Prompt-Free Diffusion: Taking "Text" out of Text-to-Image Diffusion Models

Xingqian Xu[1,5], Jiayi Guo[1,2], Zhangyang Wang[3,5], Gao Huang[2], Irfan Essa[4], Humphrey Shi[1,4,5]
[1]SHI Labs @ UIUC, Georgia Tech & Oregon, [2]Tsinghua University, [3]UT Austin, [4]Georgia Tech, [5]Picsart AI Research (PAIR)
https://github.com/SHI-Labs/Prompt-Free-Diffusion

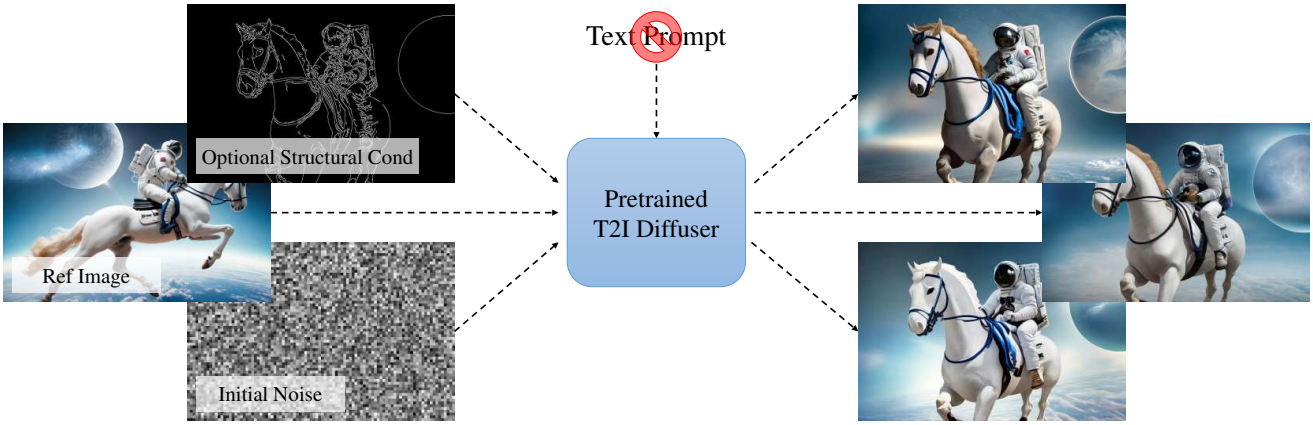


Figure 1. Given a pre-trained Text-to-Image (T2I) diffusion model, **Prompt-Free Diffusion** modifies it to intake a reference image as "context", an *optional* image structural conditioning (i.e. canny edge in this case), and an initial noise. No text prompt is needed: instead, the reference image governs the semantics and appearance, and the optional conditioning provides additional control over the structure. Note that the generated image instances are **precisely controlled** (with only **small variations**, as larger variations would mean less control precision), to faithfully reflect the desired style (reference image) and structure (conditioning).

Abstract

*Text-to-image (T2I) research has grown explosively in the past year, owing to the large-scale pre-trained diffusion models and many emerging personalization and editing approaches. Yet, **one pain point persists: the text prompt engineering**, and searching high-quality text prompts for customized results is more art than science. Moreover, as commonly argued: "an image is worth a thousand words" - the attempt to describe a desired image with texts often ends up being ambiguous and cannot comprehensively cover delicate visual details, hence necessitating more additional controls from the visual domain. In this paper, we take a bold step forward: taking "Text" out of a pre-trained T2I diffusion model, to reduce the burdensome prompt engineering efforts for users. Our proposed framework, **Prompt-Free Diffusion**, relies on **only visual inputs to generate new images**: it takes a reference image as "context", an optional image structural conditioning, and an initial noise, with absolutely no text prompt. The core architecture behind the scene is **Se**mantic Context **Encoder** (**SeeCoder**), substituting the commonly used CLIP-based or LLM-based text encoder. The reusability of SeeCoder also makes it a convenient drop-in component: one can also pre-train a SeeCoder in one T2I model and reuse it for another. Through extensive experiments, Prompt-Free Diffusion is experimentally found to (i) outperform prior exemplar-based image synthesis approaches; (ii) perform on par with state-of-the-art T2I models using prompts following the best practice; and (iii) be naturally extensible to other downstream applications such as anime figure generation and virtual try-on, with promising quality. Our code and models will be open-sourced.*

1. Introduction

The high demand for personalized synthetic results has promoted several text-to-image (T2I) related technologies, including model finetuning, prompt engineering, and controllable editing, to a new level of importance. A considerable number of recent works such as [16, 20, 23, 24, 36, 37, 47, 59, 62, 70] also emerge to address personalization

tasks from different angles. By far, it remains an open question of what is the most convenient way to achieve such personalization. One straightforward approach is to fine-tune a T2I model with exemplar images. Personalized tuning techniques [23, 36, 47, 67] such as DreamBooth [47] have shown promising quality by finetuning model weights. Their downsides are yet obvious: finetuning a model remains to be resource-costly for average users, despite the growing effort of more efficient tuning [23]. Prompt-engineering [20, 60] serves as a lite alternative for personalizing T2I models. It has been widely adopted in the industry due to its excellent cost margin, *i.e*. improving output quality from almost zero cost. Nevertheless, searching high-quality text prompts for customized results is more art than science. Moreover, as commonly argued: "an image is worth a thousand words" - the attempt to describe a desired image with texts often ends up being ambiguous, and cannot comprehensively cover delicate visual details.

