SingleDream: Attribute-Driven T2I Customization from a Single Reference Image

Ye Wang[1] Ruiqi Liu[1] Tieru Wu[1,3] Zili Yi[2*] Rui Ma[1,3*]

[1]School of Artificial Intelligence, Jilin University

[2]School of Intelligence Science and Technology, Nanjing University

[3]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China

{yewang22, liurq24}@mails.jlu.edu.cn, wutr@jlu.edu.cn, yi@nju.edu.cn ruim@jlu.edu.cn

Abstract

Personalized or customized text-to-image (T2I) models not only produce lifelike and varied visuals but also allow users to tailor the images to fit their personal taste. These customization techniques can grasp the essence of a concept through a collection of images, or adjust a pre-trained text-to-image model with a specific image input for subject-driven. Yet, accurately capturing the distinct visual attributes of a single image poses a challenge for these methods. To address this issue, we introduce SingleDream, a novel parameter-efficient fine-tuning method which only utilizes a single reference image for attribute-driven T2I customization. A novel hypernetwork-enhanced attribute-aware fine-tuning approach is employed to achieve the precise learning of various attributes, including style, appearance and shape, from the reference image. Comparing to existing image customization methods, our method shows significant superiority in attribute-driven T2I customization generation.

Keywords: Attribute-driven generation, attribute disentanglement, text-to-image customization, hypernetwork.

1. Introduction

Over the recent years, significant progress has been observed in the area of customized T2I generation [7, 19, 13, 11, 8, 29, 20, 5, 23, 12, 2, 30, 9]. These advancements in generative models have not only facilitated the generation of images that are both realistic and varied, but have also empowered users to shape these images to align with their personal visual preferences, i.e., image customization. Recent customization methods have followed two main approaches: one focuses on extracting the essence of a unified concept from image sets to generate new images via text prompts [19, 13, 7, 23], while the other integrates specific subjects into pre-trained text-to-image diffusion models using image encoders, enabling one-shot subject-aware customization [29, 9, 24]. Moreover, increasing attention has been directed towards attribute-driven customization techniques [7, 28, 24, 35], which enable fine-grained customization of specific visual attributes, thereby enhancing user control and flexibility in generated content.

Specifically, Gal et al. [7] observe that the shallow layers of the denoising U-net structures within diffusion models tend to generate colors and materials, while the deep layers provide semantic guidance. They use only 3-5 images to learn a user-provided concept and represent it using new "words" in the embedding space of a frozen text-to-image model. P+ [28] extends a single text prompt into multiple prompts and injects them into different cross-attention layers of U-net to decouple visual attributes like style, color and structure. However, P+ [28] requires multiple reference images of a specific subject, which can be hard to collect. On the other hand, StyleDrop [24] allows one-shot style-aware customization of text-to-image synthesis, by tuning a specific subset of parameters of a diffusion model. ProSpect [35] discovered that attribute generation is closely related to different stages of the sampling process. By dividing the sampling phase into 10 sub-stages, it was observed that the early stages primarily focus on learning the layout, while the later stages are dedicated to learning appearance, texture, and other finer details.

However, these attribute-driven approaches still face several limitations. First, methods like P+ [28] require multiple reference images, which can be challenging for users to collect when targeting specific attributes. Second, methods like ProSpect [35] face challenges in disentangling attributes, as the sampling stages for different attributes may overlap, causing multiple attributes to become entangled. These limitations significantly hinder the accurate capture and learning of diverse visual attributes, resulting in suboptimal attribute customization outcomes.

In this paper, our goal is to achieve efficient and well-disentangled attribute-driven customization by fine-tuning

*Corresponding author.

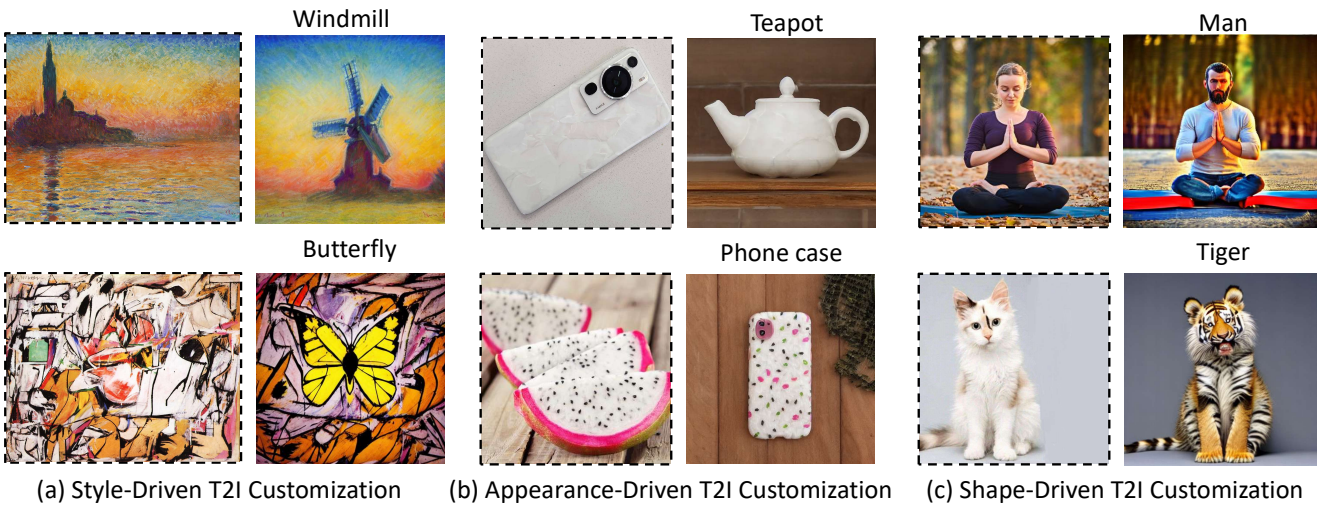

Figure 1: Our method enables text-to-image customization driven by style (see a), appearance (see b), and shape (see c), all using just a single reference image, as demonstrated within the dashed frame.

