UNIVERSITY OF CALIFORNIA

Los Angeles

Towards Training-Free Controllable Text-to-Image Generation

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Sicheng Mo

2024

ABSTRACT OF THE THESIS

Towards Training-Free Controllable Text-to-Image Generation

by

Sicheng Mo
Master of Science in Computer Science
University of California, Los Angeles, 2024
Professor Bolei Zhou, Chair

Recent large-scale text-to-image (**T2I**) diffusion models [58, 34, 21, 53] have achieved remarkable success, enabling the generation of complex and realistic images from any text prompt that describes the target concept. Despite the significant advantages, the T2I diffusion model suffers from poor spatial controllability solely from text description. This thesis focuses on improving the pre-trained T2I diffusion models with additional support to take spatial reference.

The first part of this thesis proposed FreeControl, a training-free and guidance-based approach for controllable T2I generation that supports multiple conditions, architectures, and checkpoints simultaneously. FreeControl enforces structure guidance to facilitate the global alignment with a guidance image, and appearance guidance to collect visual details from images generated without control. Extensive qualitative and quantitative experiments demonstrate the superior performance of FreeControl across a variety of pre-trained T2I models. In particular, FreeControl enables convenient training-free control over many different architectures and checkpoints, allows the challenging input conditions on which most of the existing training-free methods fail, and achieves competitive synthesis quality compared to training-based approaches.

The second part of this thesis presents Ctrl-X, a training-free and guidance-free method that supports structure and appearance customization from a large spectrum of image modalities. Ctrl-X designs feed-forward structure control to enable the structure alignment with a structure image and semantic-aware appearance transfer to facilitate the appearance transfer from a user-input image. Extensive qualitative and quantitative experiments illustrate the superior performance of Ctrl-X on various condition inputs and model checkpoints.

The thesis of Sicheng Mo is approved.

Aditya Grover

Guy Van Den Broeck

Nanyun Peng

Bolei Zhou, Committee Chair

University of California, Los Angeles

2024

To my family and friends, whose unwavering support

has shaped me into the person I am today.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my advisor, Bolei Zhou. His unwavering support, wisdom, and guidance have been pivotal to my academic success and personal growth over the years. Professor Zhou has always provided crucial insights and clear direction, helping me navigate through complex research problems and academic obstacles. His mentorship has not only enriched my knowledge but also instilled in me a profound sense of confidence and perseverance.

I am also profoundly thankful to my Master's committee members, Professors Guy Van Den Broeck, Aditya Grover, and Nanyun Peng. Their commitment to reviewing this thesis has been essential in refining and improving the quality of my work. Their thoughtful feedback and constructive critiques have significantly enhanced the depth and breadth of my research.

I am immensely grateful to my collaborator, Dr. Fangzhou Mu. His invaluable advice, steadfast support, and extensive experience have been indispensable in overcoming the challenges we faced together in research projects. Working alongside Dr. Mu has been an enlightening experience, and his dedication and expertise have greatly contributed to our successful collaborations.

My heartfelt thanks extend to my friends at UCLA: Jiawei Yang, Kuan Heng Lin, Zhizheng Liu, Zhenghao Peng, and Yiran Wang. Their constructive discussions and companionship have been a continuous source of motivation and inspiration. Our shared experiences and intellectual exchanges have not only enriched my academic journey but also provided a sense of mutual support that I deeply cherish.

Lastly, I want to express my deepest appreciation to my parents, Jinwen Mo and Qi Chen. Their unwavering support and unconditional love have been my foundation throughout the years. They have been my bedrock, offering constant encouragement and understanding, without which my achievements would not have been possible.

CHAPTER 1

Introduction

Recent large-scale text-to-image (**T2I**) diffusion models [58, 34, 21, 53] have achieved remarkable success, enabling the generation of complex and realistic images from any text prompt that describes the target concept. Despite these advancements, generating images with specific desired layouts and structures solely from text descriptions remains a significant challenge. This limitation arises because text descriptions often lack the precise spatial information needed to dictate the exact arrangement of elements within an image. Therefore, incorporating spatial references into the diffusion process is essential for enhancing the capability of T2I generation models to produce images that not only match the textual descriptions but also adhere to the intended spatial configurations.

Recent advances, such as ControlNet [75], enable spatial control of pre-trained T2I diffusion models, allowing users to specify the desired image composition by providing a guidance image from pre-defined modalities (e.g., depth map, human-pose map) alongside the text description. These methods [75, 36, 41, 78, 66, 7] achieve superior generation results by integrating additional spatial information, yet they require training an additional module specific to each spatial condition type. Given the vast array of potential control signals, the continuously evolving model architectures, and the increasing number of customized model checkpoints (such as Stable Diffusion [53] fine-tuned for Disney characters or user-specified objects [55, 27]), this repetitive training for every new model and condition type becomes highly costly and uneconomical. This process demands considerable computational resources and time, creating a barrier to the widespread adoption and scalability of these advanced controllable generation techniques. Consequently, there is a pressing need for more efficient approaches that can leverage existing models and adapt to new conditions without the necessity for extensive retraining.

Besides the high training cost and poor scalability, controllable T2I diffusion methods face several drawbacks stemming from their current training schemes. These methods are typically trained to output a target image given a spatially-aligned control condition, which is computed from the same image using an off-the-shelf model (e.g., MiDaS [52] for depth maps, OpenPose [12] for human poses). This approach inherently limits the use of many desirable control signals that are difficult to infer directly from an image, such as meshes or point clouds. Additionally, the reliance on these pre-aligned conditions introduces a bias in the model, causing it to prioritize spatial conditions over textual descriptions. This bias occurs because the model can exploit the close spatial alignment of input-output image pairs as a shortcut, leading to less effective integration of the textual content. As a result, the flexibility and versatility of these models are significantly compromised, limiting their practical applications. There is a clear need for developing new strategies that can handle a broader range of control signals and better balance the importance of both spatial and textual inputs, thus enhancing the overall performance and applicability of controllable T2I diffusion methods.

To address the aforementioned limitations, a possible approach is to leverage the strong generalizability of the pre-trained diffusion models to extract pixel-level structure information from given reference images in any modality. With this approach, no additional training will be required, and it can be adapted to any new diffusion network without re-training. The research in this thesis fulfills the missing study in this possible direction.

1.1 Thesis Outline

This thesis is organized into four chapters, each focusing on different facets of our research on training-free controllable generation in pre-trained text-to-image models with additional condition signals. The chapters are outlined as follows:

Chapter 1 introduces the field of controllable text-to-image generation and the motivation behind the research presented in this thesis. It sets the stage for the study and provides the context for the subsequent chapters.

In Chapter 2, we present FreeControl, a training-free approach for controllable T2I generation that supports multiple conditions, architectures, and checkpoints simultaneously. FreeControl enforces structure guidance to facilitate the global alignment with a guidance image, and appearance guidance to collect visual details from images generated without control. Extensive qualitative and quantitative experiments demonstrate the superior performance of FreeControl across a variety of pre-trained T2I models. In particular, FreeControl enables convenient training-free control over many different architectures and checkpoints, allows the challenging input conditions on which most of the existing training-free methods fail, and achieves competitive synthesis quality compared to training-based approaches.

In Chapter 3, we present *Ctrl-X*, a simple framework for T2I diffusion controlling structure and appearance without additional training or guidance. Ctrl-X designs feed-forward structure control to enable the structure alignment with a structure image and semantic-aware appearance transfer to facilitate the appearance transfer from a user-input image. Extensive qualitative and quantitative experiments illustrate the superior performance of Ctrl-X on various condition inputs and model checkpoints. In particular, Ctrl-X supports novel structure and appearance control with arbitrary condition images of any modality, exhibits superior image quality and appearance transfer compared to existing works, and provides instant plug-and-play to any T2I and text-to-video (T2V) diffusion model.

Chapter 4 provides a comprehensive summary of the research conducted in this thesis, highlighting key findings and contributions to the field of controllable text-to-image generation. It also discusses the limitations encountered and suggests potential directions for future research to advance the domain further.

CHAPTER 2

FreeControl: Training-Free Spatial Control of Any Text-to-Image Diffusion Model with Any Condition

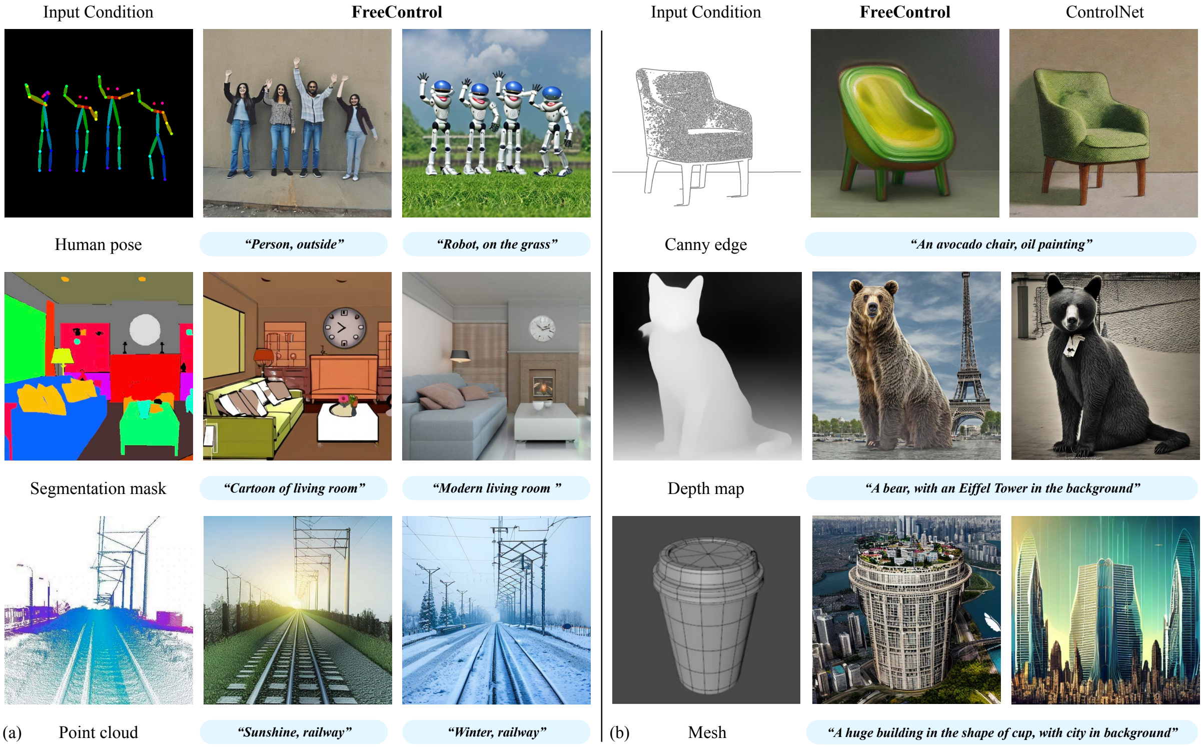


Figure 2.1: **Training-free conditional control of Stable Diffusion [53].** (a) FreeControl enables zero-shot control of pretrained text-to-image diffusion models given various input control conditions. (b) Compared to ControlNet [75], FreeControl achieves a good balance between spatial and image-text alignment, especially when facing a conflict between the guidance image and text description. Additionally, FreeControl supports several condition types (e.g.„ 2D projections of point clouds and meshes in the bottom row), where it is difficult to construct training pairs.

2.1 Introduction

Text-to-image (T2I) diffusion models [51, 4] have achieved tremendous success in high-quality image synthesis, yet a text description alone is far from enough for users to convey their preferences and intents for content creation. Recent advances such as ControlNet [75] enable spatial control of pretrained T2I diffusion models, allowing users to specify the desired image composition by providing a guidance image (e.g.,, depth map, human pose) alongside the text description. Despite their superior generation results, these methods [75, 36, 41, 78, 66, 7] require training an additional module specific to each spatial condition type. Considering the large space of control signals, constantly evolving model architectures, and a growing number of customized model checkpoints (e.g.,, Stable Diffusion [53] fine-tuned for Disney characters or user-specified objects [55, 27]), this repetitive training on every new model and condition type is costly and uneconomical.

Besides the high training cost and poor scalability, controllable T2I diffusion methods face drawbacks that stem from their training scheme: they are trained to output a target image given a spatially-aligned control condition computed from the same image using an off-the-shelf model (e.g.,, MiDaS [52] for depth maps, OpenPose [12] for human poses). This limits the use of many desired control signals that are difficult to infer from an image (e.g.,, mesh, point cloud). Further, the trained models tend to prioritize spatial condition over text description, likely because the close spatial alignment of input-output image pairs exposes a shortcut. This is illustrated in Figure 3.1(b), where there is a conflict between the guidance image and text prompt (e.g.,, an edge map of a sofa chair v.s.“an avocado chair”).

To address the aforementioned limitations, we present FreeControl, a versatile training-free method for controllable T2I diffusion. Our key motivation is that feature maps in T2I models during the generation process already capture the spatial structure and local appearance described in the input text. By modeling the subspace of these features, we can effectively steer the generation process towards a similar structure expressed in the guidance image, while preserving the appearance of the concept in the input text. To this end, FreeControl includes an analysis stage and a synthesis stage. In the analysis stage, FreeControl queries a T2I model to generate as few as one seed image and then constructs a linear feature subspace from the generated images. In the synthesis stage, FreeControl employs guidance in the subspace to facilitate structure alignment with a guidance image, as well as appearance alignment between images generated with and without control.

FreeControl offers significant strength over training-based methods by eliminating the need for additional training on a pretrained T2I model, while adeptly adhering to concepts outlined in the text description. It supports a wide range of control conditions, model architectures and customized checkpoints, achieves high-quality image generation with robust controllability in comparison to prior training-free methods [38, 23, 64, 45], and can be readily adapted for text-guided image-to-image translation. We conduct extensive qualitative and quantitative experiments and demonstrate the superior performance of our method. Notably, FreeControl excels at challenging control conditions on which prior training-free methods fail. In the meantime, it attains competitive image synthesis quality compared to training-based methods while providing stronger image-text alignment and supporting a broader set of control signals.

Our contributions. (1) We present FreeControl, a novel method for training-free controllable T2I generation via modeling the linear subspace of intermediate diffusion features and employing guidance in this subspace during the generation process. (2) Our method presents the first universal training-free solution that supports multiple control conditions (sketch, normal map, depth map, edge map, human pose, segmentation mask, natural image and beyond), model architectures (e.g.,, SD 1.5, 2.1, and SD-XL 1.0), and customized checkpoints (e.g.,, using DreamBooth [55] and LoRA [27]). (3) Our method demonstrates superior results in comparison to previous training-free methods (e.g.,, Plug-and-Play [64]) and achieves comparable performance with prior training-based approaches (e.g.,, ControlNet [75]).

2.2 Related Work

Text-to-image diffusion. Diffusion models [60, 24, 62] bring a breakthrough in text-to-image (T2I) generation. T2I diffusion models formulate image generation as an iterative denoising task guided by a text prompt. Denoising is conditioned on textual embeddings produced by language encoders [50, 49] and is performed either in pixel space [42, 51, 58, 8] or latent space [53, 21, 48], followed by cascaded super-resolution [25] or latent-to-image decoding [18] for high-resolution image synthesis. Several recent works show that the internal representations of T2I diffusion models capture mid/high-level semantic concepts, and thus can be repurposed for image recognition tasks [71, 34]. Our work builds upon this intuition and exploits the feature space of T2I models to guide the generation process.

Controllable T2I diffusion. It is challenging to convey human preferences and intents through text description alone. Several methods thus instrument pre-trained T2I models to take an additional input condition by learning auxiliary modules on paired data [75, 36, 41**?** , 66, 7]. One significant drawback of this training-based approach is the cost of repeated training for every control signal type, model architecture, and model checkpoint. On the other hand, training-free methods leverage attention weights and features inside a pre-trained T2I model for the control of object size, shape, appearance and location [46, 11, 70, 17, 20]. However, these methods only take coarse conditions such as bounding boxes to achieve precise control over object pose and scene composition. Different from all the prior works, FreeControl is a training-free approach to controllable T2I diffusion that supports any spatial condition, model architecture, and checkpoint within a unified framework.

Image-to-image translation with T2I diffusion. Controlling T2I diffusion becomes an image-to-image translation (I2I) task [29] when the control signal is an image. I2I methods map an image from its source domain to a target domain while preserving the underlying structure [29, 44, 57]. T2I diffusion enables I2I methods to specify target domains using text. Text-driven I2I is often posed as conditional generation [75, 41**?** , 10, 30, 77]. These methods finetune a pretrained model to condition it on an input image. Alternatively, recent training-free methods perform zero-shot image translation [38, 23, 64, 45] and is most relevant to our work. This is achieved by inverting the input image [61, 40, 67], followed by manipulating the attention weights and features throughout the diffusion process. A key limitation of these methods is they require the input to have rich textures, and hence they fall short when converting abstract layouts (e.g.,depth) to realistic image. By contrast, our method attends to *semantic* image structure by decomposing features into principal components, thereby it supports a wide range of modalities as layout specifications.

