ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models

YUXIN ZHANG and WEIMING DONG, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China
FAN TANG, Institute of Computing Technology, CAS, China
NISHA HUANG, School of Artificial Intelligence, UCAS, China and MAIS, Institute of Automation, CAS, China
HAIBIN HUANG and CHONGYANG MA, Kuaishou Technology, China
TONG-YEE LEE, National Cheng-Kung University, Taiwan
OLIVER DEUSSEN, University of Konstanz, Germany
CHANGSHENG XU, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China

Fig. 1. Attribute-aware image generation results using *ProSpect*. Given a single input image or text prompts, our method can intuitively control visual attributes such as material, style, content, and layout to generate a new image with the learned textual conditionings. Real image credits (from left to right): {Vojtech Okenka, Taisuke usui, Pixabay}/Pexels (Free to use) [Pexels 2023], Paul Cezanne/The Art Institute of Chicago (CC0) [Art Institute of Chicago 2023], Georges Seurat/The Barnes Foundation (CC0) [The Barnes Foundation 2023], {Rov Camato, Chevanon Photography}/Pexels (Free to use) [Pexels 2023].

Corresponding author: Weiming Dong.
Authors' addresses: Yuxin Zhang; Weiming Dong, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China, zhangyuxin2020@ia.ac.cn, weiming.dong@ia.ac.cn; Fan Tang, Institute of Computing Technology, CAS, China, tangfan@ict.ac.cn; Nisha Huang, School of Artificial Intelligence, UCAS, China and MAIS, Institute of Automation, CAS, China, huangnisha2021@ia.ac.cn; Haibin Huang; Chongyang Ma, Kuaishou Technology, China, huanghaibin03@kuaishou.com, chongyangma@kuaishou.com; Tong-Yee Lee, National Cheng-Kung University, Taiwan, tonylee@ncku.edu.tw; Oliver Deussen, University of Konstanz, Germany, oliver.deussen@uni-konstanz.de; Changsheng Xu, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China, csxu@nlpr.ia.ac.cn.

Personalizing generative models offers a way to guide image generation with user-provided references. Current personalization methods can invert an object or concept into the textual conditioning space and compose new natural sentences for text-to-image diffusion models. However, representing and editing specific visual attributes such as material, style, and layout remains a challenge, leading to a lack of disentanglement and editability. To address this problem, we propose a novel approach that leverages the step-by-step generation process of diffusion models, which generate images from low to high frequency information, providing a new perspective on representing, generating, and editing images. We develop the Prompt Spectrum Space $\mathcal{P}^*$, an expanded textual conditioning space, and a new image

0730-0301/2023/12-ART246
https://doi.org/10.1145/3618342

representation method called *ProSpect*. *ProSpect* represents an image as a collection of inverted textual token embeddings encoded from per-stage prompts, where each prompt corresponds to a specific generation stage (i.e., a group of consecutive steps) of the diffusion model. Experimental results demonstrate that $\mathcal{P}^*$ and *ProSpect* offer better disentanglement and controllability compared to existing methods. We apply *ProSpect* in various personalized attribute-aware image generation applications, such as image-guided or text-driven manipulations of materials, style, and layout, achieving previously unattainable results from a single image input without fine-tuning the diffusion models. Our source code is available at https://github.com/zyxElsa/ProSpect.

CCS Concepts: • **Computing methodologies** → **Image processing**.

Additional Key Words and Phrases: Image generation; Diffusion models; Attribute-aware editing; Model personalization.

ACM Reference Format:
Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Trans. Graph.* 42, 6, Article 246 (December 2023), 14 pages. https://doi.org/10.1145/3618342

1 INTRODUCTION

If we consider photography and painting as visual languages, we can understand that each image encapsulates a unique perspective or way of seeing. By harnessing the power of pre-trained diffusion models designed for text-to-image generation, we obtain a versatile method for influencing the synthesis process using natural language commands. The utilization of these advanced generative models not only allows for the creation of realistic and diverse images but also enables users to personalize the output according to their visual preferences. Recent personalization methods [Gal et al. 2023a; Huang et al. 2023c; Kumari et al. 2023a; Ruiz et al. 2023] learn the textual conditioning of a common concept from a set of images and then use text prompts to create new scenarios that incorporate the concept. However, representing specific visual attributes of a single image remains a challenging problem for these concept-level personalization methods.

We believe that each visual attribute (e.g., style, material, layout, etc.) within an image has its own unique features. Attribute-aware image generation, therefore, involves the representation, disentanglement, and recombination of these visual attributes to guide image synthesis and editing. The primary challenge lies in disentangling the specific attributes of a single image, as they often appear in combination. Additionally, recombining the attributes without causing conflicts or distortions is difficult when performing image attribute transfer tasks. By projecting image references into a conditioned textual space (defined as $\mathcal{P}$ in Gal [2023a], see Fig. 2(a)), text-to-image generation methods can conduct concept-level image editing. However, generating single textual embedding across all diffusion steps and U-Net structures limits the ability for visual attribute disentanglement. In line with Gal et al. [2023a], Voynov et al. [2023] 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. In this work, we conduct a detailed analysis of how textual conditioning influences the generation process of diffusion models. We present various visualization results to demonstrate that diffusion models generate images in the order of *layout* → *content* → *material/style*. Our further analysis reveals that the generation order in a diffusion model is correlated to the signal frequency of the corresponding attribute, which is progressed from low to high. This insight paves the way for obtaining better disentanglement of visual attributes in diffusion models.

Inspired by this observation, we introduce Prompt Spectrum Space $\mathcal{P}^*$ (see Fig. 2(c)), an expanded conditioning space of $\mathcal{P}$ that provides a new insight on the diffusion generation process from the perspective of *steps*. Instead of treating all diffusion steps as a whole, we consider several groups of consecutive steps as different generation *stages*. Each stage corresponds to a unique textual condition p_i. We further propose a novel inversion and condition method *ProSpect*, which learns token embeddings P in $\mathcal{P}^*$ from a single image. Unlike previous methods that consider the concept or image as a whole, *ProSpect* provides a new way to represent an image in the perspective of frequency, which improves flexibility and editability. Various visual attributes can be separated from P, enabling attribute-aware generation. Specifically, we group the textual token embeddings p_i into three classes, *i.e.*, material/style (high-frequency), content (medium-frequency), and layout (low-frequency). By replacing them with embeddings of other images, we can achieve attribute transfer, as shown in the 2$^{\text{nd}}$ row of Fig. 1. Compared to previous personalization approaches, *ProSpect* offers better transferability of diverse image visual attributes. Notably, in the context of attribute-aware image-to-text generation tasks, *ProSpect* demonstrates superior editability and fidelity, achieving results that were previously difficult to obtain, as shown in the 3$^{\text{rd}}$ row of Fig. 1. Figs. 2(b) and 2(d) show the differences between different personalization methods applying to material controlling tasks, including Textual Inversion [Gal et al. 2023a], DreamBooth [Ruiz et al. 2023], and our *ProSpect*. Textual Inversion loses most of the fidelity. Due to the lack of separation of content and material, DreamBooth tends to generate cat-like objects in each image. *ProSpect* separates content and material in the learning and conditioning process and can generate a new image that is only loosely related to the content of the reference image. Extensive experiments and evaluations demonstrate the effectiveness of $\mathcal{P}^*$ and *ProSpect*.

To summarize, our contributions are:

- We introduce a novel Prompt Spectrum Space $\mathcal{P}^*$ that enables the disentanglement of visual attributes from a single image. We also reveal that the generation process of diffusion models depends on the frequency of visual signals.
- We present Prompt Spectrum (*ProSpect*), a novel image representation and manipulation method that offers better controllability and flexibility when processing visual attributes.
- Our experimental results demonstrate the effectiveness of $\mathcal{P}^*$ and *ProSpect* in various attribute-aware image generation tasks.