pre-trained text-to-image diffusion models with only a single reference image. We focus on three important visual attributes in an image: global-level style attributes (see Figure 1.a), object-level appearance attributes (see Figure 1.b), and structure-related shape attributes (see Figure 1.c). To this end, we propose SingleDream, a simple yet effective hypernetwork-enhanced attribute-aware fine-tuning method for attribute-driven T2I customization. Initially, we propose an attribute-aware fine-tuning approach, which differs from the conventional methods in [13, 19] by selectively fine-tuning layers associated with specific attributes. To identify the layers relevant to each attribute in the U-Net, we performed a detailed analysis of U-Net for attribute-aware fine-tuning. Our findings show that the encoder primarily captures structure-related information, such as shape, while the decoder is more sensitive to appearance and style attributes. However, applying the attribute-aware fine-tuning method to a single reference image still results in severe overfitting and the failure of attribute learning. To address this, we propose a hypernetwork-enhanced attribute-aware fine-tuning approach, which employs a lightweight hypernetwork to modulate and refine the U-Net's weights. This strategy not only ensures smoother parameter updates and reduces the risk of overfitting, but also effectively identifies and represents the desired attributes from the reference image.

Our method enables T2I customization driven by three distinct attributes: style, appearance, and shape, resulting in diverse image outputs. Additionally, it supports flexible control over attribute customization strength through simple hyperparameter adjustments. Finally, our approach allows for attribute mixing, such as style mixing, this further highlights the enhanced generative diversity and creative potential of our method.

We evaluate our method against existing approaches on a dataset specifically curated for attribute-driven T2I customization. Both quantitative and qualitative results demonstrate the superiority of our approach.

In summary, the key contributions of our work are outlined as follows:

- We introduce SingleDream, a streamlined and highly efficient approach that uses only a single reference image for attribute-driven T2I customization.
- We propose an attribute-aware fine-tuning method and analyze the distinct roles of the encoder and decoder in the diffusion U-Net for attribute learning. Additionally, we incorporate a lightweight hypernetwork to mitigate the risk of overfitting during single-image fine-tuning while enabling precise attribute learning.
- Through comprehensive quantitative and qualitative evaluation, we show that our method significantly outperforms existing image customization methods on attribute-driven T2I customization.

2. Related Works

Personalized T2I Generation. Recent studies [25, 16, 4, 26, 32, 31, 10] have pivoted towards using visual exemplars as a cornerstone for image generation to navigate the inherent vagueness and unpredictability associated with text-based prompts. This methodology emphasizes the use of one or more reference images as a primary guide, moving away from the exclusive dependence on textual descriptions for synthesizing images. Nonetheless, these approaches tend to concentrate on capturing the general essence of the

reference image, such as its objects or subjects, without the capacity for attribute-driven T2I customization. Furthermore, several methodologies [32, 10] are characterized by their substantial training demands, requiring extensive fine-tuning across vast datasets to enable the use of visual images as conditional inputs for Stable Diffusion. In contrast, our proposed method seeks to overcome the limitations associated with extracting multiple visual attributes from a single reference image.

Parameter Efficient Fine Tuning (PEFT). PEFT represents an innovative approach in the refinement of deep learning models, emphasizing the adjustment of a subset of parameters rather than the entire model. These parameters are identified as either specific subsets from the originally trained model or a minimal number of newly introduced parameters during the fine-tuning phase. PEFT has been applied in text-to-image diffusion models [22, 18] through techniques such as LoRA [21] and adapter tuning [15, 33, 29, 6, 14]. To facilitate the adaptation of pre-trained T2I generators to visual inputs, ELITE [29] fine-tunes the attention layer parameters, while UMM-Diffusion [14] introduces a visual mapping layer, keeping the pre-trained generator's weights unchanged. SuTI [6] enables personalized image generation without the need for test-time fine-tuning by leveraging a vast dataset of images created by subject-specific expert models. Unlike these approaches, our method utilizes a lightweight hypernetwork to adjust and refine an attribute-specific subset of pre-trained parameters within the Diffusion U-net.

Many-Shot T2I Customization. Many-shot techniques [3, 27, 34] necessitate the training of either the diffusion model itself or its conditioning branch to facilitate customized T2I generation, relying on extensive datasets or a handful of examples for training. DreamBooth [19] introduces a methodology for embedding a new subject into the existing model architecture without compromising the model's original capabilities, by training the diffusion model with reference samples. In contrast, SuTI [6] begins by assembling a substantial dataset of input images paired with their recontextualized counterparts generated via the standard DreamBooth procedure. InstantBooth [23] devises a novel conditioning branch within the diffusion model, enabling customization with a limited set of images to produce tailored outputs across various styles. FastComposer [30] employs an image encoder to derive subject-specific embeddings, addressing the challenge of identity preservation when generating images with multiple subjects. Diverging from these many-shot strategies, our research concentrates on achieving attribute-driven T2I customization with a one-shot approach.

3. Method

Given a single reference image, our goal is to distinguish, separate and learn different visual attributes, including style, appearance and shape, and to facilitate the generation of attribute-driven T2I customization. To achieve this goal, we propose SingleDream, as illustrated in Figure 2.

