2021b) has used conditional GAN synthesized samples to predict generalization performance of image classifiers in the source domain. However, it remains unclear how to evaluate the generalization performance on unseen domains, and apply it on dense prediction tasks. Given the fact that our masked noise encoder can transfer styles even from novel domains, we utilize this attractive property to generate a stylized proxy validation set, i.e., combining styles from the target domain with the contents from the source domain training samples. For getting their styles, exemplars from the target domain do not need to be labelled. The existing ground-truth label maps of the training samples in the source domain are reused as the ground-truth annotations of the stylized proxy validation set. Visual examples of transferring ACDC style using one sample from each weather condition are provided in Fig. 3.13.

Experimental Setup. We investigate the generalization performance of 95 semantic segmentation models trained on Cityscapes, where 54 models are obtained from MMSegmentation (Contributors, 2020) model zoo and the others are trained by ourselves. The models cover both CNN-based architectures, e.g., HRNet (Wang et al., 2021b), DeepLab (Chen et al., 2017), DANet (Fu et al., 2019), and transformer-based model, e.g., SegFormer (Xie et al., 2021), SETR (Zheng et al., 2021). Besides, the models are trained using different strategies, e.g., various learning rate schedule, cropping size and data augmentation. We consider generalization performance on both source and target domain for the correlation study. Specifically, we use the Cityscapes validation set as the source test set, ACDC and BDD100K validation sets as the target test data. To verify the generalization performance on the source domain, we apply intra-source style augmentation on the Cityscapes train-

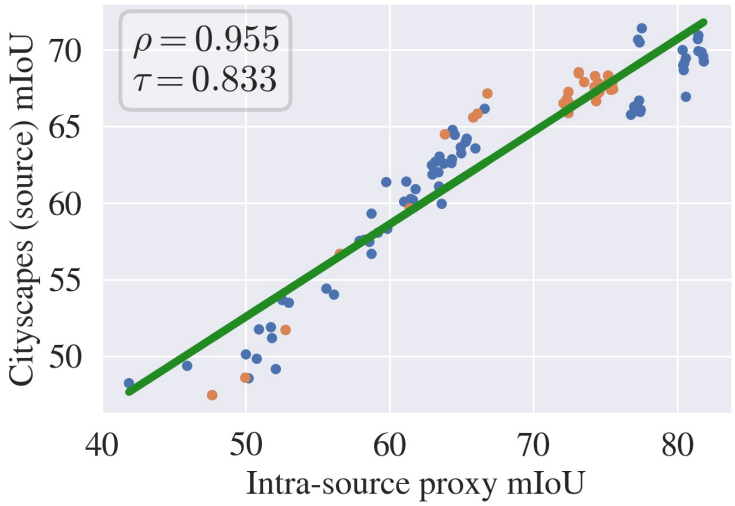


Figure 3.14: Correlation between real Cityscapes test performance and intra-source style augmented proxy performance for 95 models. Spearman's Rank Correlation coefficient (ρ) and Kendall Rank Correlation Coefficient (τ) are computed to quantitatively measure correlation strength. Blue and orange dots represent CNN- and transformer-based backbones, respectively. We observe that there is a strong correlation between the real test mIoU and proxy mIoU.

ing set and use it as the proxy validation set. For the verification of target domain generalization performance, we build a proxy set by transferring styles from the corresponding target test dataset. Further, we study the correlation between the real test performance and performance on the proxy data.

Correlation Metrics. We compute Spearman's Rank Correlation coefficient (ρ) and Kendall Rank Correlation Coefficient (τ) to quantitatively measure the correlation strength. The value of the correlation coefficient varies from $[-1, 1]$. A value closer to ± 1 indicates strong positive/negative association between the two variables. As the coefficient goes towards 0, the association becomes looser. Both correlation coefficients are non-parametric, i.e., no strict assumptions on the data distribution, and the assessment is based on the ranking of the data.

Observations. In Fig. 3.14, we show the correlation of performance on the intra-source style augmented proxy set and real Cityscapes test set across different network architectures. We clearly observe a strong correlation ($\rho > 0.95$), indicating that ISSA proxy set can serve as a good indicator for generalization in the source

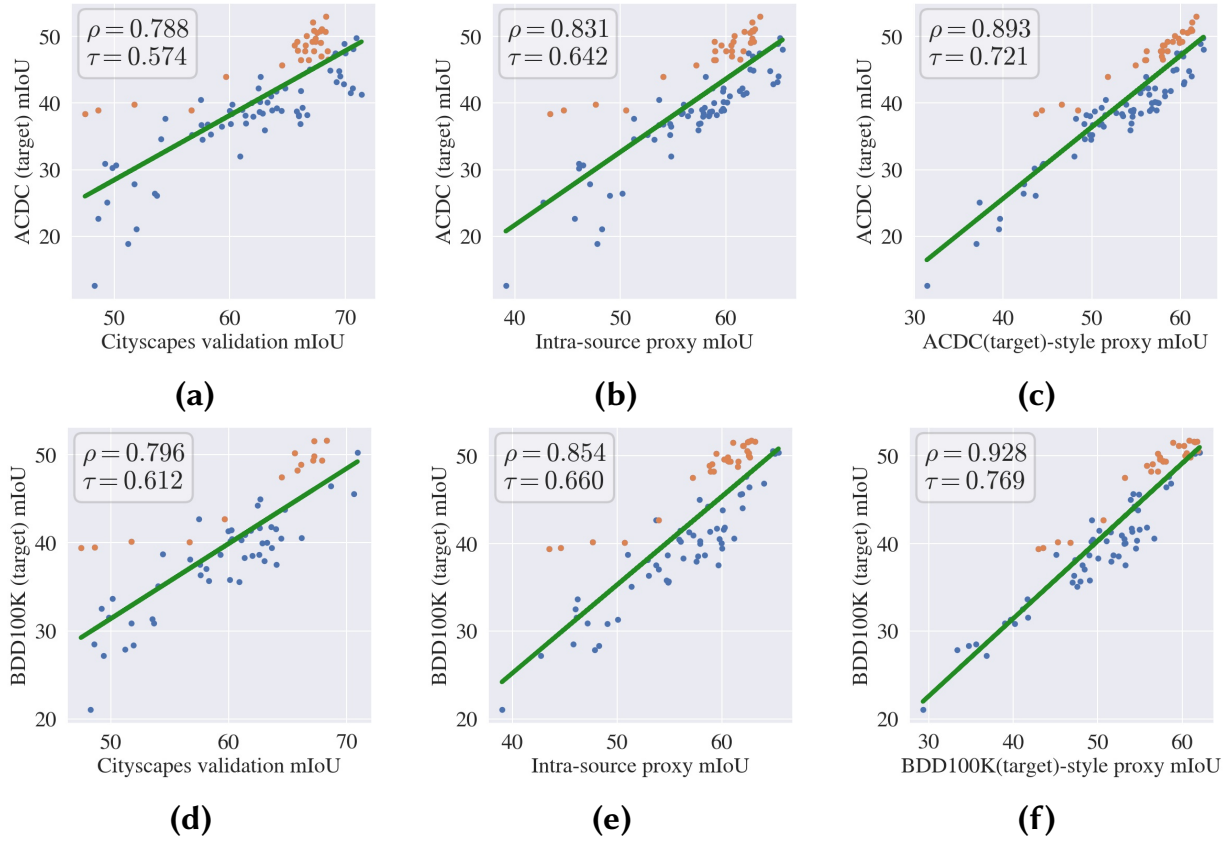


Figure 3.15: Correlation between test performance and proxy performance for 95 models. We compute Spearman's Rank Correlation coefficient (ρ) and Kendall Rank Correlation Coefficient (τ) to quantitatively measure correlation strength. Blue and orange dots represent CNN- and transformer-based backbones, respectively. In each row, we investigate the correlation between the real test performance, i.e., mIoU of ACDC and BDD100K, and mIoU of different proxy sets. We observe that Figs. 3.15(c) and 3.15(f) achieve the strongest correlation for each scenario, which indicates that it is beneficial to build a proper proxy set using styles of the corresponding test dataset.

domain.

Furthermore, we report the correlation results of target domain generalization on two datasets, i.e., ACDC and BDD100K in each row of Fig. 3.15. We compare three different choices of the proxy set in each column, namely the original Cityscapes validation set, intra-source style augmented Cityscapes validation set and target data style augmented validation set. Blue and orange dots represent CNN- and transformer-based backbones, respectively. Quantitatively, the correlation coefficients of Figs. 3.15(a) and 3.15(d) are rather low. Also from Fig. 3.15(a),

some blue points in the upper right corner has stronger performance on Cityscapes validation set compared to the orange points, but worse on ACDC test data. This suggests that evaluation of the original Cityscapes (source) validation set cannot properly reflect the generalization performance on the target domain. Therefore, this raises the concern that by following the traditional way, selecting the best model based on the source validation performance could be problematic when the deploying environment involves data of unknown target domains. By applying intra-source style augmentation on the Cityscapes validation set, the correlation coefficient has been improved (see Figs. 3.15(b) and 3.15(e)). We hypothesize that style mixing results in better data coverage and thus can better represent model's generalization ability under style shifts. Furthermore, whenever it is possible to have access to images of the target domain, even though without annotation, we can utilize styles of the unlabeled target data and achieve the strongest correlation in Figs. 3.15(c) and 3.15(f). In addition to the correlation metric, in general models have higher mIoU on the Cityscapes validation set, compared with the intra-source style and target domain style augmented proxy set. And the mIoU range on the intra-source proxy set is closer to the one of using target styles, which also justifies our hypothesis above.

Additionally, we also observe an interesting phenomenon from Fig. 3.15: all transformer-based models (orange dots) are above the linear fit. This suggests that transformer-based models present better generalization ability under natural shifts compared with CNN-based models (blue dots). This is consistent with the acknowledgement on transformers from prior works (Naseer et al., 2021, Bai et al., 2021, Zhang et al., 2022).

To sum up, we present a new use case of proposed exemplar-based style synthesis pipeline, and demonstrate that stylized samples can be used as a proxy validation set and a strong indicator for model's generalization capability without introducing additional annotation efforts. Based on this observation, we can better utilize existing annotated data together with our exemplar-based style synthesis pipeline, to select models in practice especially when deployment in an open-world environment, where unknown target data commonly exists.

3.4 Conclusion

In this paper, we propose a GAN inversion based style synthesis pipeline for domain generalization in semantic segmentation. The key enabler for our pipeline is the masked noise encoder, which is capable of preserving fine-grained content details and allows style mixing between images without affecting the semantic content. In particular, we employ intra-source style augmentation (ISSA) for learning domain generalized semantic segmentation using restricted training data from a single source domain. Extensive experimental results verify the effectiveness of ISSA on domain generalization across different datasets and network architectures. We further demonstrate the plug-n-play ability of the proposed pipeline. Without requiring retraining the encoder and generator, our model can be used directly on extra-source exemplars such as web-crawled images, enabling extra-source style augmentation (ESSA). It also opens up applications beyond data augmentation for improved domain generalization. Specifically, we show that the intra- & extra-source exemplar-based style synthesis pipeline can be used for creating proxy validation sets to compare the generalization capability of diverse models on both the source and target domain without extra data annotation effort.

Limitation and future work. One limitation of ISSA is that our style mixing is a global transformation, which cannot specifically alter the style of local objects, e.g., adjusting vehicle color from red to black, though when changing the image globally, local areas are inevitably modified. Also compared to simple data augmentation such as color transformation, our pipeline requires higher computational complexity for training. It takes around 7 days to train the masked noise encoder on 256×512 resolution using 2 GPUs. A similar amount of time is required for the StyleGAN2 training. Nonetheless, for data augmentation, it only concerns the inference time of our encoder, which is much faster, i.e., 0.1 seconds, compared to optimization based methods such as PTI (Roich et al., 2022) that takes 55.7 seconds per image.

In the future, it is challenging yet interesting to extend our work with more flexible local editing. Our proposed intra- & extra-source exemplar-based style synthesis is a global transformation, which cannot specifically alter the style of local objects, e.g., adjusting vehicle color from red to black, though when changing the

image globally, local areas are inevitably modified. One potential direction is by exploiting the pre-trained language-vision model, such as CLIP (Radford et al., 2021). We can synthesize styles conditioned on text rather than an image. For instance, by providing a text condition “snowy road", ideally we would want to obtain an image where there is snow on the road and other semantic classes remain unchanged. Recent works (Bar-Tal et al., 2022, Hertz et al., 2023, Kawar et al., 2023) studied local editing conditioned on text. However, CLIP exhibits a strong bias (Bar-Tal et al., 2022) and may generate undesirable results, and the edited image may suffer from insufficient alignment with the other parts of the image. Overall, there is still large room for improvement on synthesizing images with more controls on both style and content. In Chapter 4, we will work on layout-to-image diffusion models, where additional layout information, i.e., semantic label map is leveraged and the image style can be controlled via text prompt.

4 Improved Layout-to-Image Diffusion Models Via Adversarial Supervision

In this chapter, we focus on layout-to-image (L2I) diffusion models. Despite the recent advances in large-scale diffusion models, little progress has been made on the layout-to-image synthesis task. Current L2I models either suffer from poor editability via text or weak alignment between the generated image and the input layout. This limits their usability in practice. To mitigate this, we propose

to integrate **a**dversarial supervision into the conventional training pipeline of **L**2I **d**iffusion **m**odels (ALDM). Specifically, we employ a segmentation-based discriminator which provides explicit feedback to the diffusion generator on the pixel-level alignment between the denoised image and the input layout. To encourage consistent adherence to the input layout over the sampling steps, we further introduce the multistep unrolling strategy. Instead of looking at a single timestep, we unroll a few steps recursively to imitate the inference process, and ask the discriminator to assess the alignment of denoised images with the layout over a certain time window. Our experiments show that ALDM enables layout faithfulness of the generated images, while allowing broad editability via text prompts. Moreover, we showcase its usefulness for practical applications: by synthesizing target distribution samples via text control, we improve domain generalization of semantic segmentation models by a large margin (~12 mIoU points). The code and model are available at https://github.com/boschresearch/ALDM. This work is published at the International Conference on Learning Representations (ICLR), 2024 (Li et al., 2024a).

4.1 Introduction

Layout-to-image synthesis (L2I) is a challenging task that aims to generate images with per-pixel correspondence to the given semantic label maps. Yet, due to the tedious and costly pixel-level layout annotations of images, availability of large-scale labelled data for extensive training on this task is limited. Meanwhile, tremendous progress has been witnessed in the field of large-scale text-to-image (T2I) diffusion models (Ramesh et al., 2022, Balaji et al., 2022, Rombach et al., 2022). By virtue of joint vision-language training on billions of image-text pairs, such as LAION dataset (Schuhmann et al., 2022), these models have demonstrated remarkable capability of synthesizing photorealistic images via text prompts. A natural question is: can we adapt such pretrained diffusion models for the L2I task using a limited amount of labelled layout data while preserving their *text controllability* and *faithful alignment to the layout*? Effectively addressing this question will then foster the widespread utilization of L2I synthetic data.

Recently, increasing attention has been devoted to answer this question (Zhang and Agrawala, 2023, Xue et al., 2023, Mou et al., 2024). Despite the efforts, prior

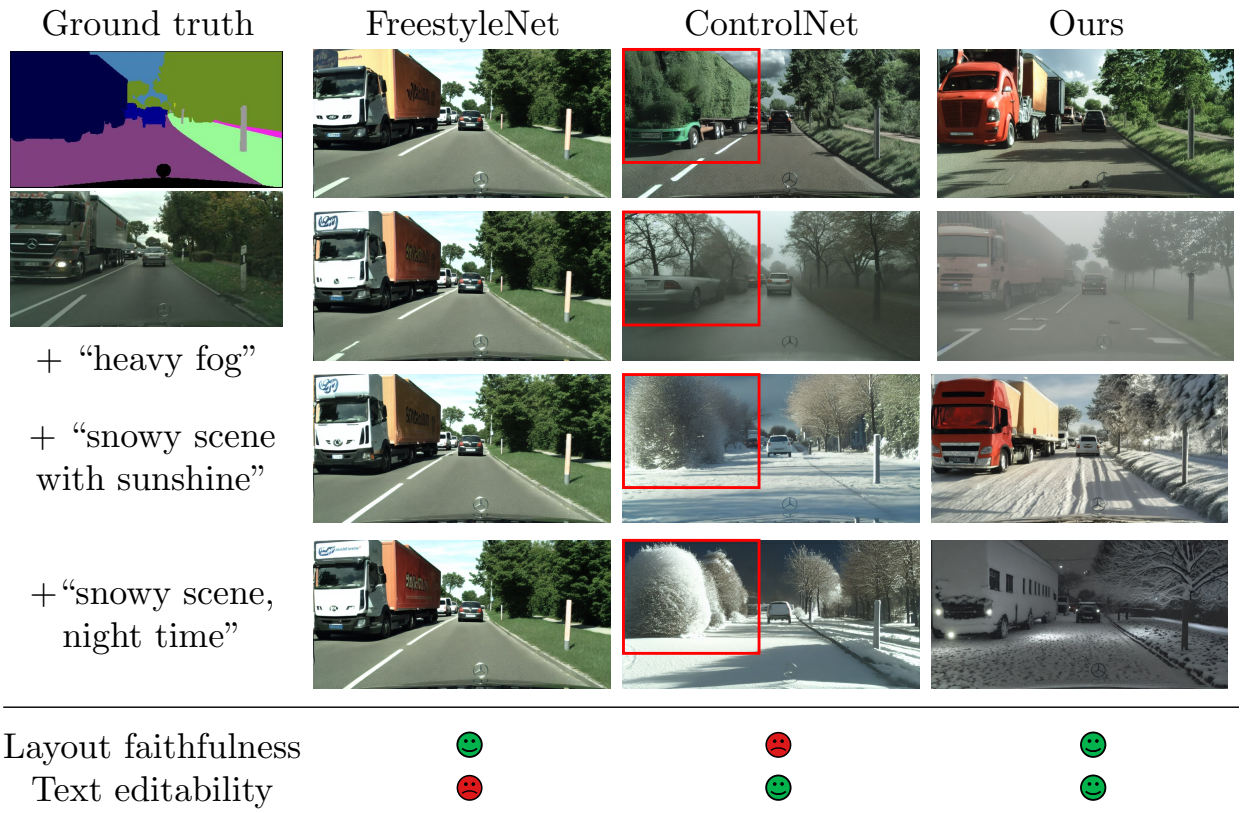


Figure 4.1: In contrast to prior L2I synthesis methods (Xue et al., 2023, Zhang and Agrawala, 2023), our **ALDM** model can synthesize faithful samples that are well aligned with the layout input, while preserving controllability via text prompt. Equipped with these both valuable properties, we can synthesize diverse samples of practical utility for downstream tasks, such as data augmentation for improving domain generalization of semantic segmentation models.

works have suffered to find a good trade-off between faithfulness to the layout condition and editability via text, which we also empirically observed in our experiments (see Fig. 4.1). When adopting powerful pretrained T2I diffusion models, e.g., Stable Diffusion (SD) (Rombach et al., 2022), for L2I tasks, fine-tuning the whole model fully as in Xue et al. (2023) can lead to the loss of text controllability, as the large model easily overfits to the limited amount of training samples with layout annotations. Consequently, the model can only generate samples resembling the training set, thus negatively affecting its practical use for potential downstream tasks requiring diverse data. For example, for downstream models deployed in an open-world, variety in synthetic data augmentation is crucial, since annotated data can only partially capture the real environment and synthetic samples should com-

plement real ones.

Conversely, when freezing the T2I model weights and introducing additional parameters to accommodate the layout information (Zhang and Agrawala, 2023, Mou et al., 2024), the L2I diffusion models naturally preserve text control of the pretrained model but do not reliably comply with the layout conditioning. In such case, the condition becomes a noisy annotation of the synthetic data, undermining its effectiveness for data augmentation. We hypothesize the poor alignment with the layout input can be attributed to the suboptimal MSE loss for the noise prediction, where the layout information is only implicitly utilized during the training process. The assumption is that the denoiser has the incentive to utilize the layout information as it poses prior knowledge of the original image and thus is beneficial for the denoising task. Yet, there is no direct mechanism in place to ensure the layout alignment. To address this issue, we propose to integrate **a**dversarial supervision on the layout alignment into the conventional training pipeline of **L**2I **d**iffusion **m**odels, which we name ALDM. Specifically, inspired by Sushko et al. (2022), we employ a semantic segmentation model based discriminator, explicitly leveraging the layout condition to provide a direct per-pixel feedback to the diffusion model generator on the adherence of the denoised images to the input layout.

Further, to encourage consistent compliance with the given layout over the sampling steps, we propose a novel multistep unrolling strategy. At inference time, the diffusion model needs to consecutively remove noise for multiple steps to produce the desired sample in the end. Hence, the model is required to maintain consistent adherence to the conditional layout over the sampling time horizon. Therefore, instead of applying discriminator supervision at a single timestep, we additionally unroll backward multiple steps over a certain time window to imitate the inference time sampling. This way the adversarial objective is designed over a time horizon and future steps are taken into consideration as well. Enabled by adversarial supervision over multiple sampling steps, our ALDM can effectively ensure consistent layout alignment, while maintaining initial properties of the text controllability of the large-scale pretrained diffusion model. We experimentally show the effectiveness of adversarial supervision for different adaptation strategies (Qiu et al., 2023, Zhang and Agrawala, 2023, Mou et al., 2024) of the SD model (Rombach et al., 2022) to the L2I task across different datasets, achieving the desired balance between lay-

out faithfulness and text editability (see Table 4.1).

Finally, we demonstrate the utility of our method on the domain generalization task, where the semantic segmentation network is evaluated on unseen target domains, whose samples are sufficiently different from the trained source domain. By augmenting the source domain with synthetic images generated by ALDM using text prompts aligned with the target domain, we can significantly enhance the generalization performance of original downstream models, i.e., $\sim$ 12 mIoU points on the Cityscapes-to-ACDC generalization task (see Table 4.6).

In summary, our main contributions include:

- We introduce adversarial supervision into the conventional diffusion model training, improving layout alignment without losing text controllability.
- We propose a novel multistep unrolling strategy for diffusion model training, encouraging better layout coherency during the synthesis process.
- We show the effectiveness of synthetic data augmentation achieved via ALDM. Benefiting from the notable layout faithfulness and text control, our ALDM improves the generalization performance of semantic segmenters by a large margin.

4.2 Method

L2I diffusion model aims to generate images based on the given layout. Its current training and inference procedure is inherited from unconditional diffusion models, where the design focus has been on how the layout as the condition is fed into the UNet for noise estimation, as illustrated in Fig. 4.2 (A). It is yet underexplored how to enforce the faithfulness of L2I image synthesis via direct loss supervision. Here, we propose novel adversarial supervision which is realized via 1) a semantic segmenter-based discriminator (Section 4.2.1 and Fig. 4.2 (B)); and 2) multistep unrolling of UNet (Section 4.2.2 and Fig. 4.2 (C)) to induce faithfulness already from early sampling steps and consistent adherence to the condition over consecutive steps.

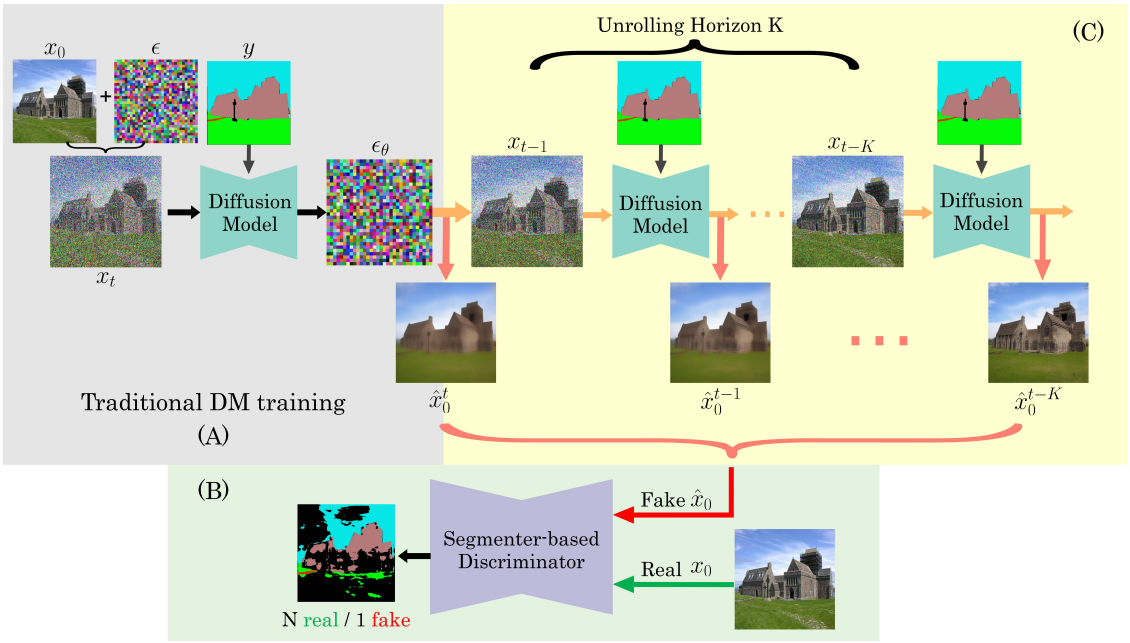


Figure 4.2: Method overview. To enforce faithfulness, we propose two novel training strategies to improve the traditional L2I diffusion model training (area (A)): adversarial supervision via a segmenter-based discriminator illustrated in area (B), and multistep unrolling strategy in area (C).