2 RELATED WORK

Text-to-image synthesis. Generative Adversarial Network (GAN)-based architectures [Goodfellow et al. 2014] are widely used in text-to-image models, which are trained on large sets of paired image-caption data [Liao et al. 2022; Tao et al. 2022; Xu et al. 2018;

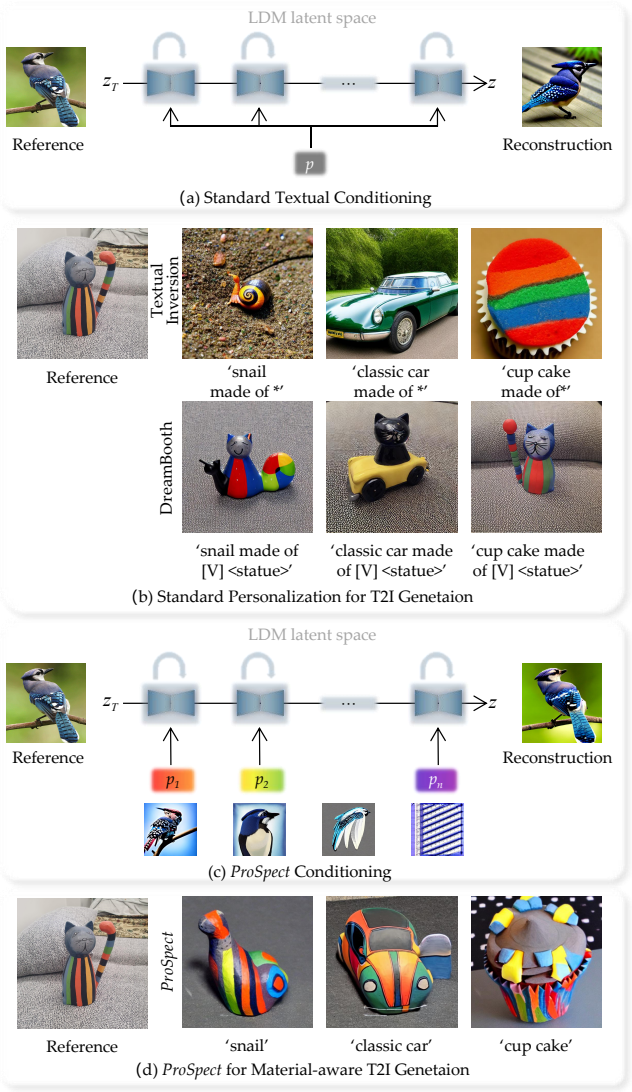


Fig. 2. Differences between (a) standard textual conditioning in $\mathcal{P}$ and (c) prompt spectrum conditioning in $\mathcal{P}^*$. Instead of learning global textual conditioning for the whole diffusion process, *ProSpect* obtains a set of different token embeddings delivered from different denoising stages. As shown in (b) standard personalization for T2I attribute-aware image generation, Textual Inversion [Gal et al. 2023a] loses some of the fidelity, and DreamBooth [Ruiz et al. 2023] generates cat-like objects in the images. (d) *ProSpect* for attribute-aware generation shows that *ProSpect* can separate content and material, and is more fit for attribute-aware T2I image generation. Reference image credit: Pixabay/Pexels (Free to use) [Pexels 2023].

Zhang et al. 2021; Zhu et al. 2019]. However, GANs have a tendency to suffer from mode collapse and their training at scale can be challenging [Brock et al. 2019; Heusel et al. 2017]. Auto-regressive models [Gafni et al. 2022; Ramesh et al. 2021; Yu et al. 2023] are inspired by the success of language models and perform the task of image generation by treating images as word sequences in a discrete latent space [Esser et al. 2021]. This scheme allows for text guidance during generation through conditioning on text-prefix or using text-to-image similarity models [Crowson et al. 2022; Gal et al. 2022; Kwon and Ye 2022] at test-time optimization. Recently, diffusion models [Dhariwal and Nichol 2021; Nichol and Dhariwal 2021] have emerged as the forefront of image generation. These models have led to significant advances in text-to-image synthesis, achieving more natural results with impressive diversity and fidelity [Balaji et al. 2022; Chang et al. 2023; Huang et al. 2022a; Nichol et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022].

Personalization of generative models. The personalization of the text-to-image generation model is the task of generating personalized content based on the pre-trained model. Gal et al. [2023a] present a textual inversion method to find a pseudo-word to describe the visual concept of a specific object. Gal et al. [2023b] further design a word-embedding encoder to predict a new pseudo-word that best describes the input concept. Li et al. [2023] invert the real image to the linear mapping network in cross-attention layers. Ruiz et al. [2023] implant a subject into the output domain of a text-to-image diffusion model to synthesize it in novel views with a unique identifier. Zhang et al. [2023b] propose an attention-based inversion style transfer method called InST. Kumari et al. [2023b] propose Custom Diffusion, which optimizes a few parameters in the conditioning mechanism and can jointly train for multiple concepts or combine several fine-tuned models. Huang et al. [2023c] propose ReVersion for relation inversion, which aims to learn a specific relation from images. Wen et al. [2023] introduce the concept of hard prompts that use hand-crafted sequences of interpretable tokens to elicit model behaviors. Voynov et al. [2023] present an extended textual conditioning space $\mathcal{P}+$ that consists of multiple textual conditions, derived from per-layer prompts, each corresponding to a layer of the denoising U-Net of the diffusion model. Tewel et al. [2023] introduce Perfusion, a mechanism that locks cross-attention keys of new concepts to their superordinate category, and a gated rank-1 approach to control the influence of a learned concept.

Most of the aforementioned methods necessitate an image set (three to five) as input or require model fine-tuning, and they aim to learn a single concept in the image or represent the overall appearance of the image. In contrast, our approach addresses the challenges of obtaining multiple visual attributes from a single image, involving the representation, disentanglement, and recombination of visual attributes.

Image editing. A variety of text-based image editing methods [Bau et al. 2021; Patashnik et al. 2021; Schaldenbrand et al. 2022] have emerged with the development of powerful multi-modal models. Enabled by diffusion models, approaches of different applications are developed, such as single-image editing [Brooks et al. 2023; Huang et al. 2023b; Kawar et al. 2023; Meng et al. 2021, 2022; Mokady et al. 2023; Valevski et al. 2023; Wu et al. 2023; Zhang et al. 2023a], style transfer [Huang et al. 2023d, 2022b; Jeong et al. 2023; Yang et al. 2023b] and inpainting [Avrahami et al. 2022; Lugmayr et al. 2022; Yang et al. 2023a]. The Composer approach [Huang et al. 2023a] is most relevant to our work. This approach introduces a generation paradigm that enables control over the output features,