In the following sections, we first introduce the preliminaries of Stable Diffusion [18] in Section 3.1. In Section 3.2, we provide a detailed explanation of SingleDream. Finally, we present the implementation details in Section 3.3.

3.1. Preliminary

Stable Diffusion. Stable Diffusion [18], a state-of-the-art T2I generation model, operates within a low-dimensional latent space. It begins by encoding an input image x into a latent representation z using a VAE encoder. Noise ϵ is then introduced at time step t to create a noisy latent z_t. To guide the generation process with text conditions, Stable Diffusion incorporates a CLIP text encoder τ to encode textual prompts c, which are integrated into the cross-attention layers for interaction with the noisy latents. Finally, a conditional U-Net backbone ϵ_θ is trained to predict the noise ϵ. The training objectives is as follows:

$$L_{SD}(\theta) := \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(c))\|^2\right]. \quad (1)$$

3.2. SingleDream Framework

As shown in Figure 2, our method architecture is streamlined and highly efficient. We propose an attribute-aware fine-tuning approach to identify and fine-tune attribute-related network layers. To further mitigate the overfitting issue during single-image fine-tuning, we incorporate a hypernetwork to refine and modulate parameter updates for enhancing attribute-aware fine-tuning. For more details, please refer to the following sections.

Attribute-Aware Fine-Tuning Traditional fine-tuning methods [19, 13] primarily focus on learning object concepts. Directly adapting these methods to attribute-driven customization tasks often results in the entanglement of various visual attributes. To address this issue, we propose an attribute-aware fine-tuning approach, where the key idea is to fine-tune only the network layers associated with specific attributes, rather than applying global fine-tuning. In order to perform attribute-aware fine-tuning, we first need to identify the network layers relevant to specific attributes. We consider that the representation spaces of different visual attributes emphasize different aspects of information and features. Style attributes capture the overall stylistic characteristics of an image, appearance attributes focus on intricate details such as texture, color, and material, while shape attributes primarily represent low-level visual features. As a

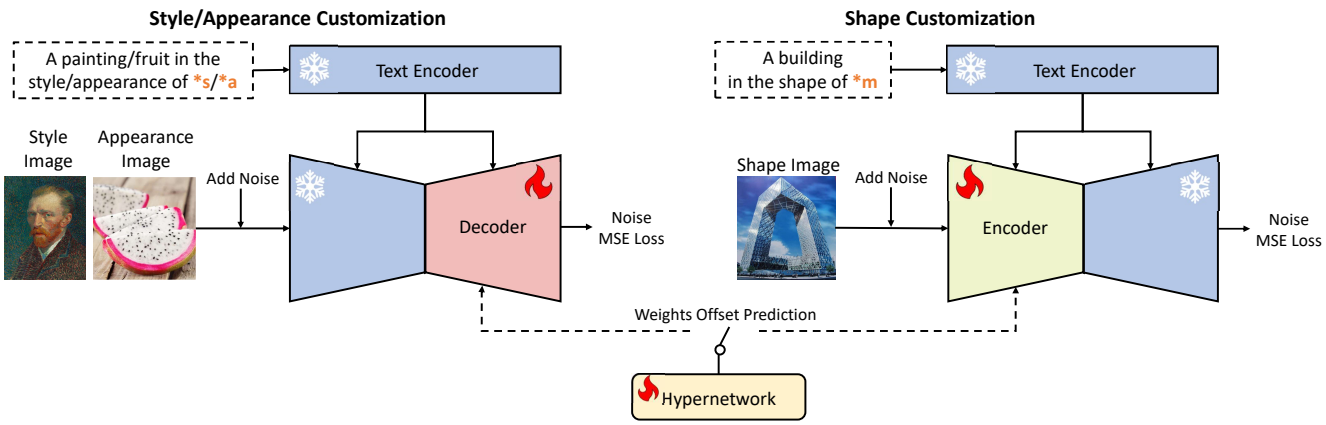

Figure 2: SingleDream pipeline. Our method requires only one reference image as input, and we introduce a hypernetwork-enhanced attribute-aware fine-tuning approach to adjust the parameters of the U-net encoder or decoder for efficient attribute-driven T2I customization.

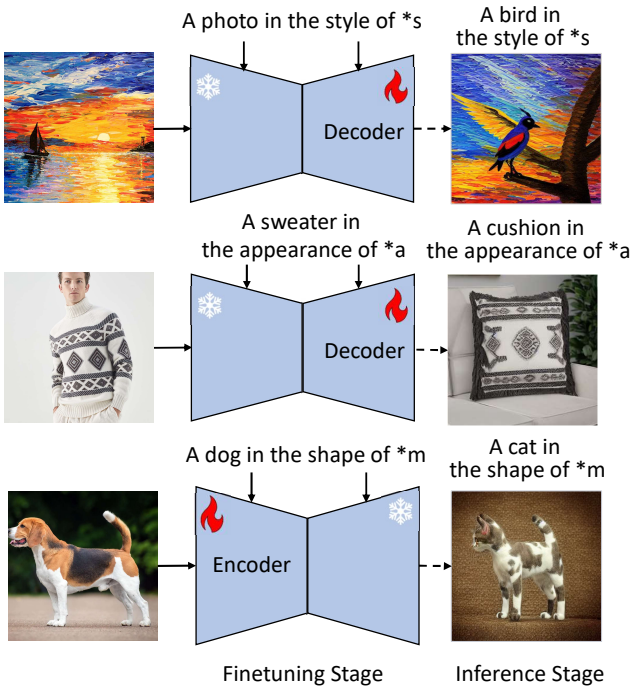

Figure 3: An illustration showing the distinct roles of the encoder and decoder in the diffusion U-Net for learning different attributes.

result, these attributes require specialized learning within distinct network modules.

