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

Sicheng Mo[†*], Fangzhou Mu[§*], Kuan Heng Lin[†], Yanli Liu[‡], Bochen Guan[‡], Yin Li[§], Bolei Zhou[†]
[†]University of California, Los Angeles, [§]University of Wisconsin-Madison, [‡]Innopeak Technology, Inc

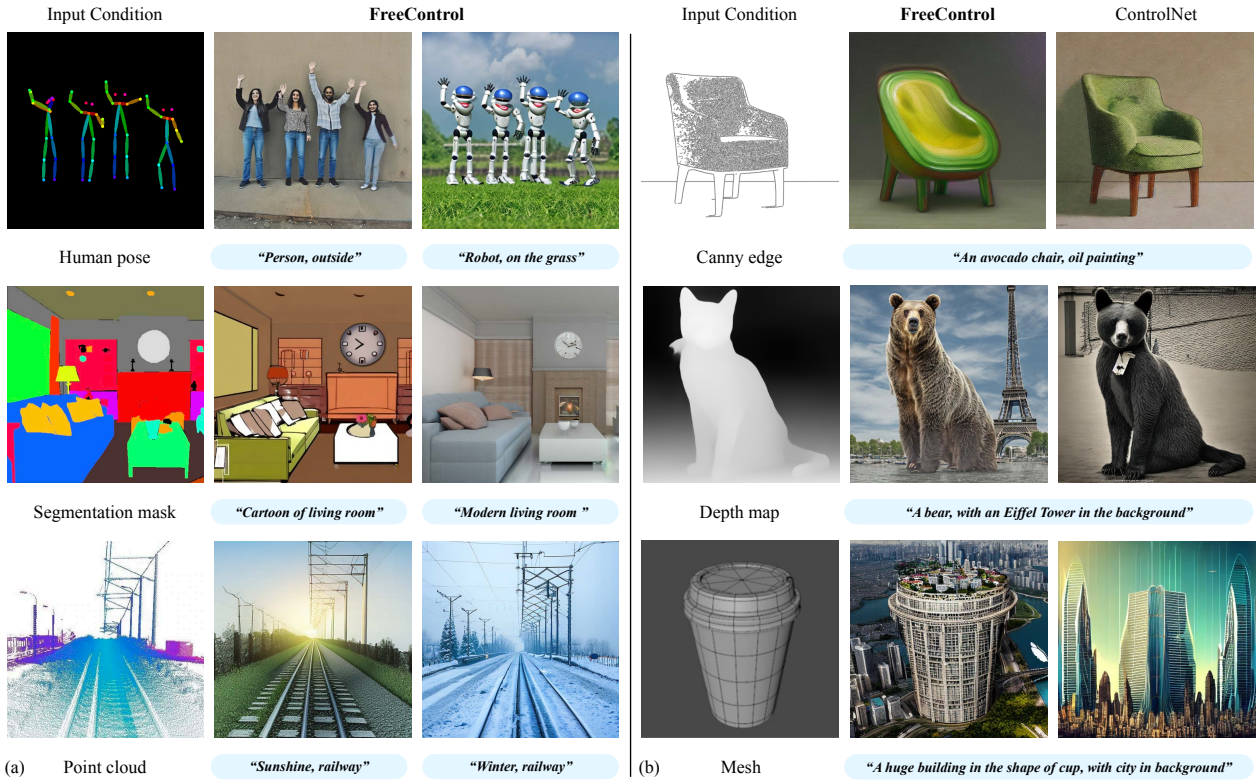

Figure 1. **Training-free conditional control of Stable Diffusion [44].** (a) FreeControl enables zero-shot control of pretrained text-to-image diffusion models given various input control conditions. (b) Compared to ControlNet [59], FreeControl achieves a good balance between spatial and image-text 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.

Abstract

Recent approaches such as ControlNet [59] offer users fine-grained spatial control over text-to-image (T2I) diffusion models. However, auxiliary modules have to be trained for each spatial condition type, model architecture, and checkpoint, putting them at odds with the diverse intents and preferences a human designer would like to convey to the AI models during the content creation process. In this work, 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. Project page: https://genforce.github.io/freecontrol/.