4.2.1 Discriminator Supervision on Layout Alignment

For training the L2I diffusion model, a Gaussian noise $\epsilon \sim N(0, I)$ is added to the clean variable x_0 with a randomly sampled timestep t, yielding x_t:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon, \tag{4.1}$$

where α_t defines the level of noise. A UNet (Ronneberger et al., 2015) denoiser ϵ_θ is then trained to estimate the added noise via the MSE loss:

$$\mathcal{L}_{noise} = \mathbb{E}_{\epsilon\sim N(0,I),y,t}\left[\|\epsilon - \epsilon_\theta(x_t, y, t)\|^2\right] = \mathbb{E}_{\epsilon,x_0,y,t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, y)\right\|^2\right]. \tag{4.2}$$

Besides the noisy image x_t and the time step t, the UNet additionally takes the layout input y. Since y contains the layout information of x_0 which can simplify the noise estimation, it then influences implicitly the image synthesis via the denoising step. From x_t and the noise prediction ϵ_θ, we can generate a denoised version of the clean

image $\hat{x}_0^{(t)}$ as:

$$\hat{x}_0^{(t)} = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, y, t)}{\sqrt{\alpha_t}}. \tag{4.3}$$

However, due to the lack of explicit supervision on the layout information y for minimizing $\mathcal{L}_{noise}$, the output $\hat{x}_0^{(t)}$ often lacks faithfulness to y, as shown in Fig. 4.3. It is particularly challenging when y carries detailed information about the image, as the alignment with the layout condition needs to be fulfilled on each pixel. Thus, we seek direct supervision on $\hat{x}_0^{(t)}$ to enforce the layout alignment. A straightforward option would be to simply adopt a frozen pre-trained segmenter to provide guidance with respect to the label map. However, we observe that the diffusion model tends to learn a mean mode to meet the requirement of the segmenter, exhibiting little variation (see Table 4.4 and Fig. 4.8).

To encourage diversity in addition to alignment, we make the segmenter trainable along with the UNet training. Inspired by Sushko et al. (2022), we formulate an adversarial game between the UNet and the segmenter. Specifically, the segmenter acts as a discriminator that is trained to classify per-pixel class labels of real images, using the paired ground-truth label maps; while the fake images generated by UNet as in (Eq. (4.3)) are classified by it as one extra "fake" class, as illustrated in area (B) of Fig. 4.2. As the task of the discriminator is essentially to solve a multi-class semantic segmentation problem, its training objective is derived from the standard cross-entropy loss:

$$L_{Dis} = -\mathbb{E}\left[\sum_{c=1}^{N} \gamma_c \sum_{i,j}^{H\times W} y_{i,j,c} \log\left(Dis(x_0)_{i,jc}\right)\right] - \mathbb{E}\left[\sum_{i,j}^{H\times W} \log\left(Dis(\hat{x}_0^{(t)})_{i,j,c=N+1}\right)\right], \tag{4.4}$$

where N is the number of real semantic classes, and $H \times W$ denotes spatial size of the input. The class-dependent weighting γ_c is computed via inverting the per-pixel class frequency

$$\gamma_c = \frac{H \times W}{\sum \mathbb{E}\left[\mathbb{1}\left[y_{i,j,c} = 1\right]\right]}, \tag{4.5}$$

for balancing between frequent and rare classes. To fool such a segmenter-based discriminator, $\hat{x}_0^{(t)}$ produced by the UNet as in (Eq. (4.3)) shall comply with the input layout y to minimize the loss

$$L_{adv} = -\mathbb{E}\left[\sum_{c=1}^{N} \gamma_c \sum_{i,j}^{H\times W} y_{i,j,c} \log\left(Dis(\hat{x}_0^{(t)})_{i,j,c}\right)\right]. \tag{4.6}$$

Such loss poses explicit supervision to the UNet for using the layout information, complementary to the original MSE loss. The total loss for training the UNet is thus

$$L_{DM} = L_{noise} + \lambda_{adv} L_{adv}, \tag{4.7}$$

where λ_{adv} is the weighting factor. The whole adversarial training process is illustrated Fig. 4.2 (B). As the discriminator is improved along with UNet training, we no longer observe the mean mode collapsing as with the use of a frozen semantic segmenter. The high recall reported in Table 4.2 confirms the diversity of synthetic images produced by our method.

4.2.2 Multistep Unrolling

Admittedly, it is impossible for the UNet to produce high-quality image $\hat{x}_0^{(t)}$ via a single denoising step as in (Eq. (4.3)), especially if the input x_t is very noisy (i.e., t is large). On the other hand, adding such adversarial supervision only at low noise inputs (i.e., t is small) is not very effective, as the alignment with the layout should be induced early enough during the sampling process. To improve the effectiveness of the adversarial supervision, we propose a multistep unrolling design for training the UNet. Extending from a single step denoising, we perform multiple denoising steps, which are recursively unrolled from the previous step:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, y, t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(x_t, y, t), \tag{4.8}$$

$$\hat{x}_0^{(t-1)} = \frac{x_{t-1} - \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_{t-1}, y, t-1)}{\sqrt{\alpha_{t-1}}}. \tag{4.9}$$

As illustrated in area (C) of Fig. 4.2, we can repeat (Eq. (4.8)) and (Eq. (4.9)) K times, yielding $\{\hat{x}_0^{(t)}, \hat{x}_0^{(t-1)}, ..., \hat{x}_0^{(t-K)}\}$. All these denoised images are fed into the segmenter-based discriminator as the "fake" examples:

$$L_{adv} = \frac{1}{K+1} \sum_{i=0}^{K} -\mathbb{E}\left[\sum_{c=1}^{N} \gamma_c y_c \log\left(Dis(\hat{x}_0^{(t-i)})_c\right)\right]. \tag{4.10}$$

By doing so, the denoising model is encouraged to follow the conditional label map consistently over the time horizon. It is important to note that while the number of unrolled steps K is pre-specified, the starting time step t is still randomly sampled.

Such unrolling process resembles the inference time denoising with a sliding window of size K. As pointed out by Fan and Lee (2023), diffusion models can be seen as control systems, where the denoising model essentially learns to mimic the ground-truth trajectory of moving from noisy image to clean image. In this regard, the proposed multistep unrolling strategy also resembles the advanced control algorithm - Model Predictive Control (MPC), where the objective function is defined in terms of both present and future system variables within a prediction horizon. Similarly, our multistep unrolling strategy takes future timesteps along with the current timestep into consideration, hence yielding a more comprehensive learning criteria.

While unrolling is a simple feed-forward pass, the challenge lies in the increased computational complexity during training. Apart from the increased training time due to multistep unrolling, the memory and computation cost for training the UNet can be also largely increased along with K. Since the denoising UNet model is the same and reused for every step, we propose to simply accumulate and scale the gradients for updating the model over the time window, instead of storing gradients at every unrolling step. This mechanism permits to harvest the benefit of multistep unrolling with a controllable increase in complexity during training.

4.2.3 Implementation Details

We apply our method to the open-source text-to-image Stable Diffusion (SD) model (Rombach et al., 2022) so that the resulting model not only synthesizes high quality images based on the layout condition, but also accepts text prompts to

change the content and style. As SD belongs to the family of latent diffusion models (LDMs), where the diffusion model is trained in the latent space of an autoencoder, the UNet denoises the corrupted latents which are further passed through the SD decoder for the final pixel space output, i.e., $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$. We employ UperNet (Xiao et al., 2018) as the discriminator, nonetheless, we also ablate other types of backbones in Table 4.4. Since Stable Diffusion can already generate photorealistic images, a randomly initialized discriminator falls behind and cannot provide useful guidance immediately from scratch. We thus warm up the discriminator firstly, then start the joint adversarial training. In the unrolling strategy, we use $K = 9$ as the moving horizon. An ablation study on the choice of K is provided in Table 4.5. Considering the computing overhead, we apply unrolling every 8 optimization steps.

4.3 Experiments

Section 4.3.1 compares L2I diffusion models in terms of layout faithfulness and text editability. Further, we provide detailed ablation studies in Section 4.3.2. In Section 4.3.3, we evaluate the use of synthesized images for data augmentation to improve domain generalization.

4.3.1 Evaluation of Layout-to-Image Synthesis

Experimental Details. We conducted experiments on two challenging datasets: ADE20K (Zhou et al., 2017) and Cityscapes (Cordts et al., 2016). ADE20K consists of 20K training and 2K validation images, with 150 semantic classes. Cityscapes has 19 classes, whereas there are only 2975 training and 500 validation images, which poses special challenge for avoiding overfitting and preserving prior knowledge of Stable Diffusion. Following ControlNet (Zhang and Agrawala, 2023), we use BLIP (Li et al., 2022b) to generate captions for both datasets.

By default, our ALDM adopts ControlNet (Zhang and Agrawala, 2023) architecture for layout conditioning and finetune Stable Diffusion v1.5 checkpoint. Nevertheless, the proposed adversarial training strategy can be combined with other L2I models as well, as shown in Table 4.1. All trainings are conducted on 512×512

Method	Cityscapes FID ↓	Cityscapes mIoU↑	ADE20K FID↓	ADE20K mIoU↑
OFT (Qiu et al., 2023)	57.3	48.9	**29.5**	24.1
+ Adversarial supervision	56.0	54.8	31.0	29.7
+ Multistep unrolling	51.3	58.8	29.7	31.8
T2I-Adapter (Mou et al., 2024)	58.3	37.1	31.8	24.0
+ Adversarial supervision	55.9	46.6	32.4	26.5
+ Multistep unrolling	51.5	50.1	30.5	29.1
ControlNet (Zhang and Agrawala, 2023)	57.1	55.2	<u>29.6</u>	30.4
+ Adversarial supervision	**50.3**	61.5	30.0	34.0
+ Multistep unrolling	<u>51.2</u>	**63.9**	30.2	**36.0**

Table 4.1: Effect of adversarial supervision and multistep unrolling on different L2I synthesis adaptation methods. Best and second best are marked in bold and underline, respectively.

resolution. For Cityscapes, we do random cropping and for ADE20K we directly resize the images. Nevertheless, we directly synthesize 512×1024 Cityscapes images for evaluation. We use AdamW optimizer and the learning rate of 1×10^{-5} for the diffusion model, 1×10^{-6} for the discriminator, and the batch size of 8. The adversarial loss weighting factor λ_{adv} is set to be 0.1. The discriminator is firstly warmed up for 5K iterations on Cityscapes and 10K iterations on ADE20K. Afterward, we jointly train the diffusion model and discriminator in an adversarial manner. We conducted all training using 2 NVIDIA Tesla A100 GPUs. For all experiments, at inference time we use DDIM sampler (Song et al., 2020) with 25 sampling steps.

Evaluation Metrics. Following Sushko et al. (2022), Xue et al. (2023), we evaluate the image-layout alignment via mean intersection-over-union (mIoU) with the aid of off-the-shelf segmentation networks. To measure the text-based editability, we use the recently proposed TIFA score (Hu et al., 2023b), which is defined as the accuracy of a visual question answering (VQA) model, e.g., mPLUG (Li et al., 2022a). Fréchet Inception Distance (FID) (Heusel et al., 2017), Precision and Recall (Sajjadi et al., 2018) are for assessing sample quality and diversity.

Main Results. In Table 4.1, we apply the proposed adversarial supervision and

Method	Cityscapes					ADE20K				
	FID ↓	mIoU↑	P.↑	R.↑	TIFA↑	FID↓	mIoU↑	P.↑	R.↑	TIFA↑
PITI	n/a	n/a	n/a	n/a	✗	**27.9**	29.4	n/a	n/a	✗
FreestyleNet	56.8	**68.8**	**0.73**	0.44	0.300	29.2	**36.1**	0.83	0.79	0.740
T2I-Adapter	58.3	37.1	0.55	0.59	**0.902**	31.8	24.0	0.79	0.81	**0.892**
ControlNet	57.1	55.2	0.61	0.60	0.822	29.6	30.4	0.84	**0.84**	0.838
ALDM (ours)	**51.2**	63.9	0.66	**0.68**	0.856	30.2	36.0	**0.86**	0.82	0.888

Table 4.2: Quantitative comparison of the state-of-the-art L2I diffusion models. Best and second best are marked in bold and underline, respectively, while the worst result is in red. Our ALDM demonstrates competitive conditional alignment with notable text editability.

multistep unrolling strategy to different Stable Diffusion based L2I methods: OFT (Qiu et al., 2023), T2I-Adapter (Mou et al., 2024) and ControlNet (Zhang and Agrawala, 2023). Through adversarial supervision and multistep unrolling, the layout faithfulness is consistently improved across different L2I models, e.g., improving the mIoU of T2I-Adapter from 37.1 to 50.1 on Cityscapes. In many cases, the image quality is also enhanced, e.g., FID improves from 57.1 to 51.2 for ControlNet on Cityscapes. Overall, we observe that the proposed adversarial training complements different SD adaptation techniques and architecture improvements, noticeably boosting their performance. By default, ALDM represents ControlNet with adversarial supervision and multistep unrolling in other tables.

In Table 4.2, we quantitatively compare our ALDM with the other state-of-the-art L2I diffusion models: PITI (Wang et al., 2022), which does not support text control; and recent SD based FreestyleNet (Xue et al., 2023), T2I-Adapter and ControlNet, which support text control. FreestyleNet has shown good mIoU by trading off the editability, as it requires fine-tuning of the whole SD. Its poor editability, i.e., low TIFA score, is particularly notable on Cityscapes. As its training set is small and diversity is limited, FreestyleNet tends to overfit and forgets about the pre-trained knowledge. This can be reflected from the low recall value in Table 4.2 as well. Both T2I-adapter and ControlNet freeze the SD, and T2I-Adapter introduces a much smaller adapter for the conditioning compared to ControlNet. Due to limited fine-tuning capacity, T2I-Adapter does not utilize the layout effectively, leading to low mIoU, yet it better preserves the editability, i.e., high TIFA score. By contrast,

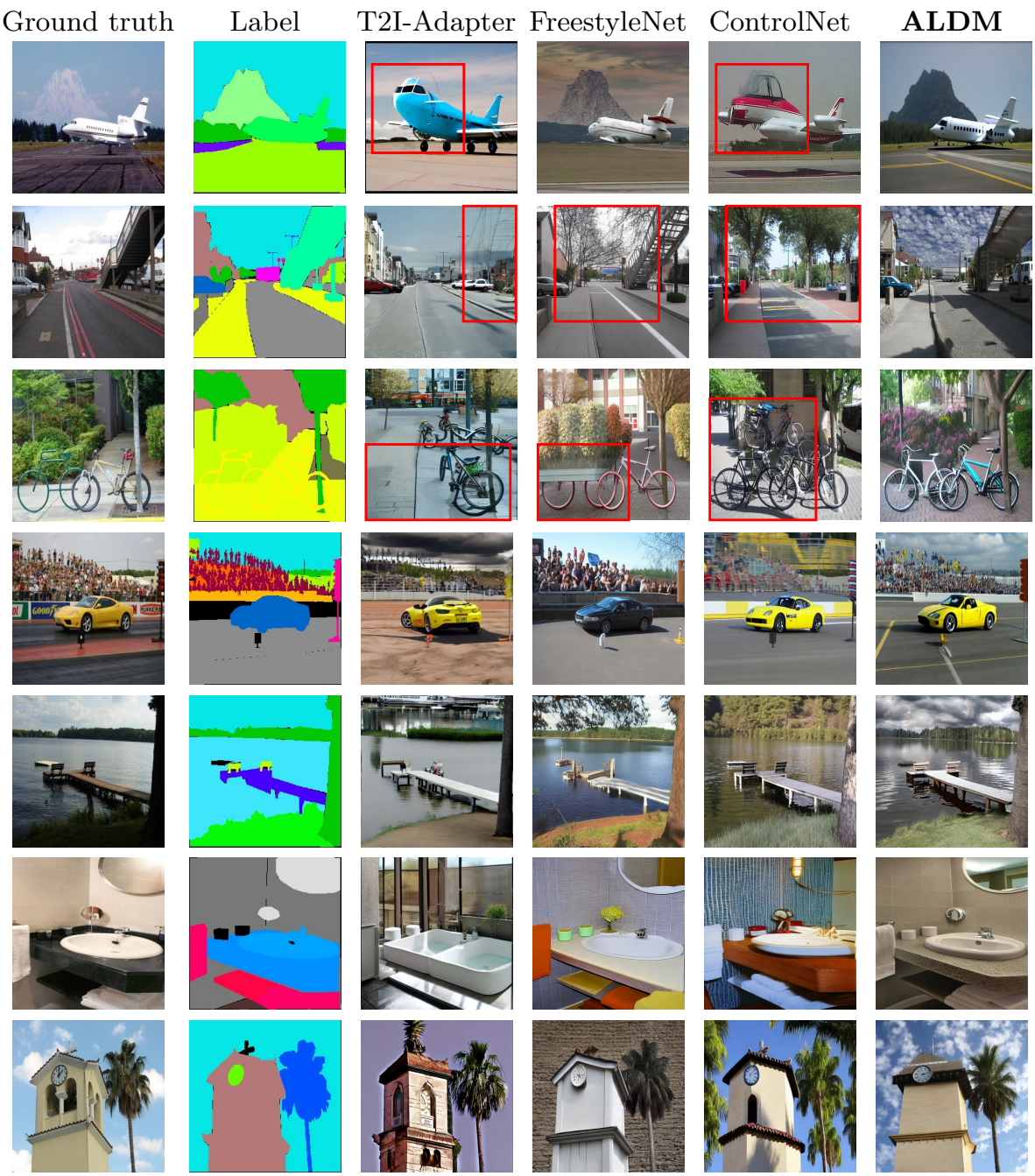


Figure 4.3: Qualitative comparison of faithfulness to the layout condition on ADE20K. Our ALDM can comply with the label map consistently, while the other may ignore the ground truth label map and hallucinate, e.g., synthesizing trees in the background (see the 2nd and 4th row).

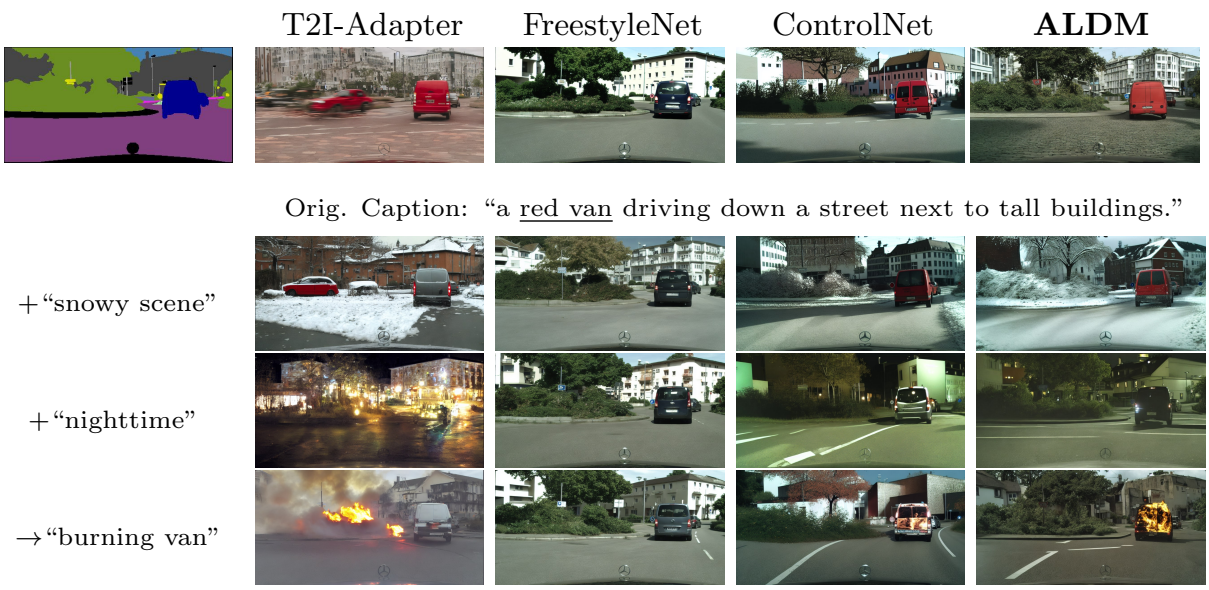


Figure 4.4: Visual comparison of text control between different L2I diffusion models on Cityscapes. Based on the image caption, we directly modify the underlined objects (indicated as →), or append a postfix to the caption (indicated as +). In contrast to prior work, ALDM can faithfully accomplish both global scene level modification (e.g., "snowy scene") and local editing (e.g., "burning van").

ControlNet improves mIoU while trading off the editability. In contrast, ALDM exhibits competitive mIoU while maintaining high TIFA score, which enables its usability for practical applications, e.g., data augmentation for domain generalization detailed in Section 4.3.3.

Qualitative comparison on the faithfulness to the label map is shown in Fig. 4.3. T2I-Adapter often ignores the layout condition (see the first row of Fig. 4.3), which can be reflected in low mIoU as well. FreestyleNet and ControlNet may hallucinate objects in the background. For instance, in the second row of Fig. 4.3, both methods synthesize trees where the ground-truth label map is sky. In the last row, ControlNet also generates more bicycles instead of the ground truth trees in the background. Contrarily, ALDM better complies with the layout in this case. Visual comparison on text editability is shown in Figs. 4.1 and 4.4. We observe that FreestyleNet only shows little variability and minor visual differences, as evidenced by the low TIFA score. T2I-Adapter and ControlNet on the other hand preserve better text control, nonetheless, they may not follow well the layout condition. In

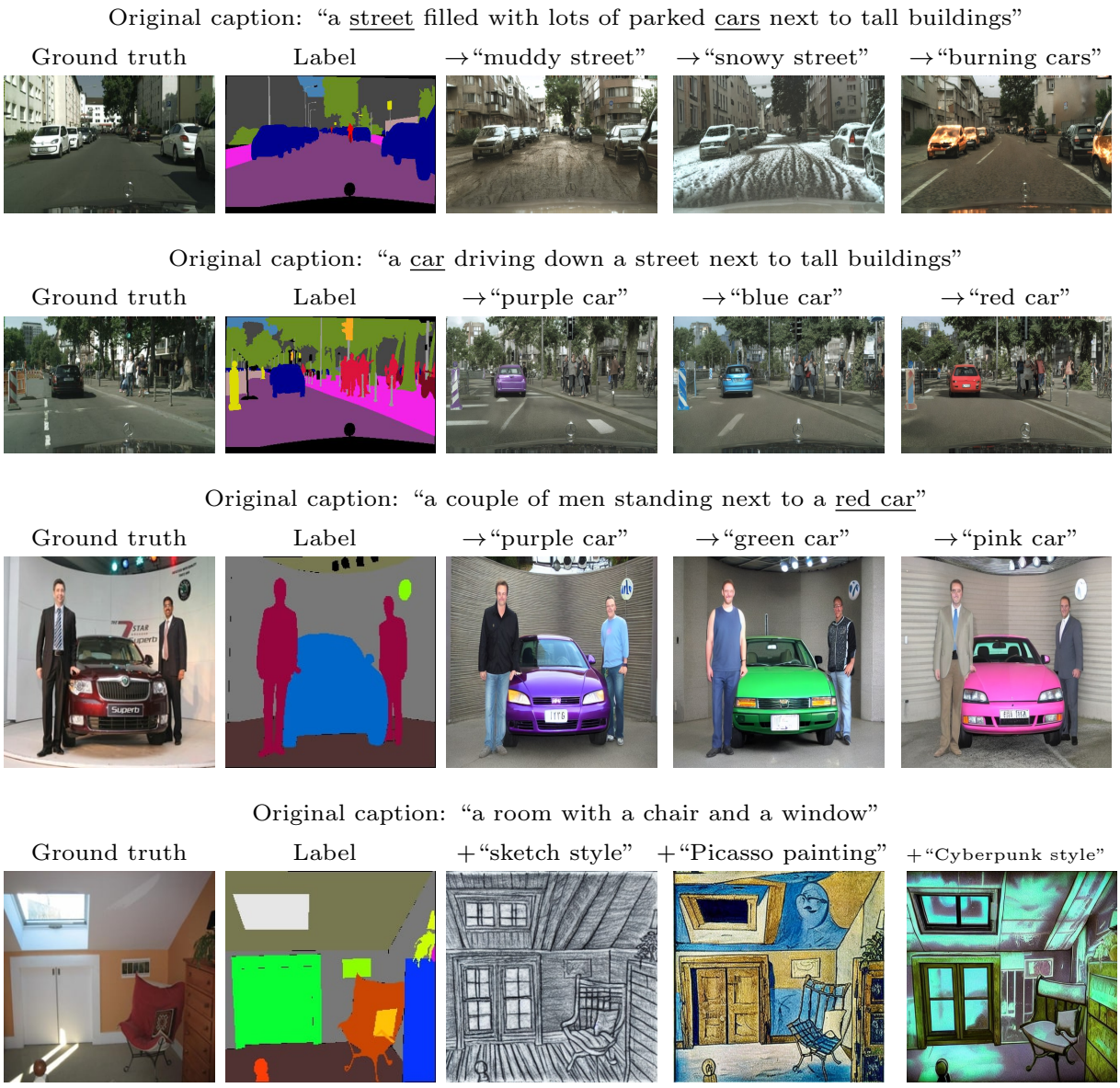


Figure 4.5: Visual examples of text controllability with our ALDM. Based on the original image captions generated by BLIP model, we can directly modify the underlined objects (indicated as →), or append a postfix to the caption (indicated as +). Our ALDM can accomplish both local attribute editing (e.g., car color) and global image style modification (e.g., sketch style).