Based on the above considerations, we conducted a simple experiment using the Stable Diffusion model [18]. As shown in Figure 3, we selected a reference image and the corresponding text prompt to fine-tune different modules of the U-net, specifically the encoder or decoder. We then used the fine-tuned model to generate images.

We observe that when fine-tuning the decoder leads to images with similar style (see Figure 3, 1st row) or appearance (see Figure 3, 2rd row), whereas fine-tuning the encoder (see Figure 3, 3nd row), the generated images exhibit similar shape to reference dog. This experiment further validates our idea that different visual attributes are learned by distinct network modules.

However, we found that attribute-aware fine-tuning tends to suffer from severe overfitting when fine-tuning on a single image. To address this issue, we propose an efficient hypernetwork-enhanced attribute-aware fine-tuning method that achieves smoother parameter updates, effectively mitigating the risk of overfitting.

Hypernetwork-Enhanced Attribute-Aware Fine-Tuning. To address the above limitations, we employ a efficient hypernetwork to enhance attribute-aware fine-tuning mechanism, where the core idea is to utilize a lightweight hypernetwork to modulate and guide the parameter updates of U-net's encoder or decoder, rather than performing direct fine-tuning. Essentially, the hypernetwork is trained to guide the updates of the main network parameters in a low-rank, smooth manner. The structure of the hypernetwork is highly lightweight, consisting of only four linear layers, as shown in Figure 4. We follow the architecture of E4T [9] weight offsets prediction module for the construction of our hypernetwork. The module takes as input a learnable constant $cons$ (default-initialized to 1) and the dimension information $[dim_r, dim_c]$ of the target weight parameters. It is then trained to predict weight offsets in the same dimensions as the target weight parameters. Here, dim_r represents the number of rows of the target weight parameters, and dim_c represents the number of columns. In detail, the learnable constant

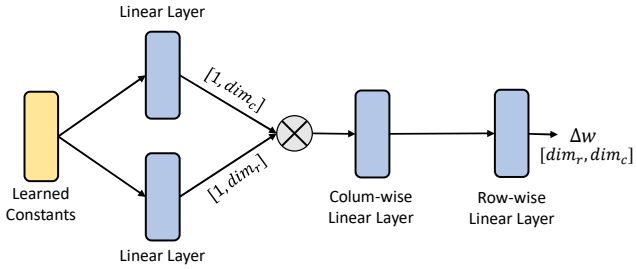

Figure 4: The architecture of hypernetwork.

passes through two linear layers, yielding outputs that are multiplied to derive the initial weight offset matrix. Row and column transformations are then applied to this matrix to obtain the final weight offset matrix Δw. As discussed in the literatures [9, 13, 29], the weights of self-attention and cross-attention play a crucial role in the process of image customization. Therefore, we utilize the hypernetwork as a weight offsets prediction module to modulate and guide the updates of attention-related weights within the encoder or decoder. The high-level parameter update process is defined as follows:

$$\Delta w = hypernetwork(cons, dim_r, dim_c), \quad (2)$$

$$w^*_{attn} = w_{attn} + \lambda * \Delta w, \quad (3)$$

the w_{attn} denotes the general term for the attention-related parameters, which includes the query matrix, key matrix and value matrix for self-attention and cross-attention layers. λ is a weight coefficient that is used to regulate the updating strength of parameters. During training, we set λ to 1.0. Once training is complete, during inference generation, we can adjust the value of λ to control the strength of attribute customization.

This efficient fine-tuning approach greatly reduces the risk of overfitting when fine-tuning on a single reference image, while also achieving high-quality, attribute-driven T2I customization generation.

Loss Function. To guide the customization and learning of attributes, we employ the original noise prediction loss function, which is expressed as:

$$L_{OSTAF}(\theta) := \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, \tau(c))\|^2\right]. \quad (4)$$

Note that θ denotes the parameters of the encoder or decoder and the corresponding hypernetwork, ϵ denotes the noise, z_t represents the latent of input image at time t, t denotes the current time step, $\tau(c)$ represents the encoding of the input text prompt c using the text encoder τ of the CLIP model.

3.3. Implementation Details

We employ Stable Diffusion 1.4 [18] as our base model. During the training process, the visual encoder and text encoder are kept frozen. We only require a single reference image without the need for any annotation information. The input text prompt is in the form of "a *class name* in the style/appearance/shape of *s/*a/*m". For style and appearance-driven T2I customization, the decoder component will be fine-tuned. To enhance the model's robustness, random cropping and horizontal flipping augmentations are applied. For shape attribute customization, the encoder module will be fine-tuned, and only resize augmentation will be applied. Our model is trained on a single NVIDIA A40 GPU with a batch size of 1 and a learning rate set to 1e-6. The fine-tuning steps and time for each reference image may vary slightly. On average, about 1000 iterations are required for attribute-focused customized generation. This process typically takes around 10 minutes to complete, compared to more than 20 minutes for Prospect [35] and P+ [28].

4. Experiments

4.1. Attribute Benchmark

Attribute Benchmark. There exists a shortage of dedicated datasets for the evaluation of attribute-driven T2I customization generation. Therefore, we collect and introduce a novel benchmark known as the "Attribute Benchmark". This benchmark consists of three sub-datasets: a style dataset with 13 images, a shape dataset with 13 images, and an appearance dataset with 12 images.