Customized T2I diffusion. Model customization is a key use case of T2I diffusion in visual content creation. By fine-tuning a pretrained model on images of custom objects or styles, several methods [55, 19, 33, 6] bind a dedicated token to each concept and insert them in text prompts for customized generation. Amid the growing number of customized models being built and shared by content creators [3, 2], FreeControl offers a *scalable* framework for zero-shot control of any model with any spatial condition.

2.3 Preliminary

Diffusion sampling. Image generation with a pre-trained T2I diffusion model amounts to iteratively removing noise from an initial Gaussian noise image $\mathbf{x}_T$ [24]. This sampling process is governed by a learned denoising network ϵ_θ conditioned on a text prompt $\mathbf{c}$. At a sampling step t, a cleaner image $\mathbf{x}_{t-1}$ is obtained by subtracting from $\mathbf{x}_t$ a noise component $\epsilon_t = \epsilon_\theta(\mathbf{x}_t; t, \mathbf{c})$. Alternatively, ϵ_θ can be seen as approximating the score function for the marginal distributions p_t scaled by a noise schedule σ_t [62]:

$$\epsilon_\theta(\mathbf{x}_t; t, \mathbf{c}) \approx -\sigma_t \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x_t}|\mathbf{c}). \tag{2.1}$$

Guidance. The update rule in Equation 2.1 may be altered by a time-dependent energy function $g(\mathbf{x_t}; t, y)$ through *guidance* (with strength s) [16, 17] so as to condition diffusion

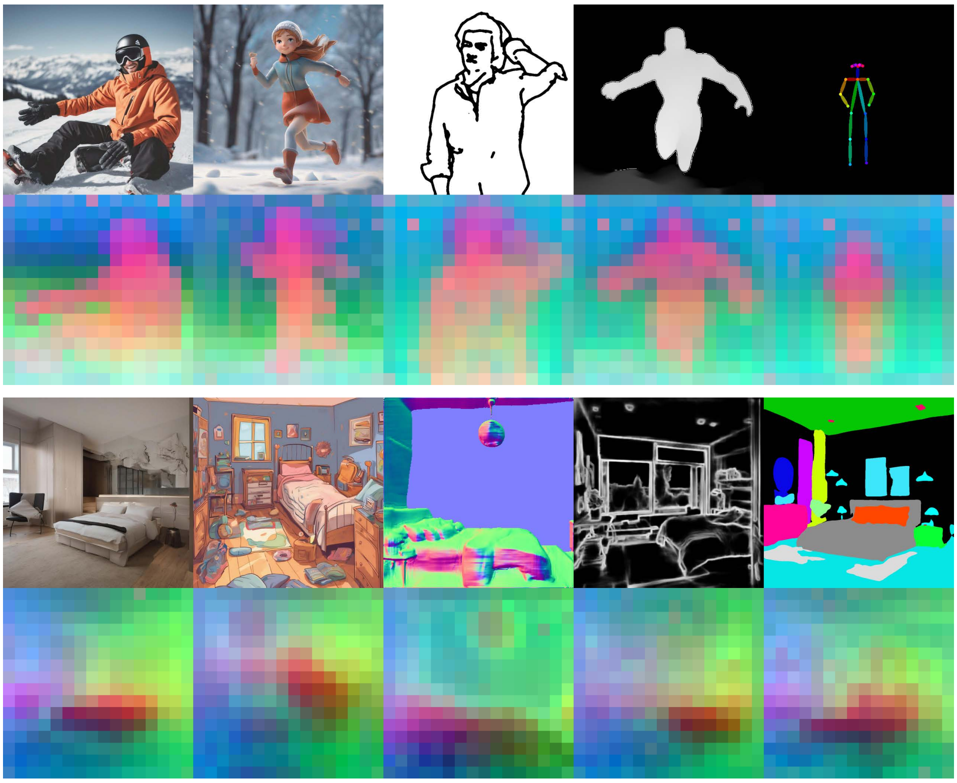


Figure 2.2: **Visualization of feature subspace given by PCA.** Keys from the first self-attention in the U-Net decoder are obtained via DDIM inversion [61] for five images in different styles and modalities (*top*: `person`; *bottom*: `bedroom`), and subsequently undergo PCA. The top three principal components (pseudo-colored in RGB) provide a clear separation of semantic components.

sampling on auxiliary information y (e.g.,, class labels):

$$\hat{\epsilon}_\theta(\mathbf{x}_t; t, \mathbf{c}) = \epsilon_\theta(\mathbf{x}_t; t, \mathbf{c}) - s\, g(\mathbf{x_t}; t, y). \tag{2.2}$$

In practice, g may be realized as classifiers [16] or CLIP scores [42], or defined using bounding boxes [70, 14], attention maps [20, 45] or any measurable object properties [17].

Attentions in ϵ_θ. A standard choice for ϵ_θ is a U-Net [54] with self- and cross-attentions [65] at multiple resolutions. Conceptually, self-attentions model interactions among spatial locations within an image, whereas cross-attentions relate spatial locations to tokens in a text prompt. These two attention mechanisms complement one another and jointly control the layout of a generated image [64, 11, 46, 20].

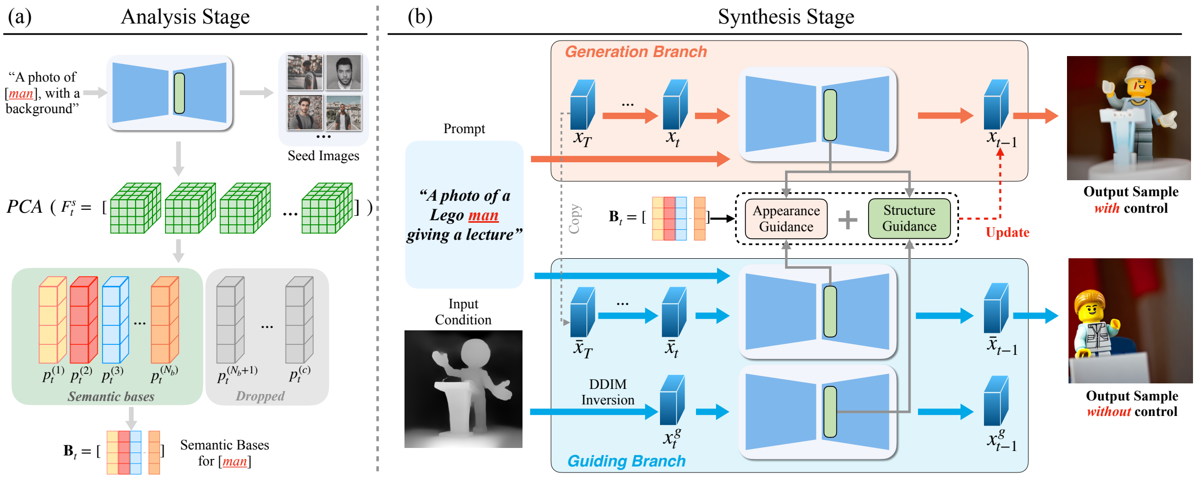


Figure 2.3: **Method overview.** (a) In the *analysis* stage, FreeControl generates seed images for a target concept (e.g.,, `man`) using a pretrained diffusion model and performs PCA on their diffusion features to obtain a linear subspace as semantic basis. (b) In the *synthesis* stage, FreeControl employs structure guidance in this subspace to enforce structure alignment with the input condition. In the meantime, it applies appearance guidance to facilitate appearance transfer from a sibling image generated using the same seed without structure control.

2.4 Training-Free Control of T2I Models

FreeControl is a unified framework for zero-shot controllable T2I diffusion. Given a text prompt $\mathbf{c}$ and a guidance image $\mathbf{I}^g$ of any modality, FreeControl directs a pre-trained T2I diffusion model ϵ_θ to comply with $\mathbf{c}$ while also respecting the semantic structure provided by $\mathbf{I}^g$ throughout the sampling process of an output image $\mathbf{I}$.

Our key finding is that the leading principal components of self-attention block features inside a pre-trained ϵ_θ provide a strong and surprisingly consistent representation of semantic structure across a broad spectrum of image modalities (see Figure 2.2 for examples). To this end, we introduce *structure guidance* to help draft the structural template of $\mathbf{I}$ under the guidance of $\mathbf{I}^g$. To texture this template with the content and style described by $\mathbf{c}$, we further devise *appearance guidance* to borrow appearance details from $\bar{\mathbf{I}}$, a sibling of $\mathbf{I}$ generated without altering the diffusion process. Ultimately, $\mathbf{I}$ mimics the structure of $\mathbf{I}^g$ with its content and style similar to $\bar{\mathbf{I}}$.

Method overview. FreeControl is a two-stage method as illustrated in Figure 3.3. It begins with an analysis stage, where diffusion features of *seed images* undergo principal component analysis (PCA), with the leading PCs forming the time-dependent bases $\mathbf{B}_t$ as our *semantic structure representation.* $\mathbf{I}^g$ subsequently undergoes DDIM inversion [61] with its diffusion features projected onto $\mathbf{B}_t$, yielding their *semantic coordinates* $\mathbf{S}_t^g$. In the synthesis stage, structure guidance encourages $\mathbf{I}$ to develop the same semantic structure as $\mathbf{I}^g$ by attracting $\mathbf{S}_t$ to $\mathbf{S}_t^g$. In the meantime, appearance guidance promotes appearance similarity between $\mathbf{I}$ and $\bar{\mathbf{I}}$ by penalizing the difference in their feature statistics.

2.4.1 Semantic Structure Representation

Zero-shot spatial control of T2I diffusion demands a unified representation of semantic image structure that is invariant to image modalities. Recent work has discovered that self-attention features (*i.e.*, keys and queries) of self-supervised Vision Transformers [63] and T2I diffusion models [11] are strong descriptors of image structure. Based on these findings, we hypothesize that manipulating self-attention features is key to controllable T2I diffusion.

A naïve approach derived from PnP [64] is to directly inject the self-attention weights (equivalently the features) of $\mathbf{I}^g$ into the diffusion process of $\mathbf{I}$. Unfortunately, this approach introduces *appearance leakage*; that is, not only the structure of $\mathbf{I}^g$ is carried over but also traces of appearance details. As seen in Figure 2.6, appearance leakage is particularly problematic when $\mathbf{I}^g$ and $\mathbf{I}$ are different modalities (e.g.,, depth v.s.natural images), common for controllable generation.

Towards disentangling image structure and appearance, we draw inspiration from Transformer feature visualization [43, 64] and perform PCA on self-attention features of semantically similar images. Our key observation is that the leading PCs form a *semantic basis*; It exhibits a strong correlation with object pose, shape, and scene composition across diverse image modalities. In the following, we leverage this basis as our *semantic structure representation* and explain how to obtain such bases in the analysis stage.

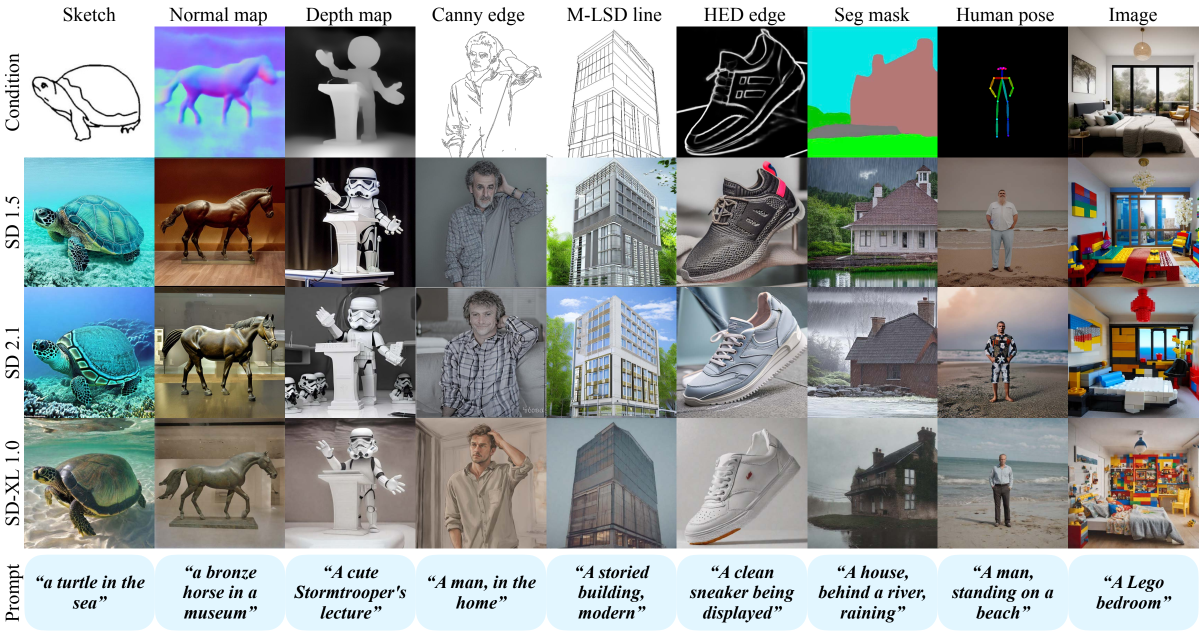

Sketch
Normal map
Depth map
Canny edge
M-LSD line
HED edge
Seg mask
Human pose
Image
Condition
SD 1.5
SD 2.1
SD-XL 1.0
Prompt
"a turtle in the sea"
"a bronze horse in a museum"
"A cute Stormtrooper's lecture"
"A man, in the home"
"A storied building, modern"
"A clean sneaker being displayed"
"A house, behind a river, raining"
"A man, standing on a beach"
"A Lego bedroom"

Figure 2.4: **Qualitative comparison of controllable T2I diffusion.** FreeControl supports a suite of control signals and three major versions of Stable Diffusion. The generated images closely follow the text prompts while exhibiting strong spatial alignment with the input images.

2.4.2 Analysis Stage

Seed images. We begin by collecting N_s images that share the target concept with $\mathbf{c}$. These *seed images* $\{\mathbf{I}^s\}$ are generated with ϵ_θ using a text prompt $\tilde{\mathbf{c}}$ modified from $\mathbf{c}$. Specifically, $\tilde{\mathbf{c}}$ inserts the concept tokens into a template that is intentionally kept generic (e.g.,, "`A photo of [] with background."`). Importantly, this allows $\{\mathbf{I}^s\}$ to cover diverse object shape, pose, and appearance as well as image composition and style, which is key to the expressiveness of *semantic bases*. We study the choice of N_s in Section 2.5.2.

Semantic basis. We apply DDIM sampling [61] to generate $\{\mathbf{I}^s\}$ and obtain time-dependent diffusion features $\{\mathbf{F}_t^s\}$ of size $N_s \times C \times H \times W$ from ϵ_θ. This yields $N_s \times H \times W$ distinct feature vectors, on which we perform PCA to obtain the time-dependent semantic bases $\mathbf{B}_t$ as the first N_b principal components:

$$\mathbf{B}_t = [\mathbf{p}_t^{(1)}, \mathbf{p}_t^{(2)}, ..., \mathbf{p}_t^{(N_b)}] \sim \mathrm{PCA}(\{\mathbf{F}_t^s\}) \quad (2.3)$$

Intuitively, $\mathbf{B}_t$ span semantic spaces $\mathbb{S}_t$ that connect different image modalities, allowing the propagation of image structure from $\mathbf{I}^g$ to $\mathbf{I}$ in the synthesis stage. We study the choice of $\mathbf{F}_t$ and N_b in Section 2.5.2 and Section B.

Basis reuse. Once computed, $\mathbf{B}_t$ can be reused for the same text prompt or shared by prompts with related concepts. The cost of basis construction can thus be amortized over multiple runs of the synthesis stage.

2.4.3 Synthesis Stage

The generation of $\mathbf{I}$ is conditioned on $\mathbf{I}^g$ through guidance. As a first step, we express the semantic structure of $\mathbf{I}^g$ with respect to the semantic bases $\mathbf{B}_t$.

Inversion of $\mathbf{I}^g$. We perform DDIM inversion [61] on $\mathbf{I}^g$ to obtain the diffusion features $\mathbf{F}_t^g$ of size $C \times H \times W$ and project them onto $\mathbf{B}_t$ to obtain their *semantic coordinates* $\mathbf{S}_t^g$ of size $N_b \times H \times W$. For local control of foreground structure, we further derive a mask $\mathbf{M}$ (size $H \times W$) from cross-attention maps of the concept tokens [20]. $\mathbf{M}$ is set to $\mathbf{1}$ (size $H \times W$) for global control.

We are now ready to generate $\mathbf{I}$ with *structure guidance* to control its underlying semantic structure.