To better handle personalized needs, several T2I controlling techniques and adaptive models are proposed [16, 24, 37, 59, 70]. Representative works such as ControlNet [70], T2I-Adapter [37], *etc*. proposed the adaptive siamese networks that take user-customized structural conditionings (*i.e*. canny edge, depth, pose, *etc*.) as generative guidances in additional to prompts. Such an approach has become one of the most popular among downstream users since it: a) disentanglement of structure from content enables more precise controls over results than prompt, b) these plug-and-run modules are reusable to most T2I models that save users from extra training stages.

Do methods like ControlNet [70] *etc*. resolve all the challenges in personalized T2I? Unfortunately, the answer is not quite. For example, it still faces the following issues: a) the current approach meets challenges in generating user-designated textures, objects, and semantics; b) searching prompt-in-need sometimes is problematic and inconvenient; and c) prompt engineering for quality purposes is still required. In conclusion, all these issues come from the fundamental knowledge gap between vision and language. Captions alone may not provide a comprehensive representation of all visual cues, and providing structural guidance does not eliminate this problem.

To overcome the aforementioned challenges, we introduced the novel *Prompt-Free Diffusion*, replacing the regular prompts input with reference images. Speaking with more details, we utilize the newly proposed *Semantic Context Encoder (SeeCoder)*, in which the pixel-based images with arbitrary resolutions can be auto-transformed into meaningful visual embeddings. Such embeddings can represent low-level information, such as textures, effects, *etc*., and high-level information, such as objects, semantics, *etc*. We then use these visual embeddings as the conditional inputs of an arbitrary T2I model, generating high-performing customized outputs on par with the current state-of-the-art. One may notice that our Prompt-Free Diffusion shares similar goals as exemplar-base image generation [10, 31, 39, 71, 72] and image-variation [45, 62]. But it stands out from prior approaches with quality and convenience. Explicitly speaking, SeeCoder is *reusable* to most open-sourced T2I models in which one can easily convert the T2I pipeline to our Prompt-Free pipeline without much effort. While this is mostly infeasible in prior works (*i.e*. exemplar-base generation, image-variation) in which they require either specific models for specific domains or finetuning models deviated from T2I purpose. In summary, our main contributions are concluded in the following:

- We proposed *Prompt-Free Diffusion*, an effective solution generating high-quality images utilizing text-to-image diffusion models without text prompts.
- Empowered by the reusability of *Semantic Context Encoder (SeeCoder)*, the proposed Prompt-Free property can be available in many other existing text-to-image models without extra training, creating a convenient pipeline for personalized image generation.
- Our method can be extended to many downstream applications with competitive quality, such as exemplar-based virtual try-on, and anime figure generation.

2. Related Works

2.1. Text-to-Image Diffusion

Diffusion models (DM) [12, 22, 45, 46, 48, 53, 54] nowadays is the *de facto* workhorsse for Text-to-Image (T2I) generation. Diffusion-based T2I models [18, 34, 38, 45, 46, 48] generate photorealistic images via iterative refinements. GLIDE [38] introduced a cascaded diffusion structure and utilized classifier-free guidance [21] for image generation and editing. DALL-E2 [45] proposed a model with several stages, encoding text with CLIP [42], decoding images from text encoding, and upsampling them from 64^2 to 1024^2. Imagen [48] discovered that scaling up the size of the text encoder [11, 42, 43] improves both sample fidelity and text-image alignment. VQ-Diffusion [18] learned T2I diffusion models on the discrete latent space of VQ-VAE [57]. The popular latent diffusion model (LDM, *i.e*. Stable Diffusion) [46] investigated the diffusion process over the latent space of pre-trained encoders, improving both training and sampling efficiency without quality degradation. Versatile Diffusion [62] further enables LDM across multimodal generation and natively supports image variation.

Although generative adversarial networks (GANs) [4, 17, 28]) and autoregressive (AR) models [5, 40, 41] show T2I capability to some extent, most of them [8, 55, 61, 68, 69, 74] generate images on specific domains instead of open-world text sets. With the advances of large-scale language encoder [6, 42, 43], GANs [27, 49, 56, 73] and AR

models [13, 15, 44, 64] recently start to handle generation with arbitrary texts with promising qualities too.

2.2. Exemplar-based Generation

Exemplar-based generation [2, 19, 33, 39, 51, 63, 65, 66, 70–72] aims to transform structural inputs, (*e.g.*, edge, pose), to photorealistic images according to exemplar images' content. Prior work SPADE [39] captured exemplars' global styles by an encoder and passed them to the spatially-adaptive normalization to synthesize images. CoCosNet [71] proposed better style controls, constructing dense semantic correspondence between a structural input and its exemplar. Zhou *et al*. [72] improve CoCosNet by leveraging a hierarchical PatchMatch method to fast correspondence. Recent progress also introduces unbalanced optimal transport [65], automatic assessment [19], contrastive learning [66], and dynamic sparse mechanism [33] for effective semantic correspondence learning. Meanwhile, techniques for diffusion models [2, 16, 24, 51, 59, 63, 70] also step into this field, in which [37, 70] setup adaptive encoders for diffuser, [24] disentangle and reassemble attributes of images, [63] use CLIP image encoding for inpainting, [16] utilize semantic masks, and [59] uses all types of conditional inputs including images.