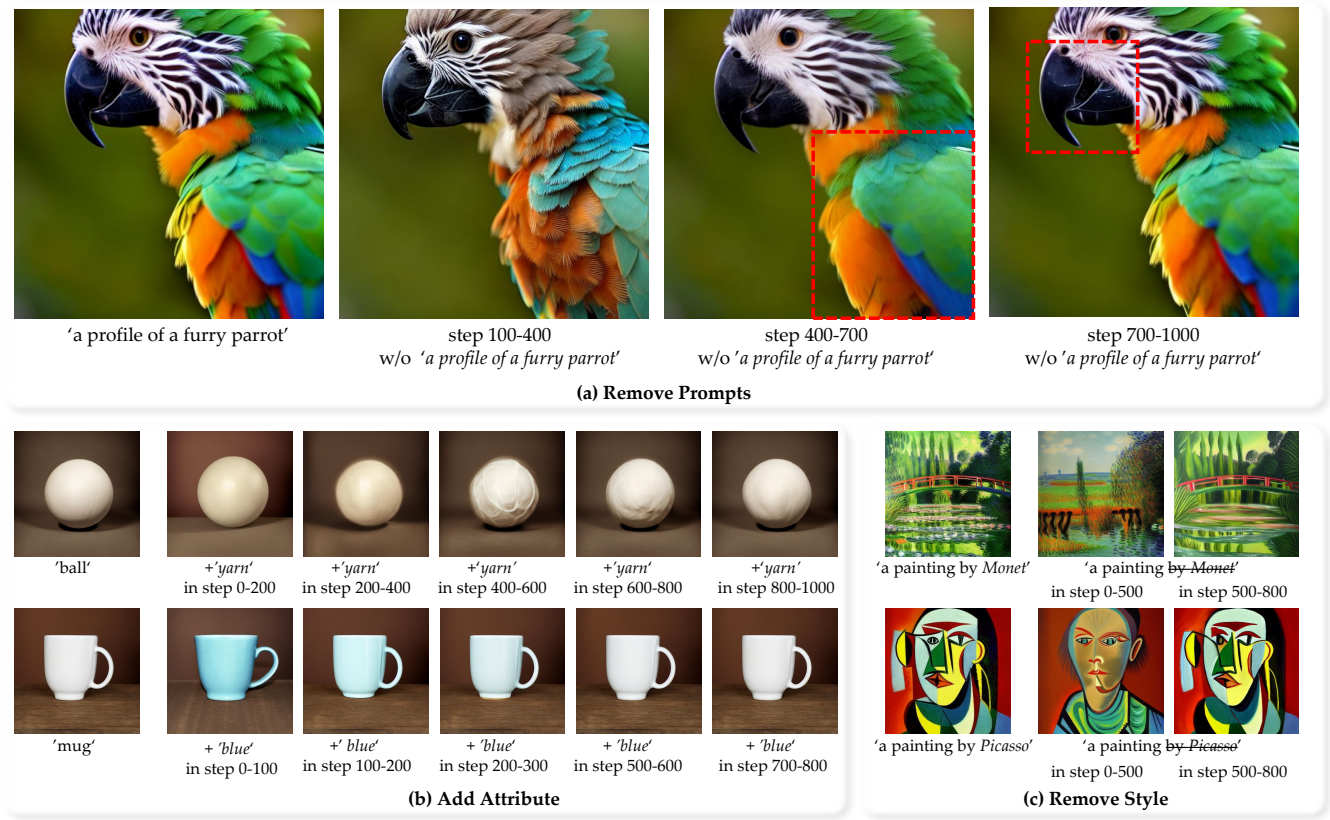


Fig. 3. Experimental results showing that different image attributes correspond to different generation steps. (a) Results of removing prompts "*a profile of a furry parrot*" of different steps. (b) Results of adding material attribute "*yarn*" and color attribute "*blue*". (c) Results of removing style attributes "*Monet*" and "*Picasso*".

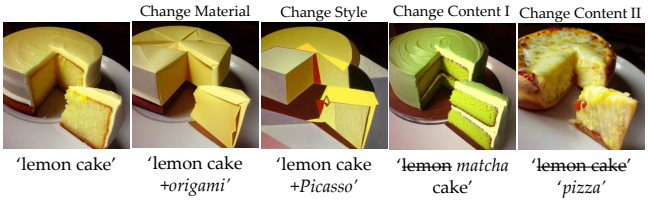


Fig. 4. Prompt-based editing results. By changing the prompts conditioning on different diffusion stages and keeping the layout-related prompts unchanged, we can achieve the effect of prompt-to-prompt editing.

while preserving synthesis quality and model creativity through decomposing images into representative factors (e.g., spatial layout and color palette) and training a diffusion model using these factors as conditions for recomposition. However, they rely on additional task-specific models to obtain image attributes, such as an edge detection model for contour extraction, a pre-trained segmentation model for extraction of instances and the corresponding masks, etc. In contrast, we exclusively use a pre-trained diffusion model to obtain the representation of corresponding attributes from the input image, which provides a neat way to disentangle and control visual attributes.

Many non-diffusion image editing methods encode images into a latent space [Lee et al. 2020; Wang et al. 2023b,a; Zhang et al. 2023c]. StyleGAN [Karras et al. 2019] consists of a mapping network, which maps latent codes to the latent space $\mathcal{W}$, and a synthesis network, which controls the feature statistics between different network layers. Fine-grained control over semantic attributes in generated images is achieved by manipulating different dimensions of the latent vectors. With the ability of generating high resolution images of high quality, StyleGAN and its followups [Gal et al. 2022; Karras et al. 2020] have become the advanced unconditional image generators. FineGAN [Singh et al. 2019] disentangles the background, object shape,and object appearance to hierarchically generate images of fine-grained object categories. MUNIT [Huang et al. 2018] decomposes the image into a domain-invariant content code and a style code that captures domain-specific properties, and achieves editing by recombining the codes. SwappingAutoencoder [Park et al. 2020] encodes an image into two independent components and enforce that any swapped combination maps to a realistic image. Differently, our approach encodes image attributes into the target text space and

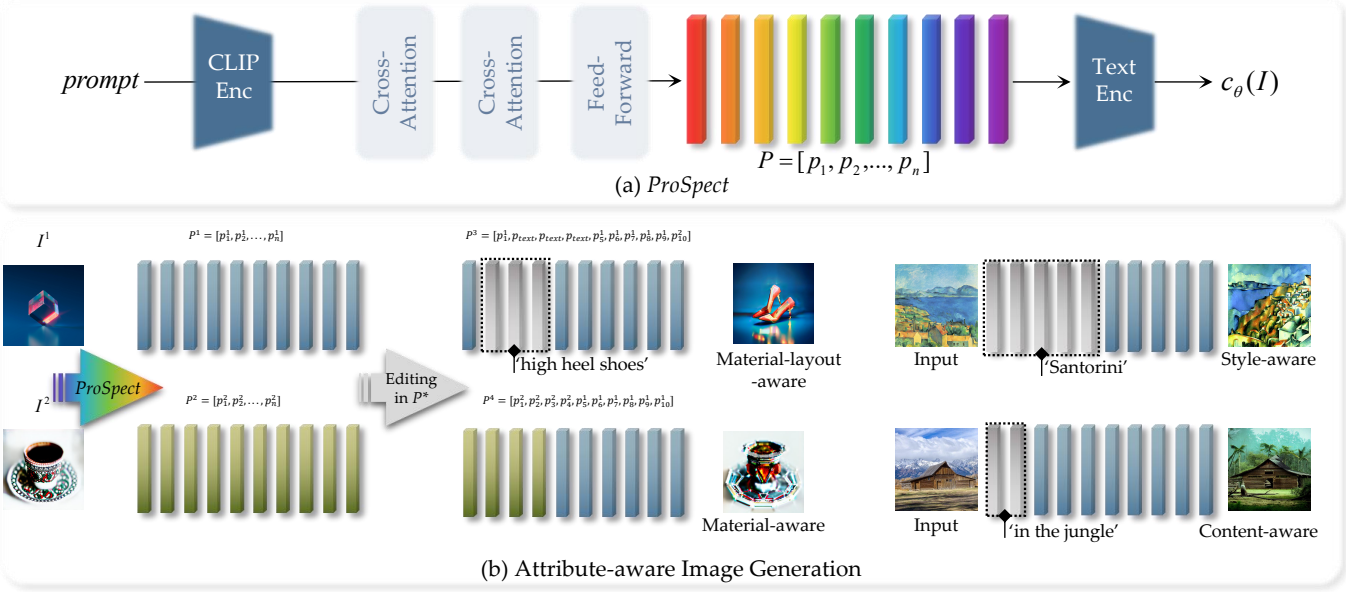


Fig. 5. (a) The pipeline of *ProSpect*, which learns a set of token embeddings $P = [p_1, p_2, ..., p_n]$. (b) Illustrations of various attribute-aware image generation tasks. Reference image credits: {Rostislav Uzunov, Lisa Fotios} /Pexels (Free to use) [Pexels 2023]. Style image credit (the 1st row): Paul Cezanne/The Art Institute of Chicago (CC0) [Art Institute of Chicago 2023].

represents attributes separately using different embeddings. Besides, the above latent space traversal is usually limited to editing within domains, in contrast, our method enables cross-domain editing.

3 METHOD

To illustrate our motivation, we start by analyzing the attribute distribution of diffusion models using text-guided image generation results. We aim to obtain multiple visual attributes from a single image, thus we need to learn the range of the steps in which different attributes are generated by the model.