Evaluation Metrics. We employ CLIP-T score and user study ratings for evaluating style-driven T2I customization, while for appearance-driven T2I customization, we utilize CLIP-T score, Gram matrix distances, and DINO similarity score. For shape-driven T2I customization, we use CLIP-T and IoU scores. The IoU score quantifies the shape consistency between binary masks extracted from generated images and reference images. The CLIP-T score measures the similarity between the generated images and textual prompts. The Gram matrix distance can assess appearance similarity, as explained in [1], while the DINO similarity score evaluates the consistency of appearance attributes between generated and reference images using DINO CLS features [10].

Comparison Methods. For style-driven T2I customization generation, we compare our method against several state-of-the-art approaches, including ProSpect [35], StyleDrop [24], and DEADiff [17]. For appearance and shape-driven T2I customization, we evaluate our method in comparison to Dreambooth [19], Custom Diffusion [13], ProSpect [35], and P+ [28].

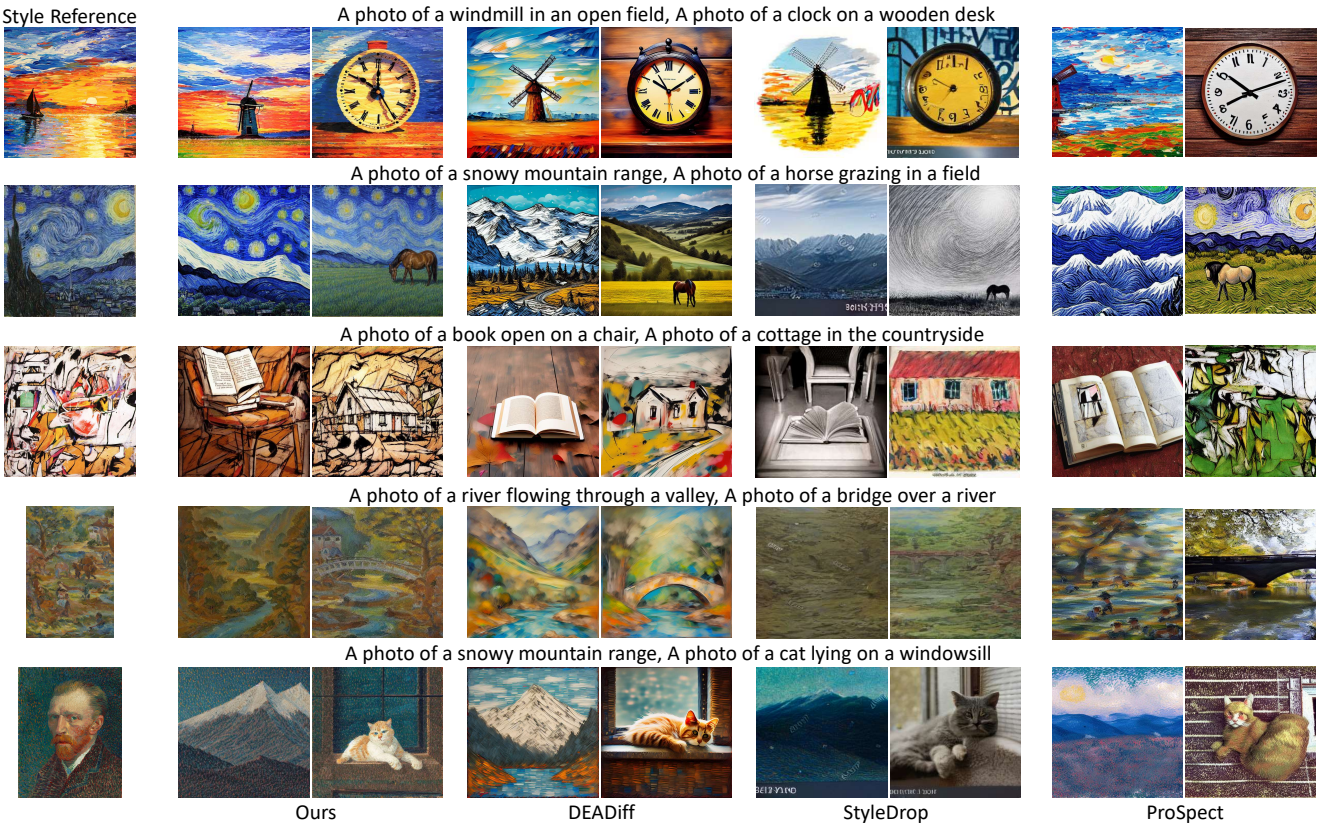

Figure 5: Qualitative comparison between our method and other approaches in terms of style-driven T2I customization.

4.2. Quantitative Experiments

Comparison on Style Customization. For each style reference image, we generated 20 text prompts covering categories such as animals, plants, objects, and scenes. Style-driven T2I customization generation was then performed based on these prompts. Each method utilized 13 style reference images, generating 20 images per style, resulting in a total of 260 images. To evaluate the text alignment between the generated images and the prompts, we computed the CLIP image-text similarity score (CLIP-T). Our method achieved the highest CLIP-T score (0.2800), demonstrating superior text consistency. Additionally, to assess style alignment, we conducted a user study, with detailed results provided in the "User Study" section.

Comparison on Shape and Appearance Customization. We individually trained DreamBooth[19], Custom Diffusion[13], Prospect[35], P+[28] and our method on the Attribute Benchmark dataset. For Dreambooth[19] and Custom Diffusion[13], we have adapted the text prompts required during training to "a *class name* in the shape/appearance of *." for attribute learning. For Prospect[35] and P+[28], we have adopted the same training and testing techniques as outlined in the original paper. For each reference image, we utilized approximately three distinct textual prompts, resulting in three generated images per text prompt. In total, each method underwent testing and produced 216 images, with 93 images dedicated to shape attribute and 123 images for appearance attribute. As shown in Table 2, Dreambooth [19] and Custom Diffusion [13] fall short in achieving excellent attribute-driven T2I customization. Furthermore, the reliance on multiple reference images for DreamBooth[19] and Custom Diffusion[13] poses a catastrophic overfitting when fine-tuning on a single reference image. In comparison with Prospect[35] and P+ [28], our method achieves more signif-

Table 1: Quantitative comparison with respect to the SOTA methods on style-driven T2I customization task.