3. Method

3.1. Preliminaries

Diffusion process $q(x_T|x_0)$ and $p_\theta(x_T|x_0)$ are T-step Markov Chains [22] that gradually degrade x_0 to x_T with random noises and recover x_T from these noises:

$$q(x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) = \prod_{t=1}^{T} \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}; \beta_t \mathbf{I})$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

in which β_t is the standard deviation of the mixed-in noise at step t, $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ are network predicted mean and standard deviation under parameter θ of the denoised signal at step t. The loss function of training is the variational bound for negative log-likelihood [22] shown as the following:

$$L = \mathbb{E}[-\log p_\theta(x_0)] \leq \mathbb{E}\left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

CLIP [42] is a double-encoder network that bridges text-image pairs by minimizing the contrastive loss (*i.e.* cosine-similarity) between embeddings. CLIP is an important prior module for modern T2I models such as DALLE-2 [45] and Stable Diffusion [46]. Also, it had been proved by prior works [44, 45] that its well-aligned cross-modal latent space is one of the core reasons for the T2I models' success.

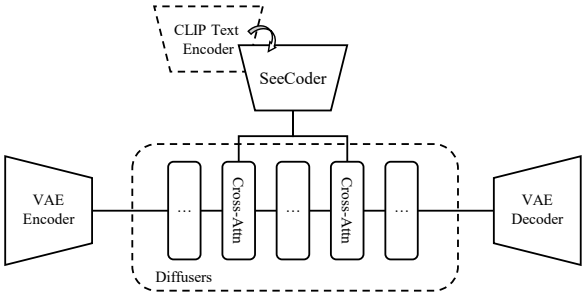


Figure 2. Graphic illustration of our Prompt-Free Diffusion with latent diffusion pipeline in which CLIP text encoder is replaced by the newly proposed SeeCoder.

Image-Variation defines a task that generates images with similar high-level semantics according to another image [45, 62]. Prompt-Free Diffusion is closer to Image-Variation than exemplar-based generation approaches, in which the former finetunes T2I models using CLIP image encoder, while the latter proposed domain-specific networks for tasks such as virtual try-on, makeup, *etc*.

3.2. Prompt-Free Diffusion

As aforementioned, we aim to propose an effective solution to handle nowadays high-demanding personalized T2I while aggressively maintaining *all merits* from prior approaches, *i.e.* high-quality, training-free, and reusable to most open-sourced models. Table 1 explains the pros and cons of varieties of approaches, in which we gauge from three angles: personalization quality, easy installation & domain adaptation, and input complexity & flexibility. The design of Prompt-Free Diffusion inherits T2I and Image-Variation models, consisting of a diffuser and a context encoder as two core modules, as well as an optional VAE that reduces dimensionality in diffusion. Particularly in this work, we have kept a precise latent diffusion structure like Stable Diffusion [46], shown in Figure 2.

Recall that text prompts are first tokenized and then encoded into N-by-C context embeddings using CLIP in common T2I. N and C represent the count and dimension of the embeddings. Later, these embeddings are fed into the diffuser's cross-attention layers as inputs. In our Prompt-Free Diffusion, we replace the CLIP text encoder with the newly proposed SeeCoder (see Figure 2). Instead of text prompts, SeeCoder is designed to take image inputs only. It captures visual cues and transforms them into compatible N-by-C embeddings representing textures, objects, backgrounds, *etc*. We then proceed with the same cross-attention layers in diffusers. Our Prompt-Free Diffusion also doesn't need any image disentanglement as priors (such as in Composer [24]) because SeeCoder can determine a proper way to encode low- and high-level visual cues in an unsupervised manner. For more details, please see Section 3.3.

Methods *vs.* Properties	Personalization Quality	Easy Installation & Domain Adaptation	Input Complexity & Flexibility
Model Finetuning (DreamBooth [47] *etc.*)	Full personalization	No easy installation & adaptation; Data and GPUs are required; Individual weights are required for individual domains	OK when special tokens are learned; Subject to input prompt quality
Prompt Searching & Engineering	Personalization is only available within a limited range, and is likely infeasible in complex cases	No installation required; Reusable to all T2I models & domains	Prompt searching & engineering is for sure required
Adaptive Layers with Structural Inputs (ControlNet [70] *etc.*)	Users can customize the output structure but still has insufficient control over other aspects such textures, styles, backgrounds, *etc.*	Easy installation; Reusable to most T2I models & domains	Structural inputs such as depth and edges are required; Subject to input prompt quality
Image-Variation (Versatile Diffusion [62] *etc.*)	Users can control only high-level semantics but has insufficient control over structures and others	Separate models are required; Not reusable to T2I models & domains.	Image inputs only; No prompts are needed
Prompt-Free Diffusion (ours)	Nearly full personalization; Users may control output structures, textures, and backgrounds using conditional image inputs	Easy installation by replacing CLIP with SeeCoder; Reusable to most T2I models & domains	Structural inputs such as depth and edges are optional; Image inputs only; No prompts are needed

Table 1. This table compares the pros (green) and cons (red) from different methodologies with our Prompt-Free Diffusion along three aspects: Personalization quality; Easy installation & domain adaptation; Input complexity & flexibility. Yellow represents neutral.