Fig. 3 shows the results of removing or adding attributes at different diffusion stages. In Fig. 3(a), removing a certain phase "a profile of a furry parrot" in some steps will cause certain changes to the generated image. Removing *steps 100-400* significantly changes the parrot's appearance, but the new image retains the details and feather layering. Removing *steps 400-700* reduces the layering of the parrot's feathers. Removing *steps 700-1000* blurs the parrot's fur and the luster of the beak is gone, while it can retain a similar overall appearance to the original image. Fig. 3(b) demonstrates the effect of adding an attribute in a specific stage. In the 1st row, the sphere's appearance remains unchanged when injected the added concept "yarn" in *steps 0-200*, but the background layout and colors are different, and adding it in *steps 200-400* blurs the sphere's outline. Injecting "yarn" in *steps 400-600* and *steps 600-800* leads to a more distinct texture. Adding "yarn" in *steps 800-100* creates a woolen texture on the sphere and reduces its reflection. The 2nd row shows that the diffusion model is color-sensitive only at certain stages. Fig. 3(c) shows the style removal results of impressionist Claude Monet and abstract painter Pablo Picasso. We remove their names at different stages, i.e., using only "a painting" to guide the generation. Removing the style in *steps 500-800* has little effect on the Picasso-guided painting, but the Monet-guided painting loses its brushstrokes. Conversely, removing *steps 0-500* changes the content of the paintings guided by "Monet", but the style is maintained, while the image guided by "Picasso" loses its style. We recommend zooming in to see experimental results of Monet's style. In conclusion, the initial generation stages of the diffusion model tend to generate overall layout and color, the middle stages tend to generate structured appearances, and the final stages tend to generate detailed textures.

Based on the above observations, we can edit the results by changing the material, style, and content while keeping the layout unchanged by changing the prompts that act on different steps. As shown in Fig. 4, keeping the prompt "lemon cake" condition in the initial stages, the image can be edited into different appearances. Prompt-to-prompt [Hertz et al. 2023] report the observation of similar effects and introduce a method that locks the corresponding attention maps.

3.1 Prompt Spectrum Space

We use Stable Diffusion [Rombach et al. 2022] as the generative backbone, which is built in the framework as Latent Diffusion Model (LDM) [Rombach et al. 2022]. LDM is a diffusion probability model that generates images by gradually denoising them.

Diffusion and denoising within an LDM typically take 1000 steps, and the text conditions the model step by step. Previously, the process of the textual conditions acting on the diffusion model is regarded as a whole. In this work, we treat them as different procedures. Specifically, we divide the 1000 steps of conditioning into ten stages on average. Each stage corresponds to a unique

Fig. 6. The visualization results of token embeddings p_i obtained by *ProSpect*. The results show that the initial generation step of the diffusion model is sensitive to structural information (e.g., bird's pose, pot's shape). As the number of steps increases, the obtained p_i gradually captures detailed information (e.g., the sideways head of the bird → bird's wing → the texture of the bird's feathers).

textual condition. The collection of textual conditions reside in the CLIP [Radford et al. 2021] text-image space, their sizes are set to $n \times 1 \times 768$ ($n = 10$ denotes the number of the stages). This way of division is designed to keep a balance between efficiency and quality.

We refer to the expanded space as *Prompt Spectrum Space*, denoted as $\mathcal{P}^*$. An illustration of how $\mathcal{P}$ and $\mathcal{P}^*$ interact with text and diffusion models is shown in Figs. 2(a) and 2(b). Thus, $\mathcal{P}^*$ is defined as:

$$\mathcal{P}^* = \{p_1, p_2, ..., p_n\}, \tag{1}$$

where p_i represents the token embedding corresponding to the conditional prompt of the ith stage of the generation process.

3.2 ProSpect

We aim to extend TI [Gal et al. 2023a] to $\mathcal{P}^*$ by extracting a *set* of textual token embeddings from an input image. To achieve this goal, we present *ProSpect*, a method that maps an image to a collection of corresponding textual token embeddings. The TI loss of LDM in $\mathcal{P}$ space is formulated as:

$$\mathcal{L}_{TI} = \mathbb{E}_{z,t,p}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, p_\theta\right)\|_2^2\right], \tag{2}$$

where p_θ is a learnable vector denoting the token embedding and $z \sim E(x), \epsilon \sim \mathcal{N}(0,1)$. Similarly, the *ProSpect* loss of LDM in $\mathcal{P}^*$ space is formulated as:

$$\mathcal{L}_{PS} = \mathbb{E}_{z,t,p}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, p_i\right)\|_2^2\right], \tag{3}$$

where $p_i = P(t)$ is a learnable vector represents the token embedding of stage i, and $P = [p_1, p_2, ..., p_n]$ is the set of textual token embeddings in $\mathcal{P}^*$ space.

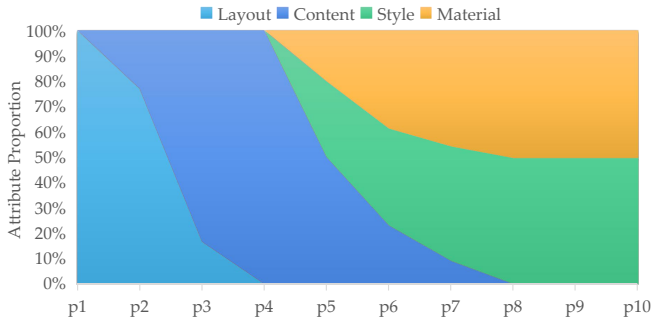


Fig. 7. Statistical results of various attribute distributions at different prompts.

As shown in Fig. 5(a), the token embedding is initialized to a frozen 1×768 text embedding with a user input text (e.g., "cup") via the CLIP text encoder. It is then fed into a randomly initialized hypernetwork and finally creates a $n \times 1 \times 768$ embedding $P = [p_1, p_2, .., p_n]$. Only the hypernetwork is trainable and the final p_i is obtained by optimizing based on Eqn. (3). The training process typically requires 1000-3000 iterations. Dropout is applied to prevent overfitting and the rate is set to 0.1.

Attribute control during inference is achieved by replacing the p_i representing different attributes with editing texts. For instance, in Fig. 5(b), content personalization involves maintaining the content-related $p_3 - p_{10}$ of image *barn* as "* in the jungle" and replacing $p_1 - p_2$ with "in the jungle" (without "*").

4 ANALYSIS OF PROMPT SPECTRUM SPACE

4.1 Visualization of Token Embeddings

We visualize the token embedding p_i obtained via *ProSpect* by using it as the condition of the entire stage of the diffusion model, i.e., $p_{1:10} = p_i$. Fig. 6 shows the corresponding visual results of p_i for four stages. It can be seen that the diffusion model acts different optimizations to token embeddings p_i at different stages to reconstruct the given image. The token embeddings that are conditioned on the initial stages are optimized to denote structure information, and then gradually represent detailed information as the generation steps increase. For instance, p_2 tends to represent the layout or content, while p_8 tends to express the textures or brushstrokes. The results indicate that different generation tendencies exist in different stages of the diffusion model.

4.2 Visualization of Attribute Distribution

To evaluate the attribute distribution, we provide 30 pairs of *attribute, object* combinations (e.g., "origami, cake"), including 10 pairs for material, style, and layout, respectively. The *object* remains unchanged while we record the impact of adding *attribute* at different p_i. Additionally, we select 10 new *objects* to replace the original *object* at different p_i and record the impact of replacement on the content. The results are shown in Fig. 7. Notably, adding attributes or replacing content at a single p_i may not significantly change the output image. To ensure a faithful evaluation, we gradually increase the intensity of the change until other attributes are affected.

Table 1. CLIP-based evaluation results. The best numbers are in **bold** and the second best results are underlined.

Metric	Text Similarity↑				Image Similarity↑			
Method	Reference	**ProSpect**	DreamBooth	TI	Reference	**ProSpect**	DreamBooth	TI
Average	0.2479	**0.3444**	0.3334	0.3115	0.9128	0.7927	**0.7987**	0.7274
Min	0.2168	**0.2869**	0.2279	0.2371	0.8771	**0.6899**	0.6450	0.4471
Max	0.2767	**0.3995**	0.3666	0.3820	0.9541	**0.929**	0.8678	0.8688
Negative Error	0.0311	0.0575	0.1055	0.0743	0.0357	0.1027	0.1537	0.2803
Positive Error	0.0288	0.0551	0.0331	0.0705	0.0412	0.1363	0.0691	0.1414

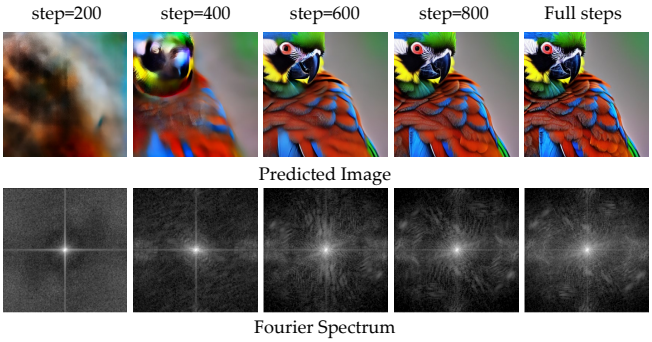


Fig. 8. Analysis of images generated at different stages in the frequency domain. The 1st row shows the predicted image obtained at different denoising steps with the text prompt "*a close-up photo of a parrot*". The 2nd row showcases the Fourier spectrum of each predicted image. As the denoising process progresses, the high-frequency information contained in the predicted image gradually increases. We enhance the contrast of the Fourier spectrum for clarity.