Fig. 4.1, ControlNet fails to generate the truck, especially when the prompt is modified. And in Fig. 4.4, the trees on the left are only sparsely synthesized. While

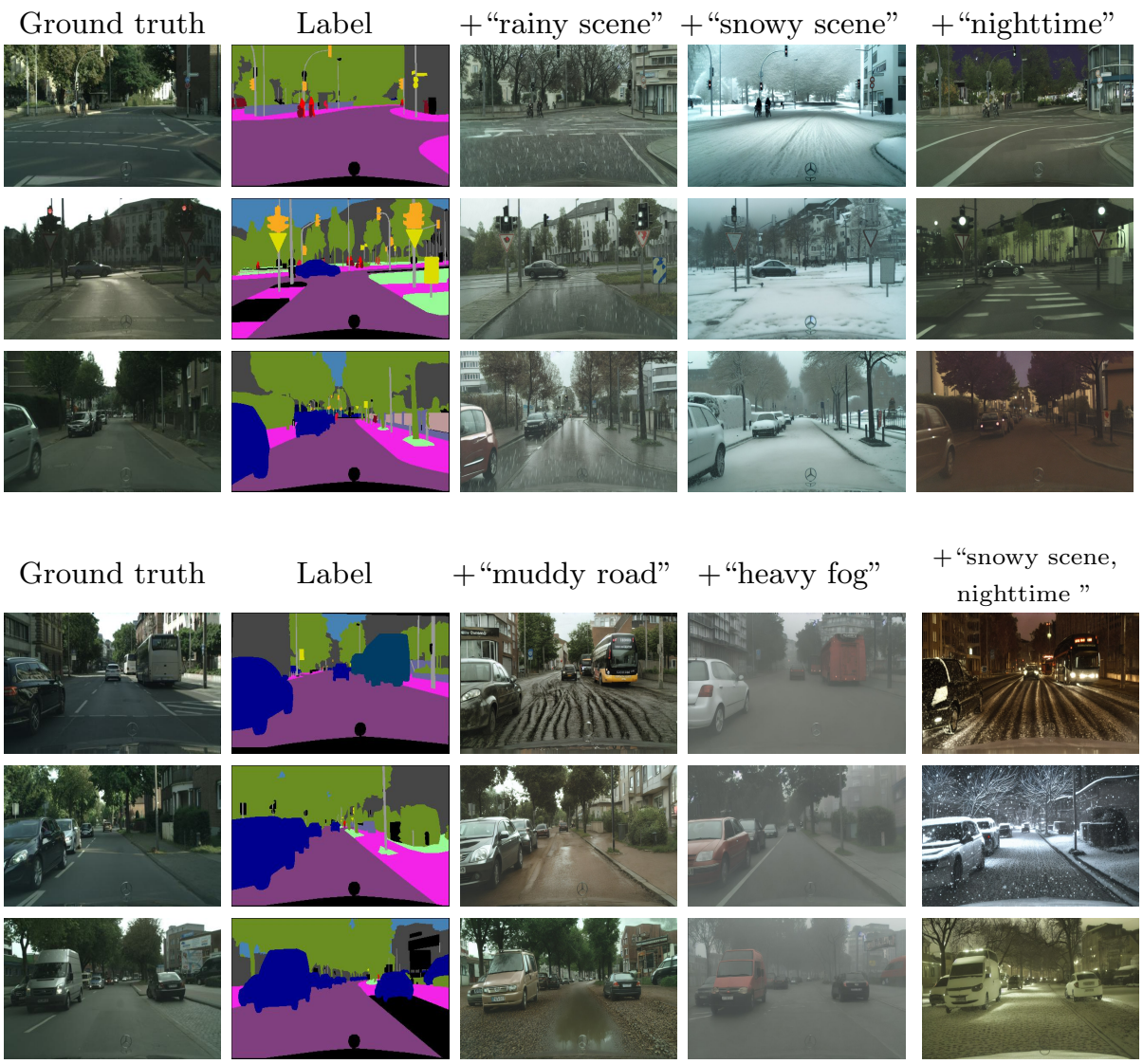


Figure 4.6: Visual examples of Cityscapes, synthesized by ALDM via various textual descriptions, which can be further utilized on downstream tasks.

ALDM produces samples that adhere better to both layout and text conditions, inline with the quantitative results. More visual editing examples are illustrated in Figs. 4.5 and 4.6.

Comparison with GAN-based L2I Methods. We additionally compare our method with prior GAN-based L2I methods in Table 4.3. It is worthwhile to mention that all GAN-based approaches do not have text controllability, thus they can only produce samples resembling the training dataset, which constrains their util-

	Method	Cityscapes FID ↓	mIoU↑	TIFA↑	ADE20K FID↓	mIoU↑	TIFA↑
GANs	Pix2PixHD (Wang et al., 2018b)	95.0	63.0	✗	81.8	28.8	✗
	SPADE (Park et al., 2019)	71.8	61.2		33.9	38.3	
	OASIS (Schönfeld et al., 2020)	47.7	69.3		28.3	45.7	
	SCGAN (Wang et al., 2021c)	49.5	55.9		29.3	41.5	
	CLADE (Tan et al., 2021b)	57.2	58.6		35.4	23.9	
	GroupDNet (Zhu et al., 2020b)	47.3	55.3		41.7	27.6	
DMs	PITI (Wang et al., 2022)	n/a	n/a	✗	27.9	29.4	✗
	FreestyleNet (Xue et al., 2023)	56.8	68.8	0.300	29.2	36.1	0.740
	T2I-Adapter (Mou et al., 2024)	58.3	37.1	0.902	31.8	24.0	0.892
	ControlNet (Zhang and Agrawala, 2023)	57.1	55.2	0.822	29.6	30.4	0.838
	ALDM (ours)	51.2	63.9	0.856	30.2	36.0	0.888

Table 4.3: Quantitative comparison results with the state-of-the-art layout-to-image GANs and diffusion models (DMs). Our ALDM demonstrates competitive conditional alignment with notable text editability.

ity on downstream tasks. On the other hand, our ALDM achieves the balanced performance between faithfulness to the layout condition and editability via text, rendering itself advantageous for the domain generalization tasks.

In Fig. 4.7, we compare our ALDM with GAN-based style transfer method ISSA (Li et al., 2024c) introduced in Chapter 3. It can be observed that ALDM produces more realistic results with faithful local details, given the label map and text prompt. In contrast, style transfer methods require two images, and mix them on the global color style, while the local details, e.g., mud, and snow may not be faithfully transferred.

4.3.2 Ablation Study

Ablation on Discriminator. We conduct the ablation study on different discriminator designs, shown in Table 4.4. Both choices for the discriminator network: CNN-based segmentation network UperNet (Xiao et al., 2018) and transformer-based Segmenter (Strudel et al., 2021), improve faithfulness of the baseline ControlNet model.

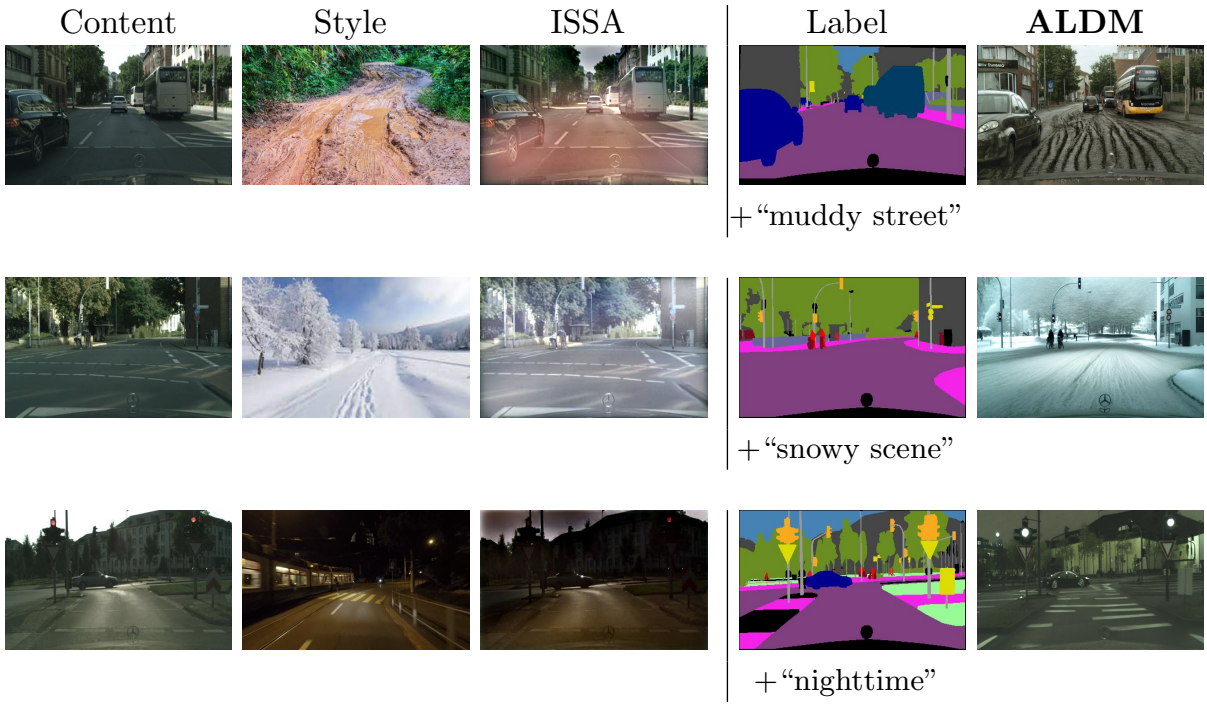


Figure 4.7: Comparison between our ALDM and GAN-based style-transfer method ISSA (Li et al., 2024c) described in Chapter 3. It can be seen that ALDM can produce more realistic results with faithful local details, given the label map and text. In contrast, style transfer methods require two images, and mix them on the global color style, while the local details, e.g., mud, and snow may not be faithfully transferred.

Instead of employing the discriminator in the pixel space, we also experiment with feature-space discriminator. Thanks to large-scale vision-language pretraining on massive datasets, Stable Diffusion (SD) (Rombach et al., 2022) has acquired rich representations, endowing it with the capability not only to generate high-quality images, but also to excel in various downstream tasks. Recent work VPD (Zhao et al., 2023b) has unleashed the potential of SD, and leveraged its representation for visual perception tasks, e.g., semantic segmentation. More specifically, they extracted cross-attention maps and feature maps from SD at different resolutions and fed them to a lightweight decoder for the specific task. Despite the simplicity of the idea, it works fairly well, presumably due to the powerful knowledge of SD. In the ablation study, we adopt the segmentation model of VPD as the feature-based discriminator. Nevertheless, different from the joint training of SD and the task-specific decoder in the original VPD implementation, we only train the newly

Method	Cityscapes FID↓	Cityscapes mIoU↑	ADE20K FID↓	ADE20K mIoU↑
ControlNet	57.1	55.2	29.6	30.4
+ UperNet	50.3	61.5	30.0	34.0
+ Segmenter	52.9	59.2	29.8	34.1
+ Feature-based	53.1	59.6	29.3	33.1
+ Frozen UperNet	-	-	50.8	40.2

Table 4.4: Ablation on the discriminator type.

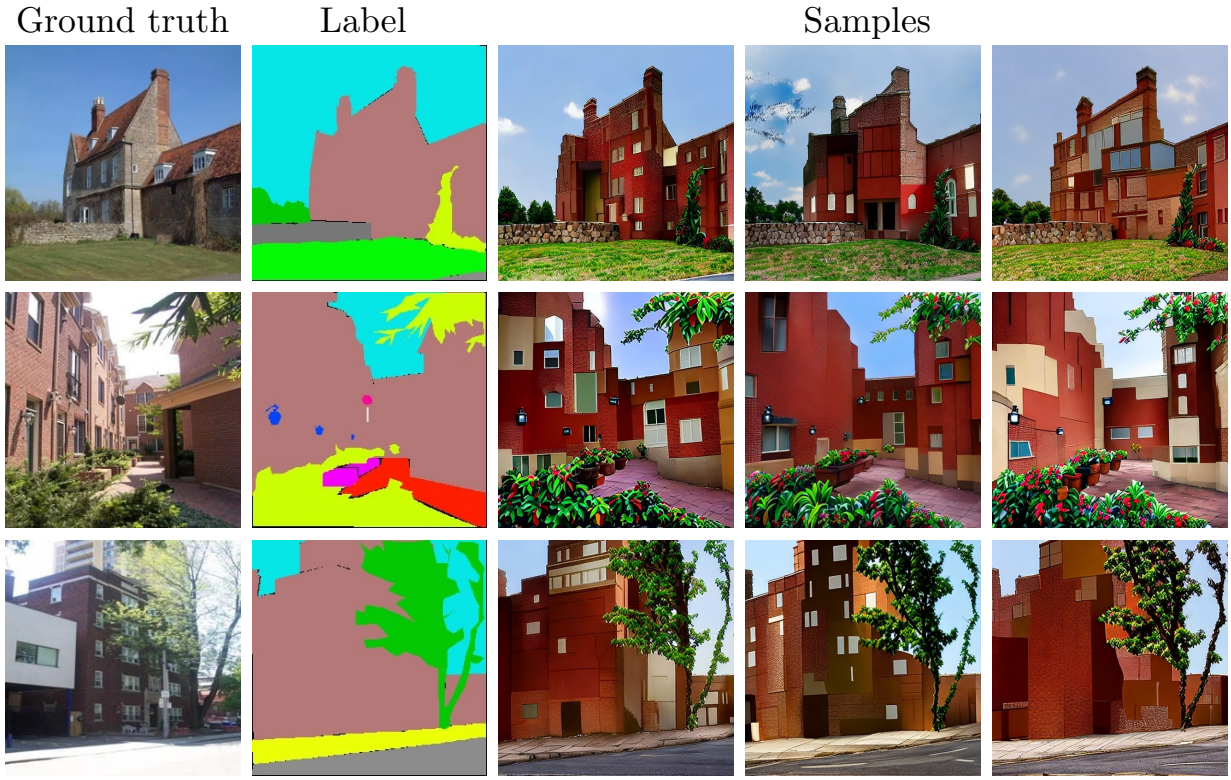


Figure 4.8: Visual results of using a *frozen* segmentation network, i.e., a pretrained UperNet (Xiao et al., 2018), to provide conditional guidance during diffusion model training. We can observe the mode collapse issue, where the diffusion model tends to learn to a mean mode and exhibits little variation in the generated samples.

added decoder, while freezing SD to preserve the text controllability as ControlNet. As shown in Table 4.4, the feature-based discriminator also works reasonably well.

Lastly, we employ a frozen semantic segmentation network to provide guidance

	Cityscapes			ADE20K			
	FID↓	mIoU↑	TIFA↑	FID↓	mIoU↑	TIFA↑	Overhead
ControlNet	57.1	55.2	0.822	29.6	30.4	0.838	0.00
K = 0	50.3	61.5	0.894	30.0	34.0	0.904	0.00
K = 3	54.9	62.7	0.856	-	-	-	1.55
K = 6	51.6	64.1	0.832	30.3	34.5	0.898	3.11
K = 9	51.2	63.9	0.856	30.2	36.0	0.888	4.65
K = 15	50.7	64.1	0.882	30.2	36.9	0.825	7.75

Table 4.5: Ablation on the unrolling step K. Overhead is measured as seconds per training iteration.

directly. Note that this case is no longer adversarial training anymore, as the segmentation model does not update itself with the generator. Despite achieving high mIoU, the generator tends to learn a mean mode of the class and produce unrealistic samples (see Fig. 4.8), thus yielding high FID. In this case, the generator can more easily find a "cheating" way to fool the discriminator as it is not updating.

Ablation on Multistep Unrolling. For the unrolling strategy, we compare different number of unrolling steps in Table 4.5. We observe that more unrolling steps is beneficial for improving the faithfulness, as the model can consider more future steps to ensure alignment with the layout condition. However, the additional unrolling time overhead also increases linearly. Therefore, we choose $K = 9$ by default in all experiments.

4.3.3 Improved Domain Generalization for Semantic Segmentation

We further investigate the utility of synthetic data generated by different L2I models for domain generalization (DG) in semantic segmentation. Namely, the downstream model is trained on a source domain, and its generalization performance is evaluated on unseen target domains. We experiment with both CNN-based segmentation model HRNet (Wang et al., 2021b) and transformer-based SegFormer (Xie et al., 2021). Quantitative evaluation is provided in Table 4.6, where all models except the oracle are trained on Cityscapes, and tested on both Cityscapes

Method	HRNet (Wang et al., 2021b)					SegFormer (Xie et al., 2021)				
	CS	Rain	Fog	Snow	ACDC	CS	Rain	Fog	Snow	ACDC
Baseline (CS)	70.47	44.15	58.68	44.20	41.48	67.90	50.22	60.52	48.86	47.04
Hendrycks-Weather	69.25	50.78	60.82	38.34	43.19	67.41	54.02	64.74	49.57	49.21
ISSA	70.30	50.62	66.09	53.30	50.05	67.52	55.91	67.46	53.19	52.45
FreestyleNet	71.73	51.78	67.43	53.75	50.27	**69.70**	52.70	68.95	54.27	52.20
ControlNet	71.54	50.07	68.76	52.94	51.31	68.85	55.98	68.14	54.68	53.16
ALDM (ours)	**72.10**	**53.67**	**69.88**	**57.95**	**53.03**	68.92	**56.03**	**69.14**	**57.28**	**53.78**
Oracle (CS+ACDC)	70.29	65.67	75.22	72.34	65.90	68.24	63.67	74.10	67.97	63.56

Table 4.6: Comparison on domain generalization, i.e., from Cityscapes (train) to ACDC (unseen). mIoU is reported on Cityscapes (CS), individual scenarios of ACDC (Rain, Fog, Snow) and the whole ACDC. Hendrycks-Weather (Hendrycks and Dietterich, 2018) simulates weather conditions in a synthetic manner for data augmentation. Oracle model is trained on both Cityscapes and ACDC in a supervised manner, serving as an upper bound on ACDC (not Cityscapes) for the other methods. ALDM can consistently improve generalization performance of both HRNet and SegFormer.

and the unseen ACDC. The oracle model is trained on both datasets.

We observe that Hendrycks-Weather (Hendrycks and Dietterich, 2018), which simulates weather conditions in a synthetic manner, brings limited benefits. ISSA (Li et al., 2024c) resorts to simple image style mixing within the source domain. For models that accept text prompts (FreestyleNet, ControlNet and ALDM), we can synthesize novel samples given the textual description of the target domain, as shown in Figs. 4.4 and 4.5. Nevertheless, the effectiveness of such data augmentation depends on the editability via text and faithfulness to the layout. FreestyleNet only achieves on-par performance with ISSA. We hypothesize that its poor text editability only provides synthetic data close to the training set with style jittering similar to ISSA's style mixing. While ControlNet allows text editability, the misalignment between the synthetic image and the input layout condition, unfortunately, can even hurt the performance. While mIoU averaged over classes is improved over the baseline, the per-class IoU shown in Table 4.7 indicates the undermined performance on small object classes, such as traffic light, rider and person. On those small objects, the alignment is noticeably more challenging to pursue than on classes with large area such as truck and bus. In contrast to it, ALDM, owing to its text editability

Method	Pole	Traf. light	Traf. sign	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bike
Baseline	48.50	59.07	67.96	72.44	52.31	92.42	70.11	77.62	64.01	50.76	68.30
ControlNet	49.53	58.47	67.37	71.45	49.68	92.30	**76.91**	**82.98**	**72.40**	50.84	67.32
ALDM	**51.21**	**60.50**	**69.56**	**73.82**	**53.01**	**92.57**	76.61	81.37	66.49	**52.79**	**68.61**

Table 4.7: Per-class IoU of Cityscapes object classes. Numbers in red indicate worse IoU compared to the baseline. The best is marked in bold. Our ALDM has demonstrated better performance on small object classes, e.g., pole, traffic light, traffic sign, person, rider, which reflects our method can better comply with the layout condition, as small object classes are typically more challenging in L2I task and pose higher requirement for the faithfulness to the layout condition.

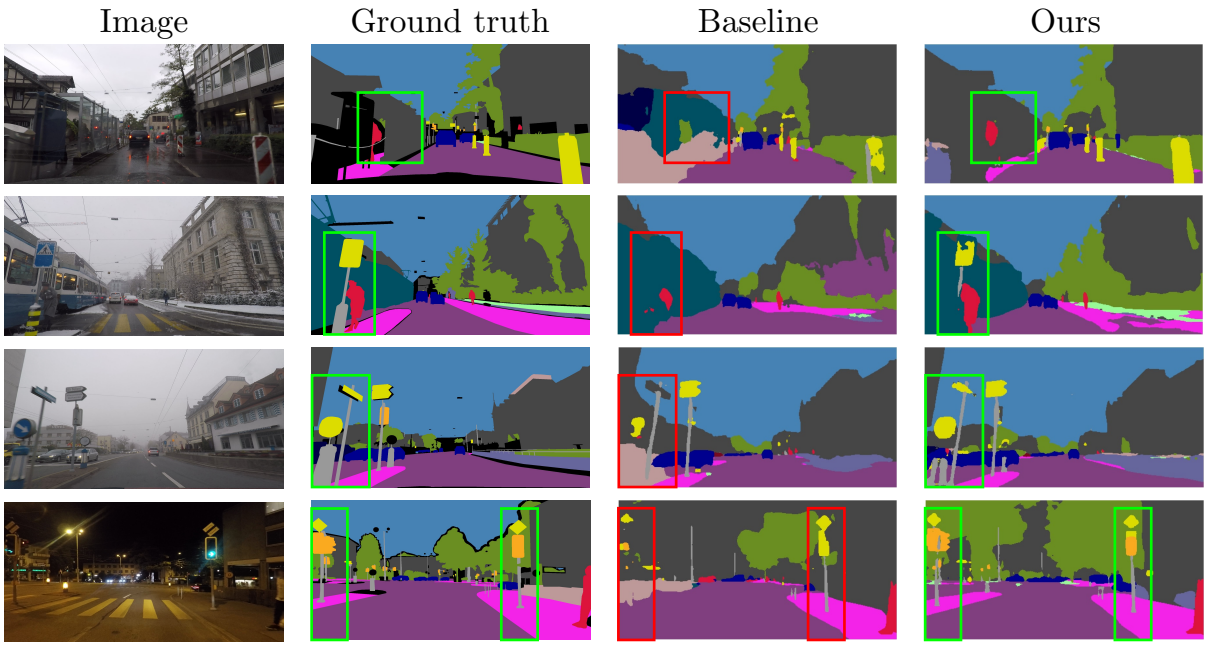


Figure 4.9: Semantic segmentation results of Cityscapes → ACDC generalization using HRNet. The HRNet is trained on Cityscapes only. Augmented with diverse synthetic data generated by our ALDM, the segmentation model can make more reliable predictions under diverse conditions under diverse unseen conditions, which is crucial for deployment in the open-world.

and faithfulness to the layout, consistently improves across individual classes and ultimately achieves pronounced gains on mIoU across different target domains, e.g., 11.6% improvement for HRNet on ACDC.

