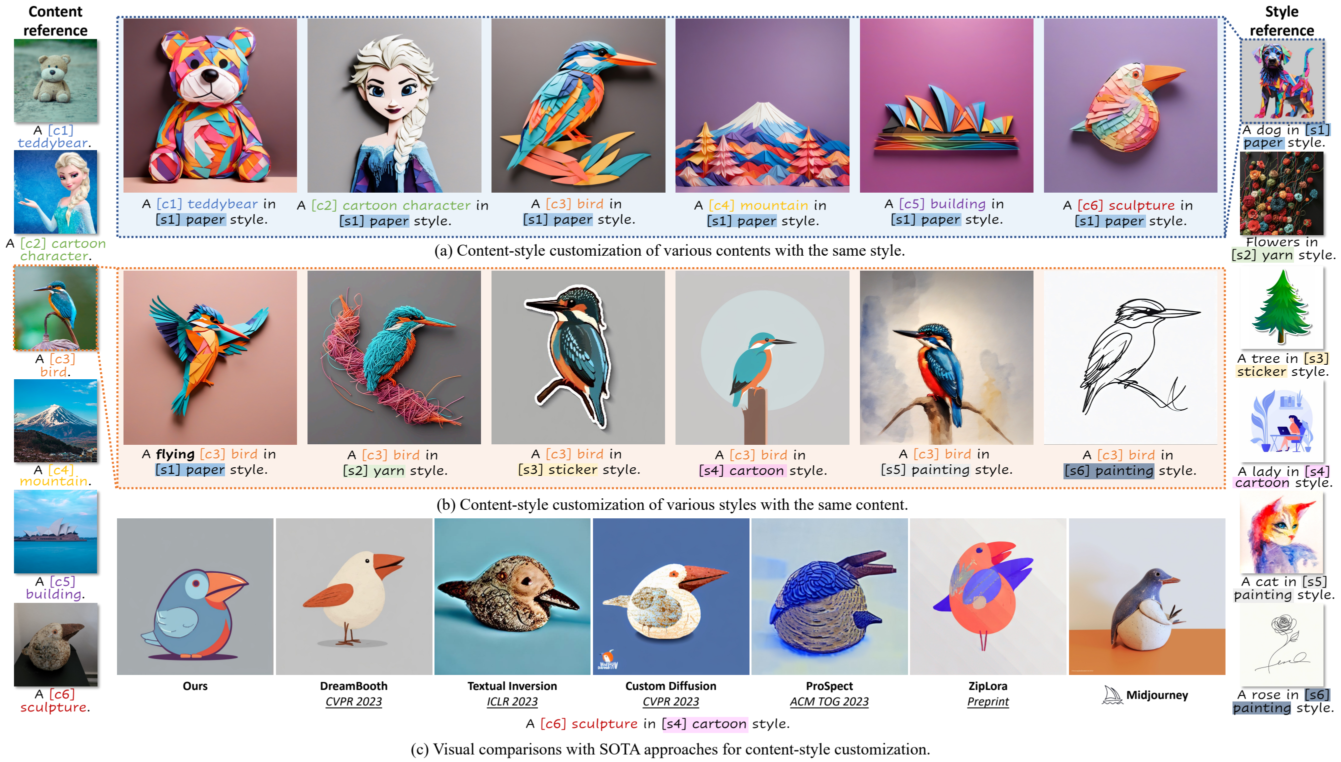
Break-for-Make: Modular Low-Rank Adaptations for Composable Content-Style Customization

Yu Xu[1,2], Fan Tang[1,2], Juan Cao[1,2], Yuxin Zhang[3,2], Oliver Deussen[4], Weiming Dong[3,2], Jintao Li[1], Tong-Yee Lee[5]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Institute of Automation, Chinese Academy of Sciences
[4]University of Konstanz [5]National Cheng-Kung University
{xuyu21b,tangfan,caojuan,jtli}@ict.ac.cn,{zhangyuxin2020,weiming.dong}@ia.ac.cn,
oliver.deussen@uni-konstanz.de,tonylee@mail.ncku.edu.tw


Content reference
A [c1] teddybear.
A [c2] cartoon character.
A [c3] bird.
A [c4] mountain.
A [c5] building.
A [c6] sculpture.
A [c1] teddybear in [s1] paper style.
A [c2] cartoon character in [s1] paper style.
A [c3] bird in [s1] paper style.
A [c4] mountain in [s1] paper style.
A [c5] building in [s1] paper style.
A [c6] sculpture in [s1] paper style.
Style reference
A dog in [s1] paper style.
Flowers in [s2] yarn style.
A tree in [s3] sticker style.
A lady in [s4] cartoon style.
A cat in [s5] painting style.
A rose in [s6] painting style.
A flying [c3] bird in [s1] paper style.
A [c3] bird in [s2] yarn style.
A [c3] bird in [s3] sticker style.
A [c3] bird in [s4] cartoon style.
A [c3] bird in [s5] painting style.
A [c3] bird in [s6] painting style.
Ours
DreamBooth
CVPR 2023
Textual Inversion
ICLR 2023
Custom Diffusion
CVPR 2023
ProSpect
ACM TOG 2023
ZipLora
Preprint
Midjourney
A [c6] sculpture in [s4] cartoon style.