4.3 Explanations

The experimental results demonstrate that a diffusion model generates images in the order of *layout* → *content* → *material/style*. A similar phenomenon has been observed in convolutional networks. Voynov et al. [2023] noted that the U-Net structure of the diffusion model has similar properties, with the shallow layer tending to generate texture and color and the deep layer generating semantic information. It is important to note that the deep receptive field size of U-Net is larger than the shallow receptive field size, making the hierarchical attribute distribution easy to comprehend. However, this size difference dose *not* exist between steps of the diffusion model, since the latent size is uniform across different stages.

The Fourier transform is a classic transformation widely used in digital image processing. It transforms a signal from the time domain into the frequency domain, facilitating the identification of subtle features and the processing of challenging components.

Fig. 8 shows the Fourier spectrum of the diffusion process. As the number of steps in the denoising process increases, the high-frequency information contained in the image predicted by the diffusion model gradually increases. This indicates that the model tends to generate structural information at the beginning of the denoising process, with details gradually increasing as the steps increase. This phenomenon explains the generation order of the diffusion model, which is caused by the signal frequency of the corresponding attribute from low to high.

5 EXPERIMENTS

We demonstrate that *ProSpect* outperforms state-of-the-art text-to-image personalization baselines in both fidelity and editability by conducting both qualitative and quantitative evaluations. Moreover, we apply *ProSpect* to diverse applications of material transfer, style transfer, and layout transfer (as shown in Sec. 5.4), and perform qualitative comparisons with related methods.

Methods for comparison. We optimize **(1) Textual Inversion (TI)** [Gal et al. 2023a] with 5000 iterations and **(2) InST** [Zhang et al. 2023b] with 1000 iterations on Stable Diffusion 1.4 [Rombach et al. 2022], both as recommended by the authors. We train **(3) DreamBooth** [Ruiz et al. 2023] for 400 steps. The resulting images of **(4) Perfusion** [Tewel et al. 2023] and **(5) XTI** [Voynov et al. 2023] are borrowed from their papers. We use the official pre-trained models of **(6) InstructPix2Pix** [Brooks et al. 2023], **(7) JoJoGAN** [Chong and Forsyth 2022], **(8) CAST** [Zhang et al. 2022], and **(9) StyTr^2** [Deng et al. 2022].

Test dataset. For fair comparison, we use nine concepts from previous papers, including cat, teddy bear, cat statue, pot, sculpture, colorful teapot, red teapot, elephant, clock, and three concepts of faces. For each concept, we use three easy prompts (changing background) and three difficult prompts (changing pose/clothes/views/etc.). Each image-prompt pair is used to generate four results. In total, we obtain 288 images for each method.

Implementation details and timing statistics. In all of our experiments, we use Stable Diffusion 1.4 [Rombach et al. 2022] with the default hyperparameters and set a base learning rate of 0.001. We employ a DDIM sampler with diffusion steps $T = 50$ and guidance scale $w = 7.5$. We use a frozen CLIP model in Stable Diffusion as the text encoder network. The texts are tokenized into start-token, end-token, and 75 non-text padding tokens. The training process on each image takes approximately 20 minutes using an NVIDIA GeForce RTX3090 with a batch size of 1, significantly less than the more than 90 minutes required for TI. The synthesis process takes about three seconds, depending on the number of diffusion steps taken.

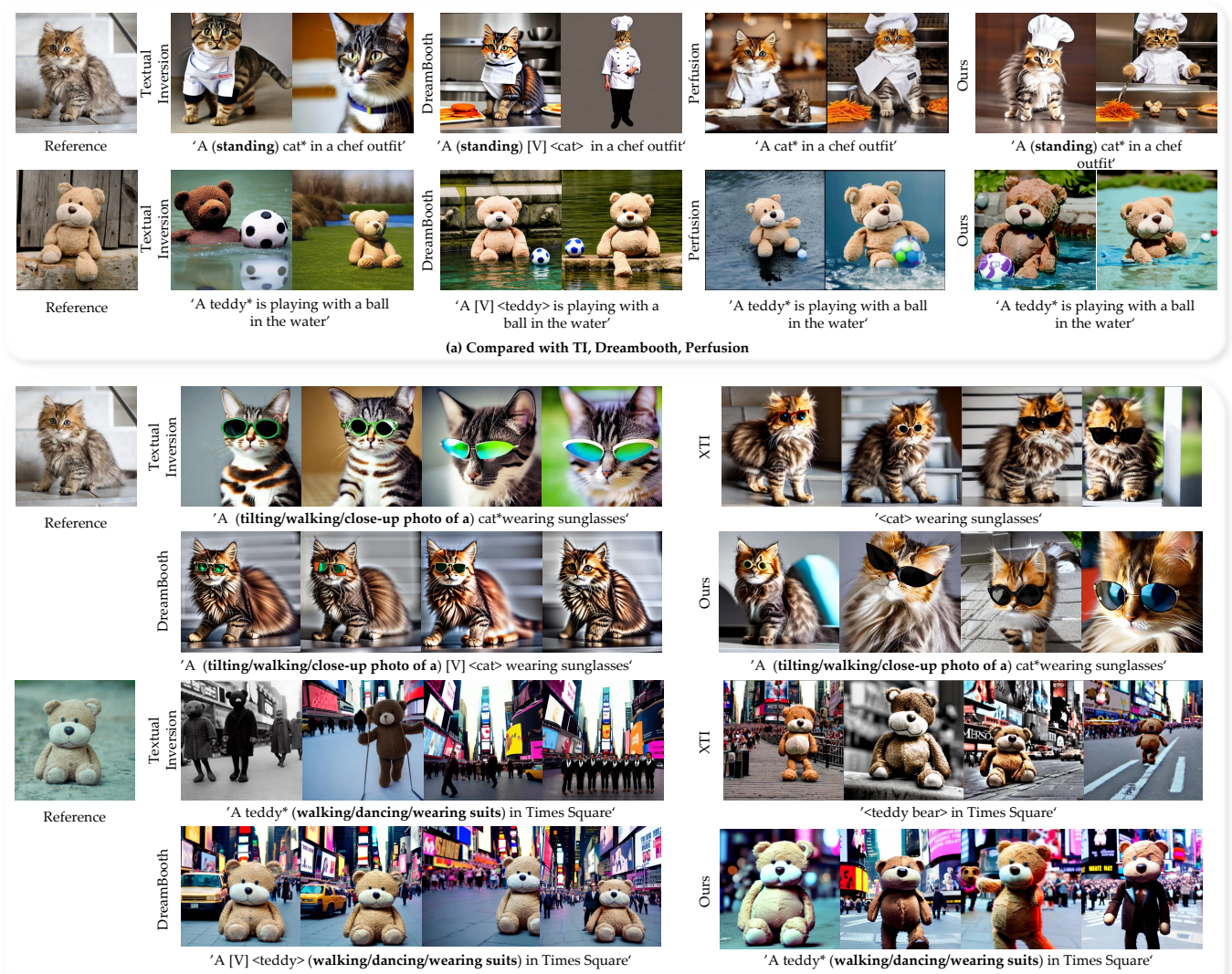


Fig. 9. Comparisons with state-of-the-art personalization methods including Textual Inversion (TI) [Gal et al. 2023a], DreamBooth [Ruiz et al. 2023], XTI [Voynov et al. 2023], and Perfusion [Tewel et al. 2023]. The **bold** words correspond to the additional concepts added to each image (e.g. the 3rd column in (a) shows the result of "*A standing cat in a chef outfit*", the 6th column in (b) shows the result of "*A tilting cat wearing sunglasses*"). XTI and Perfusion are the latest published methods and the model have not been released yet. The resulting images of XTI and Perfusion are borrowed from their paper, so the results of adding concepts are not shown. Our method can faithfully convey the appearance and material of the reference image with better controllability and diversity.