Qualitative visualization is illustrated in Fig. 4.9. The segmentation model em-

powered by ALDM can produce more reliable predictions under diverse weather conditions, e.g., improving predictions on objects such as traffic signs and person, which are safety critical cases.

4.4 Conclusion

In this work, we propose to incorporate adversarial supervision to improve the faithfulness to the layout condition for L2I diffusion models. We propose to leverage a segmenter-based discriminator to explicitly utilize the layout label map and provide a strong learning signal. Further, we propose a novel multistep unrolling strategy to encourage conditional coherency across sampling steps. Our ALDM can well comply with the layout condition, meanwhile preserving the text controllability. Capitalizing these intriguing properties of ALDM, we synthesize novel samples via text control for data augmentation on the domain generalization task, resulting in a significant enhancement of the downstream model's generalization performance.

5 Improved Generative Semantic Nursing for Text-to-Image Synthesis

Emerging large-scale text-to-image (T2I) generative models, e.g., Stable Diffusion (SD), have exhibited overwhelming results with high fidelity. Despite the magnificent progress, current state-of-the-art T2I models still struggle to generate images fully adhering to the input prompt. Prior work, Attend & Excite, has introduced the concept of Generative Semantic Nursing (GSN), aiming to optimize cross-attention during inference time to better incorporate the semantics. It demonstrates

promising results in generating simple prompts, e.g., "a cat and a dog". However, its efficacy declines when dealing with more complex prompts, and it does not explicitly address the problem of improper attribute binding. To address the challenges posed by complex prompts or scenarios involving multiple entities and to achieve improved attribute binding, we propose Divide & Bind. We introduce two novel loss objectives for GSN: a novel attendance loss and a binding loss. Our approach stands out in its ability to faithfully synthesize desired objects with improved attribute alignment from complex prompts and exhibits superior performance across multiple evaluation benchmarks. More videos can be found on the project page `https://sites.google.com/view/divide-and-bind`. This work is published as an oral paper at the British Machine Vision Conference, 2023 (Li et al., 2023a).

5.1 Introduction

In the realm of text-to-image (T2I) synthesis, large-scale generative models (Rombach et al., 2022, Ramesh et al., 2022, Saharia et al., 2022, Balaji et al., 2022, Chang et al., 2023, Yu et al., 2022, Kang et al., 2023) have recently achieved significant progress and demonstrated exceptional capacity to generate stunning photorealistic images. However, it remains challenging to synthesize images that fully comply with the given prompt input (Marcus et al., 2022, Feng et al., 2023, Chefer et al., 2023, Wang et al., 2023e). There are two well-known semantic issues in text-to-image synthesis, i.e., "missing objects" and "attribute binding". "Missing objects" refers to the phenomenon that not all objects mentioned in the input text faithfully appear in the image. "Attribute binding" represents the critical compositionality problem that the attribute information, e.g., color or texture, is not properly aligned to the corresponding object or wrongly attached to the other object. To mitigate these issues, recent work Attend & Excite (A&E) (Chefer et al., 2023) has introduced the concept of Generative Semantic Nursing (GSN). The core idea lies in updating latent codes on-the-fly such that the semantic information in the given text can be better incorporated within pretrained synthesis models.

As an initial attempt A&E (Chefer et al., 2023), building upon the powerful open-source T2I model Stable Diffusion (SD) (Rombach et al., 2022), leveraged cross-attention maps for optimization. Since cross-attention layers are the only interac-

Figure 5.1: Our **Divide & Bind** can faithfully generate multiple objects based on detailed textual description. Compared to prior state-of-the-art semantic nursing technique for text-to-image synthesis, Attend & Excite (Chefer et al., 2023), our approach exhibits superior alignment with the input prompt and maintain a higher level of realism.

tion between the text prompt and the diffusion model, the attention maps have significant impact on the generation process. To enforce the object occurrence, A&E defined a loss objective that attempts to maximize the maximum attention value for each object token. Although showing promising results on simple composition, e.g., "a cat and a frog", we observed unsatisfying outcomes when the prompt becomes more complex, as illustrated in Fig. 5.1. A&E fails to faithfully synthesize the "train" or "dog" in the first two examples, and miss one "goose" in the third one. We attribute this to the suboptimal loss objective, which only considers the single maximum value and does not take the spatial distribution into consideration. As the complexity of prompts increases, token competition intensifies. The single excitation of one object token may overlap with others, leading to the suppression of one object by another (e.g., missing "train" in Fig. 5.1) or to hybrid objects, exhibiting features of both semantic classes (e.g., mixed dog-turtle in Fig. 5.3). Similar phenomenon has been observed in Tang et al. (2023b) as well.

In this work, we propose a novel objective function for GSN. We maximize the

total variation of the attention map to prompt multiple, spatially distinct attention excitations. By spatially distributing the attention for each token, we enable the generation of all objects mentioned in the prompt, even under high token competition. Intuitively, this corresponds to *dividing* the attention map into multiple regions. Besides, to mitigate the attribute *binding* issue, we propose a Jensen-Shannon divergence (JSD) based binding loss to explicitly align the distribution between excitation of each object and its attributes. Thus, we term our method Divide & Bind. Our main contributions can be summarized as:

- We propose a novel total-variation based attendance loss enabling the presence of multiple objects in the generated image.
- We propose a JSD-based attribute binding loss for faithful attribute binding.
- Our approach exhibits outstanding capability of generating images fully adhering to the prompt, outperforming A&E on several benchmarks involving complex descriptions.

5.2 Preliminaries

Stable Diffusion (SD). We implement our method based on the open-source state-of-the-art T2I model SD (Rombach et al., 2022), which belongs to the family of latent diffusion models (LDMs). LDMs are two-stage methods, consisting of an autoencoder and a diffusion model trained in the latent space. In the first stage, the encoder $\mathcal{E}$ transforms the given image x into a latent code $z = \mathcal{E}(x)$, then z is mapped back to the image space by the decoder $\mathcal{D}$. The autoencoder is trained to reconstruct the given image, i.e. $\mathcal{D}(\mathcal{E}(x)) \approx x$. In the second stage, a diffusion model (Ho et al., 2020, Nichol and Dhariwal, 2021) is trained in the latent space of the autoencoder. During training, we gradually add noise to the original latent z_0 with time, resulting in z_t. Then the UNet (Ronneberger et al., 2015) denoiser ϵ_θ is trained with a denoising objective to predict the noise ϵ that is added to z_0:

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim N(0,I),c,t}\left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|^2\right], \tag{5.1}$$

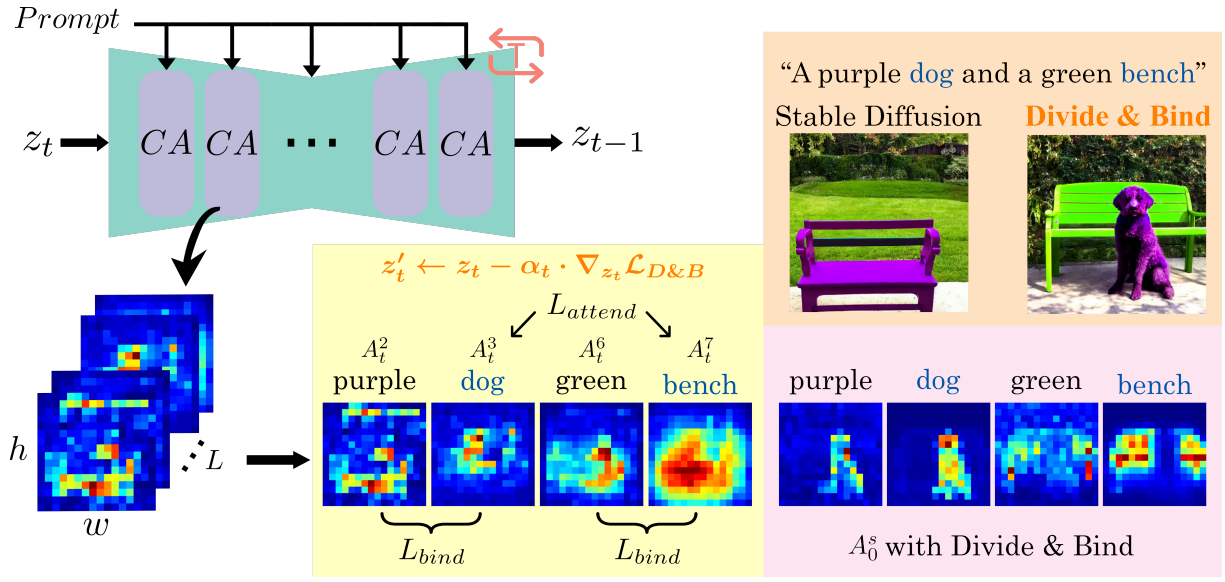


Figure 5.2: Method overview. We perform latent optimization on-the-fly based on the attention maps of the object tokens with our TV-based L_{attend} and JSD-based L_{bind}.

where c is the conditional information, e.g., text. During inference, given z_T randomly sampled from Gaussian distribution, UNet outputs noise estimation and gradually removes it, finally producing the clean version z_0.

Cross-Attention in Stable Diffusion. In SD, a frozen CLIP text encoder (Radford et al., 2021) is adopted to embed the text prompt $\mathcal{P}$ into a sequential embedding as the condition c, which is then injected into UNet through cross-attention (CA) to synthesize text-complied images. The CA layers take the encoded text embedding and project it into queries Q and values V. The keys K are mapped from the intermediate features of UNet. The attention maps are then computed by

$$A_t = Softmax(\frac{QK^T}{\sqrt{d}}), \tag{5.2}$$

where t indicates the time step, Softmax is applied along the channel dimension. The attention maps A_t can be reshaped into $\mathbb{R}^{h \times w \times L}$, where h, w is the resolution of the feature map, L is the sequence length of the text embedding. Further, we denote the cross-attention map that corresponds to the sth text token as $A_t^s \in \mathbb{R}^{h \times w}$, see an illustration in Fig. 5.2. One known issue of SD is that not all objects are present in the final image (Liu et al., 2022, Wang et al., 2023e, Chefer et al., 2023), while,

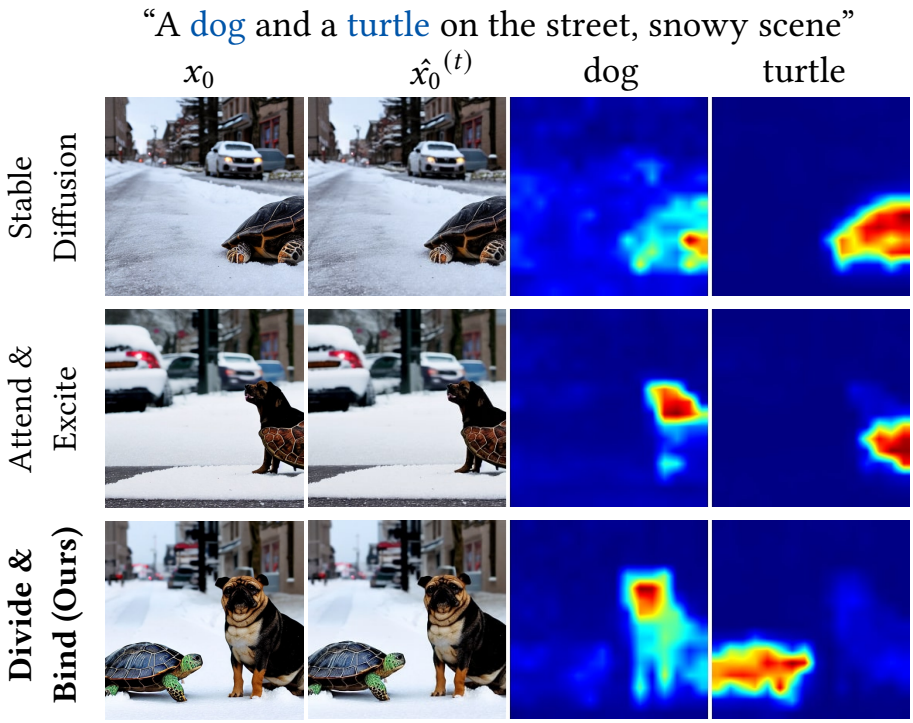


Figure 5.3: Cross-attention visualization in different timesteps for each object token and predicted clean image $\hat{x}_0^{(t)}$. Note that this is a GIF, best viewed in *Acrobat Reader*.

as shown in (Balaji et al., 2022, Hertz et al., 2023), the high activation region of the corresponding attention map strongly correlates to the appearing pixels belonging to one specific object in the final image. Hence, the activation in the attention maps is an important signal and an influencer in the semantic guided synthesis.

5.3 Method

Given the recognized significance of the cross-attention maps in guiding semantic synthesis, our method aims at optimizing the latent code at inference time to excite them based on the text tokens. We employ the generative semantic nursing (GSN) method (Section 5.3.1) for latent code optimization, and propose a novel loss formulation (Section 5.3.2). It consists of two parts, i.e. *divide* and *bind*, which encourages object occurrence and attribute binding respectively.

5.3.1 Generative Semantic Nursing (GSN)

To improve the semantic guidance in SD during inference, one pragmatic way is via latent code optimization at each time step of sampling, i.e. GSN (Chefer et al.,

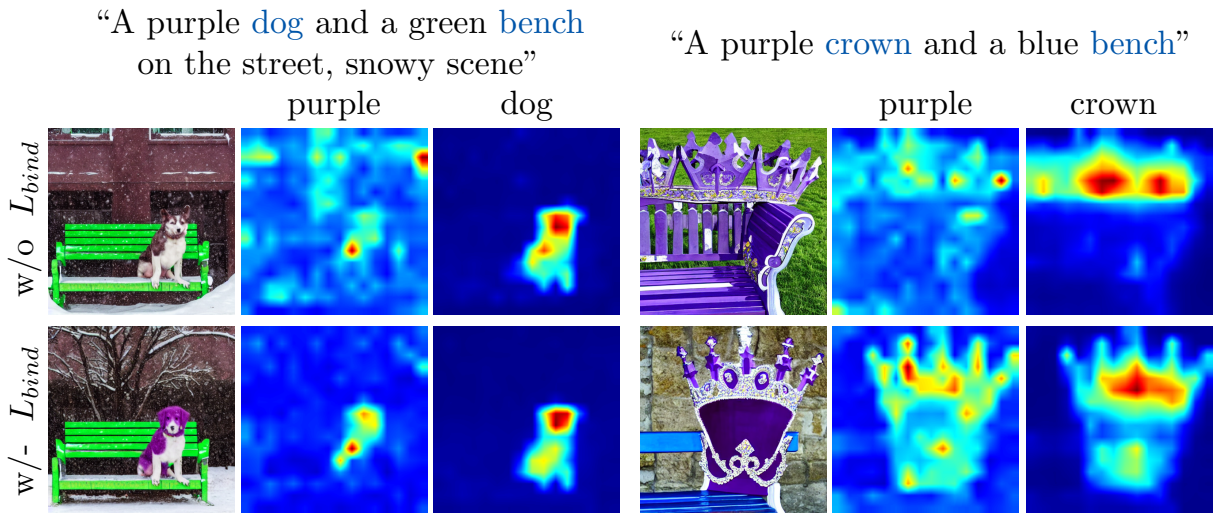


Figure 5.4: Binding loss ablation. L_{bind} aligns the excitation of attribute and object attention.

2023)

$$z'_t \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}, \tag{5.3}$$

where α_t is the updating rate and $\mathcal{L}$ is the loss to encourage the faithfulness between the image and text description, e.g. object attendances and attribute binding. GSN has the advantage of avoiding fine-tuning SD.

As the text information is injected into the UNet of SD via cross attention layers, it is natural to set the loss $\mathcal{L}$ with the cross attention maps as the inputs. Given the text prompt $\mathcal{P}$ and a list of object tokens S, we will have a set of attention maps $\{A_t^s\}$ for $s \in S$. Ideally, if the final image contains the concept provided by the object token s, the corresponding cross-attention map A_t^s should show strong activation. To achieve this, A&E (Chefer et al., 2023) enhances the single maximum value of the attention map, i.e. $L_{A\&E} = -\min_{s \in S}(\max_{i,j}(A_t^s[i,j]))$. However, it does not facilitate with multiple excitations, which is increasingly important when confronted with complex prompts and the need to generate multiple instances. As shown in Fig. 5.3, a single excitation can be easily taken over by the other competitor token, leading to missing objects in the final image. Besides, it does not explicitly address the attribute binding issue. Instead, our Divide & Bind promotes the allocation of attention across distinct areas, enabling the model to explore various regions for ob-

ject placement. Moreover, we introduce an attribute binding regularization which explicitly encourages attribute alignment.

5.3.2 Divide & Bind

Our proposed method Divide & Bind consists of a novel objective for GSN

$$\min_{z_t} \mathcal{L}_{D\&B} = \min_{z_t} \mathcal{L}_{attend} + \lambda \mathcal{L}_{bind} \tag{5.4}$$

which has two parts, the attendance loss $\mathcal{L}_{attend}$ and the binding loss $\mathcal{L}_{bind}$ that respectively enforce the object attendance and attribute binding. λ is the weighting factor. Detailed formulation of both loss terms is presented as follows.

Divide for Attendance. The attendance loss L_{attend} is to incentivize the presence of the objects, thus is applied to the text tokens associated with *objects* S,

$$\mathcal{L}_{attend} = -\min_{s \in S} TV(A_t^s),\ TV(A_t^s) = \sum_{i,j} \left|A_t^s[i+1,j] - A_t^s[i,j]\right| + \left|A_t^s[i,j+1] - A_t^s[i,j]\right| \tag{5.5}$$

where $A_t^s[i,j]$ denotes the attention value of the s-th token at the specific location $[i,j]$ and time step t. The loss formulation in Eq. (5.5) is based on the the finite differences approximation of the total variation (TV) $|\nabla A_t^s|$ along the spatial dimensions. It is evaluated for each object token and we take the smallest value, i.e., representing the worst case among the all object tokens. Taking the negative TV as the loss, we essentially maximize the TV for latent optimization in Eq. (5.4). Since TV is essentially computed as a form of summation across the spatial dimension, it encourages large activation differences across many neighboring at different spatial locations rather than a single one, thus not only having one high activation region but also many of them. Such an activation pattern in the space resembles to dividing it into different regions. The model can select some of them to display the object with single or even multiple attendances. This way, conflicts between different objects that compete for the same region can be more easily resolved. Furthermore, from an optimization perspective, it allows the model to search among different options for converging to the final solution. The loss is applied at the initial sampling steps.

As can be seen from the GIF in Fig. 5.3, for the "dog" token, regions on both left and right sides are explored in the initial phase. In the end, the left side is taken over by the "turtle" but the "dog" token covers the right side. While for SD, the "dog" token has a single weak activation, and for Attend & Excite, it only has one single high activation region on the right that is taken over by the "turtle" later.

Attribute Binding Regularization. In addition to the object attendance, the given attribute information, e.g. color or material, should be appropriately attached to the corresponding object. We denote the attention map of the object token and its attribute token as A_t^s and A_t^r, respectively. For attribute binding, it is desirable that A_t^r and A_t^s are spatially well-aligned, i.e. high activation regions of both tokens are largely overlapped. To this end, we introduce $\mathcal{L}_{bind}$. After proper normalization along the spatial dimension, we can view the normalized attention maps $\widetilde{A}_t^r$ and $\widetilde{A}_t^s$ as two probability mass functions whose sample space has size $h \times w$. To explicitly encourage such alignment, we can then minimize the symmetric similarity measure Jensen–Shannon divergence (JSD) between these two distributions:

$$\mathcal{L}_{bind} = JSD\left(\widetilde{A}_t^r \| \widetilde{A}_t^s\right). \tag{5.6}$$

Specifically, we adopt the Softmax-based normalization along the spatial dimension. When performing normalization, we also observe the benefit of first aligning the value range between the two attention maps. Namely, the original attention map of the object tokens A_t^s have higher probability values than the ones of the attribute tokens A_t^r. Therefore, we first re-scale A_t^r to the same range as A_t^s. As illustrated in Fig. 5.4, after applying L_{bind}, the attribute token (e.g. "purple") is more localized to the correct object region (e.g. "dog" or "crown").

Implementation Details. We provide the algorithm overview in Algorithm 1. Given the text prompt $\mathcal{P}$, we firstly identify the tokens of interest, e.g., object tokens and attribute tokens. This process can either be done manually or automatically with the aid of GPT-3 (Brown et al., 2020) as shown in Hu et al. (2023b). Taking advantage of the in-context learning (Brown et al., 2020, Hu et al., 2022) capability of GPT-3, by providing a few in-context examples, GPT-3 can automatically extract the desired nouns and adjectives for new input prompts. For instance, in our experiments on the COCO-Subject and COCO-Attribute benchmarks, we used the

captions of COCO images without fixed templates as the prompts, where the object and attribute tokens were selected automatically using GPT-3. Based on the token indices, we can extract attention maps and apply our $L_{B\&D}$ to update the noised latent z_t.

Algorithm 1 Simplified Algorithm Overview of Divide & Bind

Input: A text prompt $\mathcal{P}$ and a pretrained Stable Diffusion SD

Output: A noised latent z_{t-1} for the next denoising step

Determine object S and attribute R tokens by GPT with in-context learning as in TIFA (Hu et al., 2023b)
Extract attention maps for the object tokens A_t^s and attribute tokens A^r
if A^r are not None **then**
    $L_{D\&B} = L_{attend} + \lambda L_{bind}$
else
    $L_{D\&B} = L_{attend}$
end if
$z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} L_{D\&B}$
$z_{t-1} \leftarrow SD(z_t', \mathcal{P}, t)$
return z_{t-1}

We inherit the choice of optimization hyperparameters from the initial attempt for GSN - Attend & Excite (A&E) (Chefer et al., 2023). The optimization is operated on the attention map at 16×16 resolution, as they are the most semantically meaningful ones (Hertz et al., 2023). Based on the observation that the image semantics are determined by the initial denoising steps (Liew et al., 2022, Kwon et al., 2023), the update is only performed from $t = T$ to $t = t_{end}$, where $T = 50$ and $t_{end} = 25$ in all experiments. The weight of binding loss $\lambda = 1$, if the attribute information is provided. Otherwise, $\lambda = 0$, i.e., using only the attendance loss.

Evaluation Set	Description	Example	# Prompt
Animal-Animal	a [animalA] and a [animalB]	"a cat and a frog"	66
Color-Object	a [colorA] [subjectA] and a [colorB] [subjectB]	"a green backpack and a pink chair"	66
Animal-Scene	a [animalA] and a [animalB] [scene]	"a bird and a bear in the kitchen"	56
Color-Obj-Scene	a [colorA] [subjectA] and a [colorB] [subjectB] [scene]	"a black cat and a red suitcase in the library"	60
Multi-Object	more than two instances in the image	"two cats and two dogs" "three sheep standing in the field"	30
COCO-Subject	filtered COCO captions containing subject related questions from TIFA	"a dog and a cat curled up together on a couch"	30
COCO-Attribute	filtered COCO captions containing attribute related questions from TIFA	"a red sports car is parked beside a black horse"	30

Table 5.1: Description of benchmarks used for the experimental evaluation.

5.4 Experiments

5.4.1 Experimental Setup

Benchmarks. We conduct exhaustive evaluation on seven prompt sets as summarized in Table 5.1. Animal-Animal and Color-Object are proposed in Chefer et al. (2023), which simply compose two subjects and alternatively assign a color to the subject. Building on top of this, we append a postfix describing the scene or scenario to challenge the methods with higher prompt complexity, termed as Animal-Scene and Color-Obj-Scene. Further, we introduce Multi-Object which aims to produce multiple entities in the image. Note that different entities could belong to the same category. For instance, "one cat and two dogs" contains in total three entities and two of them are dogs. Besides the designed templates, we also filtered the COCO captions used in the TIFA benchmark (Hu et al., 2023b) and categorize them into COCO-Subject and COCO-Attribute. There are up to four objects without any attribute assigned in COCO-Subject and two objects with attributes COCO-Attribute, respectively. Note that the attributes in COCO-Attribute contain not only color, but also texture information, such as "a wooden bench".