Metric	Style-Driven T2I Customization	
Method	CLIP-T ↑	User Study ↑
ProSpect [35]	0.2259	0.1452
StyleDrop [24]	0.2668	0.0419
DEADiff [17]	0.2679	0.2968
Ours	**0.2800**	**0.5161**

Table 2: Quantitative comparison with respect to other methods for shape and appearance-driven T2I customization.

Metric	Appearance-Driven T2I Customization				Shape-Driven T2I Customization		
Method	CLIP-T ↑	DINO Similarity ↑	Gram Matrics ↓	User Study ↑	CLIP-T ↑	IoU Score ↑	User Study ↑
DreamBooth [19]	0.2681	0.2671	0.0864	0.0305	0.2791	0.3524	0.0307
Custom Diffusion [13]	0.2710	0.2977	0.0817	0.0111	**0.2844**	0.3691	0.0307
Prospect [35]	0.2800	0.3849	0.0842	0.1224	0.2795	0.4656	0.2665
P+ [28]	0.2791	0.3654	0.0841	0.2194	0.2831	0.4566	0.2850
Ours	**0.2822**	**0.4149**	**0.0791**	**0.6166**	0.2798	**0.4938**	**0.3871**

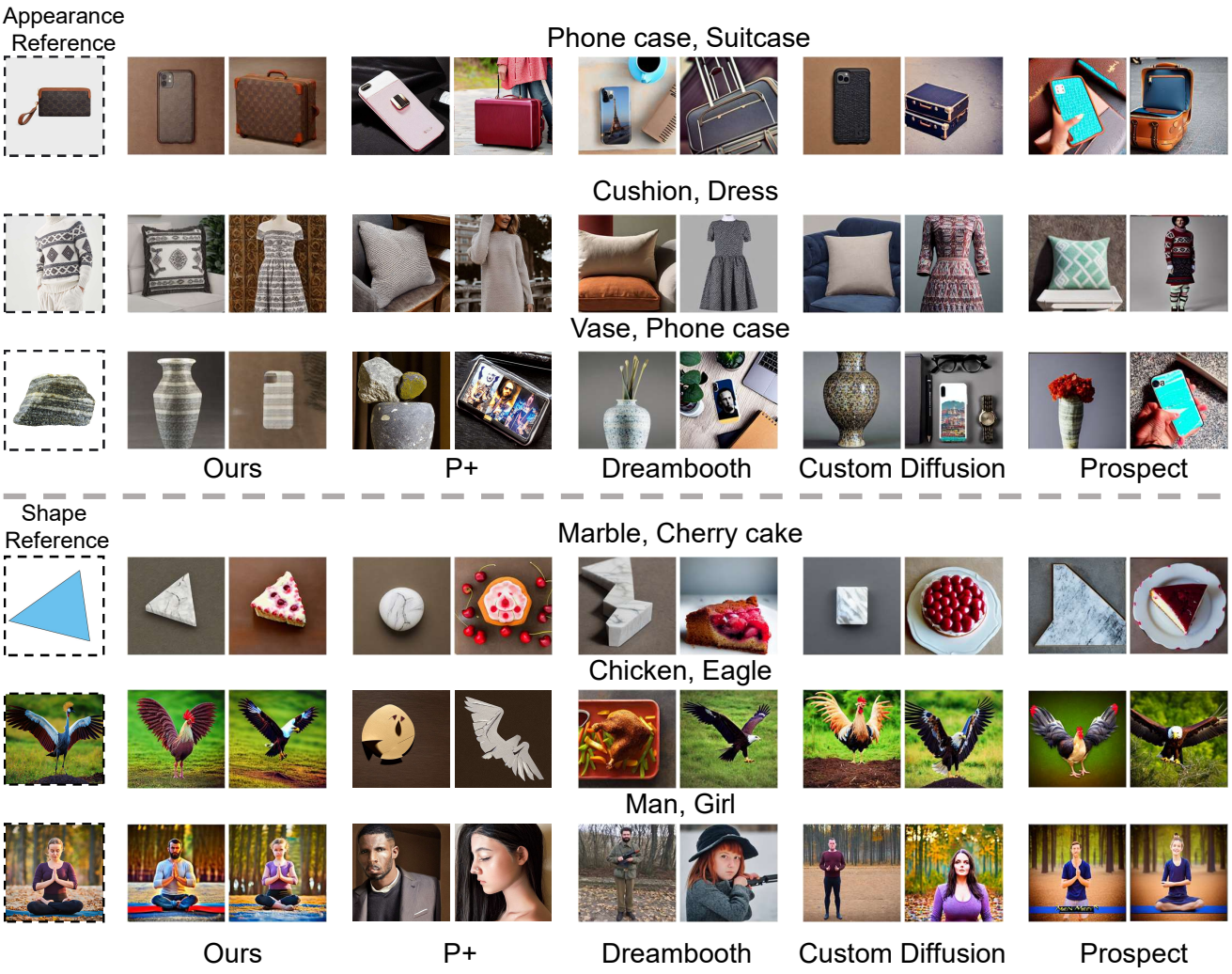

Figure 6: Qualitative comparisons with respect to existing methods for appearance and shape-driven T2I customization. Above dashed line: appearance customization. Below dashed line: shape customization.

icant improvements in DINO similarity (0.4149), Gram matrix distance (0.0791), and IoU score (0.4938), highlighting the superior attribute-driven T2I customization capabilities of our method.