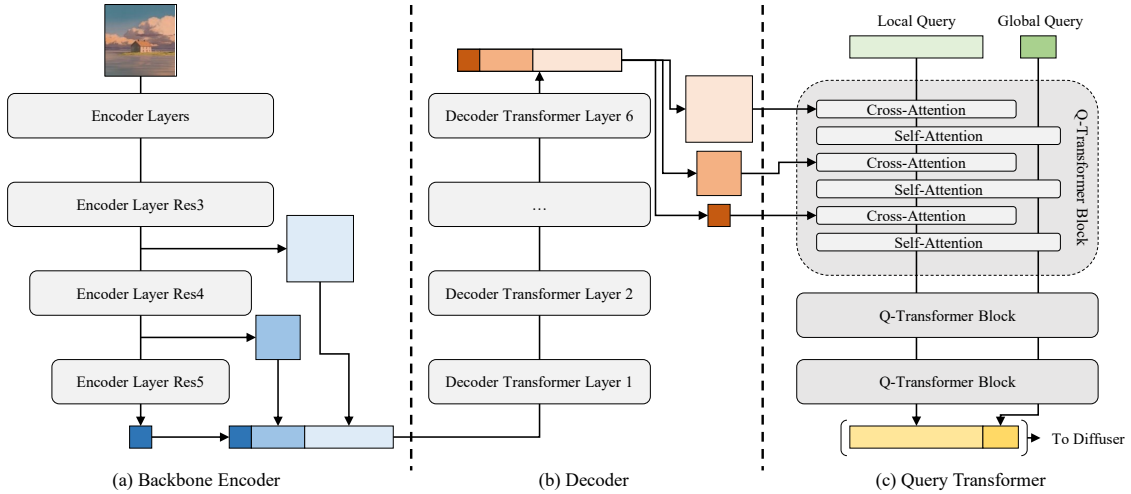


Figure 3. The structure of *Semantic Context Encoder (SeeCoder)*, which includes a *Backbone Encoder*, a *Decoder*, and a *Query Transformer*. For simplicity, we hind several detail designs: lateral connections; learnable query, level, and 2D spatial embeddings.

3.3. Semantic Context Encoder

As the core module in Prompt-Free Diffusion, our Semantic Context Encoder (SeeCoder) aims to take only image inputs and encode all visual cues into an embedding. One may notice that CLIP may also encode images. But in practice, CLIP's ViT [14] shows limited capacity because a) unable to take inputs higher than resolution 384^2; b) does not capture detail textures, objects, *etc.*; c) trained with contrastive loss making it an indirect way of processing visual cues. Therefore we propose SeeCoder, a solution better fits vision tasks than CLIP.

SeeCoder can be breakdown into three components: *Backbone Encoder*, *Decoder*, and *Query Transformer* (see Figure 3). **Backbone Encoder** uses SWIN-L [35] because it transforms arbitrary-resolution images into feature pyramids that better capture visual cues in different scales. For the **Decoder**, we proposed a transformer-based network with several convolutions. Specifically speaking, the Decoder takes features from different levels; uses convolutions to equalize channels; concatenates all flattened features; then passes it through 6 multi-head self-attention modules [58] with linear projections and LayerNorms [1]. The final outputs are split and shaped back into 2D, then sum with lateral-linked input features.

The last part of SeeCoder is **Query Transformer** which finalizes multi-level visual features into a single 1D visual embedding. The network started with 4 freely-learning global queries and 144 local queries. It holds a mixture of cross-attention and self-attention layers, one after another. The cross-attentions take local queries as Q and visual features as K and V. The self-attentions use the concatenation of global and local queries as QKV. Such design prompts a hierarchical knowledge passing in which the cross-attentions transit visual cues to local queries, and self-attentions distill local queries into global queries. Besides, the network also contains free-learned query embeddings, level embeddings, and optional 2D spatial embeddings. The optional spatial embeddings are sine-cosine encodings followed by several MLP layers. We name the network