Evaluation metrics. To quantitatively evaluate the performance of our method,

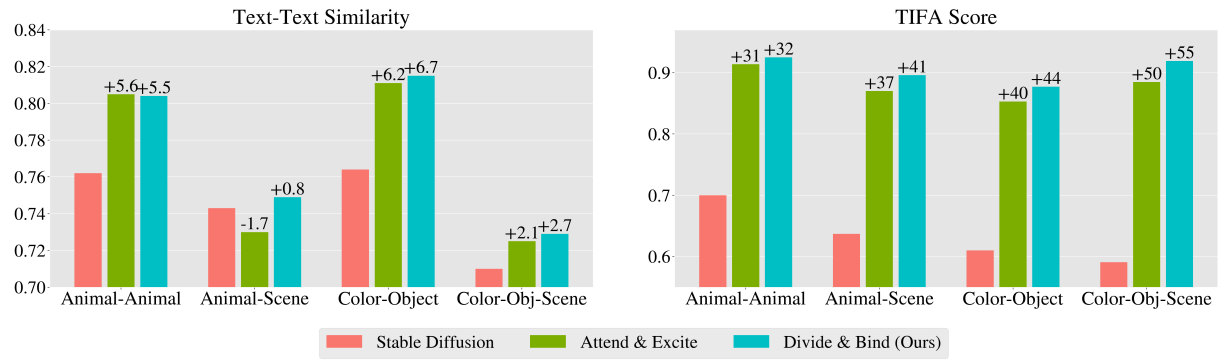


Figure 5.5: Quantitative comparison using Text-Text similarity and TIFA Score. Divide & Bind achieves comparable performance to A&E on the simple Animal-Animal and Color-Object, and shows superior results on more complex text descriptions, i.e., Animal-Scene and Color-Obj-Scene. Improvements over SD in % are reported on top of the bars.

we used the text-text similarity from Chefer et al. (2023) and the recently introduced TIFA score (Hu et al., 2023b), which is more accurate than CLIPScore (Radford et al., 2021) and has much better alignment with human judgment on text-to-image synthesis. To compute the text-text similarity, we employ the off-the-shelf image captioning model BLIP (Li et al., 2022b) to generate captions on synthesized images. We then measure the CLIP similarity between the original prompt and all captions. Evaluation of the TIFA metric is based on a performance of the visual-question-answering (VQA) system, e.g. mPLUG (Li et al., 2022a). By definition, the TIFA score is essentially the VQA accuracy. Given the text input $\mathcal{T}$, we can generate $\mathcal{N}$ multiple-choice question-answer pairs $\{Q_i, C_i, A_i\}_{i=1}^{N}$, where Q_i is a question, C_i is a set of possible choices and A_i is the correct answer. These question-answer pairs can be designed manually or automatically produced by the large-scale language model, e.g. GPT-3 (Brown et al., 2020). By providing a few in-context examples, GPT-3 can follow the instruction to generate question-answer pairs, and generalize to new text captions (Hu et al., 2022, 2023b).

5.4.2 Main Results

As shown in Fig. 5.5, we first quantitatively compare Divide & Bind with Stable Diffusion (SD) (Rombach et al., 2022) and Attend & Excite (A&E) (Chefer et al., 2023) on Animal-Animal and Color-Object, originally proposed in Chefer et al. (2023), as

Method	Multi-Object		COCO-Subject		COCO-Attribute	
	Text-Text	TIFA	Text-Text	TIFA	Text-Text	TIFA
Stable Diffusion	0.786	0.647	0.823	0.791	0.790	0.752
Attend & Excite	**0.809**	0.755	0.818	0.824	0.793	0.798
Divide & Bind	0.805	**0.785**	**0.824**	**0.840**	**0.799**	**0.805**

Table 5.2: Quantitative comparison on complex COCO-captions and Multi-Object generation. Divide & Bind surpasses the other methods when it comes to handling complex prompts.

well as our new benchmarks Animal-Scene and Color-Obj-Scene, which include scene description and has higher prompt complexity. It can be seen that Divide & Bind is on-par with A&E on Animal-Animal and achieves slight improvement on Color-Object. Due to the simplicity of the template, the potential of our method cannot be fully unleashed in those settings. In more complex prompts: Animal-Scene and Color-Obj-Scene, Divide & Bind outperforms the other methods more evidently, especially on the TIFA score (e.g., 5% improvement over A&E in Color-Obj-Scene). Qualitatively, both SD and A&E may neglect the objects, as shown in the "bird and a bear on the street, snowy scene" example in Fig. 5.6. Despite the absence of objects in the synthesized images, we found SD can properly generate the scene, while A&E tends to ignore it occasionally, e.g. the "library" and "kitchen" information in the second column of Fig. 5.6). In the "a green backpack and a pink chair in the kitchen" example, both SD and A&E struggle to bind the pink color with the chair only. In contrast, Divide & Bind, enabled by the binding loss, demonstrates a more accurate binding effect and has less leakage to other objects or background. We provide ablation on the binding loss in Section 5.4.3.

Next, we evaluate the methods on Multi-Object, where multiple entities should be generated. Visual comparison is presented in the third column of Fig. 5.6. In the "three sheep standing in the field" example, both SD and A&E only synthesize two realistic looking sheep, while the image generated by Divide & Bind fully complies with the prompt. For the "one cat and two dogs" example, SD and A&E either miss one entity or generate the wrong species. We observe that often the result of A&E resembles the one of SD. This is not surprising, as A&E does not encourage attention activation in multiple regions. As long as one instance of the corresponding object

Figure 5.6: Qualitative comparison in different settings with the same random seeds. Tokens used for optimization are highlighted in blue. Compared to others, Divide & Bind shows superior alignment with the input prompt while maintaining a high level of realism.

Figure 5.7: Qualitative comparison using novel prompts with the same random seeds. Tokens used for optimization are highlighted in blue. Compared to others, Divide & Bind can better comply with the input prompt while maintaining a high level of realism.

token appears, the loss of A&E would be low, leading to minor update. We also provide the quantitative evaluation in Table 5.2. Our Divide & Bind outperforms other methods by a large margin on the TIFA score, but only slightly underperforms A&E on Text-Text similarity. We hypothesize that this is due to the incompetence of CLIP on counting (Paiss et al., 2023), thus leading to inaccurate evaluation, as pointed out in Hu et al. (2023b) as well.

We also benchmark on real image captions, such as COCO-Subject and COCO-Attribute, where the text structure can be more complex than fixed templates. Quantitative evaluation is provided in Table 5.2, where Divide & Bind showcases its advantages on both benchmarks over SD and A&E. A visual example "a dog and a cat curled up together on a couch" is shown in Fig. 5.6. Consistent with the observation above: while A&E encourages the object occurrence, it may generate unnatural looking images. While SD, may neglect the object, its results are more realistic. Divide & Bind performs well with respect to both perspectives. We provide more visual comparison using additional novel prompts in Fig. 5.7.

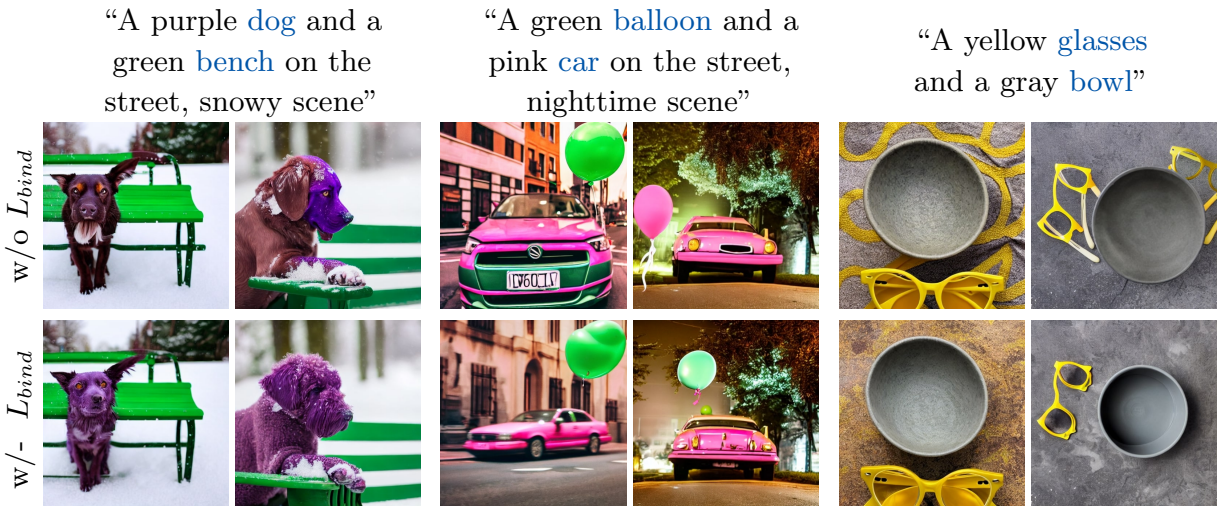


Figure 5.8: Qualitative ablation on the binding loss L_{bind}. With the binding loss, the attribute can be more accurately assigned to the corresponding object.

5.4.3 Ablation Study

We ablate the effect of the proposed binding loss L_{bind} qualitatively and quantitatively, as shown in Fig. 5.8. We observe that the binding loss introduce minor difference on the quantitative evaluation. We hypothesize that the coarse measurement of current evaluation metrics may not be able to reflect the advantage of our method and are not well aligned with human judgement (Hu et al., 2023b, Lu et al., 2024). As illustrated in Fig. 5.8, without the binding loss, the model is able to partially reflect the attribute but may mix with other attributes as well. For instance, in the second column, the front of the car is partially in green, which should be assigned to the balloon. While such imperfect results could still fool current evaluation metrics, as part of the car is indeed in pink. With L_{bind}, we can see the attributes can be more accurately localized at the corresponding object area. Therefore, we employ the binding loss by default, if the attributes are provided in the prompt.

5.4.4 Limitations

Despite improved semantic guidance, it is yet difficult to generate extremely rare or implausible cases, e.g., unusual color binding "a gray apple". Our method may generate such objects together with the common one, e.g., generating a green

Figure 5.9: Limitations: challenging rare combinations (left) and instance miscounting (right).

apple and a gray apple in the same image, see Fig. 5.9. As we use the pretrained model without fine-tuning, some data bias is inevitably inherited. Another issue is miscounting: more instances may be generated than it should. We attribute the miscounting to the imprecise language understanding limited by the CLIP text encoder (Radford et al., 2021, Paiss et al., 2023). This effect is also observed in other large-scale T2I models, e.g., Parti (Yu et al., 2022), making it an interesting case for future research.

5.5 Conclusion

In this work, we propose a novel inference-time optimization objective Divide & Bind for semantic nursing of pretrained T2I diffusion models. Targeting at mitigating semantic issues in T2I synthesis, our approach demonstrates its effectiveness in generating multiple instances with correct attribute binding given complex textual descriptions. We believe that our regularization technique can provide insights in the generation process and support further development in producing images semantically faithful to the textual input.

6 Generative Temporal Nursing for Longer Dynamic Video Synthesis

Despite tremendous progress in the field of text-to-video (T2V) synthesis, open-sourced T2V diffusion models struggle to generate longer videos with dynamically varying and evolving content. They tend to synthesize quasi-static videos, ignoring the necessary visual change-over-time implied in the text prompt. At the same time, scaling these models to enable longer, more dynamic video synthesis often

remains computationally intractable. To address this challenge, we introduce the novel concept of Generative Temporal Nursing (GTN), where we aim to alter the generative process on the fly during inference to improve control over the temporal dynamics and enable the generation of longer videos. We propose a method for GTN, dubbed VSTAR, which consists of two key ingredients: 1) **V**ideo **S**ynopsis Prompting (VSP) - automatic generation of a video synopsis based on the original single prompt leveraging LLMs, which gives accurate textual guidance to different visual states of longer videos, and 2) **T**emporal **A**ttention **R**egularization (TAR) - a regularization technique to refine the temporal attention units of the pre-trained T2V diffusion models, which enables control over the video dynamics. We experimentally showcase the superiority of the proposed approach in generating longer, visually appealing videos over existing open-sourced T2V models. We additionally analyze the temporal attention maps realized with and without VSTAR, demonstrating the importance of applying our method to mitigate neglect of the desired visual change over time. The content of this chapter corresponds to our paper "VSTAR: Generative Temporal Nursing for Longer Dynamic Video Synthesis" (Li et al., 2025), which is published at the International Conference on Learning Representations (ICLR), 2025.

6.1 Introduction

Driven by a whirlwind of activity from both published research and the open-source community, text-to-image synthesis and its natural extension to text-to-video synthesis have undergone remarkable progress in the past few years. Having transformed the idea of content creation, they are now widespread as both a research topic and an industry application. In the realm of text-to-video (T2V) synthesis specifically, recent advancements in video diffusion models (Blattmann et al., 2023, Wang et al., 2023c, Chen et al., 2023, Wang et al., 2023b, Guo et al., 2024, Chen et al., 2024a, OpenAI, 2024) have sparked promising progress, offering improved possibilities for creating novel video content from textual descriptions.

However, despite these advancements, we observe two common issues in current open-source T2V models (Wang et al., 2023c, Chen et al., 2023, Wang et al., 2023b, Guo et al., 2024, Chen et al., 2024a): limited visual changes within the video,

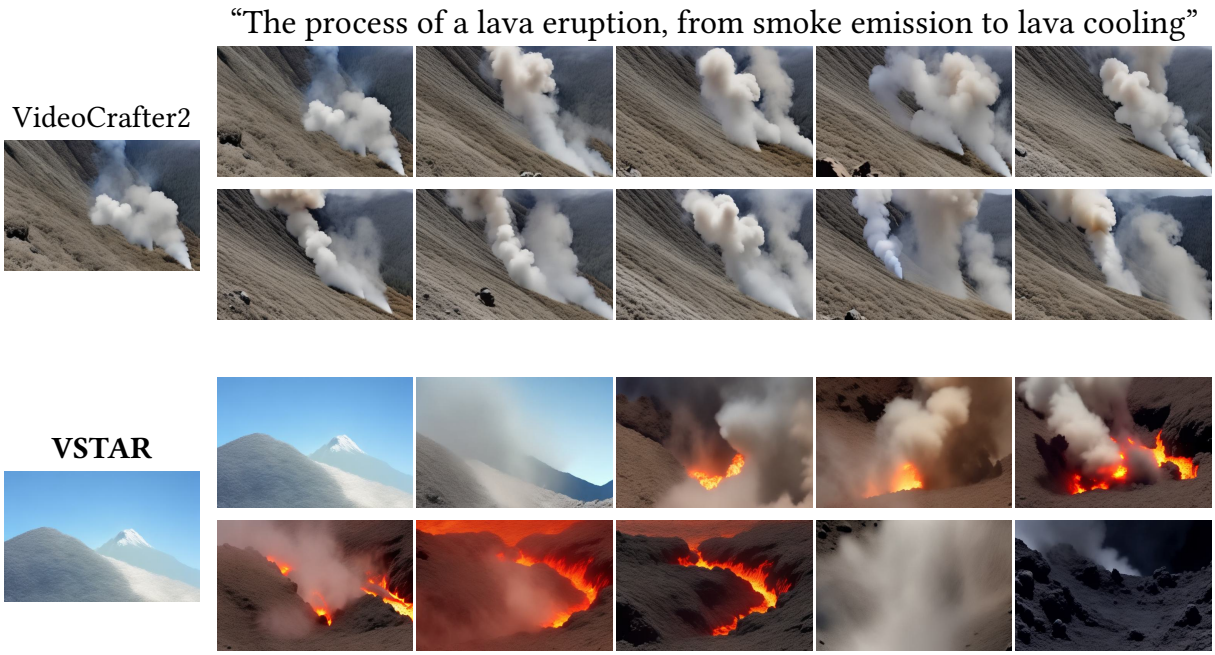


Figure 6.1: Our VSTAR can generate a 64-frame video with dynamic visual evolution in a *single* pass. Images are subsampled from the video. Note that the first column is a GIF, best viewed in *Acrobat Reader*.

and a poor ability to generate longer videos with coherent temporal dynamics. More specifically, the synthesized scenes often exhibit a high degree of similarity between frames (see Fig. 6.1), frequently resembling a static image with minor variations as opposed to a video with varying and evolving content. Additionally, these models do not generalize well to generate videos with more than the typical 16 frames in one pass (see Fig. 6.8). While several recent works attempt to generate long videos in a sliding window fashion (Wang et al., 2023a, Qiu et al., 2024), the methods not only introduce considerable overhead due to requiring multiple passes, but also face the new challenge of preserving temporal coherence throughout these passes.

To mitigate the aforementioned issues, we propose the concept of *"Generative Temporal Nursing"*(GTN), which aims to improve the temporal dynamics of (long) video synthesis on the fly during inference, without re-training T2V models, and using a single pass to not induce a high computational overhead. As a form of GTN, we propose VSTAR, consisting of **V**ideo **S**ynopsis Prompting (VSP) and **T**emporal **A**ttention **R**egularization (TAR).

Current open-sourced T2V models, such as ModelScope (Wang et al., 2023b), LaVie (Wang et al., 2023c) and VideoCrafter (Chen et al., 2023, 2024a), are built upon T2I models, and process all frames within one batch. The single text prompt is conditioned via cross-attention in the spatial transformer of the UNet and shared by all frames. However, it is challenging for the T2V models to transform the semantics from a single prompt into the required visual change across frames, especially when a video with high dynamics is desired, as shown in Fig. 6.1. For dynamic video synthesis faithful to the input prompt, the generation could benefit from a *synopsis* that describes the main events of the video, with explicit descriptions about the desired visual development over time. As a method to provide this guidance and better disseminate the single input prompt across frames, the first strategy of GTN, Video Synopsis Prompting (VSP) leverages the ability of large language models (LLMs), e.g., ChatGPT (OpenAI, 2022), to decompose the single input prompt describing a dynamic transition into several stages of visual development. More specifically, thanks to their in-context learning capability (Brown et al., 2020, Hu et al., 2022), LLMs can be instructed to perform such synopsis prompting automatically by providing a few (or even one) concrete examples. VSP can thus provide the T2V model more accurate guidance on individual visual states, encouraging diversity from the spatial perspective.

Next, we investigate the architectural units of T2V models introduced to capture the temporal interactions between frames. These units, newly incorporated into the T2I backbone, are based on temporal transformers consisting of self-attention layers (Chen et al., 2023, Wang et al., 2023c,b, Guo et al., 2024). Naturally, this temporal attention serves as a critical component in driving the dynamic aspects of video synthesis. Previous work on T2I generation has shown that cross-attention, as the only interaction between the UNet and the input text prompt, can be manipulated to steer the image generation process, e.g. control the image layout or improve attribute binding (Hertz et al., 2023, Feng et al., 2023, Chefer et al., 2023, Li et al., 2023a, Chen et al., 2024b). A resulting natural question is, *can we improve the dynamics of video synthesis by manipulating the temporal attention?* Observing the visual gap between real videos and synthesized ones leads us to compare their temporal attention maps (see Fig. 6.4). We discover that real videos have a band-matrix-like structure, indicating high temporal correlation among adjacent

frames and reduced correlation with frames further apart. Intriguingly, the attention maps of the synthesized ones are less structured, potentially explaining their inferior temporal dynamics.

Inspired by this observation, we propose a simple yet effective Temporal Attention Regularization (TAR) strategy to improve the video dynamics of generated videos. More specifically, we design a symmetric Toeplitz matrix with values along the off-diagonal direction following a Gaussian distribution. The standard deviation of this distribution can control the regularization strength, i.e., the visual variation along the temporal dimension. Adding it to the existing temporal attention maps strengthens the temporal correlation between adjacent frames, while reducing it between more distant frames. Notably, TAR is readily applicable to pre-trained T2V models and requires no optimization, thus introducing no extra inference overhead. Equipped with both strategies, our VSTAR can produce long videos with appealing visual changes in one single pass.

Finally, we analyze the temporal attention mechanisms of different T2V models, establishing valuable connections between their capability to generate longer videos and their architectures. Following the analysis, we offer several training suggestions for enhancing the generalization ability of future models.

In summary, our contributions include:

- We introduce a novel concept of "Generative Temporal Nursing", aiming to improve temporal dynamics, especially for long videos, without requiring any training or introducing high computational overhead at inference time.

- We propose VSTAR, a method for Generative Temporal Nursing, consisting of two simple yet effective strategies: Video Synopsis Prompting and Temporal Attention Regularization, which enable long video generation in a single pass with improved video dynamics.

- We are the first to provide an analysis of temporal attention within video diffusion models, and unleash its potential for controlling the video dynamics. Based on the analysis, we provide insights on how to improve the training of the next generation of T2V models.

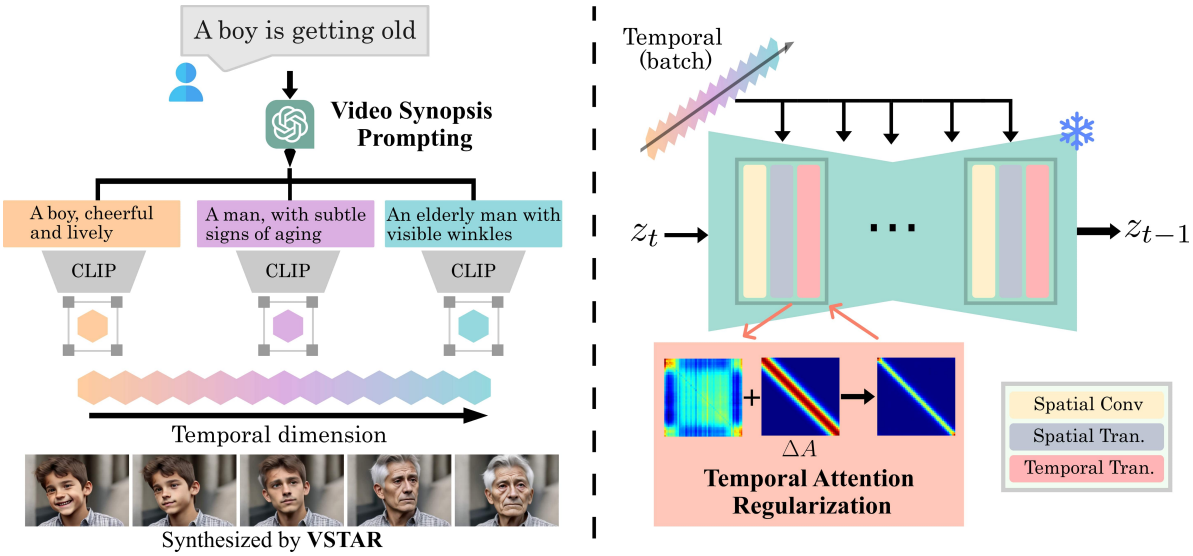


Figure 6.2: Method overview. Our VSTAR consists of two strategies: Video Synopsis Prompting (left) and Temporal Attention Regularization (right).

6.2 Method

Our concept of Generative Temporal Nursing (GTN) aims at improving the video dynamics of pre-trained T2V diffusion models. Besides the text prompt, we identify in Section 6.2.1 that the temporal attention layer is a further key component of T2V models responsible for determining video dynamics. Our first GTN strategy, Video Synopsis Prompting (Section 6.2.2), expands the initial text prompt for the whole video into a sequence of detailed descriptions that control the video progression respectively on different frames. Being inspired by the temporal attention analysis in Section 6.2.3 on real videos, we next design a simple yet effective Temporal Attention Regularization (Section 6.2.4), encouraging the temporal attention of synthetic videos to mimic the attention of real videos.