Structure guidance. At each denoising step t, we obtain the semantic coordinates $\mathbf{S}_t$ by projecting the diffusion features $\mathbf{F}_t$ from ϵ_θ onto $\mathbf{B}_t$. Our energy function g_s for structure guidance can then be expressed as

$$\begin{aligned} g_s(\mathbf{S}_t; \mathbf{S}_t^g, \mathbf{M}) = & \underbrace{\frac{\sum_{i,j} m_{ij} \|[\mathbf{s}_t]_{ij} - [\mathbf{s}_t^g]_{ij}\|_2^2}{\sum_{i,j} m_{ij}}}_{\text{forward guidance}} \\ & + w \cdot \underbrace{\frac{\sum_{i,j}(1 - m_{ij}) \| \max([\mathbf{s}_t]_{ij} - \boldsymbol{\tau}_t, 0)\|_2^2}{\sum_{i,j}(1 - m_{ij})}}_{\text{backward guidance}}, \end{aligned}$$

where i and j are spatial indices for $\mathbf{S}_t$, $\mathbf{S}_t^g$ and $\mathbf{M}$, and w is the balancing weight. The thresholds $\boldsymbol{\tau}_t$ are defined as

$$\boldsymbol{\tau}_t = \max_{i,j \text{ s.t. } m_{ij}=0} [\mathbf{s}_t^g]_{ij} \tag{2.4}$$

with max taken per channel. Loosely speaking, $[\mathbf{s}_t]_{ij} > \boldsymbol{\tau}_t$ indicates the presence of foreground structure. Intuitively, the *forward* term guides the structure of $\mathbf{I}$ to align with $\mathbf{I}^g$ in the foreground, whereas the *backward* term, effective when $\mathbf{M} \neq \mathbf{1}$, helps carve out the foreground by suppressing spurious structure in the background.

While structure guidance drives $\mathbf{I}$ to form the same semantic structure as $\mathbf{I}^g$, we found that it also amplifies low-frequency textures, producing cartoony images that lack appearance details. To fix this problem, we apply *appearance guidance* to borrow texture from $\bar{\mathbf{I}}$, a sibling image of $\mathbf{I}$ generated from the same noisy latent with the same seed yet without structure guidance.

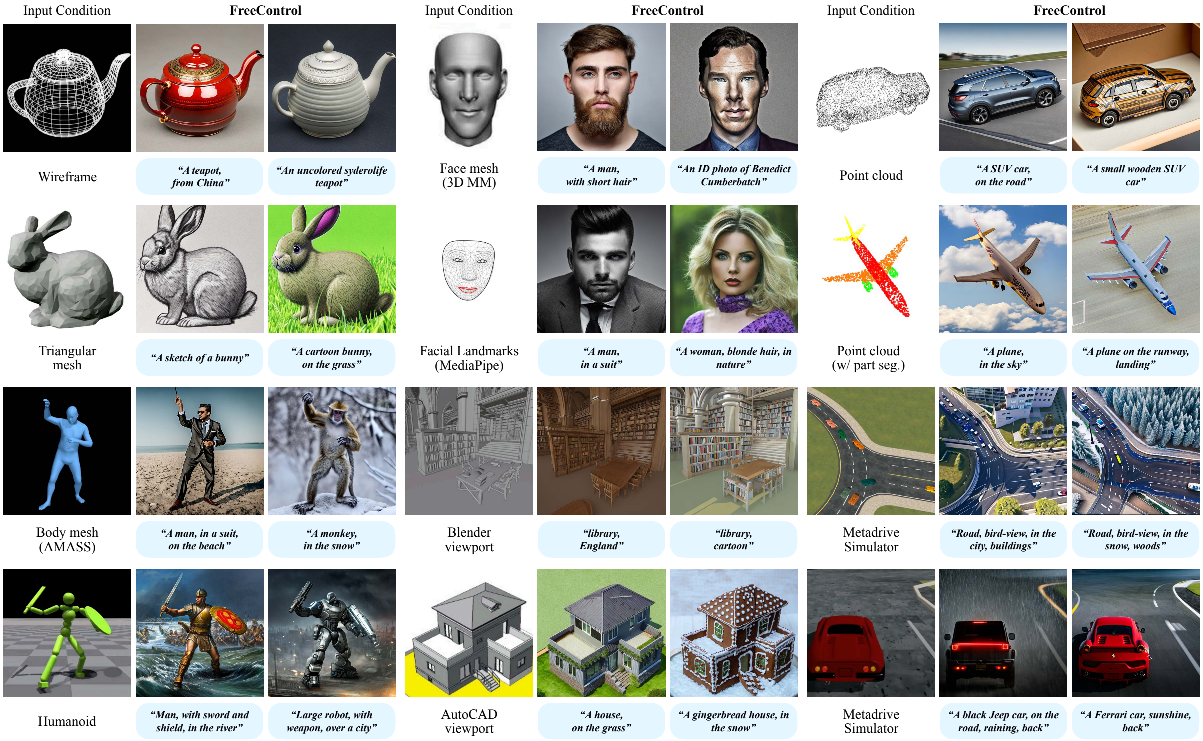

Input Condition
FreeControl
Input Condition
FreeControl
Input Condition
FreeControl
Wireframe
"A teapot, from China"
"An uncolored syderolife teapot"
Face mesh (3D MM)
"A man, with short hair"
"An ID photo of Benedict Cumberbatch"
Point cloud
"A SUV car, on the road"
"A small wooden SUV car"
Triangular mesh
"A sketch of a bunny"
"A cartoon bunny, on the grass"
Facial Landmarks (MediaPipe)
"A man, in a suit"
"A woman, blonde hair, in nature"
Point cloud (w/ part seg.)
"A plane, in the sky"
"A plane on the runway, landing"
Body mesh (AMASS)
"A man, in a suit, on the beach"
"A monkey, in the snow"
Blender viewport
"library, England"
"library, cartoon"
Metadrive Simulator
"Road, bird-view, in the city, buildings"
"Road, bird-view, in the snow, woods"
Humanoid
"Man, with sword and shield, in the river"
"Large robot, with weapon, over a city"
AutoCAD viewport
"A house, on the grass"
"A gingerbread house, in the snow"
Metadrive Simulator
"A black Jeep car, on the road, raining, back"
"A Ferrari car, sunshine, back"

Figure 2.5: **Qualitative results for more diverse control conditions.** FreeControl supports challenging control conditions not possible with training-based methods. These include 2D projections of common graphics primitives, domain-specific shape models *(point cloud, body mesh, and humanoid)*, graphics software viewports *(Blender and AutoCAD)*, and simulated driving environments *(Metadrive)*.

Appearance representation. Inspired by DSG [17], we represent image appearance as

$\{\mathbf{v}_t^{(k)}\}_{k=1}^{N_a \leq N_b}$, the weighted spatial means of diffusion features $\mathbf{F}_t$:

$$\mathbf{v}_t^{(k)} = \frac{\sum_{i,j} \sigma([s_t^{(k)}]_{ij})[\mathbf{f}_t]_{ij}}{\sum_{i,j} \sigma([s_t^{(k)}]_{ij})}, \tag{2.5}$$

where i and j are spatial indices for $\mathbf{S}_t$ and $\mathbf{F}_t$, k is channel index for $[\mathbf{s}_t]_{i,j}$, and σ is the sigmoid function. We repurpose $\mathbf{S}_t$ as weights so that different $\mathbf{v}_t^{(k)}$'s encode appearance of distinct semantic components. We calculate $\{\mathbf{v}_t^{(k)}\}$ and $\{\bar{\mathbf{v}}_t^{(k)}\}$ respectively for $\mathbf{I}$ and $\bar{\mathbf{I}}$ at each timestep t.

Appearance guidance. Our energy function g_a for appearance guidance can then be expressed as

$$g_a(\{\mathbf{v}_t^{(k)}\}; \{\bar{\mathbf{v}}_t^{(k)}\}) = \frac{\sum_{k=1}^{N_a} \|\mathbf{v}_t^{(k)} - \bar{\mathbf{v}}_t^{(k)}\|_2^2}{N_a}. \tag{2.6}$$

It penalizes difference in the appearance representations and thus facilitates appearance transfer from $\bar{\mathbf{I}}$ to $\mathbf{I}$.

Guiding the generation process. Finally, we arrive at our modified score estimate $\hat{\epsilon}_t$ by including structure and appearance guidance alongside classifier-free guidance [26]:

$$\hat{\epsilon}_t = (1+s)\,\epsilon_\theta(\mathbf{x}_t; t, \mathbf{c}) - s\,\epsilon_\theta(\mathbf{x}_t; t, \emptyset) + \lambda_s\, g_s + \lambda_a\, g_a, \tag{2.7}$$

where s, λ_s and λ_a are the respective guidance strengths, and $\emptyset$ denotes the null token input.

2.5 Experiments and Results

We report extensive qualitative and quantitative results to demonstrate the effectiveness and generality of our approach for zero-shot controllable T2I diffusion. We present additional results on text-guided image-to-image translation and provide ablation studies on key method components.

2.5.1 Controllable T2I Diffusion

Baselines. ControlNet [75] and T2I-Adapter [41] learn an auxiliary module to condition a pretrained diffusion model on a guidance image. One such module is learned for each condition type. Uni-ControlNet [78] instead learns adapters shared by all condition types for all-in-one control. Different from these training-based methods, SDEdit [38] adds noise to a guidance image and subsequently denoises it with a pretrained diffusion model for guided image synthesis. Prompt-to-Prompt (P2P) [23] and Plug-and-Play (PnP) [64] manipulate attention weights and features inside pretrained diffusion models for zero-shot image editing. We compare our method with these strong baselines in our experiments.

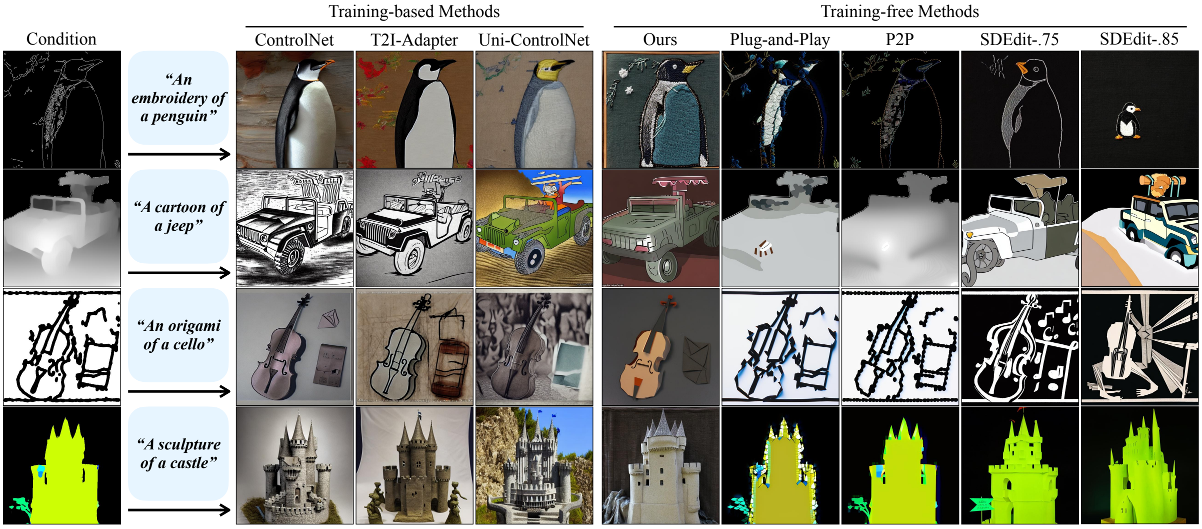


Figure 2.6: **Qualitative comparison on controllable T2I diffusion.** FreeControl achieves competitive spatial control and superior image-text alignment in comparison to training-based methods. It also escapes the appearance leakage problem manifested by the training-free baselines, producing high-quality images with rich content and appearance faithful to the text prompt.

Experiment setup. Similar to ControlNet [75], we report qualitative results on eight condition types (sketch, normal, depth, Canny edge, M-LSD line, HED edge, segmentation mask, and human pose). We further employ several previously unseen control signals as input conditions (Figure 2.5), and combine our method with all major versions of Stable Diffusion (1.5, 2.1, and XL 1.0) to study its generalization on diffusion model architectures.

Method	Canny			HED			Sketch			Depth			Normal		
	Self-Sim ↓	CLIP ↑	LPIPS ↑	Self-Sim ↓	CLIP ↑	LPIPS ↑	Self-Sim ↓	CLIP ↑	LPIPS ↑	Self-Sim ↓	CLIP ↑	LPIPS ↑	Self-Sim ↓	CLIP ↑	LPIPS ↑
ControlNet [75]	0.042	0.300	0.665	0.040	0.291	0.609	0.070	0.314	0.668	0.058	0.306	0.645	0.079	0.304	0.637
T2I-Adapter	0.052	0.290	0.689	-	-	-	0.096	0.290	0.648	0.071	0.314	0.673	-	-	-
Uni-ControlNet	0.044	0.295	0.539	0.050	0.301	0.553	0.050	0.301	0.553	0.061	0.303	0.636	-	-	-
SDEdit-0.75 [38]	0.108	0.306	0.582	0.123	0.288	0.375	0.135	0.281	0.361	0.153	0.294	0.327	0.128	0.284	0.456
SDEdit-0.85 [38]	0.139	0.319	0.670	0.153	0.305	0.485	0.139	0.300	0.485	0.165	0.304	0.384	0.147	0.298	0.512
P2P [23]	0.078	0.253	0.298	0.112	0.253	0.194	0.194	0.251	0.096	0.142	0.248	0.167	0.100	0.249	0.198
PNP [64]	**0.074**	0.282	0.417	0.098	0.286	0.271	0.158	0.267	0.221	0.126	0.287	0.268	0.107	0.286	0.347
FreeControl (Ours)	0.080	**0.322**	**0.724**	**0.078**	**0.321**	**0.561**	**0.090**	**0.322**	**0.611**	**0.090**	**0.321**	**0.576**	**0.086**	**0.322**	**0.642**

Table 2.1: **Quantitative results on controllable T2I diffusion.** FreeControl consistently outperforms all training-free baselines in structure preservation, image-text alignment and appearance diversity as measured by Self-similarity distance, CLIP score and LPIPS distance. It achieves competitive structure and appearance scores with the training-based baselines while demonstrate stronger image-text alignment.

For a fair comparison with the baselines, we adapt the ImageNet-R-TI2I dataset from PnP [64] as our benchmark dataset. It contains 30 images from 10 object categories. Each image is associated with five text prompts originally for the evaluation of text-guided image-to-image translation. We convert the images into their respective Canny edge, HED edge, sketch, depth map, and normal map following ControlNet [75], and subsequently use them as input conditions for all methods in our experiments.

Evaluation metrics. We report three widely adopted metrics for quantitative evaluation; *Self-similarity distance* [63] measures the structural similarity of two images in the feature space of DINO-ViT [13]. A smaller distance suggests better structure preservation. Similar to [64], we report self-similarity between the generated image and the dataset image that produces the input condition. *CLIP score* [49] measures image-text alignment in the CLIP embedding space. A higher CLIP score indicates a stronger semantic match between the text prompt and the generated image. *LPIPS distance* [76] measures the appearance deviation of the generated image from the input condition. Images with richer appearance details yield higher LPIPS score.

Implementation details. We adopt keys from the first self-attention in the U-Net decoder as the features $\mathbf{F}_t$. We run DDIM sampling on $N_s = 20$ seed images for 200 steps to obtain bases of size $N_b = 64$. In the synthesis stage, we run DDIM inversion on $\mathbf{I}^g$ for 1000 steps, and sample $\mathbf{I}$ and $\bar{\mathbf{I}}$ by running 200 steps of DDIM sampling. Structure and appearance guidance are applied in the first 120 steps. $\lambda_s \in [400, 1000]$, $\lambda_a = 0.2\lambda_s$, and

$N_a = 2$ in all experiments.

Qualitative results. As shown in Figure 2.4, FreeControl is able to recognize diverse semantic structures from all condition modalities used by ControlNet [75]. It produces high-quality images in close alignment with both the text prompts and spatial conditions. Importantly, it generalizes well on all major versions of Stable Diffusion, enabling effortless upgrade to future model architectures without retraining.

In Figure 2.5, we present additional results for condition types not possible with previous methods. FreeControl generalizes well across challenging condition types for which constructing training pairs is difficult. In particular, it enables superior conditional control with common graphics primitives (e.g.,, mesh and point cloud), domain-specific shape models (e.g.,, face and body meshes), graphics software viewports (e.g.,, Blender [15] and AutoCAD [1]), and simulated driving environments (e.g.,, MetaDrive [35]), thereby providing an appealing solution to visual design preview and sim2real.

Comparison with baselines. Figure 2.6 and Table 2.1 compare our methods to the baselines. Despite stronger structure preservation (*i.e.*, small self-similarity distances), the training-based methods at times struggle to follow the text prompt (e.g.,*embroidery* for ControlNet and *origami* for all baselines) and yield worse CLIP scores. The loss of text control is a common issue in training-based methods due to modifications made to the pretrained models. Our method is training-free, hence retaining strong text conditioning.

In contrast, training-free baselines are prone to appearance leakage, where the appearance of condition images is leaked to generated images, resulting in worse LIPIS scores. This is because the generated image shares latent states (SDEdit) or diffusion features (P2P & PnP) with the condition. For example, all baselines inherit the texture-less background in the *embroidery* example and the foreground shading in the *castle* example. Our method instead decouples structure and appearance, thereby avoiding the leakage.

Handling conflicting conditions. We study cases where spatial conditions have minor conflicts to input text prompts. We assume that a text prompt consists of a concept (e.g.,, batman) and a style (e.g.,, cartoon), and contrast a conflicting case with its aligned version.