5.1 Quantitative Evaluation

We use two metrics to conduct quantitative evaluations. Specifically, we compute the pair-wise CLIP cosine similarity between the reference images and the generated images as *image similarity* to evaluate content fidelity. In addition, we use the CLIP similarity between all generated images and their textual conditions as *text similarity* to evaluate the editability.

Table 1 shows the corresponding quantitative evaluation results of our method and two baseline methods. The Reference column of text similarity calculates the cosine similarity between the reference image and the various text condition, which can be regarded as the lower bound score. The Reference column of image similarity calculates the cosine similarity between the image contains the same object and the reference image, which can be regarded as the groundtruth score. TI [Gal et al. 2023a] fails to preserve object appearance, while DreamBooth tends to overfit the reference image. Though a higher fidelity score it gets, the editability is not satisfactory. Our method achieves a better balance of object fidelity and editability without fine-tuning the model.

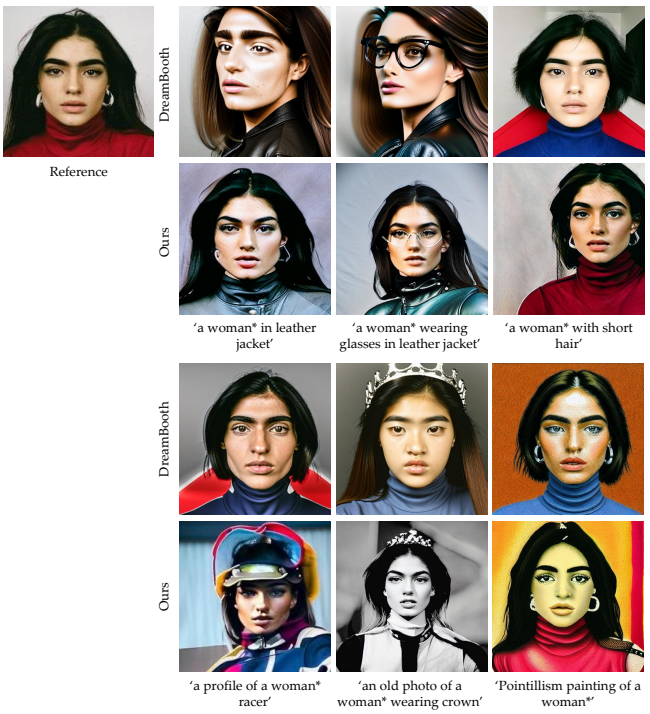


Fig. 10. Comparision with DreamBooth [Ruiz et al. 2023] on personalized one-shot portrait generation. Our inversion based method can better preserve the character identity in the input image.

5.2 Qualitative Evaluation

As shown in Fig. 9, we compare our method with four SOTA personalization methods, *i.e.*, TI [Gal et al. 2023a], DreamBooth [Ruiz et al. 2023], XTI [Voynov et al. 2023], and Perfusion [Tewel et al. 2023]. We use concepts from previous papers for fair comparison and unbiased evaluation. We add additional texts shown in bold to each set of images to demonstrate the flexibility of our method.

DreamBooth can well depict the conceptual appearance in the reference image, but tends to overfit to the reference image, resulting in a lack of editability. As shown in the results of "a (standing) cat in a chef outfit" in the second row, TI fails to maintain the object's appearance and generates normal cats. DreamBooth can generate a standing cat, but the background is blurred, and the cat's paw is confused with the human hand. Our results can generate a standing cat with a kitchen as the background and maintain the details of the cat's paws.

The results of "a (tilting/walking/close-up photo of a) cat wearing sunglasses" show that DreamBooth can generate a cat with sunglasses, but cannot change the cat's posture or zoom-in/zoom-out. Our method, shown in the third row, can generate high-fidelity concepts while maintaining diversity and flexibility. *ProSpect* not only puts sunglasses on the cat but also allows it to show its walking posture and close-up details.

In the results of "a teddy is playing with a ball in the water", Perfusion and DreamBooth can generate teddy bear, ball, and water, but they are not interacting with each other. Our method can show the posture of the teddy bear touching and throwing the ball, and the teddy bear can float on the water or half-submerge in the water.

In the results of "a teddy (walking/dancing/wearing suits) in Times Square", XTI cannot accurately maintain the appearance of the teddy bear, and DreamBooth cannot change the posture of the teddy bear. Our method can reproduce the appearance of a teddy bear while walking, dancing, and wearing a suit, always in the background of Times Square.

Our method is also capable of personalized one-shot portrait generation. Fig. 10 shows the comparison results between our method and DreamBooth [Ruiz et al. 2023]. Our method can manipulate attributes such clothing, hairstyle and artistic styles of the input portrait while preserving the identity.

5.3 User Study

We evaluate our method in attributes-aware image generation, alongside three SOTA personalization methods, *i.e.*, TI [Gal et al. 2023a], DreamBooth [Ruiz et al. 2023], and InST [Zhang et al. 2023b]. A total of 66 participants took part in the survey, including 42 researchers in computer graphics or computer vision (CGCV), 24 university students (others). The user study is divided into three parts, including personalized objects, material guidance, and style guidance.

User Study I. In the content-aware image generation survey, TI and DreamBooth are used as the baseline methods. The same 12 concepts in quantitative evaluation, each with two different prompts are used. The objective of the personalization task, which is to generate a new image with the same concept as the reference image while also matching the provided text condition, is introduced to the participants. For each question, the participants are shown a reference image and a text condition (e.g., "a photo of the same cat wearing sunglasses") and are asked to choose the option that best matches the task objective from three randomly ordered options, each corresponding to a method. *ProSpect* receives 51.97% (CGCV 52.14%, Others 51.67%) of the preferences, while TI acquires 10.30% (CGCV 9.76%, Others 11.25%), and DreamBooth obtains 37.72% (CGCV 38.09%, Others 37.08%). Thus, *ProSpect* exhibits better performance in human preference when compared to the two baseline methods.

User Study II. In the material-aware image generation survey, DreamBooth is used as the baseline method, and the participants are introduced that the objective of the task is to generate a new image composed of materials from the reference image while matching the provided text conditions. Eight material references with three results each are used. For each question, the participants are shown reference images and corresponding text conditions (e.g., "a snail made of the material in this image") and are asked to select one of two options that best matches the task objective. *ProSpect* receives 66.36%'s preference (CGCV 68.57%, Others 62.50%) and DreamBooth obtains 33.64% (CGCV 31.42%, Others 37.50%).

User Study III. The SOTA style transfer method InST [Zhang et al. 2023b] is the baseline method in the style-aware image generation survey. Eight style references are used for this study, each with one style transfer result and one T2I result. We evaluate both the

Fig. 11. Material-aware image generation results. We compare *ProSpect* with two state-of-the-art methods for this task, *i.e.*, a personalized image generation approach DreamBooth [Ruiz et al. 2023] and an image editing approach InstructPix2Pix [Brooks et al. 2023]. Our method shows better fidelity and editability than those two alternative baselines.