* indicates equal contribution

1. Introduction

Text-to-image (T2I) diffusion models [4, 42] 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 [59] 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 [6, 30, 33, 55, 59, 62] 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 [44] fine-tuned for Disney characters or user-specified objects [24, 46]), 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 [43] for depth maps, OpenPose [10] 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 1(b), where there is a conflict between the guidance image and text prompt (*e.g.*, an edge map of a sofa chair *vs*. "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 [20, 31, 37, 53], 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 [46] and LoRA [24]). (3) Our method demonstrates superior results in comparison to previous training-free methods (*e.g.*, Plug-and-Play [53]) and achieves comparable performance with prior training-based approaches (*e.g.*, ControlNet [59]).

2. Related Work

Text-to-image diffusion. Diffusion models [22, 49, 51] 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 [40, 41] and is performed either in pixel space [7, 34, 42, 48] or latent space [19, 39, 44], followed by cascaded super-resolution [23] or latent-to-image decoding [16] 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 [28, 58]. 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 [6, 30, 33, 55, 59, 62]. 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 fea-

tures inside a pre-trained T2I model for the control of object size, shape, appearance and location [9, 15, 18, 38, 57]. 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 [25] 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 [25, 36, 47]. T2I diffusion enables I2I methods to specify target domains using text. Text-driven I2I is often posed as conditional generation [8, 26, 33, 59, 61, 62]. These methods fine-tune a pretrained model to condition it on an input image. Alternatively, recent training-free methods perform zero-shot image translation [20, 31, 37, 53] and is most relevant to our work. This is achieved by inverting the input image [32, 50, 56], 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 [5, 17, 27, 46] 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 [2, 3], FreeControl offers a *scalable* framework for zero-shot control of any model with any spatial condition.

3. Preliminary

Diffusion sampling. Image generation with a pre-trained T2I diffusion model amounts to iteratively removing noise from an initial Gaussian noise image $\mathbf{x}_T$ [22]. 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 [51]:

$$\epsilon_\theta(\mathbf{x}_t; t, \mathbf{c}) \approx -\sigma_t \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x_t}|\mathbf{c}). \tag{1}$$

Guidance. The update rule in Equation 1 may be altered by a time-dependent energy function $g(\mathbf{x_t}; t, y)$ through *guidance* (with strength s) [14, 15] so as to condition diffusion sampling on auxiliary information y (*e.g.*, class labels):

$$\hat{\epsilon}_\theta(\mathbf{x}_t; t, \mathbf{c}) = \epsilon_\theta(\mathbf{x}_t; t, \mathbf{c}) - s\, g(\mathbf{x_t}; t, y). \tag{2}$$

In practice, g may be realized as classifiers [14] or CLIP scores [34], or defined using bounding boxes [12, 57], attention maps [18, 37] or any measurable object properties [15].

Attentions in ϵ_θ. A standard choice for ϵ_θ is a U-Net [45] with self- and cross-attentions [54] 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 [9, 18, 38, 53].

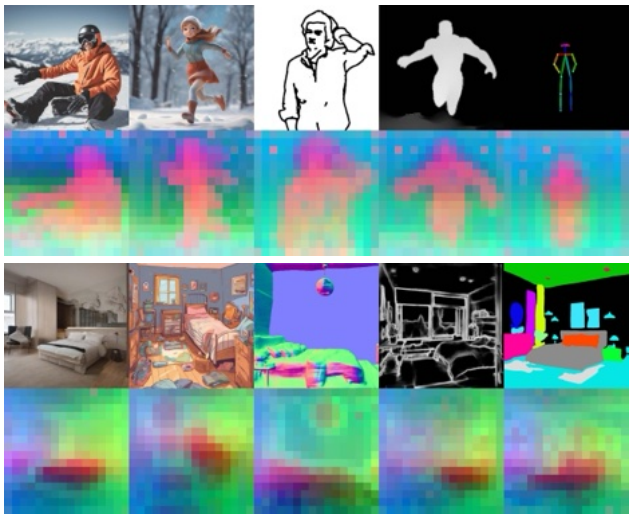

Figure 2. **Visualization of feature subspace given by PCA.** Keys from the first self-attention in the U-Net decoder are obtained via DDIM inversion [50] for five images in different styles and modalities (*top*: `person`; *bottom*: `bedroom`), and subsequently undergo PCA. The top three principal components (pseudo-colored in RGB) provide a clear separation of semantic components.