6.2.1 Preliminary: Text-to-Video Diffusion Model

Current open-sourced text-to-video (T2V) diffusion models (Wang et al., 2023b,c, Chen et al., 2023, 2024a) share a similar high-level design, even if training strategies and specific implementations vary. Based on the text-to-image (T2I) latent diffusion model, e.g., Stable Diffusion (SD) (Rombach et al., 2022), two main changes

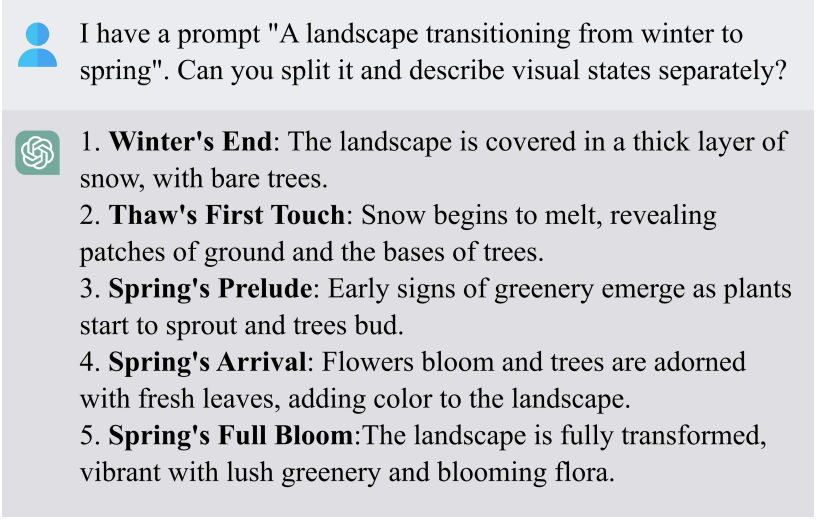


Figure 6.3: An illustration example of VSP. With the aid of LLMs, we can obtain more descriptive video synopsis for key stages.

are introduced for video diffusion models: inflating the 2D UNet to a 3D UNet and adding temporal transformers to capture the requisite temporal relationship found between video frames. With the addition of a temporal axis to the 2D convolutional kernels of SD, the resulting pseudo-3D convolutional layers can handle the input video latent $z \in \mathbb{R}^{N \times C \times H \times W}$, where N is the number of frames and C, H, W represent the channel and spatial dimension of each frame in the latent space, respectively. To generate a video of N frames given a text prompt, current T2V methods (Wang et al., 2023b,c, Chen et al., 2023, 2024a, Guo et al., 2024) process all N frames within one batch, and simply repeat the same prompt embedding for all frames. Inherently, the provided text prompt is conditioned via cross-attention of the spatial transformer in the UNet. The temporal transformer consists of several self-attention layers that operate along the temporal axis. More specifically, the spatial dimension of the intermediate features is merged into the batch dimension, resulting in a shape of $(B \times h \times w, N)$. Since the spatial layers inherited from SD can only handle each frame independently, the temporal attention layers thus play a crucial role for modeling the video dynamics.

6.2.2 Video Synopsis Prompting (VSP)

Similar to T2I models, T2V models shall generate the desired content information based on the text prompt. T2I models already struggle with handling compli-

cated text prompts, particularly when required to properly compose a scene and correctly place relative content spatially (Yang et al., 2023, Feng et al., 2023, Yang et al., 2024a, Wang et al., 2024c). The lack of semantic understanding, reasoning, and planning of the synthesis models results in low quality outputs. The issue becomes more critical when moving from image to video synthesis, as the evolution of the scenes must also now be considered. For example, the text prompt "A landscape transitioning from winter to spring" is highly abstract; the seasonal change from winter to spring inherently can consist of several visual states. As shown in Fig. 6.8, the SOTA T2V model VideoCrafter2 (Chen et al., 2024a) fails to generate such dynamic changes.

Inspired by the creation of long dynamic videos in real life, we propose to offload the task of interpreting the text prompt, reasoning about it, and creating a video synopsis to LLMs. This task can be effectively managed in the language space, where LLMs have presented strong generalization across various tasks. When we ask ChatGPT (OpenAI, 2022) to parse the same text prompt, i.e., "A landscape transitioning from winter to spring", into a sequence of text descriptions that well describe the dynamics, the result is more convincing and semantically informative as shown in Fig. 6.3. This can be done by leveraging the in-context learning capability (Brown et al., 2020, Hu et al., 2022) of LLMs, where we guide them to perform the video synopsis prompting task automatically through prompting with a single concrete example. For instance, we can instruct ChatGPT (OpenAI, 2022) with the following prompt:

> I have a prompt "A landscape transitioning from winter to spring" for video generation. Can you split the process and describe the states separately? Each state is described in only one sentence and please consider the coherency between sub-prompts. Please be straightforward and do not use a narrative style. For example, for prompt "a boy is getting old", it can be divided into two states, e.g., "a young boy" and "an old man".
> Based on this example, can you provide the description? The number of states is not limited to two.

Subsequently, ChatGPT can provide a detailed video synopsis that includes multiple visual states. Once the LLM has learned such a task, we can then simply

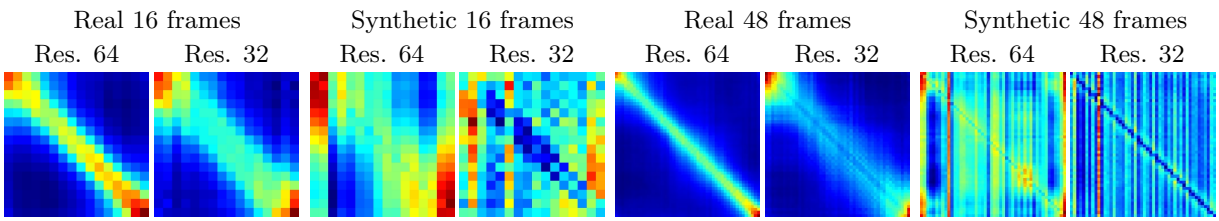


Figure 6.4: Temporal attention visualization of real and synthetic videos of 16 and 48 frames. Attention of real videos exhibits a band-matrix like structure, indicating high correlation with adjacent frames. Synthetic videos exhibit less-structured attention maps, especially for 48 frames, which explains the low quality of long video generation.

Figure 6.5: Per-layer temporal attention analysis. We replace the temporal attention maps at different resolutions with a diagonal matrix (1st row) and an all-ones matrix (2nd row), which leads to a more dynamic or a more static video, respectively. We observe that high resolution attention has a larger impact on the video dynamics. Note that this is a GIF, best viewed in *Acrobat Reader*.

prompt it to execute the task without reiterating the examples:

> I have a prompt "A peony starts to bloom, in the field". Can you split the process and describe the states separately?

It is sufficient to generate text descriptions for the main event changes in a video rather than for each frame. A text encoder e.g., CLIP text decoder (Radford et al., 2021), is then applied to extract the text embeddings of these descriptions, which are then interpolated to guide each frame's synthesis via cross attention as illustrated in Fig. 6.2. This process yields more accurate guidance for transitioning visual stages, while ensuring smooth conditioning without abrupt changes between frames.

6.2.3 Temporal Attention Analysis

To properly synthesize videos that capture the dynamics conveyed in the input prompt, we delve into the synthesis model itself. An examination of the components new to T2V models, beyond the common building blocks already used in T2I models, leads naturally to the temporal attention layers. These new modules are crucial for facilitating proper video synthesis, i.e., generating sequential frames with dynamic yet consistent content that reflect the input text information. We hypothesize that the ineffectiveness of current T2V models arises from unstructured interactions among frames in the same video within the temporal attention layers. To verify our hypothesis, we conduct a systematic analysis comparing the attention maps of real and synthetic videos. Specifically, the attention map A is expressed as:

$$A = Softmax\left(\phi\left(Q, K\right)\right) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right) \in \mathbb{R}^{N\times N}, \tag{6.1}$$

where Q and K represent the query and key of the self-attention layer, and d is the latent dimension. This attention matrix essentially depicts the pairwise correlation between the N frames of one video. For real videos, their attention maps can be obtained by adding noise to their clean latent and extracting the attention during the denoising process. For synthetic videos, we can read out their attention maps directly during their synthesis passes.

As shown in Fig. 6.4, for both 16 and 48 frame real videos, the attention matrix manifests as a band-matrix-like structure. Intuitively, closer frames should have a higher correlation with each other to maintain temporal coherency. Compared to real videos, attention matrix of the synthetic ones is less structured, especially for 48 frames. That explains why the model generalizes even worse to longer videos. High correlation is spread across a wide range of frames, resulting in a harmonized sequence with similar appearances.

Further, we conducted a per-resolution ablation as shown in Fig. 6.5. We replace the attention map at each individual resolution, i.e., 64, 32, 16, and 8, while keeping the other resolution untouched. We experiment with two extreme cases: using the Identity matrix (I_N) and the all-ones matrix (J_N). The former regularizes the frames to be mutually independent, while the latter oppositely requires full correlation, i.e.,

static sequence. The observations from Fig. 6.5 are highly consistent. When utilizing I_N to encourage independence among frames, the temporal coherence of the synthesized frames is indeed compromised. Conversely, employing J_N can significantly diminish the video dynamics, leading to a quasi-static video. This controlled experiment clearly demonstrates how the temporal attention layer impacts the dynamics of the video synthesis model.

Finally, we investigate the effect of the interplay between attention and resolution on the content dynamics of videos. As also shown in Fig. 6.5, the replacement at the higher resolutions of 64 and 32 has a more evident effect than at lower resolutions. Applying the changes jointly at both resolutions, 64 & 32, further amplifies the effect. In contrast, the videos are much less responsive to the attention replacement at resolution 8. Likely, the low resolution features encode high-level semantics, while with higher resolution features there is more capacity for representing varying local details in the scene; such details are necessary for reflecting coherent change over frames.

Based on these controlled experiments, we can conclude that manipulating temporal attention allows us to alter the video dynamics, i.e., making the visual process either more static or more dynamic. In particular, adjustments at higher resolutions, e.g. 64 & 32, are more effective.

6.2.4 Temporal Attention Regularization (TAR)

From the experiments above, we have clearly observed the role of temporal attention layers in determining the dynamics of videos. Naturally, the attention matrices of synthetic videos should be similar to that of real videos. Therefore, we propose a simple regularization technique applied on the temporal attention layers for pretrained T2V model. Note that, our proposal is directly applied to pretrained T2V models without requiring re-training, and incurs no additional optimization costs during inference.

As illustrated in Fig. 6.4, the attention correlation of the real video resembles a band-matrix-like structure, with high correlation between neighboring frames and lower correlation the larger the frame offset. To approximate such a structure, we design a symmetric Toeplitz matrix as the regularization matrix ΔA, with its values

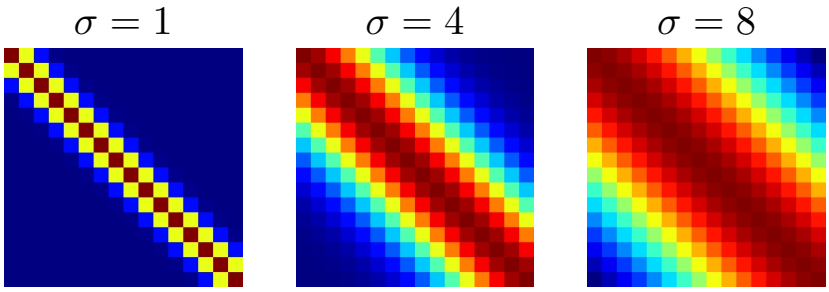


Figure 6.6: Visualization of regularization matrix ΔA with different standard deviation σ. A Smaller σ can enhance the effect of regularization.

along the off-diagonal direction following the Gaussian distribution:

$$\Delta A_{i,j} = e^{-\frac{1}{2}(\frac{j-i}{\sigma})^2}, \tag{6.2}$$

where $i, j \in \{1, ..., N\}$ represent the entry index of the attention regularization map, and σ is the standard deviation of the normal distribution. Regularization matrices with different σ are visualized in Fig. 6.6. As indicated in Fig. 6.14, the standard deviation σ can control the regularization strength, i.e. larger σ leading to less visual variations along the temporal dimension. This regularization matrix is then added to the original attention matrix in (6.1), i.e.

$$A' \leftarrow Softmax\left(\phi(Q,K) + \max[\phi(Q,K)] \cdot \Delta A\right). \tag{6.3}$$

To balance both terms, we additionally introduce $\max[\phi(Q,K)]$, which weights ΔA based on maximum in the attention matrix $\phi(Q,K)$. As illustrated in Fig. 6.2, the regularized attention map A' will be inserted back for further processing.

With the combination of both VSP and TAR, our VSTAR can effectively provide temporal nursing for video generation, enabling the synthesis of long videos with appealing visual evolution using pretrained T2V models, while also introducing no optimization overhead. We find temporal attention analysis to be a powerful tool for understanding the temporal modeling of video diffusion models and leverage it to analyze other T2V models in the next section. We establish valuable connections to their architecture designs, and provide guidance for the future training of T2V models for long video generation.

6.3 Experiments

Experimental setting. To demonstrate the effectiveness of VSTAR in creating more dynamic videos, we run experiments and ablations on prompts, generated by ChatGPT (OpenAI, 2022), that describe various visual transitions. By default, we employ the state-of-the-art open-sourced T2V model VideoCrafter2 (Chen et al., 2024a) with 320×512 resolution as our base model, which is combined with the proposed video synopsis prompting and temporal attention regularization. We refer to this combination as our method or VSTAR throughout the experiments.

6.3.1 Main Results

Comparison with other T2V methods. In Fig. 6.7 and Fig. 6.8 we compare our VSTAR with other commonly used T2V models, namely, ModelScope (Wang et al., 2023b), LaVie (Wang et al., 2023c) and AnimateDiff (Guo et al., 2024), for both 16 and 32 frame generation. For a fair comparison, we use the base model of LaVie without its cascaded components, e.g., the video super-resolution model. Although all methods are able to generate meaningful results for 16-frame videos (see Fig. 6.7), the videos created by the other T2V models do not properly reflect the visual content specified by the input prompt. Given "A Ferrari driving on the road, starts to snow", the other methods tend to focus on one particular state, e.g., the snowy scene, lacking dynamic progression throughout the video. In contrast, our VSTAR appropriately captures the weather transition from a clear day to a snowy one.

When generating 32 frames in one pass, as shown in Fig. 6.8, our method exhibits even greater advantages. The comparison methods yet again fail to generate content corresponding to the given prompt, but this time to the extent that the visual quality of the individual frames is also greatly compromised. In contrast, our VSTAR is able to generate long videos with dynamic visual evolution. More qualitative results synthesized by VSTAR are provided in Fig. 6.9. Based on these results, with a desire to further understand why other T2V models generalize poorly to long video generation, we analyze the temporal attention of these models, as detailed in the following paragraph.

Comparison on inter-frame similarity with real videos. To quantitatively as-

Figure 6.7: Comparison with other T2V models on 16 frames generation. Our VSTAR can synthesize desired visual development from a clear day to snowy scene, while the others tend to generate the final visual state, i.e., snowy day. Note that the first column is a GIF, best viewed in *Acrobat Reader*.

sess inter-frame similarity in a video, we calculate the perceptual similarity between every pair of frames using the recently proposed metric DreamSim (Fu et al., 2023), which has been demonstrated to align closely with human judgment. In Fig. 6.10, we plot the similarity matrices of real videos and those synthesized by VideoCrafter2 and our VSTAR; the values in the matrix are normalized across all methods. VideoCrafter2 exhibits very high similarity across all frames, suggesting minimal visual dynamics, which is aligned with qualitative results. Our VSTAR on the other hand mimics the perceptual similarity correlation of real videos, affirming the effectiveness of our proposal for nursing the video dynamics.

Observing the resemblance between the temporal attention maps of the real videos and their similarity matrices, we attempt to directly employ a DreamSim-based similarity matrix as ΔA for regularization. As shown in Fig. 6.11, this improves the temporal dynamics, leading to a gradual appearance of the rainbow.

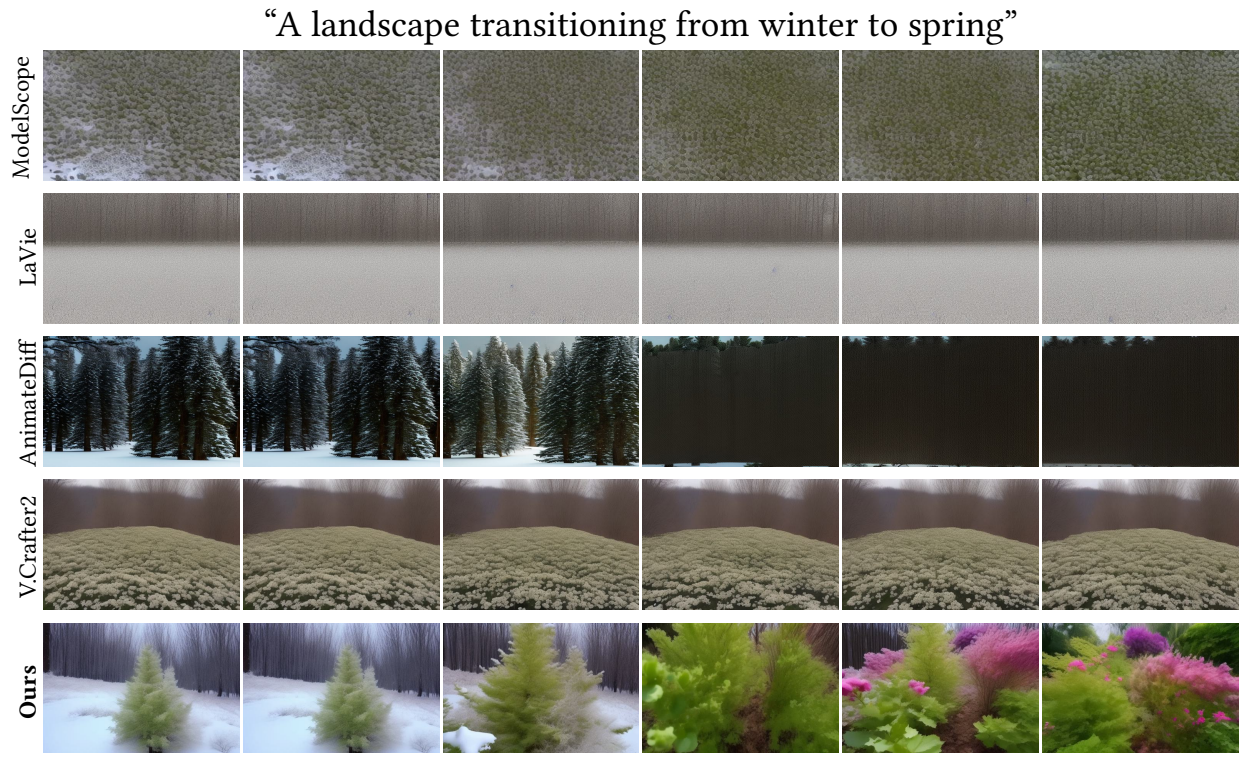


Figure 6.8: Comparison with other T2V models on 32 frames generation, which is double the length of the default option. Our VSTAR can generate long videos with desired dynamics, while the others struggle to synthesize faithful results.

Temporal attention analysis of other T2V models. In Fig. 6.12, we visualize the temporal attention layers of ModelScope (Wang et al., 2023b), LaVie (Wang et al., 2023c) and AnimateDiff (Guo et al., 2024). It can be seen that ModelScope exhibits similar attention behavior to VideoCrafter (see Fig. 6.4), in that the temporal correlation significantly deteriorates when generating longer videos. This is noticeable even for videos of 32 frames, twice the length of the standard option, and aligns with the qualitative comparison in Fig. 6.8. AnimateDiff (Guo et al., 2024) and LaVie (Wang et al., 2023c) demonstrate different temporal attention behavior, due to the incorporation of Rotary Positional Encoding (Touvron et al., 2023) in the former and Sinusoidal Positional Encoding in the latter. With the positional encoding, the models learn better temporal correlation among neighboring frames for 16 frames, showing a band-matrix structure more closely resembles that of real videos. However, when generating videos longer than its training capacity, the

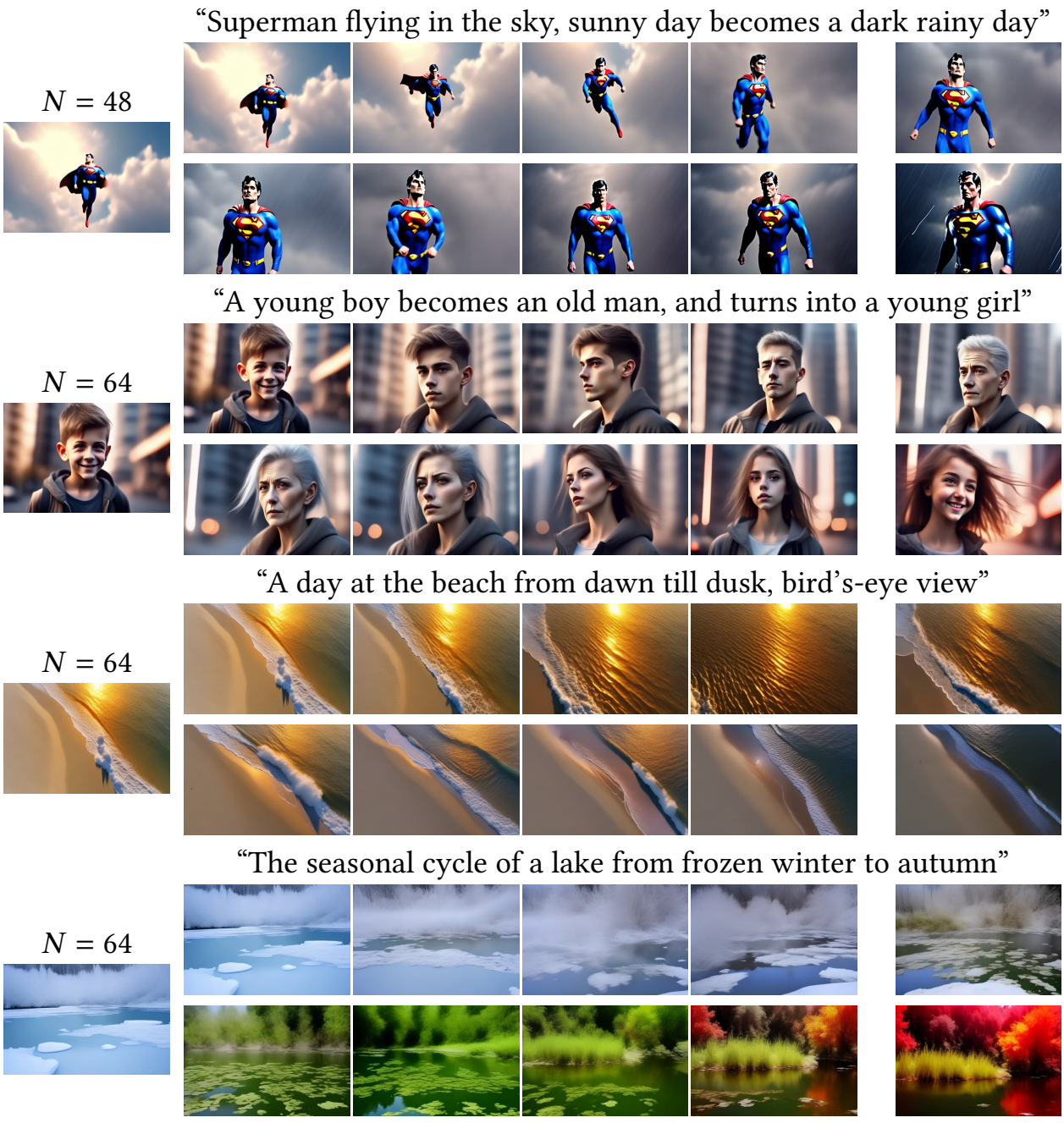


Figure 6.9: Qualitative results of videos with 48 and 64 frames synthesized by VSTAR. Images are sub-sampled from the sequence. Note that the first column is a GIF, best viewed in *Acrobat Reader*.

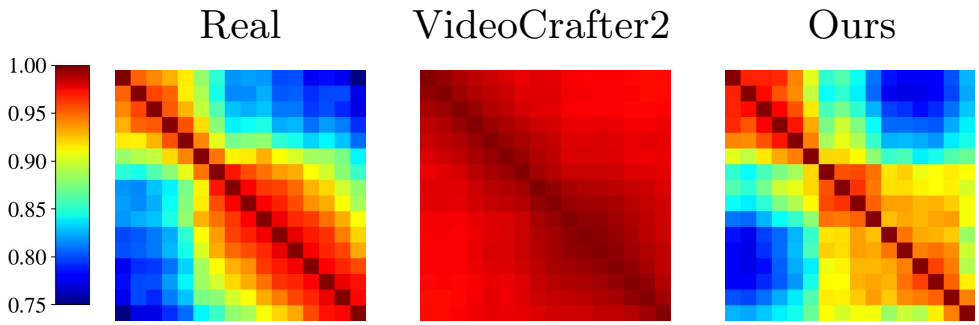


Figure 6.10: Inter-frame perceptual similarity matrix based on DreamSim Fu et al. (2023), where values are normalized across *all* methods. VideoCrafter2 has high similarity across nearly all frames, which is aligned with the visual results lacking variation. In contrast, our synthesized videos highly resemble the real ones, indicating desired dynamics.