Specifically, a conflicting case includes (a) a text prompt with a feasible combination of concept and style; and (b) a spatial condition (*i.e.*an edge map) derived from real images without the text concept. The corresponding aligned case contains a similar text prompt, yet using a spatial condition from real images with the same concept. We input those cases into ControlNet, T2I-Adapter, and FreeControl, using a set of pre-trained and customized models.

Figure 2.7 shows the results. Our training-free FreeControl consistently generates high quality images that fit the middle ground of spatial conditions and text prompts, across all test cases and models. T2I-Adapter sometimes fails even with an aligned case (see *Batman* examples), not to mention the conflicting cases. Indeed, T2I-Adapter tends to disregard the condition image, leading to diminished controllability, as exemplified by *Emma Watson* example (conflicting). ControlNet can generate convincing images for aligned cases, yet often fall short in those conflicting cases. A common failure mode is to overwrite the input text concept using the condition image, as shown by *skeleton bike* or *house in a bubble* examples (conflicting).

Extension to Image-to-Image Translation FreeControl can be readily extended to support image-to-image (I2I) translation by conditioning on a detailed/real image. A key challenge here is to allow FreeControl to preserve the background provided by the condition, *i.e.*, the input content image. To this end, we propose two variants of FreeControl. The first removes the mask $\mathbf{M}$ in structure guidance (*i.e.*, w/o mask), and the second generates from the inverted latent $\mathbf{x}_T^g$ of the condition image (*i.e.*, fixed seed). We find that removing the mask helps extract and maintain the background structure, and starting inference from $\mathbf{x}_T^g$ retains the appearance from the condition image.

Figure 2.8 evaluates FreeControl and its two variants for text-guided I2I, and compares to strong baselines for the I2I task including PnP [64], P2P [23], pix2pix-zero [45] and SDEdit [38]. The vanilla FreeControl, as we expect, often fails to preserve the background. However, our two variants with simple modification demonstrate impressive results as compared to the baselines, generating images that adhere to both foreground and

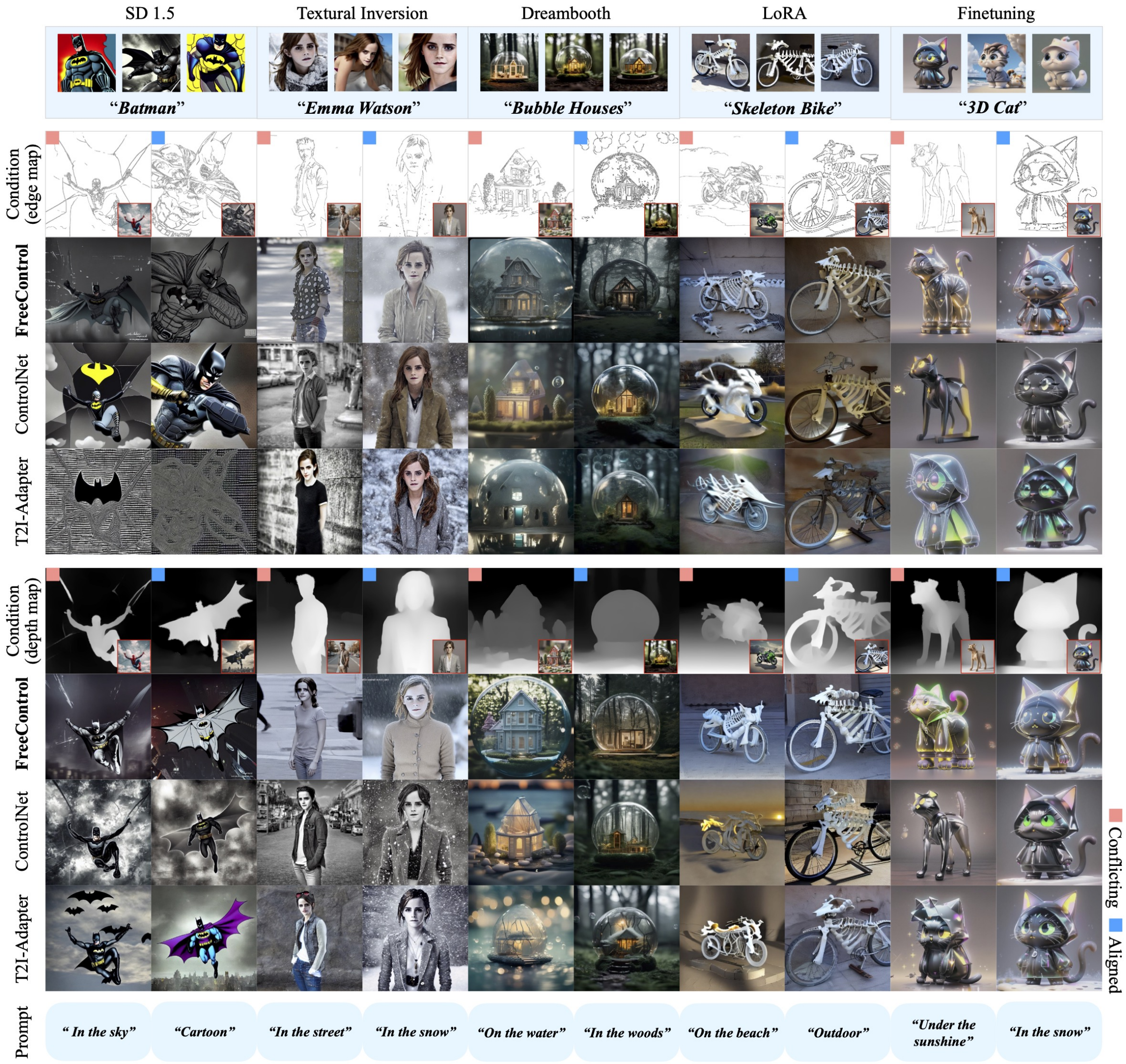


Figure 2.7: **Controllable T2I generation of custom concepts.** FreeControl is compatible with major customization techniques and readily supports controllable generation of custom concepts without requiring spatially-aligned condition images. By contrast, ControlNet fails to preserve custom concepts given conflicting conditions, whereas T2I-Adapter refuses to respect the condition image and text prompt.

background of the input image.

Further, we evaluate the *self-similarity distance* and *CLIP score* of FreeControl, its variants, and our baselines on the ImageNet-R-TI2I dataset. The results are summarized in Figure 2.8. Variants of FreeControl outperform all baselines with significantly improved structure preservation and visual fidelity, following the input text prompts.

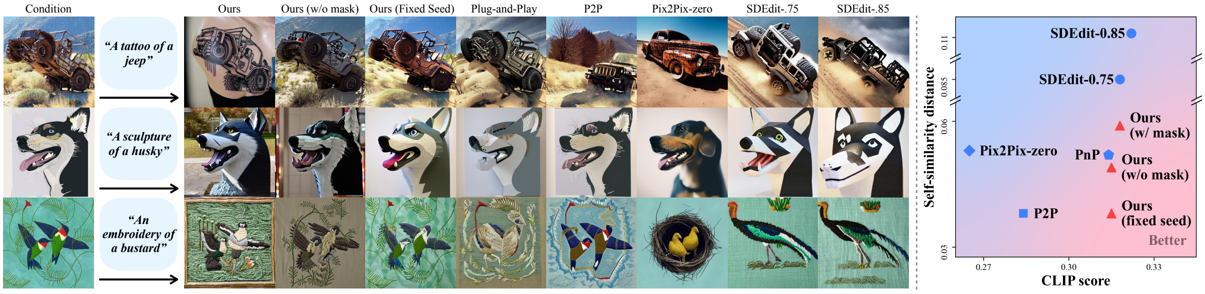


Figure 2.8: **Qualitative and quantitative comparison on text-guided image-to-image translation.** FreeControl enables flexible control of image composition and style through guidance mask $\mathbf{M}$ and random seed (*left*). It strikes a good balance between structure preservation (self-similarity distance) and image-text alignment (CLIP score) in comparison to the baselines (*right*, better towards bottom right).

Continuous control. Real-world content creation is a live experience, where an idea develops from a sketch into a more refined and finished piece of work. The intermediate states throughout this process may be interpreted as continuously evolving control signals. Figure 2.10 illustrates how FreeControl may assist an artist in his or her content creation experience. It produces spatially accurate and smoothly varying outputs guided by constantly changing conditions, thus serving as a source of inspiration over the course of painting.

Compositional control. By combining structure guidance from multiple condition images, FreeControl readily supports compositional control without altering the synthesis pipeline. Figure 2.11 presents our results using different combinations of condition types. The generated images are faithful to all input conditions while respect the text prompt.

Combination with ControlNet. Figure 2.9 demonstrates the results of combining FreeControl and ControlNet(canny), using the wireframe of a teapot and the mesh of a bunny as the condition. We use FreeContorl to denoise the latent for 30 steps, ControlNet for the next 70 steps, and the vanilla Stable Diffusion for the rest 100 steps. This hybrid approach improves the structural alignment of FreeContorl, unlocks the appearance customization, improves textual alignment, and accommodates un-trained conditions for ContorlNet.

Inference efficiency. We further study the inference cost of our method in comparison

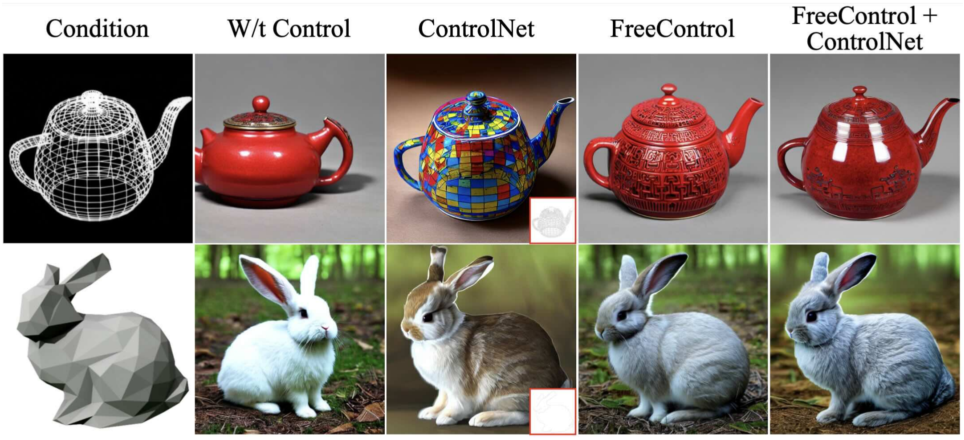


Figure 2.9: **Qualitative results of combining ControlNet and FreeControl.** Top: *"A Chinese teapot, red"*; Bottom: *"A bunny, in the forest"*.

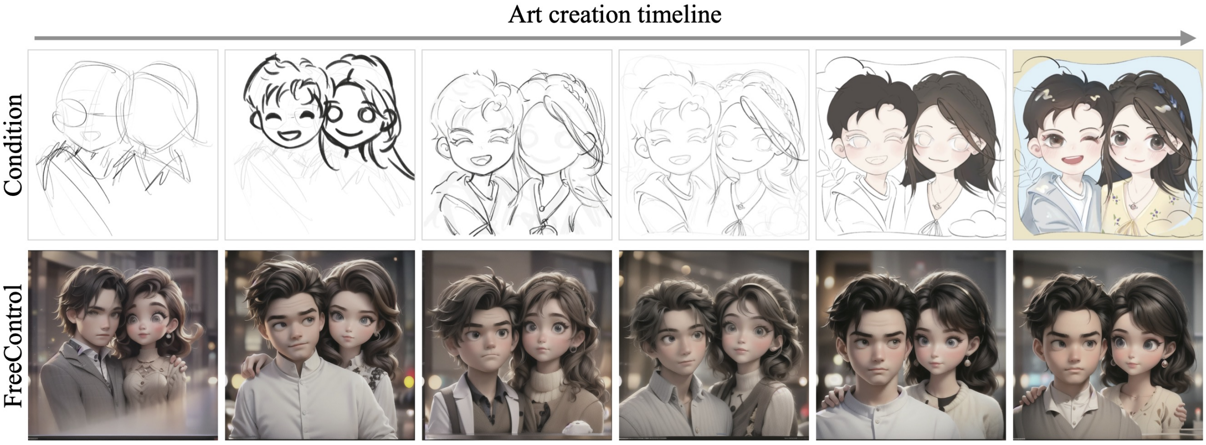


Figure 2.10: **Controllable generation over the course of art creation.** Images are generated from the same seed with the prompt *"a photo of a man and a woman, Pixar style"* with a customized model from [2]. FreeControl yields accurate and consistent results despite evolving control conditions throughout the art creation timeline.

to training-free baselines. Table 2.2 reports the average inference time using a single Nvidia A6000 GPU. The inference has three stages: (1) *Pre-processing stage*, where category-level information is extracted (analysis stage in FreeControl and the computation of edit direction in Pix2Pix-zero) ; (2) *Inversion stage*, for extracting the image-level latent representation from the input condition; and (3) *Sampling stage*, for generating the target image. FreeControl is slower than PnP (**4.2**×) and P2P (**1.8**×), yet much faster than Pix2Pix-zero (**0.14**×). When considering the reused basis and thus only counting inversion and inference time, FreeControl can achieve **1.1**× that of PnP, **0.5**× that of

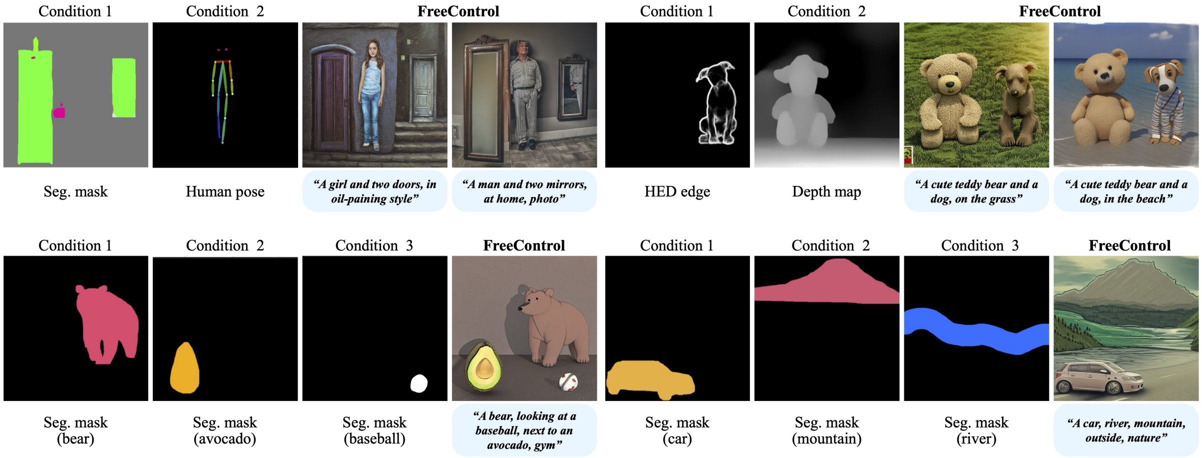


Figure 2.11: **Qualitative results on compositional control**. FreeControl allows compositional control of image structure using multiple condition images of potentially different modalities.

	FreeControl	PnP	Pix2Pix-zero	P2P+NTI
Pre-processing	127.00	0	1236.00	0
Inversion	25.36	31.96	32.57	87.51
Sampling	23.95	10.09	33.03	11.51
Total	176.31	42.05	1301.60	99.02

Table 2.2: **Runtime for training-free methods**

P2P, and **0.75**× that of Pix2Pix-zero, yet still generate diverse images.

2.5.2 Ablation Study

Effect of guidance. As seen in Figure 2.12, structure guidance is responsible for structure alignment ($-g_s$ v.s.Ours). Appearance guidance alone has no impact on generation in the absence of structure guidance ($-g_a$ v.s.$-g_s$, $-g_a$). It only becomes active after image structure has shaped up, in which case it facilitates appearance transfer ($-g_a$ v.s.Ours).

Choice of diffusion features $\mathbf{F}_t$. Figure 2.13 compares results using self-attention keys, queries, values, and their preceding Conv features from up_block.[1,2] in the U-Net decoder. It reveals that up_block.1 in general carries more structural cues than up_block.2, whereas keys better disentangle semantic components than the other features.

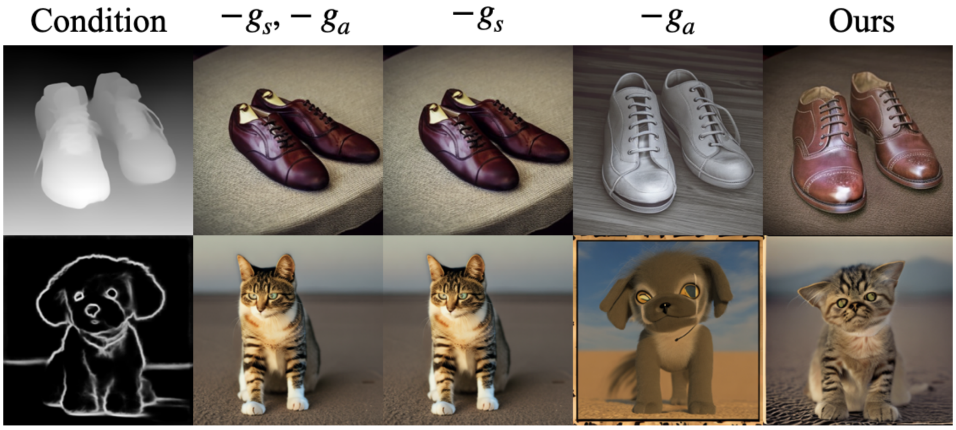


Figure 2.12: **Ablation on guidance effect.** Top: *"leather shoes"*; Bottom: *"cat, in the desert"*. g_s and g_a stand for structure and appearance guidance, respectively.