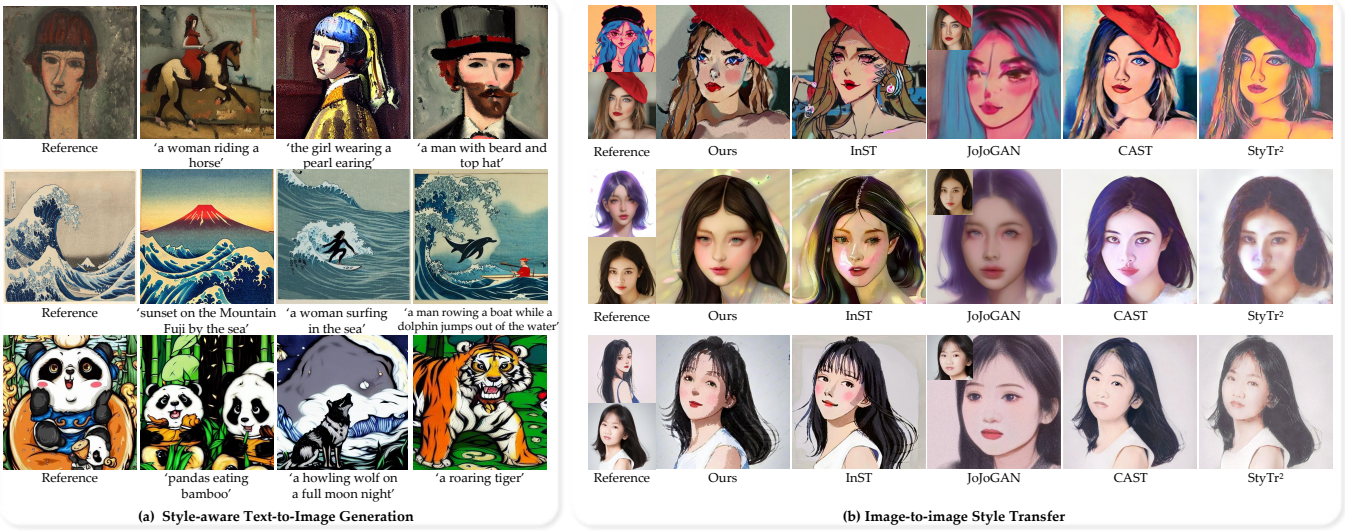
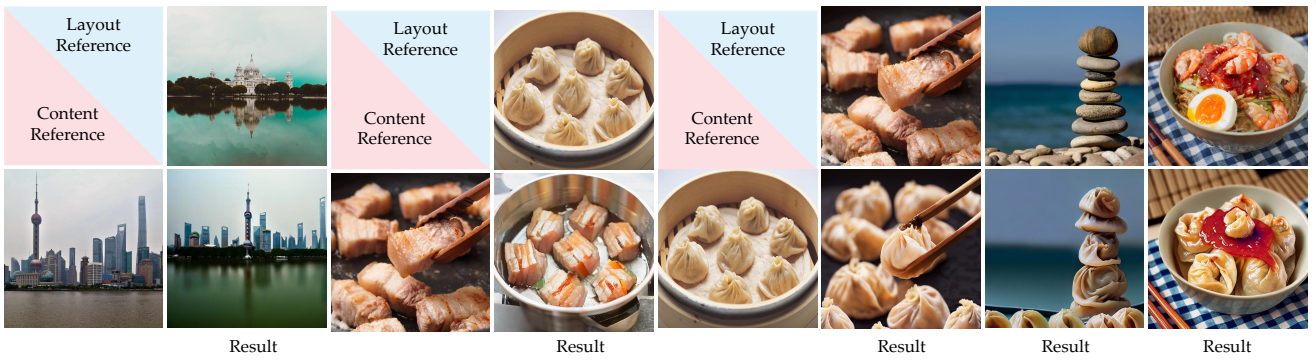


Fig. 12. Style-aware image generation results. We compare *ProSpect* with four state-of-the-art style transfer methods, including InST [Zhang et al. 2023b], JoJoGAN [Chong and Forsyth 2022], CAST [Zhang et al. 2022], and StyTr^2 [Deng et al. 2022]. Our method better preserves the identity information of the content image than the diffusion-based method InST while generating better brush strokes than other GAN-based and encoder-based methods. Style image credits (the 1st and 2nd rows in (a)): {Amedeo Modigliani, Katsushika Hokusai}/The Art Institute of Chicago (CC0) [Art Institute of Chicago 2023].

style-guided text-to-image generation task and the style transfer task. The participants are introduced that the objective of the task is to generate a new image consistent with the style of the reference artistic image while also being consistent with the content of the provided textual condition/content image. For each question, the participants are presented with either a style image and a corresponding text condition (e.g., "a painting of Einstein drawn in the style of the reference image") or a pair of style and content images, and are asked to select one of two options that best matches the task objective. *ProSpect* outperforms InST by receiving 61.67% (CGCV

(a) Layout-aware Text-to-Image Generation



(b) Image-to-image Layout Transfer

Fig. 13. Layout-aware image generation results. *ProSpect* can generate an image with the same layout of an layout reference image by using a text prompt or a content reference image.

61.19%, Others 62.50%) the preference of compared with InST's 38.33% (CGCV 38.80%, Others 37.50%).

5.4 Applications

In this section, we demonstrate the effectiveness of our approach in various attribute-aware image generation tasks, including material-aware image generation, style-aware image generation, as well as layout-aware image generation.

Material-aware image generation. Our approach is well-suited for material-aware image generation tasks, including material transfer between images, image material-guided text-to-image generation, and image material editing with text. Results shown in Fig. 11 demonstrate the high visual quality and flexibility of our method. Fig. 11(a) shows the results of material transfer, where our method can transfer materials between semantically unrelated objects (e.g., gears and teacups, apples, and dandelions). Fig. 11(b) shows the material-guided text-to-image generation using a reference image, which we compare with a state-of-the-art personalization method DreamBooth [Ruiz et al. 2023]. DreamBooth requires both prompt learning and model fine-tuning, making it prone to overfitting on specific images and lacking flexibility with single-image input. Our method, however, can guide image generation using references with unrelated materials (e.g., rings and snails, teapot, and beetle), demonstrating superior editability. Fig. 11(c) shows the results of modifying an image's material with natural language. We compare our method with a state-of-the-art image editing method InstructPix2Pix [Brooks et al. 2023], which works on semantically related images (e.g., hummingbird to peacock feather) but fails on semantically unrelated modifications (e.g., teddy to origami). Unlike InstructPix2Pix, our method can edit images into completely unrelated materials while retaining their overall appearance and background.

Style-aware image generation. Our method is also effective for generating artistic images. The material in a realistic image reflects high-frequency information, while strokes and shapes reflect the same in an artistic image. Using a similar approach to material transfer, we can perform style transfer and style-guided text-to-image generation. Fig. 12(a) shows the results of style-guided text-to-image generation, where our method learns the style from a single artistic image and generates new images that are semantically different (e.g., "an astronaut landing on a planet") or more vivid in content (e.g., "a man rowing a boat while a dolphin jumps out of the water"), while accurately reproducing the reference image's style. Fig. 12(b) shows

Fig. 14. Results of multi-attribute-aware image generation with *ProSpect*. (a) Each reference offers one kind of visual attribute, and we combine them progressively to generate joint results by mixing the triplet references. (b) Each reference indicates two kinds of visual attributes, and we mix two references by taking the material/layout/style attribute from individual references and scaling the range of content conditions.

the results of style transfer, comparing it with the state-of-the-art diffusion-based style transfer method InST [Zhang et al. 2023b], the GAN-based method JoJoGAN [Chong and Forsyth 2022], encoder-decoder-based method CAST [Zhang et al. 2022], and ViT-based method StyTr^2 [Deng et al. 2022]. Since InST considers the overall appearance of an image as a condition and lacks disentanglement of style and content, the generated image often lacks identity. JoJoGAN needs to align the face key points of the content image and style image, so some special styles may cause artifacts and distortions (as shown in the 1st row), and the generated images may have content in-consistency (as shown in the 2nd row). CAST and StyTr^2 fail to transfer the shape changes and large brushstrokes. Our method produces more realistic strokes (e.g., the hair in 1st and 3rd rows), fewer artifacts (e.g. the 2nd row), and better-maintained identity.

Layout-aware image generation. Layout is a core element of photography that determines the quality of a photo. The low-frequency information of an image reflects its layout. By learning this information, our method can use the layout of a single given image to guide text-to-image generation and transfer the layout of an image to another image. Fig. 13(a) shows the results of layout-guided text-to-image generation, where our method learns complex composition (e.g., "a spoon of strawberry cupcake") and guides the generation of semantically unrelated content (e.g., strawberry cupcake and rock) from a reference image. Fig. 13(b) displays the results of layout transfer for landscape and still-life images. Our method can transfer the "centering" and "reflection" features of a photo to another landscape image (see the second column in Fig. 13(b)) and transfer complex object layouts to another still-life image.

Fig. 15. Comparison of results by training with a small number of images.