Figure 6.11: Regularization with inter-frame DreamSim matrix of one real reference video.

model faces considerable difficulty in preserving the desired temporal dynamics, resulting in inferior synthesis quality, as depicted in Fig. 6.8. The Rotary Positional Encoding employed in LaVie is a form of relative positional encoding, i.e., it depends on the relative offsets of frames, which could explain the periodic pattern seen in the attention maps. While the Sinusoidal Positional Encoding used in AnimateDiff is based on the absolute frame index, leading to the model failing completely for indices unseen during training (past 16). These observations concerning T2V models are interestingly aligned with prior studies regarding Positional Encoding on length generalization in Transformers (Kazemnejad et al., 2023) in the context of LLMs.

This comparison offers valuable insights into improving the training of the next generation of T2V models. For instance, omitting positional encoding can improve

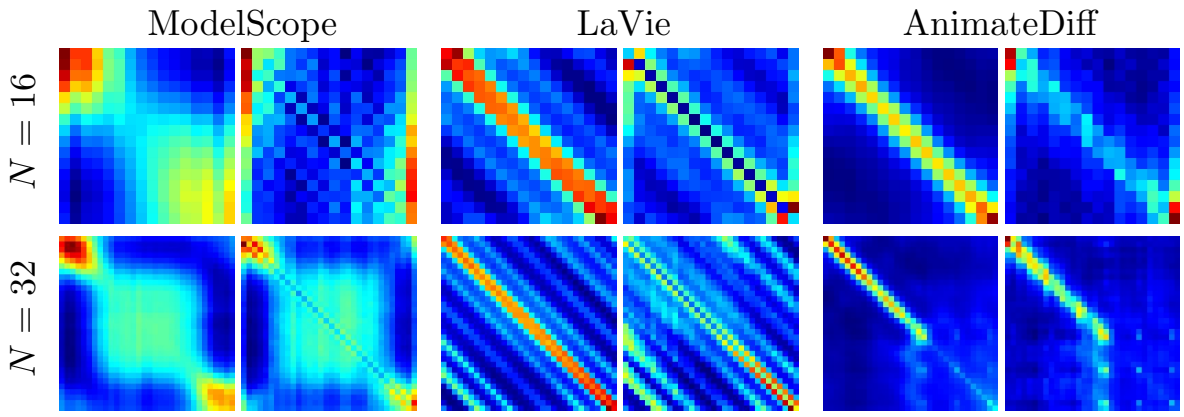


Figure 6.12: Temporal attention visualization of other T2V Models for the default 16 frame and longer 32 frame videos. ModelScope has similar issues to VideoCrafter2 (see Fig. 6.4), i.e., high correlation spread across many frames, especially for $N = 32$. LaVie and AnimateDiff incorporate positional encoding of frame indices, thus naturally do not generalize well to long video generation beyond trained 16 frames.

generalization capability, and incorporating a regularization loss on the temporal attention maps can help to enforce the desired temporal dynamics. Alternatively, one can employ a better combination of data format and positional encodings, as explored in the recent work (Zhou et al., 2024), which achieves improved length generalization. For instance, Randomized Positional Encoding (Ruoss et al., 2023) can help to avoid overfitting on the position indices, and mixing up subsampled video sequences can further strengthen local correlations. Combining such techniques may improve the generalization to long video generation.

6.3.2 Ablation Study

Ablation on the effect of TAR and VSP. We investigate the effects of the proposed Temporal Attention Regularization and Video Synopsis Prompting individually in Fig. 6.13, where we generate videos of 48 frames in one inference pass based on the prompt "Spiderman on the beach from morning to evening", using the same initial noise. The synthesized video clips are presented in the first column as GIFs; the other images are subsampled from the full sequence. The baseline model VideoCrafter2 struggles to synthesize a video faithful to the input prompt, generating a sequence of highly similar frames, with a stride-like texture in the

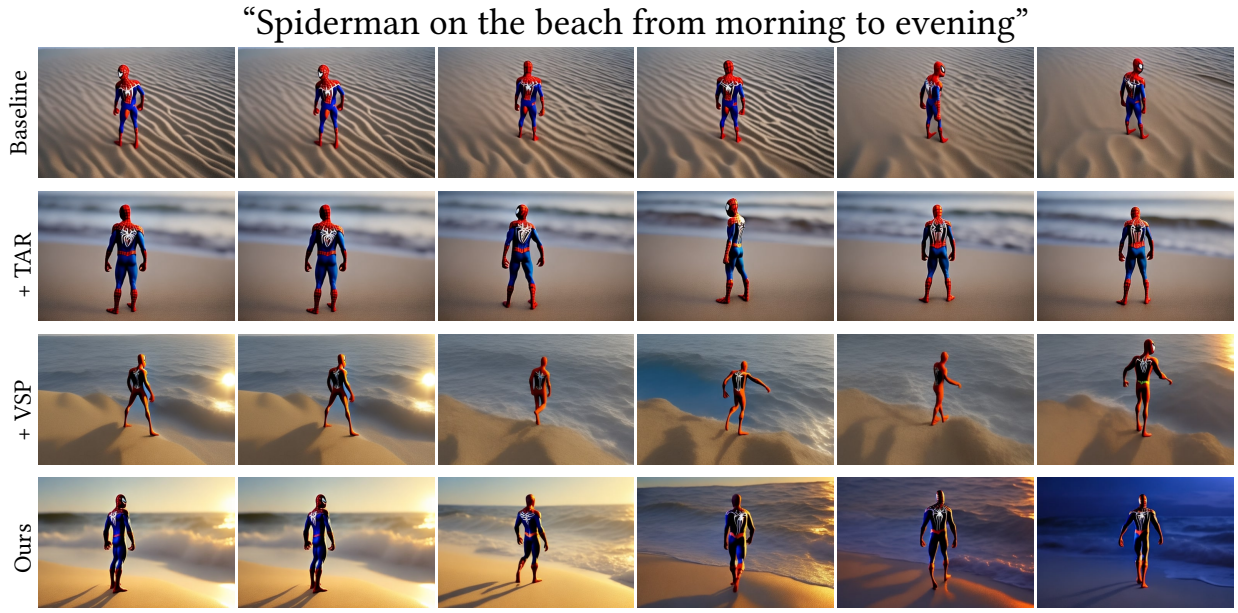


Figure 6.13: Ablation on the effect of Video Synopsis Prompting (VSP) and Temporal Attention Regularization (TAR). Subsampled from 48 frames. Combination of TAR and VSP effectively enables long video generation with desired visual evolution. While individual strategy improves upon the baseline, there still lacks of desired dynamics.

background, that fail to depict the time-lapse video. When employing the TAR, the model generates a more realistic sequence, however without the desired visual evolution; the single plain prompt is insufficient to describe the scene changes. Interestingly, while VSP provides a more descriptive summary of different visual states, without TAR, the temporal attention remains strongly correlated. The model then attempts to depict the provided textual description, however with limited visual variation. When combining both strategies, our VSTAR can effectively synthesize the desired visual content, exhibiting improved dynamics with a more appealing time-lapse effect.

Ablation on regularization matrix. We further ablate by investigating the effect of using a different standard deviation σ in the regularization matrix ΔA, shown in Fig. 6.14. We start from applying regularization at the highest temporal resolution i.e., 64, since high-resolution temporal attention more greatly influences the video dynamics, as demonstrated in the temporal attention analysis in Section 6.2.3. The results show that decreasing σ results in a stronger regularization effect, inducing

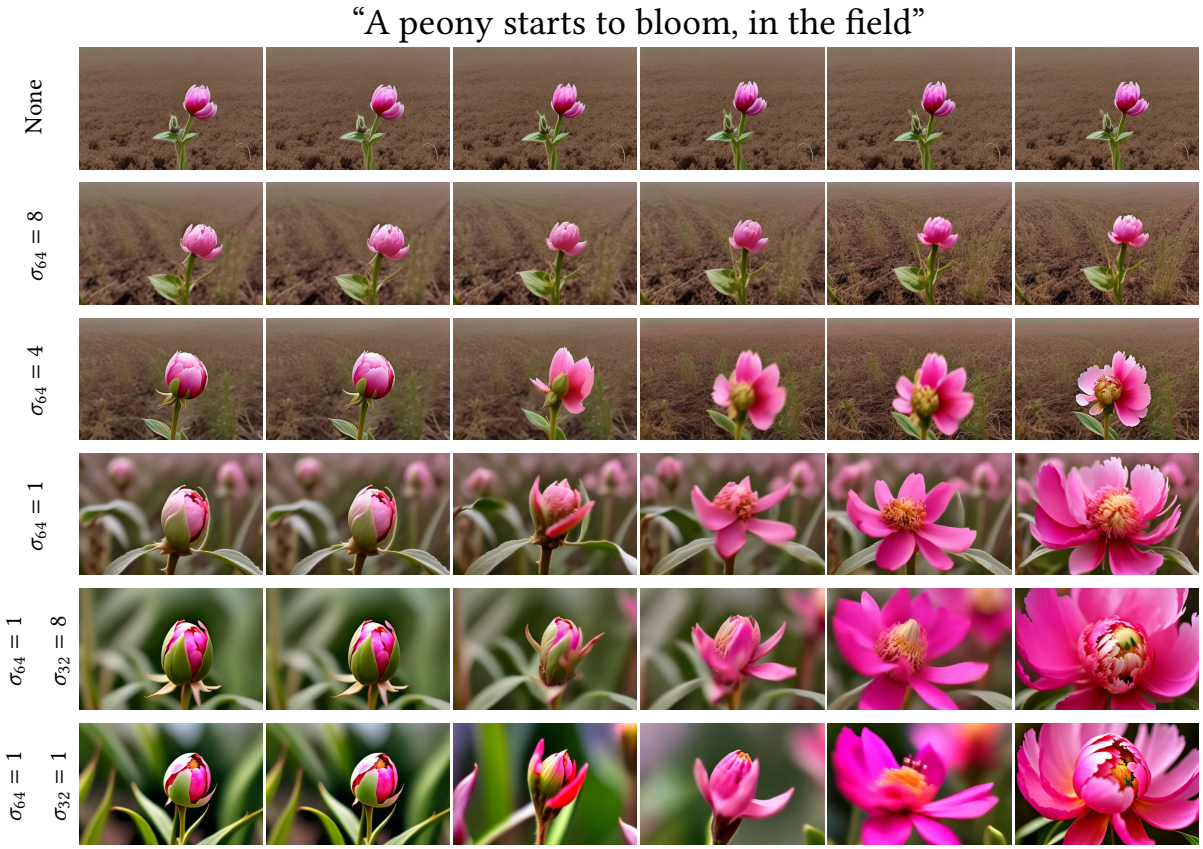


Figure 6.14: Ablation of attention regularization matrix ΔA. Smaller σ induces a stronger regularization effect, leading to increasing temporal dynamics. When applying regularization at both 64 & 32, the video becomes more dynamic, i.e., the peony is fully bloomed. Yet, excessive regularization, i.e., $\sigma_{64} = \sigma_{32} = 1$, can leave the impression of temporal incoherency. Contains GIFs, best viewed in *Acrobat Reader*.

more pronounced visual changes throughout the video (e.g. compare row 2 to row 4, and notice the extent of the blooming of the flower). Going one step further, applying regularization also at a resolution of 32 results in the peony reaching its fullest bloom. However, when equally strong regularization is applied at both a resolution of 64 and 32, i.e., $\sigma_{64} = \sigma_{32} = 1$, the visual changes can be too excessive, leaving the impression of poor temporal coherency across frames. Empirically, we find that applying $\sigma_{64} = 1$ strikes a good balance between dynamic changes and temporal coherency.

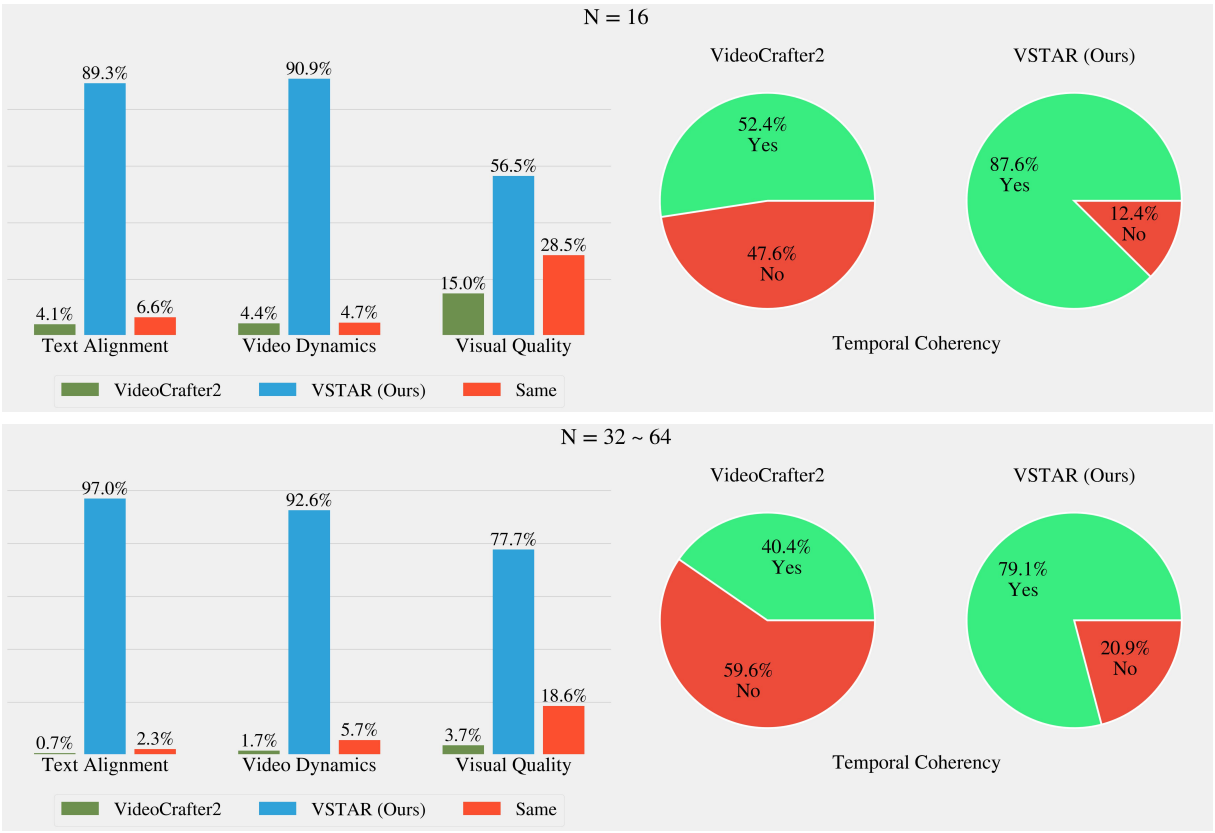


Figure 6.15: User study on both standard 16 frames and longer videos with 32 ~ 64 frames. For the first three aspects, participants review pairs of videos, choosing between them or rating them as the same. For temporal coherency, the numbers are the absolute probability that a participant perceives the video from the respective method as having smooth temporal progression.

6.3.3 User Study

For further evaluation, we conducted a user study to compare our VSTAR with the SOTA T2V model VideoCrafter2 (Chen et al., 2024a). 110 individuals with diverse backgrounds participated in the user study, working in fields such as computer vision, reinforcement learning, natural language processing, art design, medical engineering, mechanical engineering, and administrative management, among others. We assess the videos across four dimensions: text alignment, video dynamics, visual quality and temporal coherency. Text alignment concerns whether the synthesized results properly reflect the input text prompt. Video dynamics examines the dynamic visual changes within the progression of the video. A higher visual quality indicates fewer artifacts and distortions, leading to a more visually pleasing

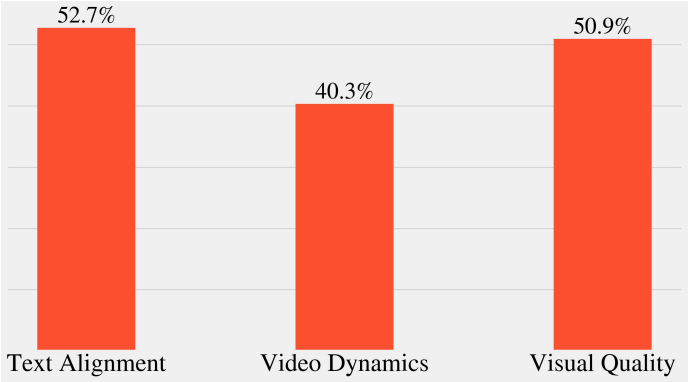


Figure 6.16: User study on paired of videos, both generated by our VSTAR, to verify the consistency of our method's improvement, making it challenging for users to make a clear choice. Indeed, a large number of participants perceived both videos as identical across all three aspects. The rest had diverse preferences between the two videos. This demonstrates the consistency of our synthesis results and their closely matched quality.

result. Temporal coherency evaluates if the result is temporally smooth, i.e., there are no abrupt or unexplained changes that could disrupt the viewing experience. For the first three aspects, participants are presented with paired results to evaluate, selecting one over the other or deeming them equivalent. Regarding temporal coherency, we pose a simple yes-or-no question, asking whether the participants perceive the video as being temporally smooth.

The outcome is summarized in Fig. 6.15. Our VSTAR emerges as the preferred choice across various frame lengths from all aspects, with its advantages becoming more pronounced in the generation of longer videos with $N = 32 \sim 64$. Importantly, our method not only enhances video dynamics but also preserves temporal coherency. A majority of participants confirmed that our results exhibit smooth temporal transitions, with 87.6% for standard-length videos and 79.1% for longer videos agreeing to this assessment. This favorable reception surpasses the baseline VideoCrafter2, possibly as a result of its less engaging content.

Additionally, we included pairs of videos, both generated by VSTAR, to verify the consistency of our method's improvement, making it challenging for users to make a clear choice. As shown in Fig. 6.16, participants indeed often found it difficult to differentiate, with 52.7%, 40.3% and 50.9% of them rating both videos as equal in terms of text alignment, visual dynamics, and visual quality, respectively. The remaining participants were divided in their preference between the two videos.

This indicates that our synthesis results are consistent and display a narrow gap between them.

6.3.4 Discussion

Limitations. VSTAR offers a simple yet effective solution for improving pretrained T2V models, however, there are fundamental issues of pretrained models that may not be completely resolved via generative nursing at inference time only. Although our VSTAR has eased the process of reasoning prompts that involve dynamic evolution, the model can still struggle with responding to the decomposed open-world prompts, resulting in visuals that are not aligned with the prompt, potentially due to limited capability of the text encoder (Podell et al., 2024, Liu et al., 2024). Nevertheless, several recent works (Chefer et al., 2023, Agarwal et al., 2023) as well as our approach (Li et al., 2023a) described in Chapter 5 have employed on-the-fly latent optimization to improve the textual alignment of a frozen T2I model. One may explore the combination of VSTAR with such techniques for further improvement.

Potential negative societal impact. Given the imbalanced nature of large-scale datasets, pretrained T2V models may inherit certain data biases, inaccurately representing the diversity of the overall population. These biases can potentially reinforce existing societal stereotypes and inequalities. Therefore, it is advisable to undertake proactive steps to identify and mitigate such biases, which may include the involvement of human reviewers in sensitive contexts.

6.4 Conclusion

In this paper, we contribute two simple concepts, Video Synopsis Prompting (VSP) and Temporal Attention Regularization (TAR), that, when employed together, facilitate the generation of longer (e.g. 64 frames), temporally coherent videos with improved dynamics. We show the benefit of both VSP and TAR on diverse prompts and in comparison to the state of the art, and ablate on the employed TAR regularization matrix. Besides motivating TAR, our analysis of temporal correlation in real videos may offer valuable insights towards improving design and training of the next generation of T2V models. For example, some form of positional encoding appears to be hampering generalization capability, while the incorporation of

a regularization loss on temporal attention maps can help to enforce temporal dynamics. While VSTAR is readily applied to pretrained T2V models, future work may incorporate it during training for improved procedural dynamics, such as complex activities on respective data.

7 Conclusion and Future Perspectives

This thesis has delved into the rapidly evolving field of generative modeling, with a focus on image and video synthesis using GANs and diffusion models. In this work, we have addressed key challenges related to alignment and controllability during the generation process, proposing several novel approaches that push the boundaries of AI-generated content. In this chapter, we begin by summarizing our contributions in Section 7.1, where we highlight the advancements made through

our research. Following this, we discuss the future perspectives of generative modeling and their broader impact in Section 7.2.

7.1 Summary

In this work, we have proposed several novel techniques, aiming at improving the alignment and controllability of generative models, such as GANs and diffusion models. These advancements have been applied across four different tasks: GAN inversion, layout-to-image, text-to-image, and text-to-video synthesis. Our research pushes the boundaries of visual synthesis, allowing greater control over generated outcomes and fostering the practical integration of synthetic data into real-world applications. In what follows, we summarize the individual contributions of each chapter.

7.1.1 Exemplar-Based Synthesis with Content-Style Disentanglement

In Chapter 3, we worked on GAN inversion for more complex scene-centric datasets, which enables many real-life applications of great practical value. Our method contains two key enablers: spatial noise prediction and random noise masking. Rather than encoding the given image solely into latent vectors, we also predict noise maps with spatial dimensions, which greatly enhances reconstruction quality. However, these noise maps can be so expressive that they render the latent vectors redundant, leading to poor editability. To address this, we randomly mask noise during encoder training, which encourages the model to rely on the latent vectors for reconstruction. The proposed Masked Noise Encoder not only achieves superior fidelity but also features style mixing capabilities, with style and content information respectively encoded into the latents and the noise map. Notably, our encoder exhibits excellent plug-n-play ability and can be readily applied to unseen data. In addition to delivering impressive visual results, we demonstrated the effectiveness of our pipeline in real-world applications, such as enhancing domain generalization and model validation for various semantic segmentation networks.

7.1.2 Improved Layout-To-Image Diffusion Models via Adversarial Supervision

In Chapter 4, we focused on layout-to-image (L2I) diffusion models. Prior GAN-based methods generally suffer from the mode collapse issue, characterized by the limited diversity of their outputs. Large-scale pretrained diffusion models, such as Stable Diffusion (Rombach et al., 2022), have demonstrated astonishing capabilities for synthesizing diverse data. Consequently, one would naturally consider adopting them for the L2I task. Prior to our work, attention was primarily focused on architecture design to incorporate the layout condition into the diffusion model. Our work, however, explored training strategies for the L2I diffusion model, which are independent of the architecture. We observed that L2I diffusion models often align poorly with the input layout condition, ignoring the provided semantic class information. We attribute this to the suboptimal training objective that fails to explicitly consider the layout condition. We designed ALDM, which consists of two novel training strategies to enhance alignment: adversarial supervision and multistep unrolling. First, we employed a segmentation network-based discriminator to provide per-pixel guidance, explicitly leveraging the ground-truth label map condition. Furthermore, we introduced a multistep unrolling mechanism, which involves unrolling backward multiple steps over a specified time window to imitate the inference time sampling. In this way, ALDM promotes consistent adherence to the conditional layout across the sampling time horizon, resulting in significantly improved alignment in the output. We applied ALDM to various adaptation methods, e.g., ControlNet, T2I-Adapter, and demonstrated consistent improvement over the baselines, underscoring the effectiveness of the proposed training strategies. Remarkably, our work (Li et al., 2024a) provides a powerful data generator that is beneficial for training downstream models in various fields such as robotic grasping (Li et al., 2024b).