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 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

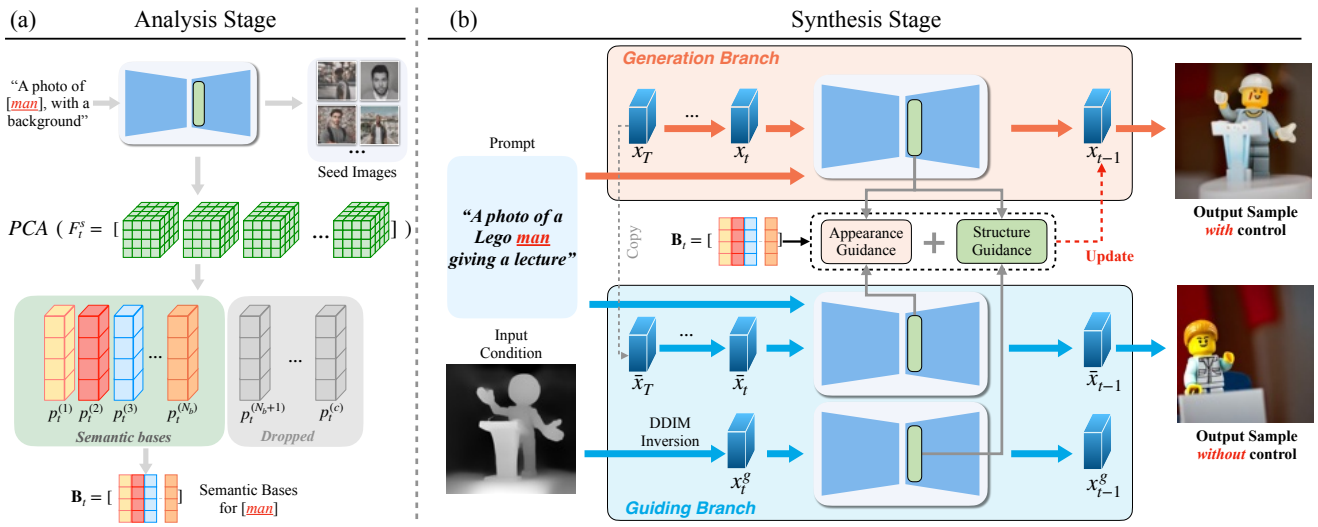

Figure 3. **Method overview.** (a) In the *analysis* stage, FreeControl generates seed images for a target concept (*e.g.*, `man`) using a pretrained diffusion model and performs PCA on their diffusion features to obtain a linear subspace as semantic basis. (b) In the *synthesis* stage, FreeControl employs 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.

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. 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 [50] 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.

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 [52] and T2I diffusion models [9] 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 [53] 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 6, appearance leakage is particularly problematic when $\mathbf{I}^g$ and $\mathbf{I}$ are different modalities (*e.g.*, depth *vs.* natural images), common for controllable generation.

Towards disentangling image structure and appearance, we draw inspiration from Transformer feature visualization [35, 53] 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.

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 5.2.