(a) Content-style customization of various contents with the same style.

(b) Content-style customization of various styles with the same content.

(c) Visual comparisons with SOTA approaches for content-style customization.

Figure 1: By separately learning content and style in "partly learnable projection" (PLP), our method is able to generate images of customized content and style aligned with various prompts while successfully disentangling content and style and maintaining high fidelity of them.

ABSTRACT

Personalized generation paradigms empower designers to customize visual intellectual properties with the help of textual descriptions by tuning or adapting pre-trained text-to-image models on a few images. Recent works explore approaches for concurrently customizing both content and detailed visual style appearance. However, these existing approaches often generate images where the content and style are entangled. In this study, we reconsider the customization of content and style concepts from the perspective of parameter space construction. Unlike existing methods that utilize a shared parameter space for content and style, we propose a learning framework that separates the parameter space to facilitate individual learning of content and style, thereby enabling disentangled content and style. To achieve this goal, we introduce "partly learnable projection" (**PLP**) matrices to separate the original adapters into divided sub-parameter spaces. We propose "**break-for-make**" customization learning pipeline based on PLP, which is simple yet effective. We **break** the original adapters into "up projection" and "down projection", train content and style PLPs individually with the guidance of corresponding textual prompts in the separate adapters, and maintain generalization by employing a multi-correspondence projection learning strategy. Based on the adapters broken apart for separate training content and style, we then **make** the entity parameter space by reconstructing the content and style PLPs matrices, followed by fine-tuning the combined adapter to generate the target object with the desired appearance. Experiments on various styles, including textures, materials, and artistic style, show that our method outperforms state-of-the-art single/multiple concept learning pipelines in terms of content-style-prompt alignment.

KEYWORDS

Customize generation, content-style fusion, text-to-image generation.

1 INTRODUCTION

Text-to-image (T2I) models based on diffusion technology [Ho et al. 2020; Ho and Salimans 2022; Song et al. 2020] have demonstrated remarkable proficiency in generating high-quality images, expanding the imaginative capabilities of humans through textual descriptions. Represented by Stable Diffusion [Rombach et al. 2022] and Midjourney [Midjourney 2023], various diffusion models and platforms have been widely applied in the field of creativity design or digital content generation. Despite the outstanding generalization ability of T2I models, it is challenging for users to generate specific visual concepts using only textual descriptions.

Customized generation approaches have thus been proposed for subject-driven generation by techniques such as tuning the base model with regularization [Ruiz et al. 2023], learning additional parameters as pseudo words [Alaluf et al. 2023; Gal et al. 2022; Voynov et al. 2023] or low-rank adaptations [Hu et al. 2021]. Most of these approaches, however, only support generating images depicting a single concept (e.g., objects, textures, materials, art style, etc.), leaving the customized generation of multi-concept (e.g. specific content with a specific style) a challenging task. For example, designers may wish to render specific objects with different textures or materials to examine various effects. Similarly, artists may want to render specific objects in their own distinctive styles.

Multi-concept generation approaches [Avrahami et al. 2023a; Kumari et al. 2023] are first proposed to learn and generate different contents by manipulating or constraining cross-attentions mechanisms. Different objects would be distinguished on the cross-attention maps using corresponding textual descriptions [Hertz et al. 2022]. However, the intricate nature of visual style, which is often entangled with content, poses challenges in effectively decoupling content and style concepts due to the shared parameter space and the lack of disentanglement strategies employed by these methods. As a result, previous approaches cannot be well applied to jointly learn content and style concepts.

To address the content-style customization problem, Zhang et al. [2023b] recently propose a step-wise pseudo words generation pipeline, which supports combining content and style concepts. However, relying on step-wise diffusion priors limits ProSpect's generability across different types of visual styles. More intuitively, ZipLoRA [Shah et al. 2023] merges two independently fine-tuned content and style adaptations using a loss function based on cosine similarity to alleviate the entanglement between content and style. Nevertheless, the merging process often leads to interference between the parameters of different adapters [Ortiz-Jimenez et al. 2023]. This oversight in failing to optimally align the integrated parameters can result in a notable performance degradation of the merged model, leading to ineffective preservation of the distinct qualities of both content and style [Yadav et al. 2023]. Therefore, a method that decouples the learning of content and style, and recombines them in the generation process without interference, is necessary.