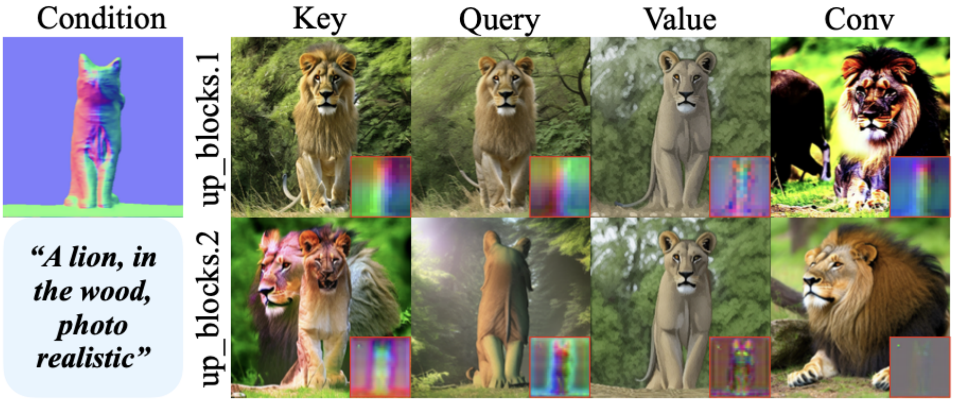


Figure 2.13: **Ablation on feature choice.** Keys from self-attention of up_block.1 in the U-Net decoder expose the strongest controllability. PCA visualization of the features are in the insets.

Size of semantic bases N_b. Figure 2.14 presents generation results over the full spectrum of N_b. A larger N_b improves structure alignment yet triggers the unintended transfer of appearance from the input condition. Hence, a good balance is achieved with N_b's in the middle range.

Number of seed images N_s. Figure 2.15 suggests that N_s has minor impact on image quality and controllability, allowing the use of *as few as* 1 *seed image* in the analysis stage. Large N_s diversifies image content and style, which helps perfect structural details (e.g.,, limbs) in the generated images.

Choice of threshold τ_t. Figure 2.16 demonstrates that no *hard* threshold within the

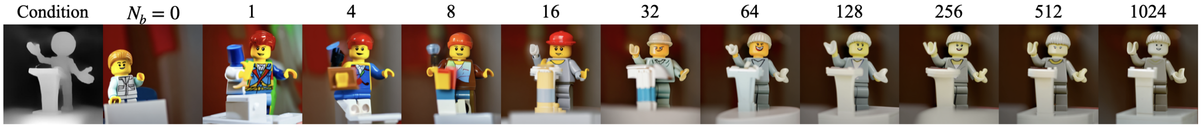


Figure 2.14: **Ablation on size of semantic bases N_b.** Images are generated using the prompt *"a Lego man giving a lecture"*. They illustrate an inherent tradeoff between structure and appearance quality. A good balance can be achieved with N_b's in the middle range.

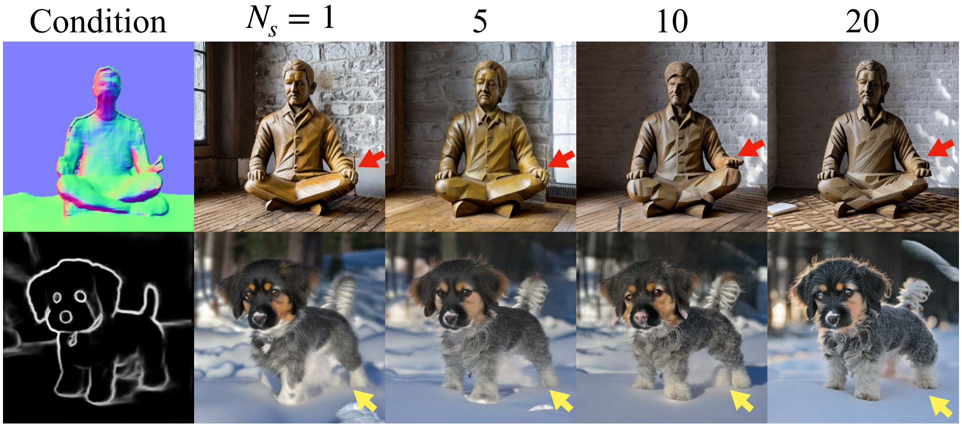


Figure 2.15: **Ablation on number of seed images N_s.** Top: *"wooden sculpture of a man"*; Bottom: *"dog, in the snow"*. Larger N_s brings minor improvement on structure alignment.

range of $[0, 1]$ can fully eliminate spurious background signal while ensure a foreground structure consistent with the condition image. By contrast, our *dynamic* thresholding scheme, implemented as a per-channel `max` operation, allows FreeControl to accurately carve out the foreground without interference from the background.

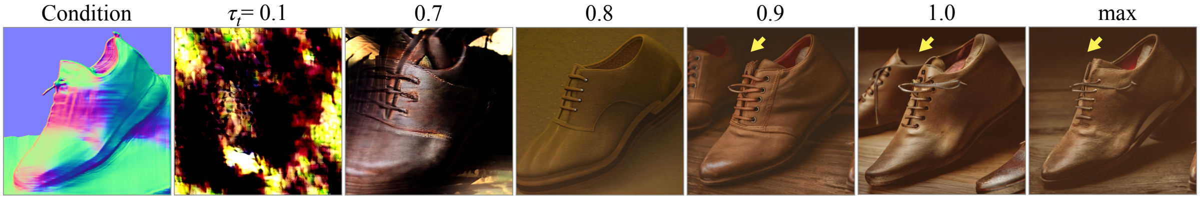


Figure 2.16: **Ablation on threshold τ_t.** Images are generated using the prompt *"leather shoe on the table"*. Our dynamic threshold (max) encourages more faithful foreground structure and cleaner background in comparison to various hard thresholds (e.g.„ 0.1).

Number of guidance steps. Figure 2.17 reveals that the first 40% sampling steps are key to structure and appearance formation. Applying guidance beyond that point has little to no impact on generation quality.

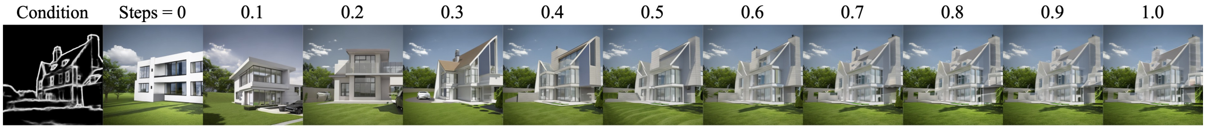


Figure 2.17: **Ablation on the number of guidance steps.** Images are generated using the prompt *"a modern house, on the grass, side look"*. Applying guidance beyond the first 40% diffusion steps (0.4) has little to no impact on the generation result.

Choice of guidance weights λ_s and λ_a. Figure 2.19 confirms that FreeControl produces strong results within a wide range of guidance strengths. In particular, the output images yield accurate spatial structure when $\lambda_s \geq 400$ and rich appearance details when $\lambda_a \geq 0.2\lambda_s$. We empirically found that these ranges work for all examples in our experiments.

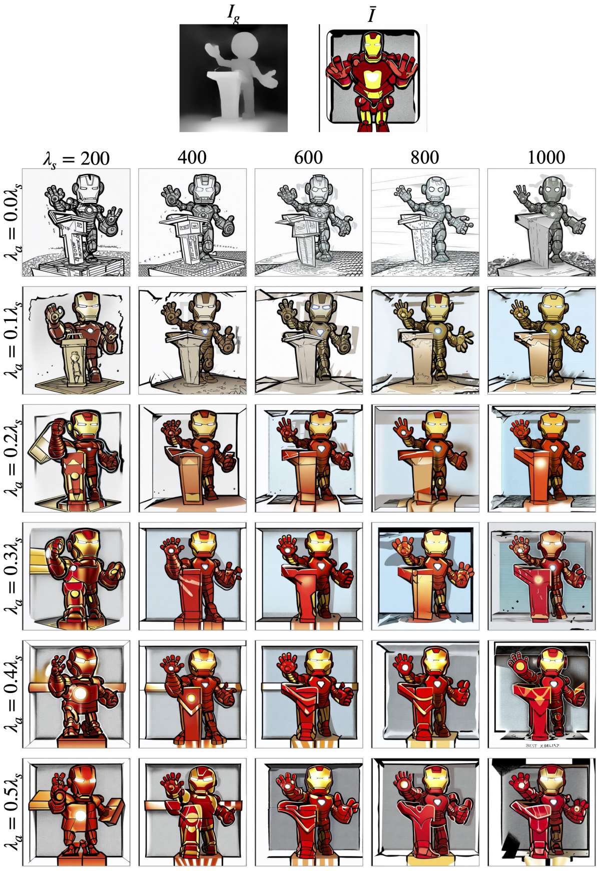

I_g
$\bar{I}$
$\lambda_s = 200$
400
600
800
1000
$\lambda_a = 0.0\lambda_s$
$\lambda_a = 0.1\lambda_s$
$\lambda_a = 0.2\lambda_s$
$\lambda_a = 0.3\lambda_s$
$\lambda_a = 0.4\lambda_s$
$\lambda_a = 0.5\lambda_s$

Figure 2.18: **Ablation on guidance weights λ_s and λ_a.** Images are generated with the prompt *"an iron man is giving a lecture"*. FreeControl yields strong results across guidance weights.

Basis reuse across concepts. Once computed, the semantic bases $\mathbf{S}_t$ can be reused for the control of semantically related concepts. Figure 2.19 provides one such example, where $\mathbf{S}_t$ derived from seed images of `man` generalize well on other mammals including `cat`, `dog` and `monkey`, yet fail for the semantically distant concept of `bedroom`.

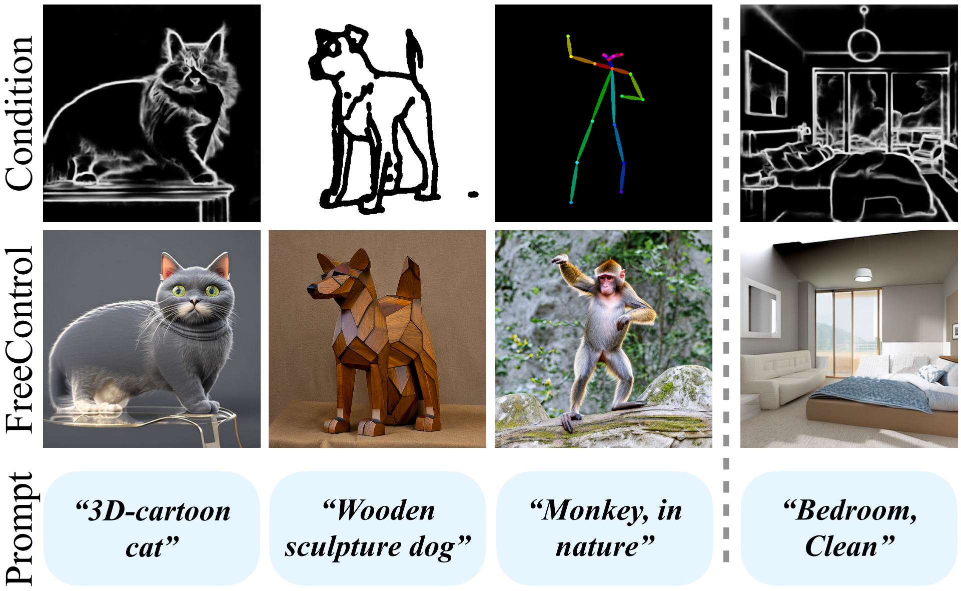


Figure 2.19: **Ablation on basis reuse.** The semantic bases computed for *"man"* enable the controllable generation of semantically related concepts (cat, dog, and monkey) while falling short for unrelated concepts (bedroom).

2.6 Conclusion

We present FreeControl, a training-free method for spatial control of any T2I diffusion model with any condition. FreeControl exploits the feature space of pretrained T2I models, facilitates convenient control over many architectures and checkpoints, allows various challenging input conditions on which most of the existing training-free methods fail, and achieves competitive synthesis quality with training-based approaches. One limitation is that FreeContorl relies on the DDIM inversion process to extract intermediate features of the guidance image and compute additional gradients during the synthesis stage, resulting in increased inference time. We hope our findings and analysis can shed light on controllable visual content creation.

CHAPTER 3

Ctrl-X: Controlling Structure and Appearance for Text-To-Image Generation Without Guidance

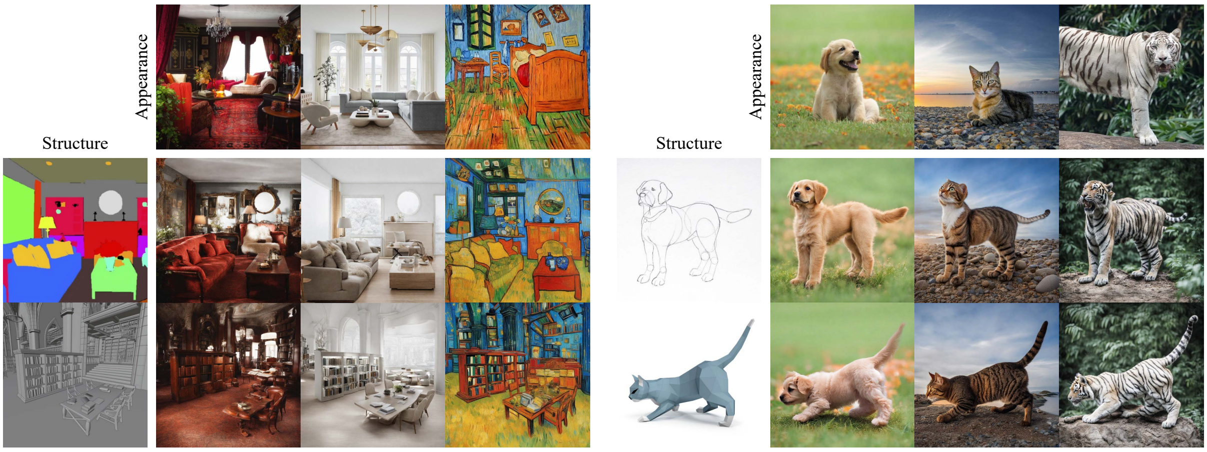


Figure 3.1: **Guidance-free structure and appearance control of Stable Diffusion XL (SDXL) [48]**. Ctrl-X enables training-free and guidance-free zero-shot control of pretrained text-to-image diffusion models given any structure conditions and appearance images.

3.1 Introduction

The rapid advance of large text-to-image (T2I) generative models has made it possible to generate high-quality images with just one text prompt. However, it remains challenging to specify the exact concepts that can accurately reflect human intents using only textual descriptions. Recent approaches like ControlNet [75] and IP-Adapter [74] have enabled controllable image generation upon pretrained T2I diffusion models regarding structure and appearance, respectively. Despite the impressive results in controllable generation, these approaches [75, 41, 78, 36] require fine-tuning the entire generative model or training auxiliary modules on large amounts of paired data.

Training-free approaches [17, 39, 9] have been proposed to address the high overhead associated with additional training stages. These methods optimize the latent embedding across diffusion steps using specially designed score functions to achieve finer-grained control than text alone with a process called guidance. Although training-free approaches avoid the training cost, they significantly increase computing time and required GPU memory in the inference stage due to the additional backpropagation over the diffusion network. They also require sampling steps that are 2–20 times longer. Furthermore, as the expected latent distribution of each time step is predefined for each diffusion model, it is critical to tune the guidance weight delicately for each score function; Otherwise, the latent might be out-of-distribution and lead to artifacts and reduced image quality.

To tackle these limitations, we present *Ctrl-X*, a simple *training-free* and *guidance-free* framework for T2I diffusion with structure and appearance control. We name our method "Ctrl-X" because we reformulate the controllable generation problem by 'cutting' (and 'pasting') two tasks together: Spatial structure preservation and semantic-aware stylization. Our insight is that diffusion feature maps capture rich spatial structure and high-level appearance from early diffusion steps sufficient for structure and appearance control without guidance. To this end, Ctrl-X employs feature injection and spatially-aware normalization in the attention layers to facilitate structure and appearance alignment with user-provided images. By being guidance-free, Ctrl-X eliminates additional optimization overhead and sampling steps, resulting in a 40-fold increase in inference speed compared to guidance-based methods. Figure 3.1 shows some generation results. Moreover, Ctrl-X supports arbitrary structure conditions beyond natural images and can be applied to any T2I and even text-to-video (T2V) diffusion models. Extensive quantitative and qualitative experiments demonstrate the superior image quality and appearance alignment of our method over prior works.