User Study. We conducted 30 user studies for each task: style-driven, appearance-driven, and shape-driven T2I customization. Participants were shown a series of reference images alongside generated images from various methods, each accompanied by the corresponding text descriptions. For style customization, we randomly selected 10 style reference images and generated one image for each using a randomly chosen text prompt. Users were instructed to select the image that best matched the visual style of the reference while remaining consistent with the text. In ap-

Figure 7: The diverse generation results of our method. Top: style-driven T2I customization. Middle: appearance-driven T2I customization. Bottom: shape-driven T2I customization.

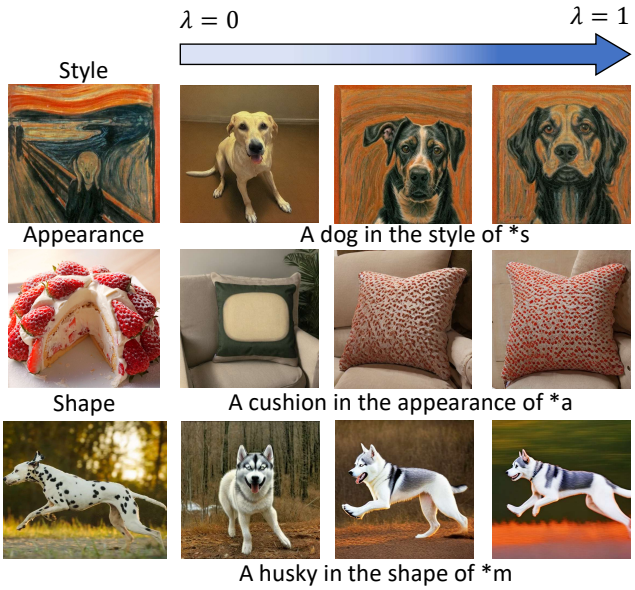

Figure 8: The generation results of our method under different λ settings.

pearance customization, participants evaluated all reference images and identified the image that best matched the reference in appearance, ensuring consistency with the accompanying text. For shape customization, we again used all reference images. Users selected the image that most closely aligned with the reference in terms of shape, while also matching the textual description and displaying realistic content. As shown in Table 1 and Table 2, our method achieved the highest user study scores in style, appearance, and shape-driven T2I customization compared to other methods. This demonstrates that our approach outperforms others in terms of human preference for attribute-

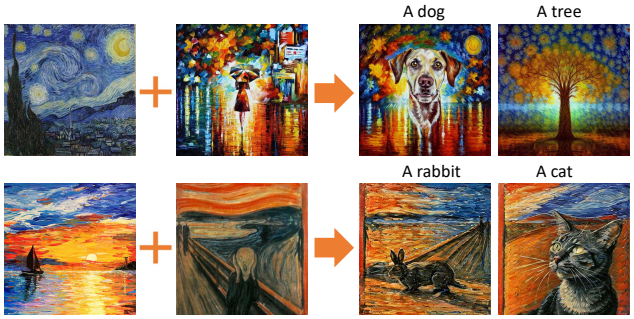

Figure 9: Our method supports style mixing of two different styles for customized T2I generation.

driven T2I customization.

4.3. Qualitative Experiments

Comparison on Style Customization. We present qualitative results of style-driven T2I customization in Figure 5. ProSpect struggles with insufficient decoupling of style attributes due to multiple attributes sharing a sampling stage, leading to incomplete separation of style features (see the results in the fourth row of Figure 5). DEADiff generates results with overly smooth styles, limited to color transfer, while failing to capture other important stylistic elements such as brushstrokes and lines. This suggests that extracting style features via Q-Former is insufficient, resulting in suboptimal style transfer outcomes. Although StyleDrop demonstrates a solid understanding of text semantics, it performs poorly in preserving style. In contrast, our method not only effectively aligns with the semantic content of the text prompts but also retains key stylistic information, highlighting its advantages in style-driven T2I customization generation.

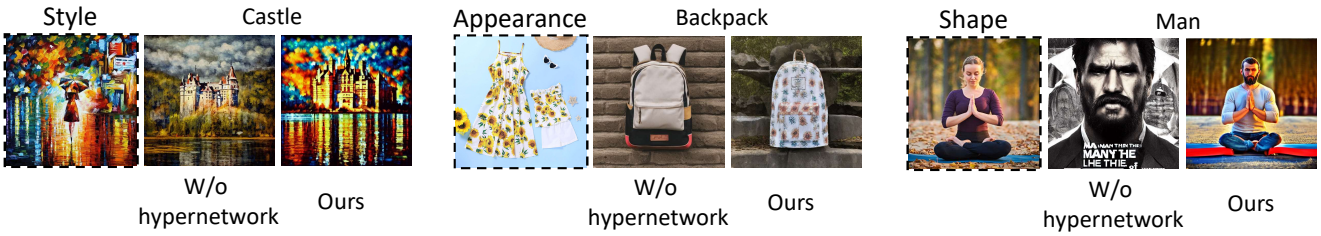

Figure 10: The ablation results of hypernetwork-driven fine-tuning strategy.