7.1.3 Improved Generative Semantic Nursing for Text-To-Image Synthesis

In Chapter 5, we focused on text-to-image (T2I) synthesis, aiming at mitigating semantic alignment issues of a pretrained T2I model. One of the primary challenges

in T2I synthesis is the frequent occurrence of object missing and attribute binding issues, where the generated image fails to include all the objects mentioned in the textual prompt or incorrectly assigns attributes to objects, leading to visual inconsistencies and a lack of fidelity in the generated content. To address these issues, we proposed two novel optimization objectives, namely, *divide* and *bind* loss, for generative semantic nursing, where the latent is updated on the fly to improve the faithfulness of the outcome. The divide loss aims to incentivize the presence of objects by maximizing the total variation of the attention map, encouraging the model to distribute attention across regions, thereby allowing it to generate different objects. The bind loss explicitly aligns the spatially normalized attention distribution of the subject token with its corresponding attribute token using Jensen–Shannon divergence, which significantly improves the accuracy of the attribute assignment. Through extensive experiments, we demonstrated that our Divide & Bind can effectively generate multiple instances with the desired attributes, resulting in significant improvements in semantic alignment and the overall quality of the generated images, particularly when dealing with complex textual descriptions. This enhanced precision opens up new possibilities in applications like art design, where the ability to accurately translate intricate ideas into visual content is crucial.

7.1.4 Generative Temporal Nursing for Longer Dynamic Video Synthesis

In Chapter 6, we targeted at generating longer dynamic videos using a pretrained text-to-video (T2V) model. As a newly emerging and increasingly popular field, current open-source T2V models are still in their early stages, with limited capabilities and some notable limitations. Among them, we focused on two key issues: the limited variation within generated clips and the difficulty in generalizing these models to produce longer video sequences. More specifically, the synthesized scenes often exhibit high similarity between frames, resembling a static image with minor variations, despite the text prompt specifying an evolving process. Additionally, these models by default can only generate 16 frames in one single inference pass, as they were trained on short video clips. They cannot straightforwardly produce longer videos by simply setting a larger length.

To tackle these challenges, we introduce the novel concept of "Generative Temporal Nursing" (GTN), which aims to intervene in the video generation process on the fly during inference to improve the temporal dynamics of the results, without requiring any training or inducing high computational overhead at inference time. As a form of GTN, we propose VSTAR, consisting of Video Synopsis Prompting (VSP) and Temporal Attention Regularization (TAR). Leveraging the outstanding reasoning capability of LLMs, VSP can provide more descriptive textual guidance along the temporal dimension by breaking the single text prompt into several key visual states. As our core contribution, TAR is a simple yet effective regularization technique designed to enhance the temporal dynamics of the output, inspired by our thorough analysis of the temporal attention units in T2V models. We observed that the attention map of real videos exhibits a band-matrix-like structure, whereas that of synthesized ones is much less organized and more dispersed across different frames, particularly in longer videos. To bridge this gap, TAR employs a symmetric Toeplitz matrix as the regularization matrix, with values along the off-diagonal direction following a Gaussian distribution. TAR effectively enhances the dynamic aspects of the video and is crucial for enabling long video synthesis. With the combination of VSP, VSTAR can generate visually appealing evolutions using a pretrained T2V model in a single inference pass, without requiring additional training. Beyond the visually appealing results, our work (Li et al., 2025) is the first to conduct a detailed temporal attention analysis in T2V models. Our study provides valuable insights into improving the training of the next generation of T2V models.

7.2 Future Perspectives

This thesis has presented several contributions aimed at improving controllability and alignment in GANs and diffusion models. As the field of generative modeling continues to evolve, promising directions remain that could further enhance the capabilities and applications of these models. In this section, we explore future perspectives on advancing generative modeling and its broader impact. For a more detailed discussion of the limitations and outlook concerning individual contributions, we refer the reader to the respective chapters.

7.2.1 More Fine-Grained Control

In Chapters 4 and 5, we improved the controllability of the image generation process conditioned on the text prompt and semantic label map. Nevertheless, such conditioning types may not be able to specify every aspect of the image. For instance, the label map does not contain any geometry constraints such as object pose and orientation. On one hand, further conditions can be explored. Wang et al. (2024a) proposed a new type of conditioning descriptor "neural layout", which essentially comprises neural features extracted from pretrained foundation models, e.g., DINO (Caron et al., 2021, Oquab et al., 2024). Given that they are trained on massive datasets, foundation models have demonstrated rich semantic and geometric knowledge (Zhang et al., 2023, Tang et al., 2023a, El Banani et al., 2024), making them advantageous candidates for providing conditioning information. On the other hand, different conditions can complement each other. Several studies (Qin et al., 2023, Zhao et al., 2023a) have made initial attempts using embedding concatenation and task-aware weight modulation. It remains an interesting and active research question on how to effectively and flexibly utilize multiple conditioning inputs to better steer the synthesis process. Moreover, some fine-grained aspects such as explicit lighting control (Kocsis et al., 2024, Zeng et al., 2024a), and material properties of the object (Sharma et al., 2024, Zeng et al., 2024b), are also crucial for photorealistic image synthesis, especially when physical rendering rules are of concern.

Controllable video generation is inherently more complicated due to the additional temporal dimension, and exploration of this topic is still in its early stages. Some studies (Guo et al., 2023, Zhang et al., 2024, Xing et al., 2024, Lin et al., 2024) have adapted the image control paradigm to videos, allowing for the manipulation of spatial layouts and semantic aspects of the content. However, the temporal and dynamic aspects of video introduce new complexities, particularly in the control of camera movements and object motion. These aspects present unique challenges that differ significantly from those encountered in static image manipulation. Recent works attempt to utilize explicit camera parameters for camera control (Wang et al., 2024d, Yang et al., 2024b, He et al., 2024), which are not straightforward to provide at inference time. A point-based trajectory map has been explored to specify object movement (Yin et al., 2023, Wang et al., 2024d). Despite showing promising

results, these methods primarily operate in simple scenarios involving one or two objects. Future work could focus on exploring more expressive yet user-friendly motion conditioning information to better describe movement in the scene.

7.2.2 Personalized Control

Different people often have unique preferences and tastes, meaning that the same input condition could lead to vastly different "ideal" outputs depending on the individual. This variability underscores the importance of personalization in generative modeling. Personalization can serve as a powerful tool to reduce ambiguity, ensuring that the generated content aligns more closely with the user's specific desires and expectations. Previous methods have explored subject-driven synthesis (Ruiz et al., 2023, Kumari et al., 2023, Liu et al., 2023) and portrait face-driven synthesis (Wang et al., 2024b, Kim et al., 2024). However, the subject is not the only aspect that requires attention. There are more fine-grained concepts and details that matter. For instance, one might prefer certain attributes of an object, such as its material, or specific aspects of the example, like the lighting conditions. In video synthesis, a user might want to customize the specific subject motion (Yatim et al., 2024) or camera movement pattern. Therefore, it is an interesting research direction to investigate more fine-grained personalization. For instance, by leveraging the rapid development in (region-level) vision-language models (Rasheed et al., 2024, Ma et al., 2024), one can extract high-level features or textual descriptions for more precise conditioning.

7.2.3 Multimodal World Models

Recent advancements in video generation models, such as Sora (OpenAI, 2024), have not only demonstrated astonishing visual results, but also highlighted the potential of large-scale video generation models as a promising path toward building general-purpose simulators of the physical world. The comprehensive understanding of environmental dynamics and physical constraints by world models can provide significant value across various industries, such as media production, autonomous driving, and the development of autonomous agents. World models (Hu et al., 2023a, Bruce et al., 2024, Xiang et al., 2024) are generally multimodal models

that can interact with various conditions, e.g., natural language, images, and actions. With the development and availability of such models, many exciting future research opportunities arise. For instance, the world model can serve as a neural simulator for testing and validating model performance. Autonomous agents, e.g., autonomous vehicles, robots, can directly interact with the world model, which will adapt based on the maneuvers received from the agents. This paradigm reduces the need for extensive real-world data collection and can address more corner cases, thereby mitigating safety concerns. Besides serving as a data source, the world model processes general knowledge of the physical world, and thus its representation can be utilized or distilled for other tasks as well. Prior works (Zhao et al., 2023b, Kaplan et al., 2024) have demonstrated the potential of using Stable Diffusion (Rombach et al., 2022) as a feature extraction backbone for various visual perception tasks. We believe that this powerful world model can further enhance the performance of a wide range of downstream tasks.

In summary, visual synthesis is a fascinating and rapidly evolving field with the potential to revolutionize a wide range of real-world applications. Our work has contributed to addressing key challenges in alignment and controllability. The advancements discussed in this outlook have the potential to further enhance how we create and interact with digital content, enabling more precise, tailored, and immersive experiences across various industries. More fine-grained control allows industries such as film production and game development to achieve higher levels of precision and creativity, enabling creators to generate complex visuals that closely align with their vision. Personalized control takes this even further, enabling applications in e-commerce, virtual experiences, and interactive content to deliver uniquely tailored outputs based on individual user preferences. Additionally, the development of multimodal world models opens new possibilities for areas like autonomous driving and robotics, where the ability to simulate complex environments with accurate physical and causal relationships is essential. Ultimately, advanced generative models can go beyond simply producing better visual content; they have the potential to profoundly influence our everyday lives through various real-world applications.

List of Figures

List of Tables

Abbreviations

AE Autoencoder

DM Diffusion Model

ELBO Evidence Lower Bound

FID Fréchet Inception Distance

GAN Generative Adversarial Network

GSN Generative Semantic Nursing

GTN Generative Temporal Nursing

L2I Layout-to-Image

LDM Latent Diffusion Model

LLM Large Language Model

mIoU Mean Intersetion-over-Union

MNE Masked Noise Encoder

SD Stable Diffusion

SIS Semantic Image Synthesis

T2I Text-to-Image

T2V Text-to-Video

VAE Variational Autoencoder

VQA Visual Question Answering

Bibliography

R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019.

R. Abdal, Y. Qin, and P. Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020.

R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ToG*, 2021.

A. Agarwal, S. Karanam, K. Joseph, A. Saxena, K. Goswami, and B. V. Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *ICCV*, 2023.

L. AI. Dreammachine. https://lumalabs.ai/dream-machine, 2024.

Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, 2022.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Y. Bai, J. Mei, A. L. Yuille, and C. Xie. Are transformers more robust than cnns? In *NeurIPS*, 2021.

Y. Balaji, S. Sankaranarayanan, and R. Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018.

Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022.

C. Bartz, J. Bethge, H. Yang, and C. Meinel. One model to reconstruct them all: A novel way to use the stochastic noise in StyleGAN. In *BMVC*, 2021.

Y. Bin, X. Cao, X. Chen, Y. Ge, Y. Tai, C. Wang, J. Li, F. Huang, C. Gao, and N. Sang. Adversarial semantic data augmentation for human pose estimation. In *ECCV*, 2020.

A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023.

A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

S. Burton, L. Gauerhof, and C. Heinzemann. Making the case for safety of machine learning in highly automated driving. In *SAFECOMP*, 2017.

H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

L. Chai, J.-Y. Zhu, E. Shechtman, P. Isola, and R. Zhang. Ensembling with deep generative views. In *CVPR*, 2021.

H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022.

H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023.

H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or. Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*, 2023.

H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024a.

L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018a. doi: 10.1109/TPAMI.2017.2699184.

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018b.

M. Chen, I. Laina, and A. Vedaldi. Training-free layout control with cross-attention guidance. In *WACV*, 2024b.

A. Cherepkov, A. Voynov, and A. Babenko. Navigating the gan parameter space for semantic image editing. In *CVPR*, 2021.

S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo. RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021.

E. Collins, R. Bala, B. Price, and S. Susstrunk. Editing in style: Uncovering the local semantics of gans. In *CVPR*, 2020.

M. Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

A. Creswell and A. A. Bharath. Inverting the generator of a generative adversarial network. *TNNLS*, 2019. doi: 10.1109/TNNLS.2018.2875194.

T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint*, 2017.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

T. M. Dinh, A. T. Tran, R. Nguyen, and B.-S. Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *CVPR*, 2022.

M. El Banani, A. Raj, K.-K. Maninis, A. Kar, Y. Li, M. Rubinstein, D. Sun, L. Guibas, J. Johnson, and V. Jampani. Probing the 3d awareness of visual foundation models. In *CVPR*, 2024.

P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Y. Fan and K. Lee. Optimizing DDPM sampling with shortcut fine-tuning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *ICML*, 2023.

W. Feng, X. He, T.-J. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023.

J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023.

L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

J. Gu, Y. Shen, and B. Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020.

Y. Guo, C. Yang, A. Rao, M. Agrawala, D. Lin, and B. Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023.

Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. Ganspace: Discovering interpretable gan controls. In *NeruIPS*, 2020.

H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018.

D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2019.

A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPs*, 2020.

M. Hong, J. Choi, and G. Kim. StyleMix: Separating content and style for enhanced data augmentation. In *CVPR*, 2021.

L. Hoyer, D. Dai, and L. Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022.

A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023a.

S. Hu, K. Zhang, Z. Chen, and L. Chan. Domain generalization via multidomain discriminant analysis. In *UAI*, 2020.

Y. Hu, C.-H. Lee, T. Xie, T. Yu, N. A. Smith, and M. Ostendorf. In-context learning for few-shot dialogue state tracking. In *EMNLP*, 2022.

Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023b.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

A. Jahanian, L. Chai, and P. Isola. On the" steerability" of generative adversarial networks. In *ICLR*, 2020.

L. Jiang, B. Dai, W. Wu, and C. C. Loy. Deceive d: Adaptive pseudo augmentation for gan training with limited data. In *NeurIPS*, 2021.

X. Jin, C. Lan, W. Zeng, and Z. Chen. Feature alignment and restoration for domain generalization and adaptation. *arXiv preprint*, 2020.

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

Kaishou. Kling. https://kling.kuaishou.com/en, 2024.

K. Kang, S. Kim, and S. Cho. Gan inversion for out-of-range images with geometric transformations. In *ICCV*, 2021.

M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023.

U. A. Kaplan, Y. Li, M. Keuper, A. Khoreva, and D. Zhang. Domain-aware fine-tuning of foundation models. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020a.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020b.

T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023.

A. Kazemnejad, I. Padhi, K. Natesan, P. Das, and S. Reddy. The impact of positional encoding on length generalization in transformers. In *NeurIPS*, 2023.

C. Kim, J. Lee, S. Joung, B. Kim, and Y.-M. Baek. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*, 2024.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

P. Kocsis, J. Philip, K. Sunkavalli, M. Nießner, and Y. Hold-Geoffroy. Lightit: Illumination modeling and control for diffusion models. In *CVPR*, 2024.

N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

M. Kwon, J. Jeong, and Y. Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023.

C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, and L. Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *EMNLP*, 2022a.

D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018a.

H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018b.

H. Li, Y. Wang, R. Wan, S. Wang, T.-Q. Li, and A. Kot. Domain generalization for medical imaging classification with linear-dependency regularization. In *NeurIPS*, 2020.

J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022b.

X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, and L. DUAN. Uncertainty Modeling for Out-of-Distribution Generalization. In *ICLR*, 2022c.

Y. Li, L. Yuan, and N. Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019.

Y. Li, M. Keuper, D. Zhang, and A. Khoreva. Divide & bind your attention for improved generative semantic nursing. In *BMVC*, 2023a.

Y. Li, D. Zhang, M. Keuper, and A. Khoreva. Intra-source style augmentation for improved domain generalization. In *WACV*, 2023b.

Y. Li, M. Keuper, D. Zhang, and A. Khoreva. Adversarial supervision makes layout-to-image diffusion models thrive. In *ICLR*, 2024a.

Y. Li, Z. Wu, H. Zhao, T. Yang, Z. Liu, P. Shu, J. Sun, R. Parasuraman, and T. Liu. Aldm-grasping: Diffusion-aided zero-shot sim-to-real transfer for robot grasping. *arXiv preprint arXiv:2403.11459*, 2024b.

Y. Li, D. Zhang, M. Keuper, and A. Khoreva. Intra-& extra-source exemplar-based style synthesis for improved domain generalization. *IJCV*, 2024c.

Y. Li, W. Beluch, M. Keuper, D. Zhang, and A. Khoreva. Vstar: Generative temporal nursing for longer dynamic video synthesis. In *ICLR*, 2025.

J. H. Liew, H. Yan, D. Zhou, and J. Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022.

H. Lin, J. Cho, A. Zala, and M. Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024.

T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022.

Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, 2024.

Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao. Cones: Concept neurons in diffusion models for customized generation. In *ICML*, 2023.

Y. Lu, X. Yang, X. Li, X. E. Wang, and W. Y. Wang. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. In *NeurIPS*, 2024.

Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019.

C. Ma, Y. Jiang, J. Wu, Z. Yuan, and X. Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, 2024.

X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017.

G. Marcus, E. Davis, and S. Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

S.-J. Moon and G.-M. Park. Interestyle: Encoding an interest region for robust stylegan inversion. In *ECCV*, 2022.

C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024.

M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016.

E. Ntavelis, A. Romero, I. Kastanis, L. Van Gool, and R. Timofte. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *ECCV*, 2020.

OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022.

OpenAI. Sora: Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel. Teaching clip to count to ten. In *ICCV*, 2023.

T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021.

W. Peebles and S. Xie. Scalable diffusion models with transformers. In *CVPR*, 2023.

X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *CVPR*, 2018.

D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, et al. Unicontrol: a unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023.

H. Qiu, M. Xia, Y. Zhang, Y. He, X. Wang, Y. Shan, and Z. Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *ICLR*, 2024.

Z. Qiu, W. Liu, H. Feng, Y. Xue, Y. Feng, Z. Liu, D. Zhang, A. Weller, and B. Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *NeurIPS*, 2023.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

H. Rasheed, M. Maaz, S. Shaji, A. Shaker, S. Khan, H. Cholakkal, R. M. Anwer, E. Xing, M.-H. Yang, and F. S. Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021.

D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or. Pivotal tuning for latent-based editing of real images. *TOG*, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

Runway. Gen-3 alpha. https://runwayml.com/ai-tools/gen-3-alpha/, 2024.

A. Ruoss, G. Delétang, T. Genewein, J. Grau-Moya, R. Csordás, M. Bennani, S. Legg, and J. Veness. Randomized positional encodings boost length generalization of transformers. In *ACL*, 2023.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018.

C. Sakaridis, D. Dai, and L. V. Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019.

C. Sakaridis, D. Dai, and L. Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021.

A. Sauer, K. Schwarz, and A. Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022.

A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *ICML*, 2023.

E. Schönfeld, V. Sushko, D. Zhang, J. Gall, B. Schiele, and A. Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2020.

C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

S. Shafaei, S. Kugele, M. H. Osman, and A. Knoll. Uncertainty in machine learning: A safety perspective on autonomous driving. In *SAFECOMP*, 2018.

P. Sharma, V. Jampani, Y. Li, X. Jia, D. Lagun, F. Durand, B. Freeman, and M. Matthews. Alchemist: Parametric control of material properties with diffusion models. In *CVPR*, 2024.

Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

H. Song, Y. Du, T. Xiang, J. Dong, J. Qin, and S. He. Editing out-of-domain gan inversion via differential activations. In *ECCV*, 2022.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2020.

R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *CVPR*, 2021.

A. Šubrtová, D. Futschik, J. Čech, M. Lukáč, E. Shechtman, and D. Sỳkora. Chunky-gan: Real image inversion via segments. In *ECCV*, 2022.

V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva. Oasis: only adversarial supervision for semantic image synthesis. *IJCV*, 2022.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Z. Tan, M. Chai, D. Chen, J. Liao, Q. Chu, B. Liu, G. Hua, and N. Yu. Diverse semantic image synthesis via probability distribution modeling. In *CVPR*, 2021a.

Z. Tan, D. Chen, Q. Chu, M. Chai, J. Liao, M. He, L. Yuan, G. Hua, and N. Yu. Efficient semantic image synthesis via class-adaptive normalization. *TPAMI*, 2021b.

L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023a.

R. Tang, L. Liu, A. Pandey, Z. Jiang, G. Yang, K. Kumar, P. Stenetorp, J. Lin, and F. Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *ACL*, 2023b.

R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or. Designing an encoder for stylegan image manipulation. *TOG*, 2021. doi: 10.1145/3450626.3459838.

Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.

Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00154.

A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016.

A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, 2019.

F.-Y. Wang, W. Chen, G. Song, H.-J. Ye, Y. Liu, and H. Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023a.

J. Wang, S. Jin, W. Liu, W. Liu, C. Qian, and P. Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *CVPR*, 2021a.

J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2021b. doi: 10.1109/TPAMI.2020.2983686.

J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023b.

J. Wang, K. A. Laube, Y. Li, J. H. Metzen, S.-I. Cheng, J. Borges, and A. Khoreva. Label-free neural semantic image synthesis. In *ECCV*, 2024a.

Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018a.

T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018b.

X. Wang, T. Darrell, S. S. Rambhatla, R. Girdhar, and I. Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, 2024c.

Y. Wang, L. Qi, Y.-C. Chen, X. Zhang, and J. Jia. Image synthesis via semantic composition. In *ICCV*, 2021c.

Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023c.

Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T. S. Huang, and H. Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, 2020.

Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou. Diffusion-gan: Training gans with diffusion. In *ICLR*, 2023d.

Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 2024d.

Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *ACL*, 2023e.

T. Wei, D. Chen, W. Zhou, J. Liao, W. Zhang, L. Yuan, G. Hua, and N. Yu. E2Style: Improve the efficiency and effectiveness of stylegan inversion. *TIP*, 2022. doi: 10.1109/TIP.2022.3167305.

W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang. Gan inversion: A survey. *TPAMI*, 2022.

J. Xiang, G. Liu, Y. Gu, Q. Gao, Y. Ning, Y. Zha, Z. Feng, T. Tao, S. Hao, Y. Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

Z. Xiao, K. Kreis, and A. Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *ICLR*, 2022.

E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. SimMIM: A simple framework for masked image modeling. In *CVPR*, 2022.

J. Xing, H. Liu, M. Xia, Y. Zhang, X. Wang, Y. Shan, and T.-T. Wong. Tooncrafter: Generative cartoon interpolation. *arXiv preprint arXiv:2405.17933*, 2024.

T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.

H. Xue, Z. Huang, Q. Sun, L. Song, and W. Zhang. Freestyle layout-to-image synthesis. In *CVPR*, 2023.

L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and C. Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *ICML*, 2024a.

S. Yang, L. Hou, H. Huang, C. Ma, P. Wan, D. Zhang, X. Chen, and J. Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH*, 2024b.

Y. Yang and S. Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020.

Z. Yang, J. Wang, Z. Gan, L. Li, K. Lin, C. Wu, N. Duan, Z. Liu, C. Liu, M. Zeng, et al. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023.

X. Yao, A. Newson, Y. Gousseau, and P. Hellier. Feature-Style Encoder for Style-Based GAN Inversion. *arXiv preprint*, 2022.

D. Yatim, R. Fridman, O. Bar-Tal, Y. Kasten, and T. Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *CVPR*, 2024.

S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint*, 2015.

F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. BDD100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.

J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022.

S. Yuan, J. Huang, Y. Xu, Y. Liu, S. Zhang, Y. Shi, R. Zhu, X. Cheng, J. Luo, and L. Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *arXiv preprint arXiv:2406.18522*, 2024.

S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

C. Zeng, Y. Dong, P. Peers, Y. Kong, H. Wu, and X. Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH*, 2024a.

Z. Zeng, V. Deschaintre, I. Georgiev, Y. Hold-Geoffroy, Y. Hu, F. Luan, L.-Q. Yan, and M. Hašan. Rgb↔x: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH*, 2024b.

C. Zhang, M. Zhang, S. Zhang, D. Jin, Q. Zhou, Z. Cai, H. Zhao, X. Liu, and Z. Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *CVPR*, 2022.

H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *ICLR*, 2018a.

J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *NeurIPS*, 2023.

L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018b.

Y. Zhang, A. Carballo, H. Yang, and K. Takeda. Autonomous Driving in Adverse Weather Conditions: A Survey. *arXiv preprint*, 2021a.

Y. Zhang, A. Gupta, N. Saunshi, and S. Arora. On predicting generalization using gans. In *ICLR*, 2021b.

Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian. Controlvideo: Training-free controllable text-to-video generation. In *ICLR*, 2024.

S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023a.

W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023b.

Y. Zhao, Z. Zhong, N. Zhao, N. Sebe, and G. H. Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *ECCV*, 2022.

S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.

Z. Zheng and Y. Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 2021. doi: 10.1007/s11263-020-01395-y.

B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

F. Zhou, Z. Jiang, C. Shui, B. Wang, and B. Chaib-draa. Domain generalization with optimal transport and metric learning. *arXiv preprint*, 2020.

K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with mixstyle. In *ICLR*, 2021.

Y. Zhou, U. Alon, X. Chen, X. Wang, R. Agarwal, and D. Zhou. Transformers can achieve length generalization but not robustly. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020a.

J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.

Z. Zhu, Z. Xu, A. You, and X. Bai. Semantically multi-modal image synthesis. In *CVPR*, 2020b.

Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang. Confidence regularized self-training. In *ICCV*, 2019.