Semantic basis. We apply DDIM sampling [50] 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

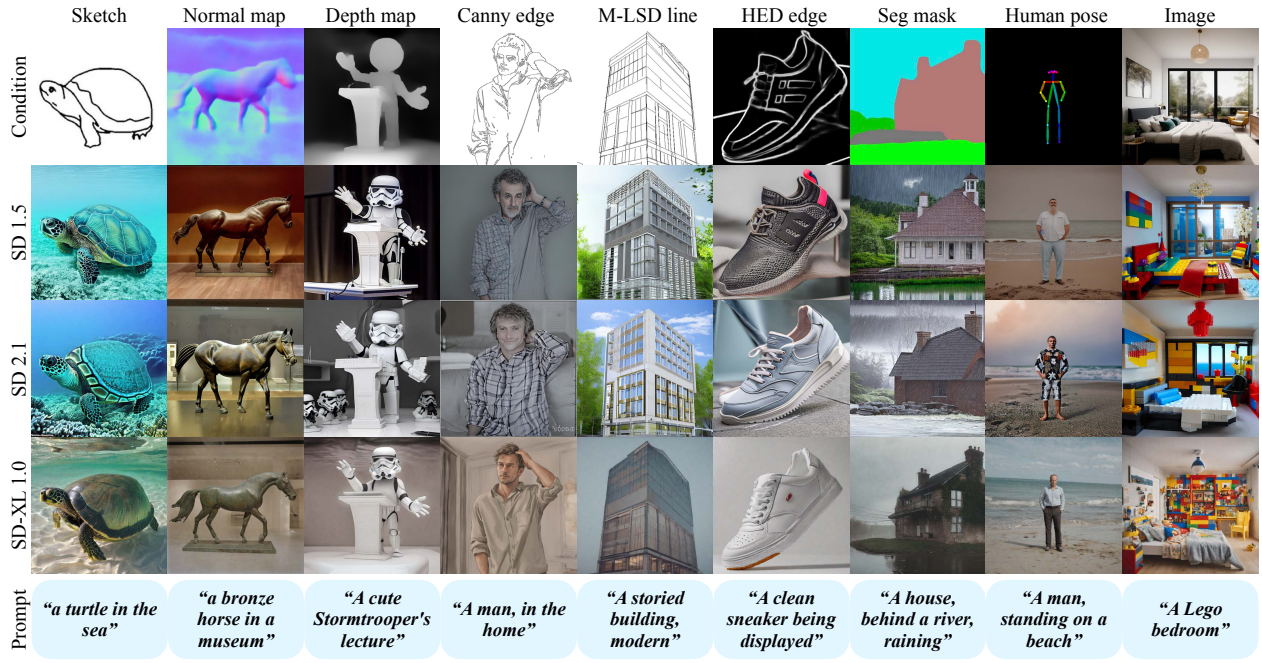

Figure 4. **Qualitative comparison of controllable T2I diffusion.** FreeControl supports a suite of control signals and three major versions of Stable Diffusion. The generated images closely follow the text prompts while exhibiting strong spatial alignment with the input images.

principal components:

$$\mathbf{B}_t = [\mathbf{p}_t^{(1)}, \mathbf{p}_t^{(2)}, ..., \mathbf{p}_t^{(N_b)}] \sim \mathrm{PCA}(\{\mathbf{F}_t^s\}) \tag{3}$$

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 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.

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 [50] 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 [18]. $\mathbf{M}$ 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

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

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

$$\boldsymbol{\tau}_t = \max_{i,j \text{ s.t. } m_{ij}=0} [\mathbf{s}_t^g]_{ij} \tag{4}$$

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.