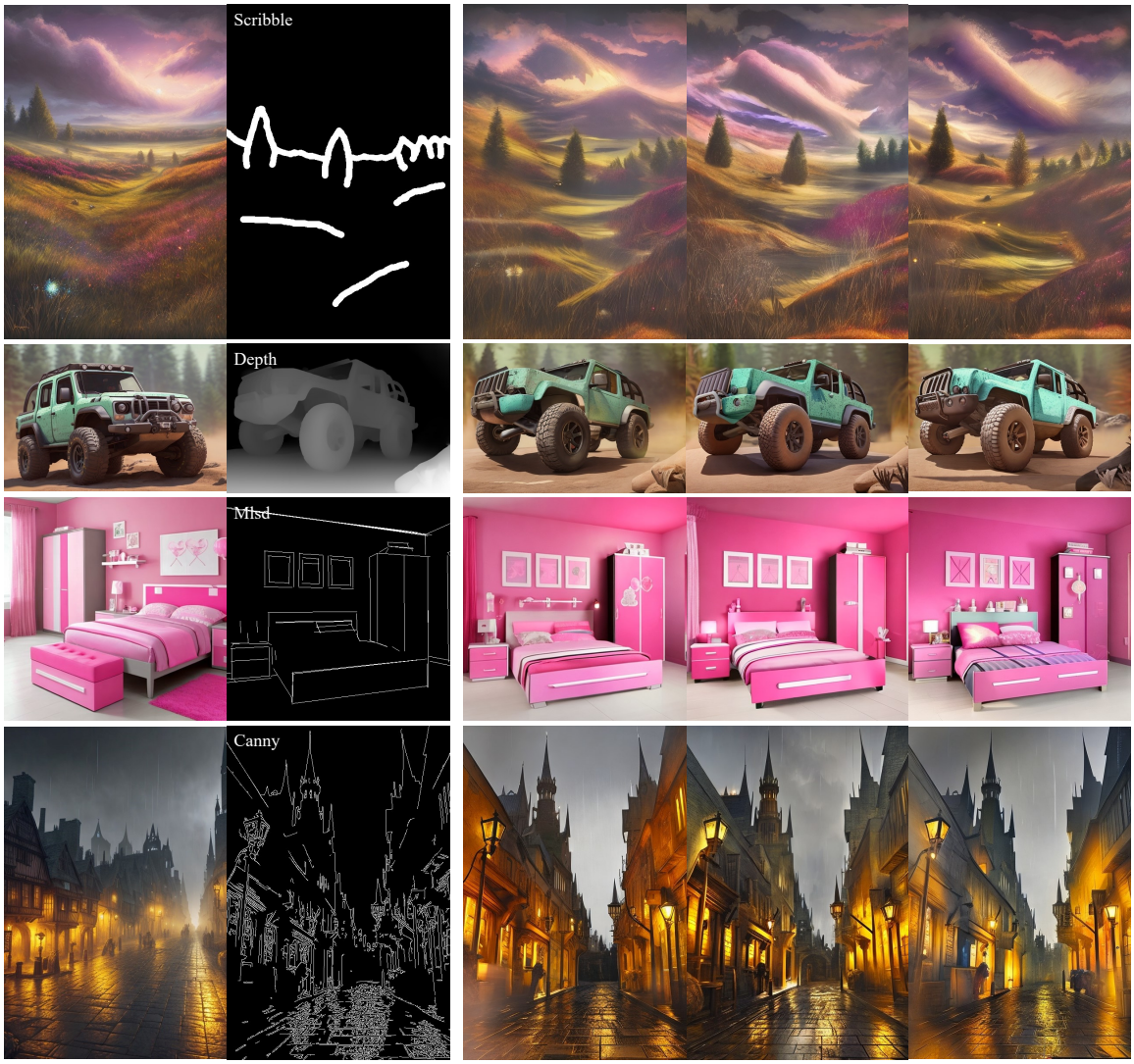


Figure 4. This figure shows results from Prompt-Free Diffusion, in which we sample three outputs for each case using one reference image and one structural conditioning (*i.e.* canny edge, depth, mlsd, and scribble) as input. No prompts are required.

SeeCoder-PA when there exist 2D spatial embeddings, where PA is short for Position-Aware. Finally, global and local queries are concatenated and passed to the diffuser for content generation. Notice that our network shares similarities with segmentation approaches, such as [9, 25, 26, 32]. Nevertheless, their end purposes vary: to capture visual embeddings for discriminative *vs*. generative tasks.