Fig. 16. Examples of failure cases. (a) Results of transferring materials between images with large domain gaps. (b) When the image background is composed of similar objects sharing the same frequency information, attribute editing may be applied to the entire image.

Multi-attribute-aware image generation. In Fig. 14, we combine attributes from multiple images to guide the generation process. In Fig. 14(a), the layout, content, and style are guided by three reference images. Results for a landscape example are shown in the left pink pyramid. The first row displays reference images, the second row displays results using dual-attribute guidance, and the bottom row shows the result using triple-attribute guidance. The bottom result maintains the relative position of the flowers and architecture in the layout image, has the three-floor building structure from the content reference, and replicates the appearance of Chinese architecture from the style reference. In the right blue pyramid, we show results for a portrait example. The result is guided by the layout of a single person in the middle, the content of a cyclist, and the style of an astronaut. Fig. 14(b) shows a different setting by mixing multiple attributes from one image.

Few-shot image generation. *ProSpect* is designed to accept a single image as input, but it can also work on a set of images, similar to DreamBooth [Ruiz et al. 2023]. As shown in Fig. 15, *ProSpect* can produce results with improved fidelity and diversity compared to prior approaches when applied to four sculpture images. In addition, *ProSpect* can also be applied to model fine-tuning methods.

5.5 Limitations

First, although *ProSpect* is faster than TI [Gal et al. 2023a], it is still not as fast as some encoder-based methods [Gal et al. 2023b], given that each iteration of optimization is calculated on a random step and *ProSpect* learns several token embeddings at different steps. Second, as shown in Fig. 16(a), *ProSpect* can achieve attribute disentanglement, but the attribute transfer between images with large domain gap may not be visually aesthetic. Finally, Fig. 16(b) shows the cases of dealing with images in which the background is composed of similar objects. Since the objects of the same category are of similar scales, sometimes the attribute modification may act on the background objects undesirably.

6 CONCLUSION AND FUTURE WORK

In this paper, we delve into the image generation process of the diffusion model from the perspective of steps. We propose an expanded textual conditioning space, denoted by $\mathcal{P}^*$, for diffusion models. Our experiments demonstrate that $\mathcal{P}^*$ has better disentanglement and controllability, allowing for generating images from different granularities. To further enable images to be represented in $\mathcal{P}^*$, we propose *ProSpect*, which inverts the text conditions of the diffusion model step by step. *ProSpect* provides more fidelity and editable image representations, paving the way for attributes-aware image generation. Using *ProSpect*, material/style/content/layout-related transfer and editing tasks can be performed. Our evaluations and experimental results demonstrate that *ProSpect* offers superior fidelity, expressiveness, and controllability for diverse image generation tasks. In the future, we plan to further develop and improve methods for attribute disentanglement, such as making a more detailed attribute division and recombination methods as well as studying the mutual impact of different textual conditions.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under no. 2020AAA0106200, in part by the National Natural Science Foundation of China under nos. 61832016, 62102162, and U20B2070, in part by Beijing Natural Science Foundation under no. L221013, in part by the National Science and Technology Council under no. 111-2221-E-006-112-MY3, Taiwan, and in part by the Deutsche Forschungsgemeinschaft (DFG) under no. 413891298.

REFERENCES

Art Institute of Chicago. 2023. https://www.artic.edu/ Last accessed on 2023-09-12.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-Driven Editing of Natural Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18208–18218.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324* (2022).

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by word. *arXiv preprint arXiv:2103.10951* (2021).

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18392–18402.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. In *International Conference on Machine Learning (ICML)*.

Min Jin Chong and David Forsyth. 2022. JoJoGAN: One Shot Face Stylization. In *European Conference on Computer Vision (ECCV)* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 128–152.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision (ECCV)*. Springer, 88–105.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. StyTr^2: Image Style Transfer with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11326–11336.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*. 8780–8794.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12873–12883.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*. Springer, 89–106.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations (ICLR)*.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023b. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–13.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM Transactions on Graphics* 41, 4, Article 141 (2022), 13 pages.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-Prompt Image Editing with Cross Attention Control. In *International Conference on Learning Representations (ICLR)*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023a. Composer: Creative and Controllable Image Synthesis with Composable Conditions. In *International Conference on Machine Learning (ICML)*.

Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. 2023b. Region-Aware Diffusion for Zero-shot Text-driven Image Editing. *arXiv preprint arXiv:2302.11797* (2023).

Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. 2022a. Draw Your Art Dream: Diverse Digital Art Synthesis with Multimodal Guided Diffusion. In *ACM International Conference on Multimedia* (Lisboa, Portugal). 1085–1094.

Nisha Huang, Yuxin Zhang, and Weiming Dong. 2023d. Style-A-Video: Agile Diffusion for Arbitrary Text-based Video Style Transfer. *arXiv preprint arXiv:2305.05464* (2023).

Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Yong Zhang, Weiming Dong, and Changsheng Xu. 2022b. DiffStyler: Controllable Dual Diffusion for Text-Driven Image Stylization. *arXiv preprint arXiv:2211.10682* (2022).

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-Image Translation. In *European Conference on Computer Vision (ECCV)*. 172–189.

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. 2023c. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv preprint arXiv:2303.13495* (2023).

Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. 2023. Training-free Style Transfer Emerges from h-space in Diffusion models. *arXiv preprint arXiv:2303.15403* (2023).

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. In *Advances in Neural Information Processing Systems (NeurIPS)*. 12104–12114.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4401–4410.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6007–6017.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023a. Multi-Concept Customization of Text-to-Image Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023b. Multi-Concept Customization of Text-to-Image Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1931–1941.

Gihyun Kwon and Jong Chul Ye. 2022. CLIPstyler: Image Style Transfer with a Single Text Condition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18062–18071.

Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. 2020. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision* 128 (2020), 2402–2417.

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2023. StyleDiffusion: Prompt-Embedding Inversion for Text-Based Editing. *arXiv preprint arXiv:2303.15649* (2023).

Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. 2022. Text to Image Generation with Semantic-Spatial Aware GAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18187–18196.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11461–11471.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*. 17359–17372.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6038–6047.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*. 8162–8171.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems* 33 (2020), 7198–7211.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2085–2094.

Pexels. 2023. https://www.pexels.com Last accessed on 2023-09-12.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22500–22510.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*. 36479–36494.

Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. 2022. StyleCLIPDraw: Coupling Content and Style in Text-to-Drawing Translation. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 4966–4972.

Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. 2019. FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6490–6499.

Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16494–16504.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. In *ACM SIGGRAPH 2023 Conference Proceedings* (Los Angeles, CA, USA) *(SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, Article 12, 11 pages.

The Barnes Foundation. 2023. https://www.barnesfoundation.org/ Last accessed on 2023-09-12.

Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. 2023. UniTune: Text-Driven Image Editing by Fine Tuning a Diffusion Model on a Single Image. *ACM Transactions on Graphics* 42, 4, Article 128 (2023), 10 pages.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. *P+*: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522* (2023).

Cong Wang, Fan Tang, Yong Zhang, Tieru Wu, and Weiming Dong. 2023b. Towards harmonized regional style transfer and manipulation for facial images. *Computational Visual Media* 9, 2 (2023), 351–366.

Zongji Wang, Yunfei Liu, and Feng Lu. 2023a. Discriminative feature encoding for intrinsic image decomposition. *Computational Visual Media* 9, 3 (2023), 597–618.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668* (2023).

Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. 2023. Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1900–1910.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1316–1324.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023a. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18381–18391.

Serin Yang, Hyunmin Hwang, and Jong Chul Ye. 2023b. Zero-Shot Contrastive Loss for Text-Guided Diffusion Image Style Transfer. *arXiv preprint arXiv:2303.08622* (2023).

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2023. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research* (2023).

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 833–842.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023b. Inversion-Based Style Transfer with Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10146–10156.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning. In *ACM SIGGRAPH 2022 Conference Proceedings*. Article 12, 8 pages.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2023c. A Unified Arbitrary Style Transfer Framework via Adaptive Contrastive Learning. *ACM Transactions on Graphics* 42, 5, Article 169 (2023), 16 pages.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. 2023a. SINE: SINgle Image Editing with Text-to-Image Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6027–6037.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5802–5810.