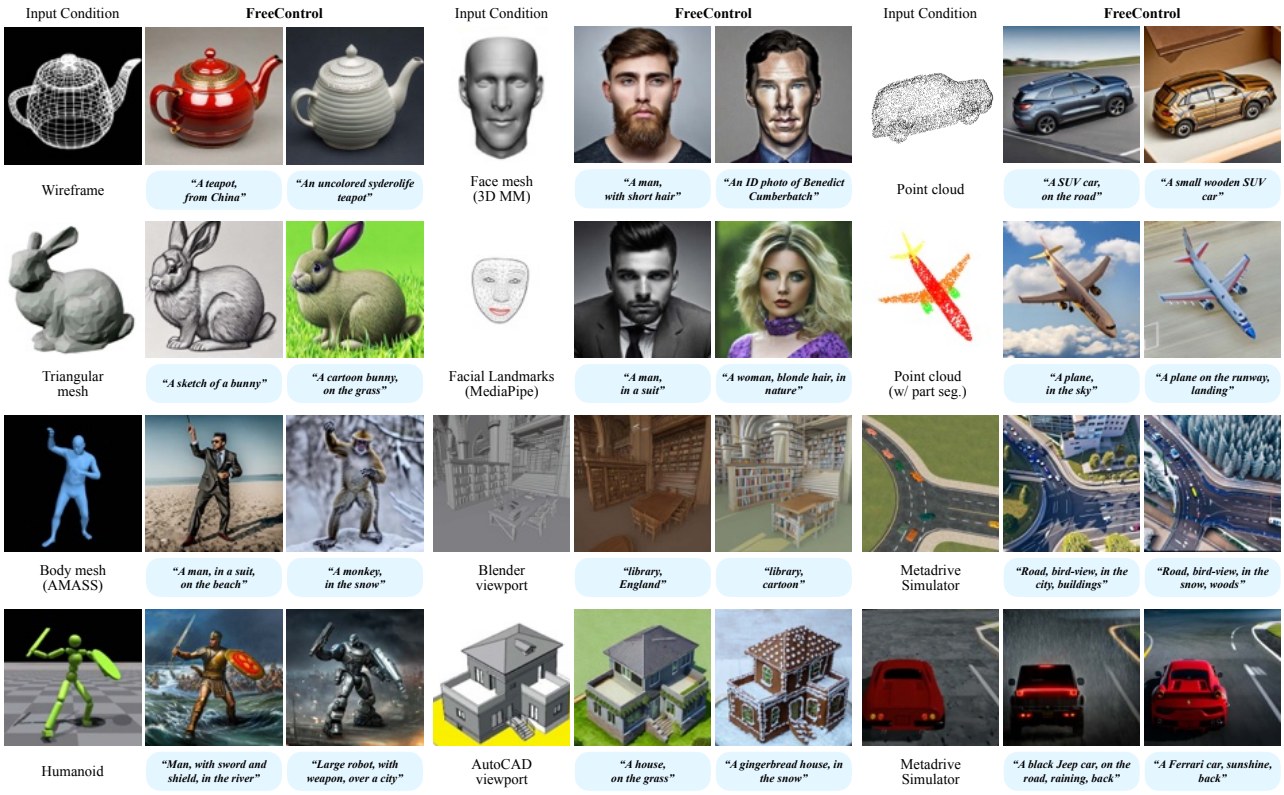

Figure 5. **Qualitative results for more diverse control conditions.** FreeControl supports challenging control conditions not possible with training-based methods. These include 2D projections of common graphics primitives, domain-specific shape models *(point cloud, body mesh, and humanoid)*, graphics software viewports *(Blender and AutoCAD)*, and simulated driving environments *(Metadrive)*.

Appearance representation. Inspired by DSG [15], we represent image appearance as $\{\mathbf{v}_t^{(k)}\}_{k=1}^{N_a \leq N_b}$, the weighted spatial means of diffusion features $\mathbf{F}_t$:

$$\mathbf{v}_t^{(k)} = \frac{\sum_{i,j} \sigma([s_t^{(k)}]_{ij})[\mathbf{f}_t]_{ij}}{\sum_{i,j} \sigma([s_t^{(k)}]_{ij})}, \tag{5}$$

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

$$g_a(\{\mathbf{v}_t^{(k)}\}; \{\bar{\mathbf{v}}_t^{(k)}\}) = \frac{\sum_{k=1}^{N_a} \|\mathbf{v}_t^{(k)} - \bar{\mathbf{v}}_t^{(k)}\|_2^2}{N_a}. \tag{6}$$

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 [21]:

$$\hat{\epsilon}_t = (1+s)\,\epsilon_\theta(\mathbf{x}_t; t, \mathbf{c}) - s\,\epsilon_\theta(\mathbf{x}_t; t, \emptyset) + \lambda_s\, g_s + \lambda_a\, g_a, \tag{7}$$

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

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.