We summarize our contributions as follows:

1. We present *Ctrl-X*, a simple plug-and-play method that builds on pretrained text-

to-image diffusion models to provide disentangled and zero-shot control of structure and appearance during the generation process requiring no additional training or guidance.

2. Ctrl-X presents the first universal guidance-free solution that supports multiple conditional signals (structure and appearance) and model architectures (e.g.,text-to-image and text-to-video).

3. Our method demonstrates superior results compared to previous training-based and guidance-based baselines (e.g.,ControlNet + IP-Adapter [75, 74] and FreeControl [39]) in terms of condition alignment, text-image alignment, and image quality.

3.2 Related work

Diffusion structure control Previous spatial structure control methods can be categorized into two types (training-based v.s.training-free) based on whether they require training on paired data.

Training-based structure control methods require paired condition-image data to train additional modules or fine-tune the entire diffusion network to facilitate generation from spatial conditions [75, 41, 36, 78, 73, 7, 79, 68, 80]. While pixel-level spatial control can be achieved with this approach, a significant drawback is needing a large number of condition-image pairs as training data. Although some condition data can be generated from pretrained annotators (e.g.,depth and segmentation maps), other condition data is difficult to obtain from given images (e.g.,3D mesh, point cloud), making these conditions challenging to follow. Compared to these training-based methods, Ctrl-X supports conditions where paired data is challenging to obtain, making it a more flexible and effective solution.

Training-free structure control methods typically focus on specific conditions. For example, R&B [69] facilitates bounding-box guided control with region-aware guidance, and DenseDiffusion [31] focuses on generating images with segmentation map conditions by controlling the attention weights. Universal Guidance [9] employs various pretrained classifiers to support multiple types of condition signals. FreeControl [39] analyzes semantic correspondence in the sub-space of diffusion features and harnesses it to support spatial control from any visual condition. While these approaches do not require training data, they usually need to compute the gradient of the latent to lower an auxiliary loss, which requires substantial computing time and GPU memory. In contrast, Ctrl-X requires no guidance at the inference stage and controls structure via direct feature injections, enabling faster and more robust image generation with spatial control.

Diffusion appearance control Existing appearance control methods that build upon pretrained diffusion models can also similarly be categorized into two types (training-based v.s.training-free).

Training-based appearance control methods can be divided into two categories: Those trained to handle any image prompt and those overfitting to a single instance. The first category [75, 41, 74, 68] trains additional image encoders or adapters to align the generated process with the structure or appearance from the reference image. The second category [55, 27, 19, 6, 47, 56] is typically applied to customized visual content creation by finetuning a pretrained text-to-image model on a small set of images or binding special tokens to each instance. The main limitation of these methods is that the additional training required makes them unscalable. However, Ctrl-X offers a scalable solution to transfer appearance from any instance without training data.

Training-free appearance control methods generally follow two approaches: One approach [5, 11, 72] manipulates self-attention features using pixel-level dense correspondence between the generated image and the target appearance, and the other [17, 39] extracts appearance embeddings from the diffusion network and transfers the appearance by guiding the diffusion process towards the target appearance embedding. A key limitation of these approaches is that a single text-controlled target cannot fully capture the details of the target image, and the latter methods require additional optimization steps. By contrast, our method exploits the spatial correspondence of self-attention layers to achieve

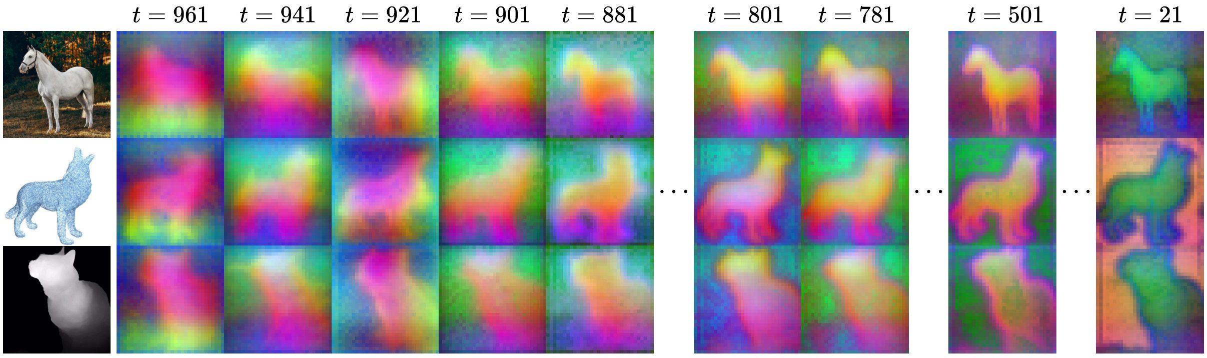


Figure 3.2: **Visualizing early diffusion features** Using 20 real, generated, and condition images of animals, we extract Stable Diffusion XL [48] features right after decoder layer 0 convolution. We visualize the top three principal components computed for each time step across all images. $t = 961$ to 881 corresponds to inference steps 1 to 5 of the DDIM scheduler with 50 time steps. We obtain $\mathbf{x}_t$ by directly adding Gaussian noise to each clean image $\mathbf{x}_0$ via the diffusion forward process.

semantically-aware appearance transfer without targeting specific subjects.

3.3 Preliminaries

Diffusion models are a family of probabilistic generative models characterized by two processes: The *forward process* iteratively adds Gaussian noise to a clean image $\mathbf{x}_0$ to obtain $\mathbf{x}_t$ for time step $t \sim [1, T]$, which can be reparameterized in terms of a noise schedule α_t where

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1 - \alpha_t}\epsilon \tag{3.1}$$

for $\epsilon \sim \mathcal{N}(0, \mathbf{I})$; The *backward process* generates images by iteratively denoising an initial Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, also known as diffusion sampling [24]. This process uses a parameterized denoising network ϵ_θ conditioned on a text prompt $\mathbf{c}$, where at time step t we obtain a cleaner $\mathbf{x}_{t-1}$

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\hat{\mathbf{x}}_0 + \sqrt{1 - \alpha_{t-1}}\epsilon_\theta(\mathbf{x}_t \mid t, \mathbf{c}), \qquad \hat{\mathbf{x}}_0 := \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t}\epsilon_\theta(\mathbf{x}_t \mid t, \mathbf{c})}{\sqrt{\alpha_t}}. \tag{3.2}$$

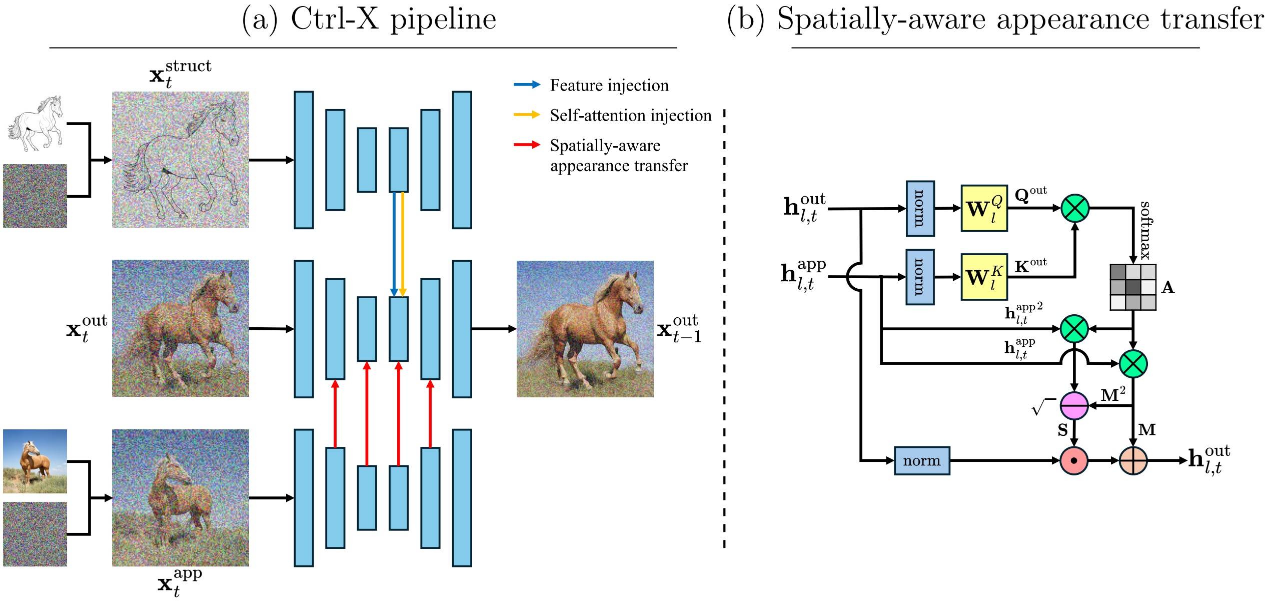
Formally, $\epsilon_\theta(\mathbf{x}_t \mid t, \mathbf{c}) \approx -\sigma_t \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t \mid t, \mathbf{c})$ approximates a score function scaled by a noise schedule σ_t that points towards high density of data, i.e., $\mathbf{x}_0$, at noise level t [62].

Guidance The iterative inference of diffusion enables us to guide the sampling process on auxiliary information. *Guidance* modifies Equation 3.2 to compose additional score functions that point towards richer and specifically conditioned distributions [9, 17], expressed as

$$\hat{\epsilon}_\theta(\mathbf{x}_t \mid t, \mathbf{c}) = \epsilon(\mathbf{x}_t \mid t, \mathbf{c}) - s\,\mathbf{g}(\mathbf{x}_t \mid t, y), \tag{3.3}$$

where $\mathbf{g}$ is an energy function and s is the guidance strength. In practice, $\mathbf{g}$ can range from classifier-free guidance (where $\mathbf{g} = \epsilon$ and $y = \varnothing$, *i.e.*the empty prompt) to improve image quality and prompt adherence for T2I diffusion [26, 53], to arbitrary gradients $\nabla_{\mathbf{x}_t}\ell(\epsilon(\mathbf{x}_t \mid t, \mathbf{c}) \mid t, y)$ computed from auxiliary models or diffusion features common to guidance-based controllable generation [9, 17, 39]. Consequently, though guidance provides great customizability on the type and variety of conditioning for controllable generation, as it only requires any loss that can be backpropagated to $\mathbf{x}_t$, this backpropagation requirement often translates to slow inference time and high memory usage. Moreover, as guidance-based methods often compose multiple energy functions, tuning the guidance strength s for each $\mathbf{g}$ may be finicky and present robustness issues. Thus, Ctrl-X avoids guidance and provides instant applicability to larger T2I and T2V models with minor hyperparameter tuning.

Diffusion U-Net architecture Many pretrained T2I diffusion models are text-conditioned U-Nets, which contains an encoder and decoder that downsamples and then upsamples the input $\mathbf{x}_t$ to predict ϵ, with long skip connections between matching encoder and decoder resolutions [24, 53, 48]. Each encoder/decoder block contains convolution layers, self-attention layers, and cross-attention layers: The first two both control structure and appearance, and the last injects textual information. Thus, many training-free controllable generation methods utilize these layers, whether through direct manipulation [23, 64, 32, 5, 72] or for computing guidance losses [17, 39], with self-attention commonly used: Let $\mathbf{h}_{l,t} \in \mathbb{R}^{(hw)\times c}$ be the diffusion feature with height h, width w, and channel size



(a) Ctrl-X pipeline

(b) Spatially-aware appearance transfer

$\mathbf{x}_t^{\text{struct}}$
Feature injection
Self-attention injection
Spatially-aware appearance transfer
$\mathbf{x}_t^{\text{out}}$
$\mathbf{x}_{t-1}^{\text{out}}$
$\mathbf{x}_t^{\text{app}}$
$\mathbf{h}_{l,t}^{\text{out}}$
$\mathbf{h}_{l,t}^{\text{app}}$
norm
norm
$\mathbf{W}_l^Q$
$\mathbf{W}_l^K$
$\mathbf{Q}^{\text{out}}$
$\mathbf{K}^{\text{out}}$
softmax
$\mathbf{A}$
$\mathbf{h}_{l,t}^{\text{app}2}$
$\mathbf{h}_{l,t}^{\text{app}}$
$\sqrt{\ }$
$\mathbf{M}^2$
$\mathbf{S}$
$\mathbf{M}$
norm
$\mathbf{h}_{l,t}^{\text{out}}$

Figure 3.3: **Overview of Ctrl-X** (a) At each sampling step t, we obtain $\mathbf{x}_t^{\text{s}}$ and $\mathbf{x}_t^{\text{a}}$ via the forward diffusion process, feeding them into the T2I diffusion model to obtain their convolution and self-attention features. Then, we inject convolution and self-attention features from $\mathbf{x}_t^{\text{s}}$ and leverage self-attention correspondence to transfer spatially-aware appearance statistics from $\mathbf{x}_t^{\text{a}}$ to $\mathbf{x}_t^{\text{o}}$. (b) Details of our spatially-aware appearance transfer, where we exploit self-attention correspondence between $\mathbf{x}_t^{\text{o}}$ and $\mathbf{x}_t^{\text{a}}$ to compute weighted feature statistics $\mathbf{M}$ and $\mathbf{S}$ applied to $\mathbf{x}_t^{\text{o}}$.

c at time step t right before attention layer l. Then, the self-attention operation is

$$\begin{gathered}\mathbf{Q} := \mathbf{h}_{l,t}\mathbf{W}_l^Q \quad \text{and} \quad \mathbf{K} := \mathbf{h}_{l,t}\mathbf{W}_l^K \quad \text{and} \quad \mathbf{V} := \mathbf{h}_{l,t}\mathbf{W}_l^V, \\ \mathbf{h}_{l,t} \leftarrow \mathbf{A}\mathbf{V}, \qquad \mathbf{A} := \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right),\end{gathered} \tag{3.4}$$

where $\mathbf{W}_l^Q, \mathbf{W}_l^K, \mathbf{W}_l^V \in \mathbb{R}^{c \times d}$ are linear transformations which produce the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$, respectively, and softmax is applied across the second (hw)-dimension. (Generally, $c = d$ for diffusion models.) Intuitively, the attention map $\mathbf{A} \in \mathbb{R}^{(hw)\times(hw)}$ encodes how each pixel in $\mathbf{Q}$ corresponds to each in $\mathbf{K}$, which then rearranges and weighs $\mathbf{V}$. This correspondence is the basis for Ctrl-X's spatially-aware appearance transfer.

3.4 Guidance-free structure and appearance control

Ctrl-X is a general framework for training-free, guidance-free, and zero-shot T2I diffusion with structure and appearance control. Given a structure image $\mathbf{I}^{\mathrm{s}}$ and appearance image $\mathbf{I}^{\mathrm{a}}$, Ctrl-X manipulates a pretrained T2I diffusion model ϵ_θ to generate an output image $\mathbf{I}^{\mathrm{o}}$ that inherits the structure of $\mathbf{I}^{\mathrm{s}}$ and appearance of $\mathbf{I}^{\mathrm{a}}$.

Method overview Our method is illustrated in Figure 3.3 and is as follows: Given clean structure and appearance latents $\mathbf{I}^{\mathrm{s}} = \mathbf{x}_0^{\mathrm{s}}$ and $\mathbf{I}^{\mathrm{a}} = \mathbf{x}_0^{\mathrm{a}}$, we first directly obtain noised structure and appearance latents $\mathbf{x}_t^{\mathrm{s}}$ and $\mathbf{x}_t^{\mathrm{a}}$ via the diffusion forward process, then extracting their U-Net features from a pretrained T2I diffusion model. When denoising the output latent $\mathbf{x}_t^{\mathrm{o}}$, we inject convolution and self-attention features from $\mathbf{x}_t^{\mathrm{s}}$ and leverage self-attention correspondence to transfer spatially-aware appearance statistics from $\mathbf{x}_t^{\mathrm{a}}$ to $\mathbf{x}_t^{\mathrm{o}}$ to achieve structure and appearance control.