4. Experiments

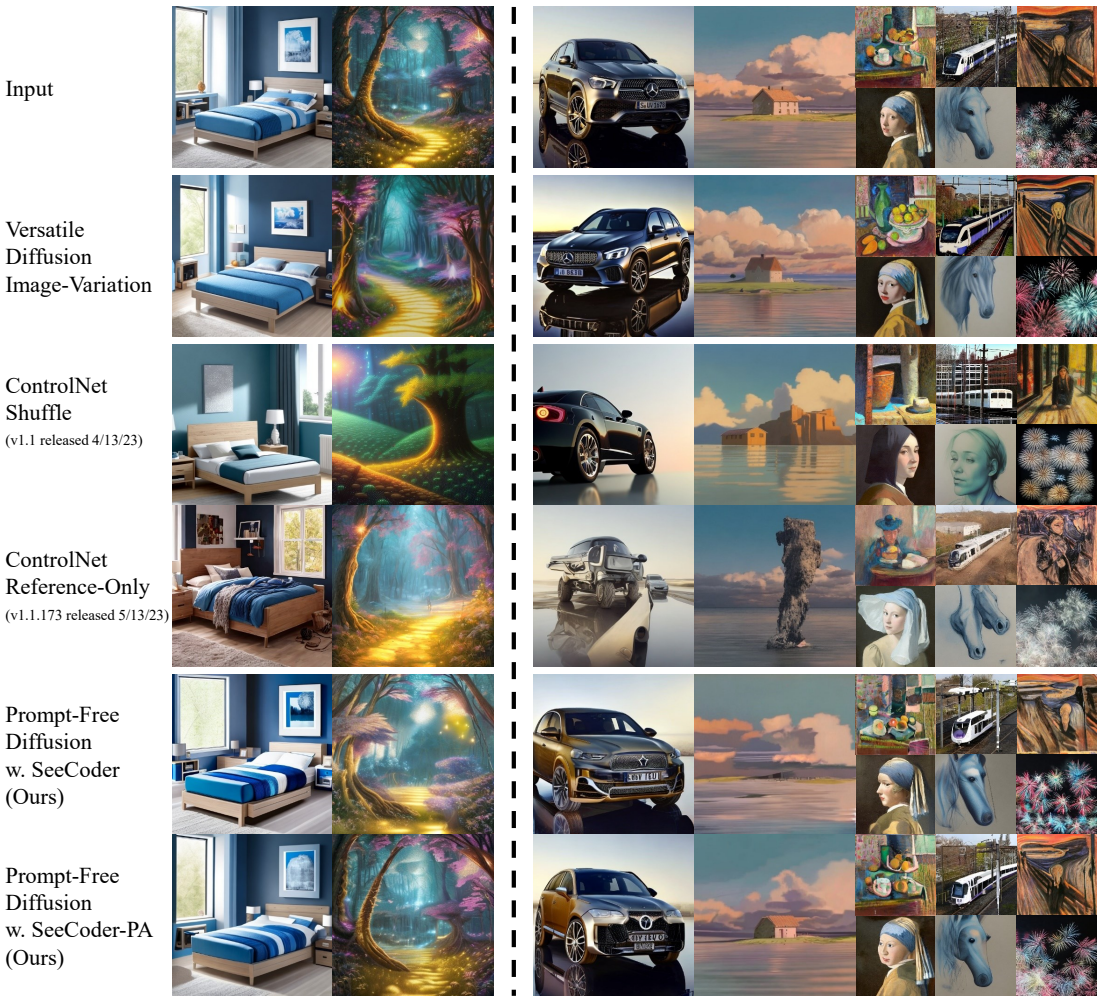
4.1. Data

Aligned with many prior works [45, 46, 62], we adopted Laion2B-en [50] and COYO-700M [7] as our training data. Laion2B and COYO are large-scale image-text pairs that contain 2 billion and 700 million web-collected samples. Both datasets were frequently used in T2I research. Since Prompt-Free Diffusion requires no prompts, we actually only used these datasets' image collections for model training and evaluation.

4.2. Training

The pretrained models' selection impacts Prompt-Free Diffusion's final model due to its unique training procedure with frozen weights. As mentioned in Section 3.3, we used SWIN-L [35] as SeeCoder's Backbone Encoder, SD2.0's [46] VAE, and an in-house SD1.5-based T2I diffuser as the pretrained diffuser. Despite involving an in-house model, we demonstrated that a well-trained SeeCoder

Figure 5. This figure gathers the performance of ControlNet+T2I using prompts with progressive complexity. It starts with the prompt *"medieval streets, yellow and blue, rain and fog"* which gives basic semantics and styles. Later we improve the performance with semantic decorative prompts such as *"dystopian"* and *"wide shot"*, style prompts such as *"fantasy"*, and common prompt engineering such as *"realistic"* and *"8k"* (displayed in green). Our Prompt-Free Diffusion matches the quality of the top two most sophisticated prompts.



(a) In-domain samples (b) Out-of-domain samples

Figure 6. Image-Variation comparison between VD [62], ControlNet [70], and Prompt-Free Diffusion. Testing samples s are categorized into a) in-domain and b) out-of-domain, meaning whether the input can be generated using the pretrained T2I diffuser. In conclusion, Prompt-Free Diffusion w. SeeCoder-PA beats ControlNet Shuffle and Reference-Only for both in-domain and out-of-domain cases.

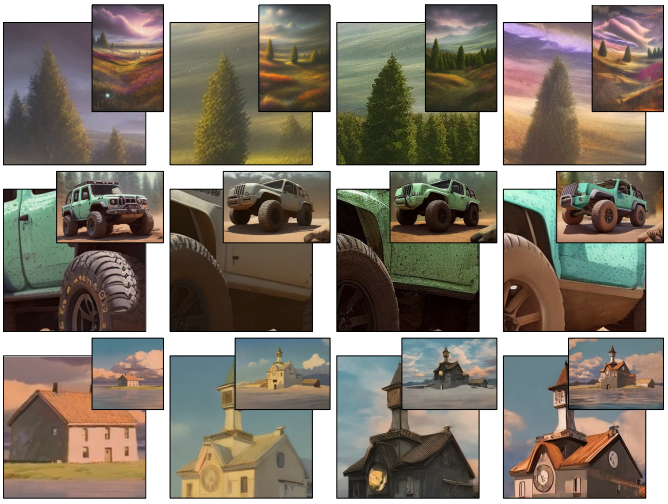


Figure 7. Performance comparison between Dual-ControlNet [70] (*i.e*. Shuffle+x and Reference-Only+x, x representing an arbitrary ControlNet other than Shuffle or Reference-Only) and our Prompt-Free Diffusion. From which we notice that our method outperforms both ControlNet solutions in terms of texture quality, color consistency, style replication, *etc*. (see the zoomed-in views).

is reusable to other open-source T2I models in Section 4.4.

Other training settings are listed as the following: We used DDPM with $T = 1000$ diffusion timesteps with linearly increased β from 8.5×10^{-5} to 1.2×10^{-2}. For each iteration, we sampled DDPM timestep $t \in T$ uniformly. We trained the model with 100k iterations, 50k with a learning rate 10^{-4}, and the other 50k with 10^{-5}. Our training batch size was set to 512, 8 samples per GPU, a total of 16 A100 GPUs across two nodes, and a gradient accumulation of 4.