5.1. Controllable T2I Diffusion

Baselines. ControlNet [59] and T2I-Adapter [33] 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 [62] instead learns adapters shared by all condition types for all-in-one control. Different from these training-based methods, SDEdit [31] adds noise to a guidance image and subsequently denoises it with a pretrained diffusion model for guided image synthesis. Prompt-to-Prompt (P2P) [20] and Plug-and-Play (PnP) [53] 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.

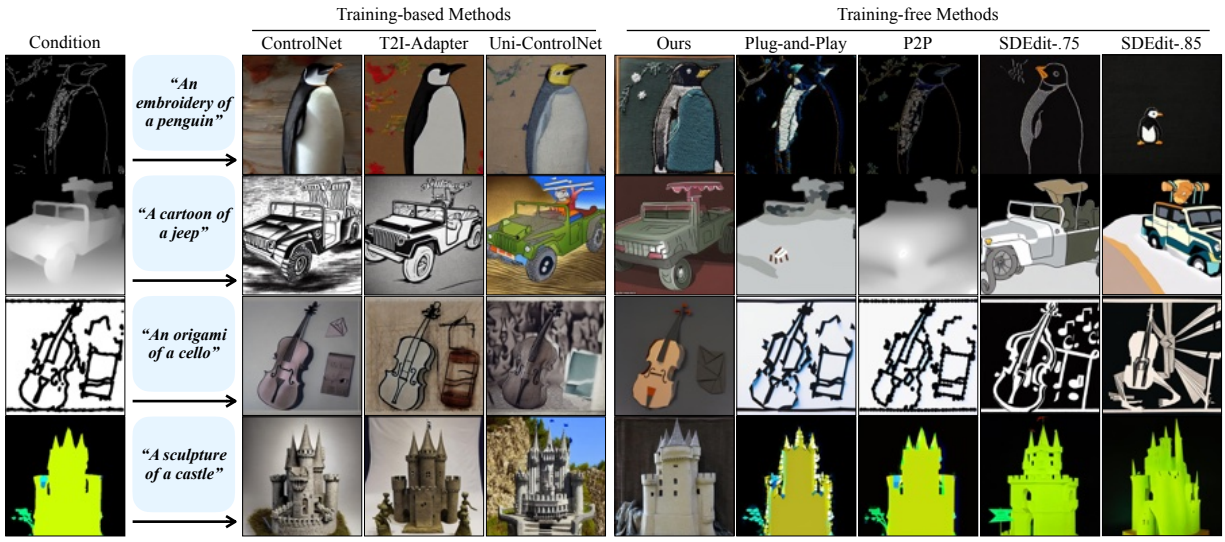

Figure 6. **Qualitative comparison on controllable T2I diffusion.** FreeControl achieves competitive spatial control and superior image-text alignment in comparison to training-based methods. It also escapes the appearance leakage problem manifested by the training-free baselines, producing high-quality images with rich content and appearance faithful to the text prompt.

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 [59]	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 [31]	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 [31]	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 [20]	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 [53]	**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
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 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.

Experiment setup. Similar to ControlNet [59], 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 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.

For a fair comparison with the baselines, we adapt the ImageNet-R-TI2I dataset from PnP [53] 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 [59], 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* [52] measures the structural similarity of two images in the feature space of DINO-ViT [11]. A smaller distance suggests better structure preservation. Similar to [53], we report self-similarity between the generated image and the dataset image that produces the input condition. *CLIP score* [40] 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* [60] 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 4, FreeControl is able to recognize diverse semantic structures from all condition modalities used by ControlNet [59]. 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 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 [13] and AutoCAD [1]), and simulated driving environments (*e.g.*, MetaDrive [29]), thereby providing an appealing solution to visual design preview and sim2real.

Comparison with baselines. Figure 6 and Table 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.

	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. **Runtime for training-free methods**

Inference efficiency. We further study the inference cost of our method in comparison to training-free baselines. Table 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 P2P, and **0.75**× that of Pix2Pix-zero, yet still generate diverse images.