3.4.1 Feed-forward structure control

Structure control of T2I diffusion requires transferring structure information from $\mathbf{I}^{\mathrm{s}} = \mathbf{x}_0^{\mathrm{s}}$ to $\mathbf{x}_t^{\mathrm{o}}$, especially during early time steps. To this end, we initialize $\mathbf{x}_T^{\mathrm{o}} = \mathbf{x}_T^{\mathrm{s}} \sim \mathcal{N}(0, \mathbf{I})$ and obtain $\mathbf{x}_t^{\mathrm{s}}$ via the diffusion forward process in Equation 3.1 with $\mathbf{x}_0^{\mathrm{s}}$ and randomly sampled $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Inspired by the observation where diffusion features contain rich layout information [64, 32, 39], we perform feature and self-attention injection as follows: For U-Net layer l and diffusion time step t, let $\mathbf{f}_{l,t}^{\mathrm{o}}$ and $\mathbf{f}_{l,t}^{\mathrm{s}}$ be features/activations after the convolution block from $\mathbf{x}_t^{\mathrm{o}}$ and $\mathbf{x}_t^{\mathrm{s}}$, and let $\mathbf{A}_{l,t}^{\mathrm{o}}$ and $\mathbf{A}_{l,t}^{\mathrm{s}}$ be the attention maps of the self-attention block from $\mathbf{x}_t^{\mathrm{o}}$ and $\mathbf{x}_t^{\mathrm{s}}$. Then, we replace

$$\mathbf{f}^{\mathrm{o}_{l,t}} \leftarrow \mathbf{f}^{\mathrm{s}_{l,t}} \quad \text{and} \quad \mathbf{A}^{\mathrm{o}_{l,t}} \leftarrow \mathbf{A}^{\mathrm{s}_{l,t}}. \tag{3.5}$$

In contrast to [64, 32, 39], we do not perform inversion and instead directly use forward diffusion (Equation 3.1) to obtain $\mathbf{x}_t^{\mathrm{s}}$. We observe that $\mathbf{x}_t^{\mathrm{s}}$ obtained via the forward diffusion process contains sufficient structure information even at *very* early/high time steps, as shown in Figure 3.2. This also reduces appearance leakage common to inversion-based methods observed by FreeControl [39]. We study our feed-forward structure control method in Sections 3.5.1 and 3.5.3.

We apply feature injection for layers $l \in L^{\text{feat}}$ and self-attention injection for layers $l \in L^{\text{self}}$, and we do so for (normalized) time steps $t \leq \tau^{\text{s}}$, where $\tau^{\text{s}} \in [0, 1]$ is the structure control schedule.

3.4.2 Spatially-aware appearance transfer

Inspired by prior works that define appearance as feature statistics [28, 37], we consider appearance transfer as a stylization task. T2I diffusion self-attention transforms the value $\mathbf{V}$ with attention map $\mathbf{A}$, where the latter represents how pixels in $\mathbf{Q}$ corresponds to pixels in $\mathbf{K}$. As observed by Cross-Image Attention [5], $\mathbf{Q}\mathbf{K}^\top$ can represent the semantic correspondence between two images when $\mathbf{Q}$ and $\mathbf{K}$ are computed from features from each, even when the two images differ significantly in structure. Thus, inspired by AdaAttN [37], we propose spatially-aware appearance transfer, where we exploit this correspondence to generate self-attention-weighted mean and standard deviation maps from $\mathbf{x}_t^{\text{a}}$ to normalize $\mathbf{x}_t^{\text{o}}$: For any self-attention layer l, let $\mathbf{h}_{l,t}^{\text{o}}$ and $\mathbf{h}_{l,t}^{\text{a}}$ be diffusion features right before self-attention for $\mathbf{x}_t^{\text{o}}$ and $\mathbf{x}_t^{\text{a}}$, respectively. Then, we compute the attention map

$$\mathbf{A} = \operatorname{softmax}\left(\frac{\mathbf{Q}^{\text{o}}\mathbf{K}^{\text{a}\top}}{\sqrt{d}}\right), \qquad \mathbf{Q}^{\text{o}} := \operatorname{norm}(\mathbf{h}_{l,t}^{\text{o}})\mathbf{W}_l^Q \quad \text{and} \quad \mathbf{K}^{\text{a}} := \operatorname{norm}(\mathbf{h}_{l,t}^{\text{a}})\mathbf{W}_l^K, \tag{3.6}$$

where norm is applied across spatial dimension (hw). Notably, we normalize $\mathbf{h}_{l,t}^{\text{o}}$ and $\mathbf{h}_{l,t}^{\text{a}}$ first to remove appearance statistics and thus isolate structural correspondence. Then, we compute mean and standard deviation maps $\mathbf{M}$ and $\mathbf{S}$ of $\mathbf{h}_{l,t}^{\text{a}}$ weighted by $\mathbf{A}$ and use them to normalize $\mathbf{h}_{l,t}^{\text{o}}$,

$$\mathbf{h}_{l,t}^{\text{o}} \leftarrow \mathbf{S} \odot \mathbf{h}_{l,t}^{\text{o}} + \mathbf{M}, \qquad \mathbf{M} := \mathbf{A}\mathbf{h}_{l,t}^{\text{a}} \quad \text{and} \quad \mathbf{S} := \sqrt{\mathbf{A}(\mathbf{h}_{l,t}^{\text{a}} \odot \mathbf{h}_{l,t}^{\text{a}}) - (\mathbf{M} \odot \mathbf{M})}. \tag{3.7}$$

$\mathbf{M}$ and $\mathbf{S}$, weighted by structural correspondences between $\mathbf{I}^{\mathrm{o}}$ and $\mathbf{I}^{\mathrm{a}}$, are spatially-aware feature statistics of $\mathbf{x}_t^{\mathrm{a}}$ which is transferred to $\mathbf{x}_t^{\mathrm{o}}$. Lastly, we perform layer l self-attention on $\mathbf{h}_{l,t}^{\mathrm{o}}$ as normal.

We apply appearance transfer for layers $l \in L^{\mathrm{app}}$, and we do so for (normalized) time steps $t \leq \tau^{\mathrm{a}}$, where $\tau^{\mathrm{a}} \in [0, 1]$ is the appearance control schedule.

Structure and appearance control Finally, we replace ϵ_θ in Equation 3.2 with

$$\hat{\epsilon}_\theta \left(\mathbf{x}_t^{\mathrm{o}} \mid t, \mathbf{c}, \{\mathbf{f}_{l,t}^{\mathrm{s}}\}_{l \in L^{\mathrm{feat}}}, \{\mathbf{A}_{l,t}^{\mathrm{s}}\}_{l \in L^{\mathrm{self}}}, \{\mathbf{h}_{l,t}^{\mathrm{a}}\}_{l \in L^{\mathrm{app}}} \right), \tag{3.8}$$

where $\{\mathbf{f}_{l,t}^{\mathrm{s}}\}_{l \in L^{\mathrm{feat}}}$, $\{\mathbf{A}_{l,t}^{\mathrm{s}}\}_{l \in L^{\mathrm{self}}}$, and $\{\mathbf{h}_{l,t}^{\mathrm{a}}\}_{l \in L^{\mathrm{app}}}$ corresponds to $\mathbf{x}_t^{\mathrm{s}}$ features for feature injection, $\mathbf{x}_t^{\mathrm{s}}$ attention maps for self-attention injection, and $\mathbf{x}_t^{\mathrm{a}}$ features for appearance transfer.

3.5 Experiments

We present extensive quantitative and qualitative results to demonstrate the structure preservation and appearance alignment of Ctrl-X on T2I diffusion. Appendix ?? contains more implementation details.

3.5.1 T2I diffusion with structure and appearance control

Baselines For training-based methods, ControlNet [75] and T2I-Adapter [41] learn an auxiliary module that injects a condition image into a pretrained diffusion model for structure alignment, and we combine them with IP-Adapter [74], a trained module for image prompting and thus appearance transfer; Splicing ViT Features [63] trains a U-Net from scratch per source-appearance image pair to minimize their DINO-ViT self-similarity distance and global `[CLS]` token loss. (For structure conditions not supported by a training-based baseline, we convert them to canny edge maps.) For guidance-based methods, FreeControl [39] enforce structure and appearance alignment via backpropagated

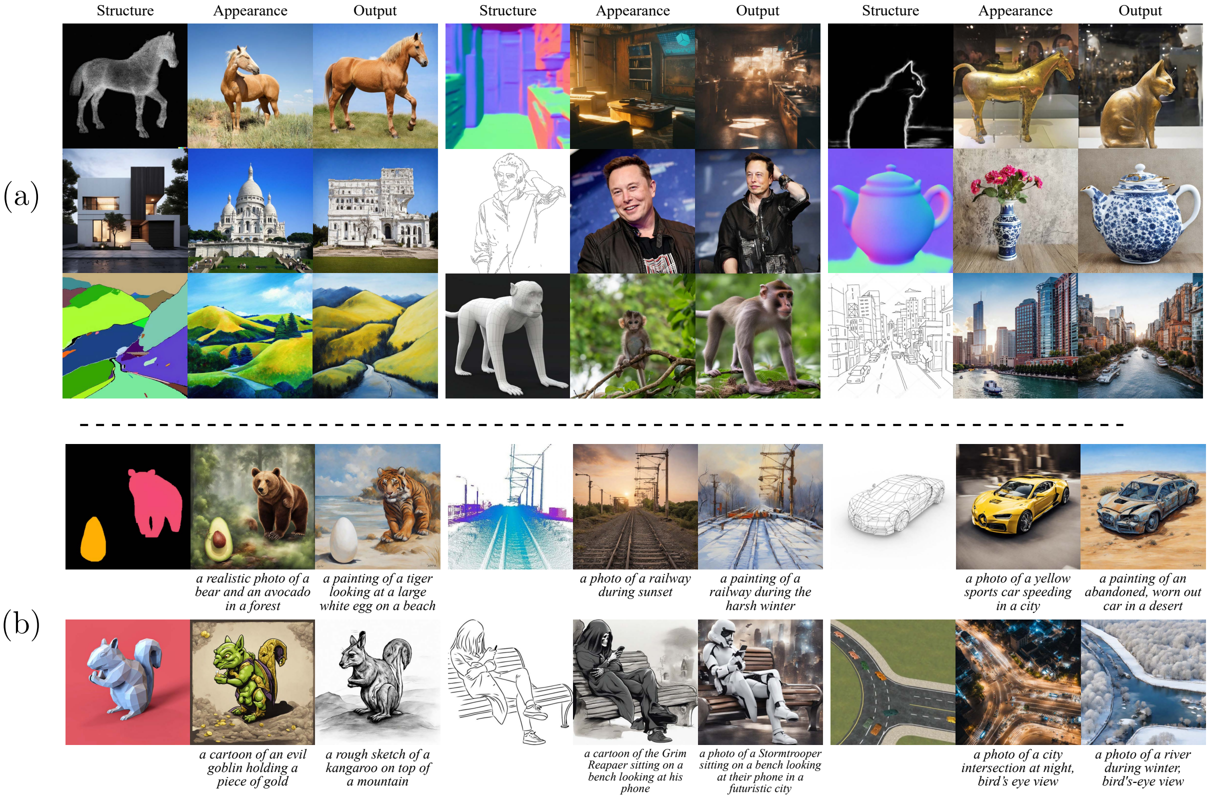

Structure
Appearance
Output
Structure
Appearance
Output
Structure
Appearance
Output
(a)
(b)
a realistic photo of a bear and an avocado in a forest
a painting of a tiger looking at a large white egg on a beach
a photo of a railway during sunset
a painting of a railway during the harsh winter
a photo of a yellow sports car speeding in a city
a painting of an abandoned, worn out car in a desert
a cartoon of an evil goblin holding a piece of gold
a rough sketch of a kangaroo on top of a mountain
a cartoon of the Grim Reapaer sitting on a bench looking at his phone
a photo of a Stormtrooper sitting on a bench looking at their phone in a futuristic city
a photo of a city intersection at night, bird's eye view
a photo of a river during winter, bird's-eye view

Figure 3.4: **Qualitative results for T2I diffusion structure and appearance control and conditional generation** Ctrl-X supports a diverse variety of structure images for both (a) structure and appearance controllable generation and (b) prompt-driven conditional generation.

score functions computed from diffusion feature subspaces. For guidance-free methods, Cross-Image Attention [5] manipulates attention weights to transfer appearance while maintaining structure. We run all methods on SDXL v1.0 [48] when possible and their default base models if not.

Dataset Our method supports T2I diffusion with appearance transfer and arbitrary-condition structure control. Since no benchmarks exist for such a flexible task, we create a new dataset comprising 256 diverse structure-appearance pairs. The structure images consist of 31% natural images, 49% ControlNet-supported conditions (e.g.,canny, depth, segmentation), and 20% in-the-wild conditions (e.g.,3D mesh, point cloud), and the appearance images are a mix of Web and generated images. We use templates and hand-annotation for the structure, appearance, and output text prompts.

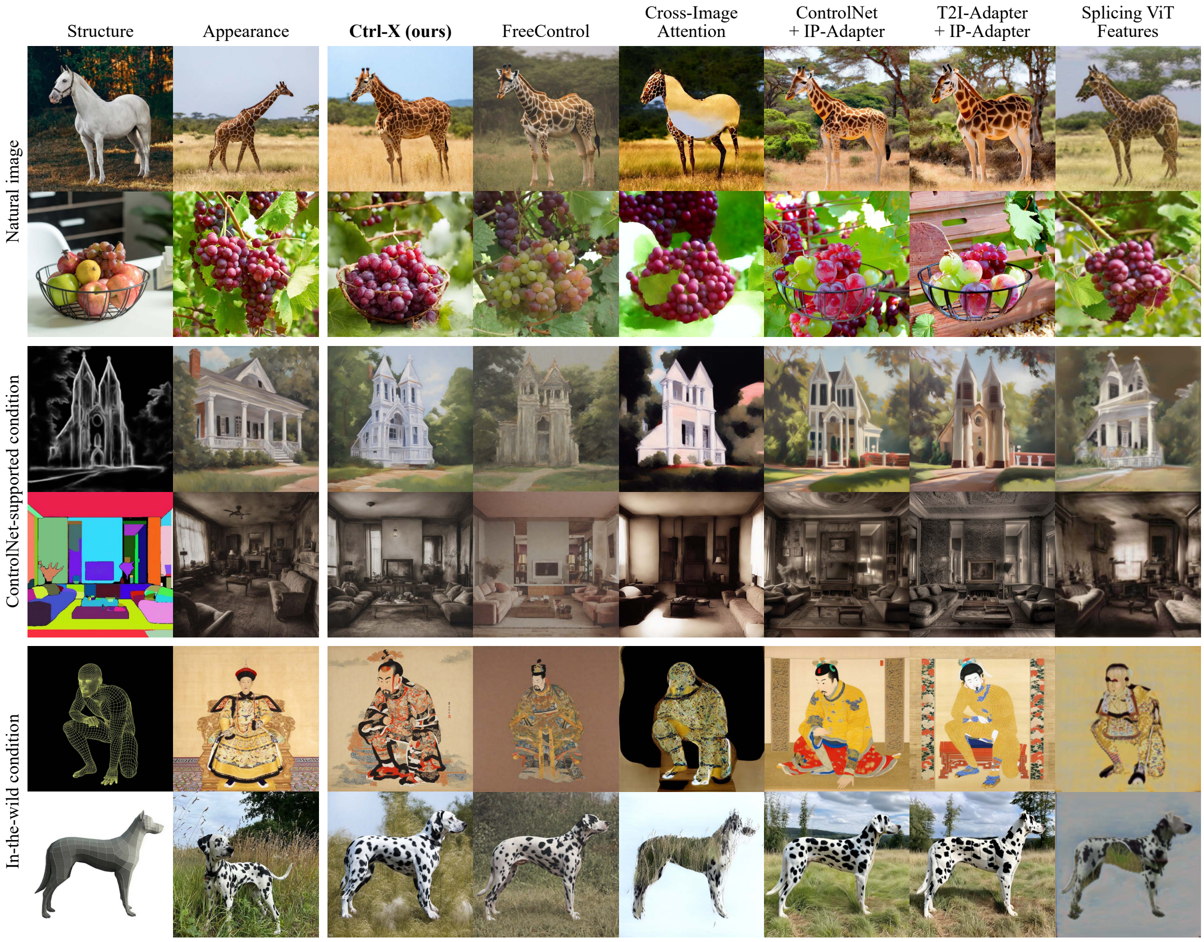


Figure 3.5: **Qualitative comparison of structure and appearance control** Ctrl-X displays comparable structure control and superior appearance transfer compared to training-based methods. It is also more robust than guidance-based and -free methods across a wide variety of structure types.

Evaluation metrics For quantitative evaluation, we report two widely-adopted metrics: *DINO Self-sim* measures the self-similarity distance [63] between the structure and output image in the DINO-ViT [13] feature space, where a lower distance indicate better structure preservation; *DINO* `CLS` measures the loss between the DINO-ViT global `[CLS]` tokens of the appearance and output image [63], where a lower loss indicate better appearance alignment.

Qualitative results As shown in Figures 3.4 and 3.5, Ctrl-X faithfully preserves structure from structure images ranging from natural images and ControlNet-supported conditions (e.g.,HED, segmentation) to in-the-wild conditions (e.g.,wireframe, 3D mesh) not possible in prior training-based methods while adeptly transfers appearance from the appearance image with semantic correspondence.

Comparison to baselines Figure 3.5 and Table 3.1 compare our method to the baselines. For training-based and guidance-based methods, despite T2I-Adapter [41] and FreeControl's [39] stronger structure preservation (smaller DINO self-similarity distances), they generally struggle to enforce faithful appearance transfer and yield worse global `CLS` losses, which is particularly visible in Figure 3.5 row 1 and 3. Since the training-based methods combine a structure control module (ControlNet [75] and T2I-Adapter) with a separately-trained appearance transfer module IP-Adapter [74], the two modules sometimes exert conflicting control signals at the cost of appearance transfer (e.g.,row 1)—and for ControlNet, structure preservation as well. For FreeControl, its appearance score function from extracted embeddings may not sufficiently capture more complex appearance correspondences, which, along with needing per-image hyperparameter tuning, results in lower contrast outputs and sometimes failed appearance transfer (e.g.,row 4). Moreover, despite Splicing ViT Features [63] having the best DINO self-similarity and `CLS` scores in Table 3.1, Figure 3.5 reveals that its output images are often blurry while displaying structure image appearance leakage with non-natural images (e.g.,row 3, 5, and 6). It benchmarks well because its per-image training minimizes these two metrics directly.