Comparison on Shape and Appearance Customization. As shown in Figure 6, we find that Dreambooth[19] and Custom Diffusion[13] exhibit less precise attribute recognition. P+ [28] struggles to recognize and learn shape and appearance attributes, only capable of generating content aligned with textual descriptions. Prospect[35] achieves some degree of attribute-aware customization, but is limited to generating categories similar to the reference image, such as woman-to-man or bird-to-chicken. In contrast, our method can accurately identify target attributes while also generating high-quality, cross-domain, text-controlled attribute-driven customization results.

Diversity. We present a range of diverse generation results to validate the diversity generation capability of our method. As shown in Figure 7, our method achieves various results for attribute-driven T2I customization, demonstrating its exceptional diversity generation capability.

Adjustable Attribute Customization Intensity. Our method provides flexible control over attribute customization intensity via the weight coefficient λ. In Figure 8, we illustrate the impact of different λ values on customization outcomes. Increasing λ results in closer alignment of the attributes between the generated and the target images. This capability enables users to finely adjust customization levels by selecting suitable weights, a feature not found in alternative approaches.

Style Mixing As shown in the Figure 9, our method supports attribute mixing, such as creating a new style by blending two existing styles. This capability not only fosters more creative applications but also demonstrates the broad adaptability and potential of our method in practical use.

4.4. Ablation Study

We performed ablation studies to validate the effectiveness of the hypernetwork-enhanced attribute-aware fine-tuning method, as illustrated in Figure 10. Without this method, the network struggled to capture the targeted image attributes, often generating content that was irrelevant to the intended attributes. However, with the hypernetwork enabled, the network successfully learned and reproduced the target attributes, producing images such as a castle in the style of *rain_princess*, a backpack that resembles a dress, and a seated man. These outcomes demonstrate the efficacy of the hypernetwork-driven attribute-aware fine-tuning approach.

5. Conclusions, Limitations and Future Work

We introduce SingleDream as a novel approach for attribute-driven T2I customization based on a single reference image. Unlike existing subject-driven customization methods, we propose an attribute-aware fine-tuning approach that focuses on adjusting the parameters of network layers related to attributes. Furthermore, we introduce a hypernetwork to predict and modulate model parameters rather than directly fine-tuning them, which not only further enhances the attribute-aware fine-tuning method but also significantly reduces the risk of overfitting when fine-tuning on a single image. Through comprehensive evaluation, our method outperforms existing solutions in the domain of attribute-driven T2I customization. While our method achieves efficiency by requiring only a single reference image for fine-tuning, the tuning time still exceeds 10 minutes, indicating room for improvement. Future work will aim to accelerate the fine-tuning process and extend our technique to video content, enabling more dynamic and detailed attribute customization.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62202199, 62406134), the Suzhou Key Technologies Project (Grant No. SYG2024136) and the Fundamental Research Funds for the Central Universities.

References

[1] Y. Alaluf, D. Garibi, O. Patashnik, H. Averbuch-Elor, and D. Cohen-Or. Cross-image attention for zero-shot appearance transfer. *arXiv preprint arXiv:2311.03335*, 2023. 5

[2] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, and D. Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 1

[3] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 3

[4] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 2

[5] W. Chen, H. Hu, Y. Li, N. Rui, X. Jia, M.-W. Chang, and W. W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 1

[6] W. Chen, H. Hu, Y. Li, N. Ruiz, X. Jia, M.-W. Chang, and W. W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[7] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[8] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, G. Chechik, and D. Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 1

[9] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, G. Chechik, and D. Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 1, 4, 5

[10] V. Goel, E. Peruzzo, Y. Jiang, D. Xu, N. Sebe, T. Darrell, Z. Wang, and H. Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 2, 3, 5

[11] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 1

[12] X. Jia, Y. Zhao, K. C. Chan, Y. Li, H. Zhang, B. Gong, T. Hou, H. Wang, and Y.-C. Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 1

[13] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 3, 5, 6, 7, 9

[14] Y. Ma, H. Yang, W. Wang, J. Fu, and J. Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 3

[15] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[16] T. Nguyen, Y. Li, U. Ojha, and Y. J. Lee. Visual instruction inversion: Image editing via visual prompting. *arXiv preprint arXiv:2307.14331*, 2023. 2

[17] T. Qi, S. Fang, Y. Wu, H. Xie, J. Liu, L. Chen, Q. He, and Y. Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. *arXiv preprint arXiv:2403.06951*, 2024. 5, 6

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4, 5

[19] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 2, 3, 5, 6, 7, 9

[20] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 1

[21] S. Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023. 3

[22] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[23] J. Shi, W. Xiong, Z. Lin, and H. J. Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 1, 3

[24] K. Sohn, N. Ruiz, K. Lee, D. C. Chin, I. Blok, H. Chang, J. Barber, L. Jiang, G. Entis, Y. Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 1, 5, 6

[25] Y. Sun, Y. Yang, H. Peng, Y. Shen, Y. Yang, H. Hu, L. Qiu, and H. Koike. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *arXiv preprint arXiv:2308.00906*, 2023. 2

[26] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 2

[27] D. Valevski, D. Lumen, Y. Matias, and Y. Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3

[28] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 1, 5, 6, 7, 9

[29] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 1, 3, 5

[30] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 1, 3

[31] X. Xu, J. Guo, Z. Wang, G. Huang, I. Essa, and H. Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023. 2

[32] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2, 3

[33] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[34] G. Yuan, X. Cun, Y. Zhang, M. Li, C. Qi, X. Wang, Y. Shan, and H. Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023. 3

[35] Y. Zhang, W. Dong, F. Tang, N. Huang, H. Huang, C. Ma, T.-Y. Lee, O. Deussen, and C. Xu. Prospect: Expanded conditioning for the personalization of attribute-aware image generation. *arXiv preprint arXiv:2305.16225*, 2023. 1, 5, 6, 7, 9