5.2. Ablation Study

Effect of guidance. As seen in Figure 7, structure guidance is responsible for structure alignment ($-g_s$ *vs.* Ours). Appearance guidance alone has no impact on generation in the absence of structure guidance ($-g_a$ *vs.* $-g_s, -g_a$). It only becomes active after image structure has shaped up, in which case it facilitates appearance transfer ($-g_a$ *vs.* Ours).

Choice of diffusion features $\mathbf{F}_t$. Figure 8 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.

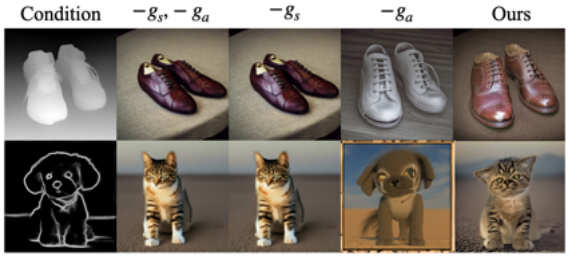

Figure 7. **Ablation on guidance effect.** Top: *"leather shoes"*; Bottom: *"cat, in the desert"*. g_s and g_a stand for structure and appearance guidance, respectively.

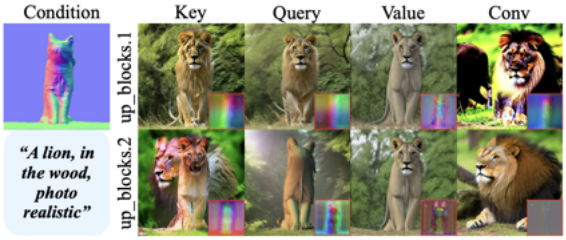

Figure 8. **Ablation on feature choice.** Keys from self-attention of up_block.1 in the U-Net decoder expose the strongest controllability. PCA visualization of the features are in the insets.

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.

Acknowledgement: This work was partially supported by NSF grant IIS-2339769, and by grants from McPherson Eye Research Institute and VCGRE at UW Madison.

References

[1] Autocad. https://www.autodesk.com/products/autocad. 8

[2] Civitai. https://civitai.com/. 3

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

[4] Midjourney. https://www.midjourney.com. 2

[5] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*, 2023. 3

[6] 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 *CVPR*, 2023. 2

[7] 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. 2

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

[9] 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 *ICCV*, 2023. 3, 4

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7

[12] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 3

[13] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 8

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[15] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023. 3, 6

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[17] 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 *ICLR*, 2023. 3

[18] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *ICCV*, 2023. 3, 5

[19] 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 *CVPR*, 2022. 2

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 2, 3, 6, 7

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[23] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 2

[24] 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 *ICLR*, 2022. 2

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[26] 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 *CVPR*, 2023. 3

[27] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[28] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023. 2

[29] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *TPAMI*, 2022. 8

[30] 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 *CVPR*, 2023. 2

[31] 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 *ICLR*, 2022. 2, 3, 6, 7

[32] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 3

[33] 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. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3, 6

[34] 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 *ICML*, 2022. 2, 3

[35] 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. *arXiv preprint arXiv:2304.07193*, 2023. 4

[36] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 3

[37] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, 2023. 2, 3

[38] 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 *ICCV*, 2023. 3

[39] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2

[40] 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 *ICML*, 2021. 2, 7

[41] 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. *JMLR*, 2020. 2

[42] 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. 2

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[46] 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 *CVPR*, 2023. 2, 3

[47] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 3

[48] 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 *NeurIPS*, 2022. 2

[49] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 4, 5

[51] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2, 3

[52] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *CVPR*, 2022. 4, 7

[53] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 2, 3, 4, 6, 7

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[55] Andrey Voynov, Kfir Aberman, and Daniel Cohen-

Or. Sketch-guided text-to-image diffusion models. In *SIGGRAPH*, 2023. 2

[56] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *CVPR*, 2023. 3

[57] 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 *ICCV*, 2023. 3

[58] 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 *CVPR*, 2023. 2

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 2, 3, 6, 7

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[61] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, 2023. 3

[62] 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 *NeurIPS*, 2023. 2, 3, 6