Guidance-free baseline Cross-Image Attention [5], in contrast, is less robust and more sensitive to the structure image's appearance, as the inverted structure latents contain strong appearance information. This causes both poorer structure alignment and frequent appearance leakage or artifacts (e.g.,row 6) from the structure to the output images, resulting in worse DINO self-similarity distances and global `CLS` losses. In practice, we find Cross-Image Attention sensitive to its masking domain and sometimes fails to produce outputs with crossmodal pairs (e.g.,wireframes to photos).

Method	Natural image		ControlNet-supported		New condition		Inference time (s)
	Self-sim ↓	DINO `CLS` ↓	Self-sim ↓	DINO `CLS` ↓	Self-sim ↓	DINO `CLS` ↓	
Splicing ViT Features [63]	0.030	0.006	0.043	0.012	0.037	0.013	4289.20
ControlNet + IP-Adapter [75, 74]	0.068	0.109	0.136	0.092	0.139	0.103	23.10
T2I-Adapter + IP-Adapter [41, 74]	**0.055**	0.119	0.118	0.118	0.109	0.131	**17.70**
Cross-Image Attention [5]	0.145	0.110	0.196	0.152	0.195	0.139	216.46
FreeControl [39]	0.058	0.132	**0.101**	0.119	**0.089**	0.139	1210.02
Ctrl-X (ours)	0.057	**0.096**	0.121	**0.084**	0.109	**0.097**	30.65

Table 3.1: **Quantitative comparison of structure and appearance control** Ctrl-X consistently outperforms both training-based and training-free methods in appearance alignment and shows comparable or better structure preservation compared to training-based and guidance-free methods, measured by DINO ViT self-similarity and global `CLS` token loss [63], respectively.

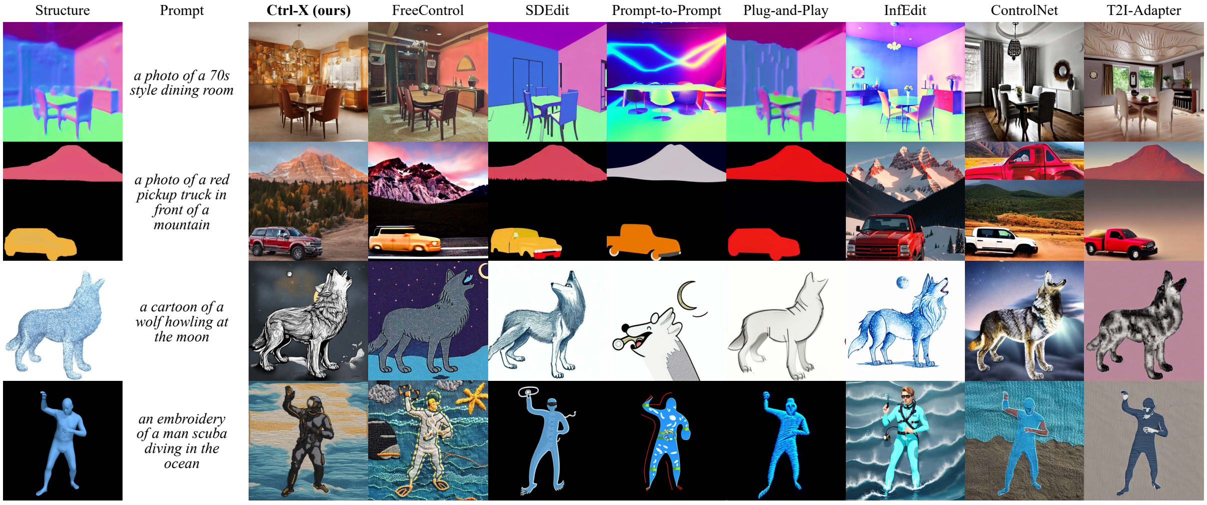

Structure
Prompt
Ctrl-X (ours)
FreeControl
SDEdit
Prompt-to-Prompt
Plug-and-Play
InfEdit
ControlNet
T2I-Adapter
a photo of a 70s style dining room
a photo of a red pickup truck in front of a mountain
a cartoon of a wolf howling at the moon
an embroidery of a man scuba diving in the ocean

Figure 3.6: **Qualitative comparison of conditional generation** Ctrl-X displays comparable structure control and superior prompt alignment to training-based methods, and it also has better image quality and is more robust than guidance-based and -free methods across different conditions.

Inference efficiency We study the inference time of our method compared to the baselines, all with base model SDXL v1.0 except Cross-Image Attention (SD v1.5) and Splicing ViT Features (U-Net). Table 3.1 reports the average inference time using a single NVIDIA A6000 GPU. Ctrl-X is slightly slower than training-based ControlNet (1.32×) and T2I-Adapter (1.73×) with IP-Adapter yet significantly faster than per-image-trained Splicing ViT (0.0071×), guidance-based FreeControl (0.025×), and guidance-free Cross-Image Attention (0.14×). Our training-free and guidance-free method achieves comparable runtimes to training-based methods, indicating its flexibility.

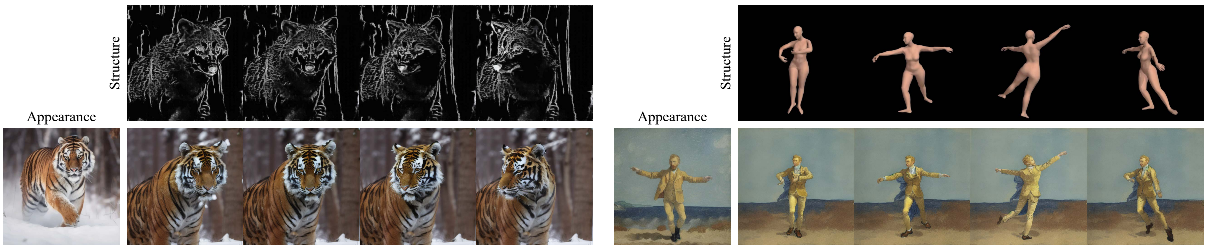


Figure 3.7: **Extension to text-to-video (T2V) models** Ctrl-X can be directly applied to T2V models [22, 59] for controllable video structure and appearance control.

Extension to prompt-driven conditional generation Ctrl-X also supports prompt-driven conditional generation, where it generates an output image complying with the given text prompt while aligning with the structure from the structure image, as shown in Figures 3.4 and 3.6. Inspired by FreeControl [39], instead of a given $\mathbf{I}^{\mathrm{a}}$, Ctrl-X can jointly generate $\mathbf{I}^{\mathrm{a}}$ based on the text prompt alongside $\mathbf{I}^{\mathrm{o}}$, where we obtain $\mathbf{x}^{\mathrm{a}}_{t-1}$ via denoising with Equation 3.2 from $\mathbf{x}^{\mathrm{a}}_{t}$ without control.

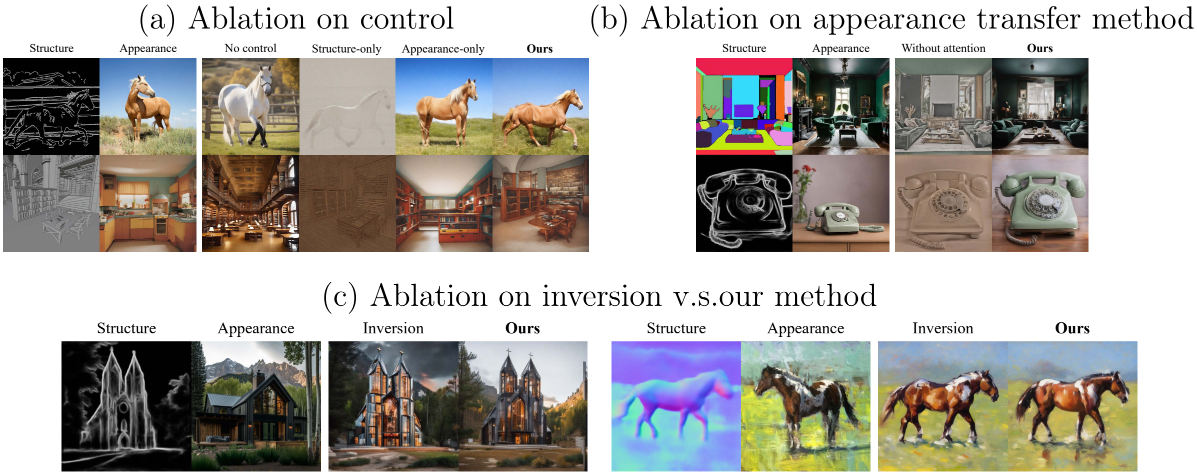
3.5.2 Extension to video diffusion models

Ctrl-X is training-free, guidance-free, and demonstrates competitive runtime. Thus we can directly apply our method to text-to-video (T2V) models, as seen in Figure 3.7. Our method closely aligns the structure between the structure and output videos while transferring temporally consistent appearance from the appearance image.

3.5.3 Ablations

Effect of control As seen in Figure 3.8(a), structure control is responsible for structure preservation (appearance-only v.s.ours). Also, structure control alone cannot isolate structure information, displaying strong structure image appearance leakage and poor-quality outputs (structure-only v.s.ours), as it merely injects structure features, which creates the semantic correspondence for appearance control.

Appearance transfer method As we consider appearance transfer as a stylization task, we compare our appearance statistics transfer with and without attention weighting.


(a) Ablation on control

(b) Ablation on appearance transfer method

(c) Ablation on inversion v.s.our method

Figure 3.8: **Ablations** We study ablations on control, appearance transfer method, and inversion.

Without attention weighting (equivalent to AdaIN [28]), the normalization is global and thus cannot consider the semantic correspondence between the appearance and output images, so the outputs look low-contrast.

Effect of inversion We compare DDIM inversion v.s.forward diffusion (ours) to obtain $\mathbf{x}_T^{\mathrm{o}} = \mathbf{x}_T^{\mathrm{s}}$ and $\mathbf{x}_t^{\mathrm{s}}$ in Figure 3.8(c). Inversion displays appearance leakage from structure images in challenging conditions (left) while being similar to our method in others (right). Considering inversion costs and additional model inference time, forward diffusion is a better choice for our method.

CHAPTER 4

Conclusion and Discussion

This final chapter concludes the discussed chapters in this thesis and summarizes their findings, limitations, and potential future works. The primary objective of this thesis is to enhance conditional text-to-image generation with pre-trained generative diffusion models without additional training. Our key question is: *How can we leverage pre-trained T2I diffusion models to take condition signals from a wide spectrum?*

In Chapter 2, we present FreeControl, a training-free approach for controllable T2I generation that supports multiple conditions, architectures, and checkpoints simultaneously. FreeControl enforces structure guidance to facilitate the global alignment with a guidance image, and appearance guidance to collect visual details from images generated without control. Extensive qualitative and quantitative experiments demonstrate the superior performance of FreeControl across a variety of pre-trained T2I models. In particular, FreeControl enables convenient training-free control over many different architectures and checkpoints, allows the challenging input conditions on which most of the existing training-free methods fail, and achieves competitive synthesis quality compared to training-based approaches.

In Chapter 3, we present *Ctrl-X*, a simple framework for T2I diffusion controlling structure and appearance without additional training or guidance. Ctrl-X designs feed-forward structure control to enable the structure alignment with a structure image and semantic-aware appearance transfer to facilitate the appearance transfer from a user-input image. Extensive qualitative and quantitative experiments illustrate the superior performance of Ctrl-X on various condition inputs and model checkpoints. In particular, Ctrl-X supports novel structure and appearance control with arbitrary condition images of any modality, exhibits superior image quality and appearance transfer compared to existing works, and provides instant plug-and-play to any T2I and text-to-video (T2V) diffusion model.

The research presented in this thesis has significantly advanced the field of controllable image synthesis, with the proposed method for conditionally controllable generation having a substantial impact on subsequent studies. While FreeControl and Ctrl-X have effectively introduced spatial and appearance controllability to several pre-trained text-to-image diffusion models without additional training, achieving pixel-level control remains challenging due to the small size of the internal activation maps in the denoiser network. By continuously pushing the boundaries of current models and fostering innovation, we can aim to further enhance the capabilities of controllable image generation.

REFERENCES

[1] Autocad. https://www.autodesk.com/products/autocad. 18

[2] Civitai. https://civitai.com/. x, 8, 22

[3] Hugging face. https://huggingface.co/. 8

[4] Midjourney. https://www.midjourney.com. 5

[5] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM Spec. Int. Grp. Comput. Graph. Int. Tech.*, 2024. 31, 33, 36, 38, 40, 41

[6] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *ACM Spec. Int. Grp. Comput. Graph. Int. Tech. Asia*, 2023. 8, 31

[7] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 5, 7, 30

[8] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 7

[9] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Int. Conf. Learn. Represent.*, 2023. 29, 31, 33

[10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 7

[11] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Int. Conf. Comput. Vis.*, 2023. 7, 9, 11, 31

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2, 5

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, 2021. 17, 39

[14] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *IEEE W. Conf. App. Comp. Vis.*, 2024. 9

[15] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 18

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2021. 8, 9

[17] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Adv. Neural Inform. Process. Syst.*, 2023. 7, 8, 9, 14, 29, 31, 33

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 7

[19] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Int. Conf. Learn. Represent.*, 2023. 8, 31

[20] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Int. Conf. Comput. Vis.*, 2023. 7, 9, 13

[21] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. ii, 1, 7

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *Int. Conf. Learn. Represent.*, 2024. xii, 42

[23] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *Int. Conf. Learn. Represent.*, 2023. 6, 8, 16, 17, 19, 33

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Adv. Neural Inform. Process. Syst.*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 7, 8, 32, 33

[25] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 2022. 7

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 15, 33

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent.*, 2022. 1, 5, 6, 31

[28] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Int. Conf. Comput. Vis.*, 2017. 36, 43

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 7

[30] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 7

[31] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7701–7711, 2023. 30

[32] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Int. Conf. Comput. Vis.*, pages 7701–7711, 2023. 33, 35

[33] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 8

[34] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Int. Conf. Comput. Vis.*, 2023. ii, 1, 7

[35] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Trans. Pattern Analy. Mach. Intelli.*, 2022. 18

[36] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 5, 7, 28, 30

[37] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Int. Conf. Comput. Vis.*, pages 6649–6658, 2021. 36

[38] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2022. 6, 8, 16, 17, 19

[39] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 29, 30, 31, 33, 35, 36, 37, 40, 41, 42

[40] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 8

[41] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. 2024. 1, 5, 7, 16, 28, 30, 31, 37, 40, 41

[42] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Int. Conf. Mach. Learn.*, 2022. 7, 9

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. 2024. 11

[44] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 7

[45] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM Spec. Int. Grp. Comput. Graph. Int. Tech.*, 2023. 6, 8, 9, 19

[46] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, 2023. 7, 9

[47] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models, 2023. 31

[48] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Int. Conf. Learn. Represent.*, 2024. xi, 7, 28, 32, 33, 38

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021. 7, 17

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 2020. 7

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 5, 7

[52] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Analy. Mach. Intelli.*, 2020. 2, 5

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. ii, viii, 1, 4, 5, 7, 33

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Med. Image Comp. Comp. Assis. Inter.*, 2015. 9

[55] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 5, 6, 8, 31

[56] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models, 2023. 31

[57] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM Spec. Int. Grp. Comput. Graph. Int. Tech.*, 2022. 7

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Adv. Neural Inform. Process. Syst.*, 2022. ii, 1, 7

[59] SG_161222. Realistic vision v5.1. `https://civitai.com/models/4201?modelVersionId=130072`. xii, 42

[60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. Mach. Learn.*, 2015. 7

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021. viii, 8, 9, 11, 12, 13

[62] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2021. 7, 8, 32

[63] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing ViT features for semantic appearance transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. xiii, 11, 17, 37, 39, 40, 41

[64] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1921–1930, 2023. 6, 8, 9, 11, 16, 17, 19, 33, 35

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 9

[66] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM Spec. Int. Grp. Comput. Graph. Int. Tech.*, 2023. 1, 5, 7

[67] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 8

[68] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation, 2024. 30, 31

[69] Jiayu Xiao, Henglei Lv, Liang Li, Shuhui Wang, and Qingming Huang. R&b: Region and boundary aware zero-shot grounded text-to-image generation. In *Int. Conf. Learn. Represent.*, 2024. 30

[70] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Int. Conf. Comput. Vis.*, 2023. 7, 9

[71] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 7

[72] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 31, 33

[73] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 30

[74] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 28, 30, 31, 37, 40, 41

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, 2023. viii, 1, 4, 5, 6, 7, 16, 17, 18, 28, 30, 31, 37, 40, 41

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 17

[77] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 7

[78] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Adv. Neural Inform. Process. Syst.*, 2023. 1, 5, 16, 28, 30

[79] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 30

[80] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis, 2024. 30