Besides, we also train a separate position-aware model with 2D spatial embeddings, namely SeeCoder-PA. Prompt-Free Diffusion with SeeCoder-PA performs better than SeeCoder when no structural conditionings are used (see Sec 4.3), an explainable phenomenon as the spatial embeddings partly cover the missing structural inputs. SeeCoder-PA is trained from a 50k SeeCoder checkpoint and finetuned additional 20k steps with a learning rate 5×10^{-5}.

4.3. Performance

We demonstrate the performance of Prompt-Free Diffusion in Figure 4, in which our method generates high-quality images replicating details from reference inputs. In this experiment, Prompt-Free Diffusion extensively uses ControlNets [70] to handle a variety of structural conditionings, *e.g*. canny-edge, depth, mlsd, and scribble. Also, Prompt-Free Diffusion is insensitive to resolution and aspect ratios as we show three scales: 512^2, 512×768, and 768×512. The reference dimension *does not* need to match the output dimension as well. Figure 4 shows cases with

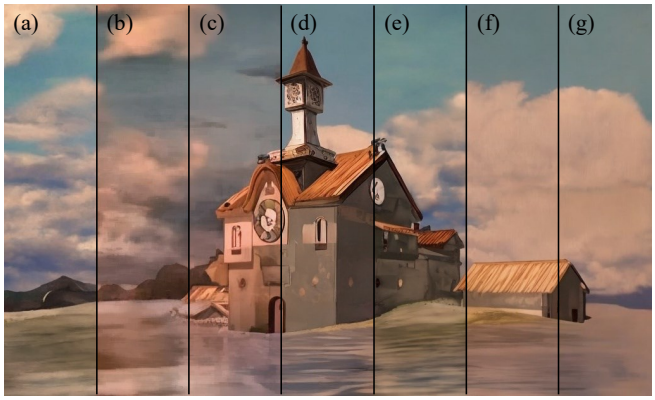


Figure 8. SeeCoder's adaptability examination on 7 open-sourced & in-house models: (a) SD1.5 [46]; (b) OAM-V2; (c) Anything-V4; (d) In-house model (*i.e*. the diffuser SEE was trained with); (e) OpenJourney-V4; (f) Deliberate-V2; (g) RealisticVision-V2

matched dimensions merely for better exhibition. Besides, the aforementioned in-house T2I diffuser is applied to generate these results. Unless separately mentioned, such a setting has been kept in all experiments.

Compare with T2I: The following experiment shows the performance comparison between our Prompt-Free Diffusion and the traditional prompt-based ControlNet+T2I [46, 70] (see Figure 5). Specifically, we use prompt inputs with progressive complexities and check which level of prompt complexity is equivalent to our Prompt-Free Diffusion in performance. As shown in Figure 5, our approach roughly reaches the top two levels: between "requiring semantic and style decorative prompt" and "requiring extra prompt engineering". We also noticed a tricky color shift using ControlNet+T2I, which we gave up fixing after numerous tries.

Image-Variation: Next, we evaluate Image-Variation (IV): generating images from other reference images. Notice that IV is a natural setup for Prompt-Free Diffusion when no ControlNet is involved. Prior baseline such as VD [62] finetunes a diffusion model for IV. Concurrently, the ControlNet team also proposed two new models, "shuffle" and "reference-only" (v1.1.173) [70], so we compared both. Testing samples are categorized into in-domain and out-of-domain, meaning whether the reference images were generated by the pre-trained T2I diffuser, or from a different source. We draw the following conclusions in this test: a) VD has the best overall performance, meaning finetuning still yields good performance; b) ControlNet Shuffle and Reference-Only have significant performance drops in out-of-domain tests. Reference-Only outperforms Shuffle on in-domain samples and vice versa on out-of-domain samples. These results reflect ControlNet's weakness in generality and show its limitation of not having quality and generality in both hands. c) Prompt-Free Diffusion better replicates the reference images (*i.e*. semantic, texture,

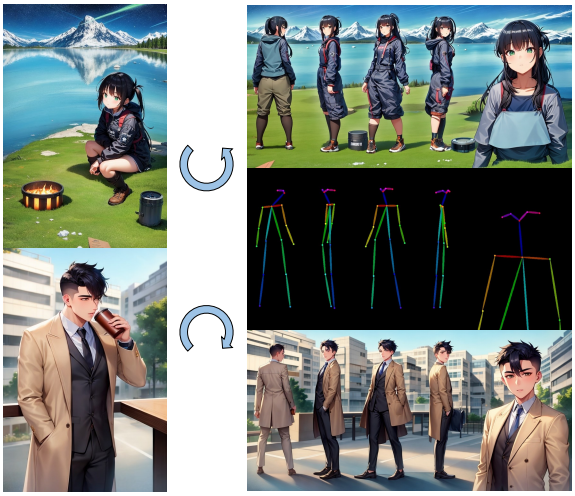


Figure 9. Demo of anime figure generation using Prompt-Free Diffusion with a reference image and conditioning pose.

Figure 10. Demo of virtual try-on using Prompt-Free Diffusion and state-of-the-art exemplar-based approaches.

background, *etc*.), which aligns with its training objective. SeeCoder-PA (*i.e*. SeeCoder with spatial encoding) outperforms SeeCoder in terms of quality (see Section 4.2). Both models also do well on in-domain samples and have quality gaps on out-of-domain samples. Overall, our approach beats ControlNet for Image-Variation.

Compare with Dual-ControlNet: We also compare our structural-guided performance (*i.e*. Prompt-Free Diffusion with Single ControlNet) with Dual-ControlNet setup. Figure 7 shows the overall performance with zoomed-in details, from which we demonstrate that our approach has strength in replicating textures, colors, styles, backgrounds, *etc*. from reference.

4.4. Reusability

A critical property of our SeeCoder is that: once trained, SeeCoder can be plug-and-use for other T2I diffusers. Therefore, one can easily customize their own T2I model to Prompt-Free Diffusion by replacing CLIP with SeeCoder. We test such property by **directly plug the same pre-trained SeeCoder** with six other open-source T2I models, including the base model *SD1.5* [46]; art-focused models *OpenJourney-V4* and *Deliberate-V2*; anime models *OAM-V2* and *Anything-V4*; and the photorealistic model *RealisticVision-V2*. We emphasize that such an experiment requires no training at all. The results are shown in Figure 8, proving SeeCoder is highly reusable to other T2I models.

4.5. Downstream Applications

Image-based Anime Figure Generation is one of the practical uses of Prompt-Free Diffusion in anime or game design. In Figure 9, we generate anime figures based on SeeCoder-captured visual cues and conditional poses. To better accommodate anime data that is "out-of-domain", SeeCoder here is finetuned on OAM-V2 and its synthesized images for 30k iterations. From the results, Prompt-Free Diffusion demonstrates promising results for this task.

Virtual Try-on is a traditional exemplar-based task target by prior works such as [10] and [31]. Our Prompt-Free Diffusion can also handle this task with minor modifications: We obtain cloth masks from SAM [30] and inputs these masks as conditions of ControlNet. Then like other applications, we use SeeCoder's visual embedding to guide our generation. Results are shown in Figure 10. Other applications such as model cognitive study [67] and video generation [3, 29, 52] can be future applications supported by SeeCoder.

5. Conclusion and Ethical Discussion

In this article, we propose Prompt-Free Diffusion, a novel pipeline that generates personalized outputs based on exemplar images, not text prompts. Through experiments, we show that our core module, SeeCoder, can generate high-quality results and can easily plug-and-use in various well-established T2I pipelines through CLIP replacement.

While our proposed Prompt-Free Diffusion can assist artists and designers in creative content generation, it is important to acknowledge that the misuse or abuse of our system may cause negative social impacts and ethical concerns similar to other controllable image synthesis approaches. In addition, open-source pretrained text-to-image models used in conjunction with Prompt-Free Diffusion may contain harmful bias, stereotypes, *etc*. As a crucial step to address these concerns, we encourage users to deploy and utilize our approach in a responsible way with ethical regulations and enhanced transparency.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, 2023. 3

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *arXiv preprint arXiv:2304.08818*, 2023. 8

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[7] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 5

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 2

[9] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 5

[10] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021. 2, 8

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[13] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 3

[16] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 1, 2, 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[18] 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

[19] Jiayi Guo, Chaoqun Du, Jiangshan Wang, Huijuan Huang, Pengfei Wan, and Gao Huang. Assessing a single image in reference-guided image synthesis. In *AAAI*, 2022. 3

[20] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Shiji Song, Humphrey Shi, and Gao Huang. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, 2023. 1, 2

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[23] 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. *arXiv preprint arXiv:2106.09685*, 2021. 1, 2

[24] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 1, 2, 3

[25] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. *arXiv preprint arXiv:2112.12782*, 2021. 5

[26] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, 2023. 5

[27] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. *arXiv preprint arXiv:2303.05511*, 2023. 2

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[29] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant

Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 8

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 8

[31] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, 2022. 2, 8

[32] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022. 5

[33] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. DynaST: Dynamic sparse transformer for exemplar-guided image generation. In *ECCV*, 2022. 3

[34] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *WACV*, 2023. 2

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 5

[36] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *CVPR*, 2023. 1, 2

[37] 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. 1, 2, 3

[38] 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2, 3

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[42] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3

[43] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21 (140):1–67, 2020. 2

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[45] 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, 3, 5

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5, 7, 8

[47] 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. 1, 2, 4

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[49] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 2

[50] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[51] Junyoung Seo, Gyuseong Lee, Seokju Cho, Jiyoung Lee, and Seungryong Kim. MIDMs: Matching interleaved diffusion models for exemplar-based image translation. *arXiv preprint arXiv:2209.11047*, 2022. 3

[52] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 8

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[55] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 2

[56] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. GALIP: Generative adversarial clips for text-to-image synthesis. *arXiv preprint arXiv:2301.12959*, 2023. 2

[57] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[58] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4

[59] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*, 2023. 1, 2, 3

[60] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022. 2

[61] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2

[62] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 1, 2, 3, 4, 5, 6, 7

[63] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by Example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023. 3

[64] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[65] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *CVPR*, 2021. 3

[66] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, and Changgong Zhang. Marginal contrastive correspondence for guided image generation. In *CVPR*, 2022. 3

[67] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2211.08332*, 2023. 2, 8

[68] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2

[69] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 2018. 2

[70] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 2, 3, 4, 6, 7

[71] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, 2020. 2, 3

[72] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. CoCosNet v2: Full-resolution correspondence learning for image translation. In *CVPR*, 2021. 2, 3

[73] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021. 2

[74] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019. 2