In this work, we introduce a two-stage learning approach for customized content-style generation, namely "break-for-make". In the first stage, we propose "partly learnable projection" (**PLP**) matrices to train content and style in separated sub-parameter spaces of low-rank adapters. Specifically, we freeze certain parameters in both the "up projection" and "down projection" matrices, allowing separate training of content and style within their respective trainable parameter subsets. To avoid interference between content and style after matrix multiplication by frozen parameters, we initialize the frozen rows and columns within the projection matrices to approximate orthogonal bases. To maintain the generalization of the learned content/style PLPs, we utilize a "multi-correspondence projection" (**MCP**) learning strategy to learn unbiased content and style parameter spaces. Specifically, we train customized content in "up projection" matrices with diverse reference styles in "down projection" matrices and vice versa. This approach avoids one-to-one binding between content and style, thereby mitigating the overfitting of content/style PLPs when composing with other corresponding PLPs. In the second stage, we reconstruct the unified parameter space using the content and style PLP matrices trained in the first stage, then fine-tune the combined adapter to obtain content-style customized results. As the specific content and style are learned separately and in a generalized manner during the first stage, fine-tuning (approximately a few dozen steps) is required for the combined adapter to generate images that better align with the content and style references, as shown in Fig. 1(a) and (b).

Our contributions can be summarized as follows:

- We separate the parameter space of low-rank adapters for disentangling the content and style representations and introduce a content-style customization learning pipeline.
- We propose a Partly Learnable Projection (PLP) with an orthogonal frozen parameters strategy that enables the disentanglement of content and style. During training, a Multi-Corresp-ondence Projection (MCP) mechanism is proposed to maintain generalization.
- Extensive qualitative and quantitative experiments validate the superior effectiveness of our approach over current baseline methods, particularly in the realms of content and style disentanglement and the preservation of content-style fidelity.

2 RELATED WORK

2.1 Text-to-Image Customization

Diffusion models [Ho et al. 2020] have demonstrated the capability to produce high-quality images in text-to-image generation [Betker et al. 2023; Chang et al. 2023; Rombach et al. 2022; Saharia et al. 2022]. Text-to-image customization aims to inject specific concepts or styles into diffusion models to generate diverse images, including different views, poses, scenes, and more [Chen et al.

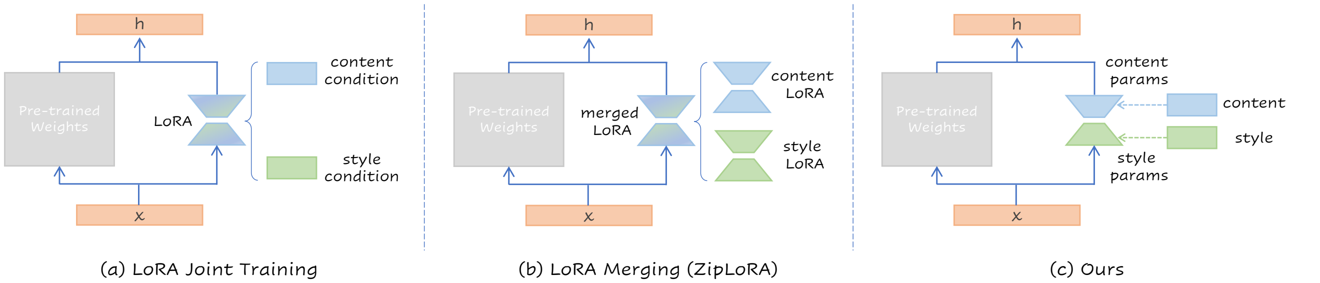


Figure 2: Frameworks of the two main approaches and ours for customized content-style image generation. LoRA joint training incorporates image-text pairs to fine-tuning the overall model parameters. ZipLoRA [Shah et al. 2023] effectively merging independently trained content and style LoRAs, then add to per-trained weights to generaet images of the customized content and style. Our method trains content and style in separated parameter subspaces of LoRA, results in disentanglement of content and style while maintaining high level of fidelity.

2024; Gal et al. 2022, 2023; Huang et al. 2024; Ruiz et al. 2023; Wei et al. 2023; Zhang et al. 2023c]. To achieve this, numerous approaches have been proposed across various aspects. Textual Inversion [Gal et al. 2022] employs inherent parameter space to describe specific concepts and inverts training images back to text embeddings. DreamBooth [Ruiz et al. 2023] fine-tunes backbone models with specific token-images pairs and a prior preservation loss. Custom diffusion [Kumari et al. 2023] optimizes a few diffusion model parameters to represent new concepts/styles while enabling fast tuning for multiple concepts jointly. LoRA [Hu et al. 2021], a parameter-efficient fine-tuning approach first revealed for large language models, has proven effective for customization by adapting only a few adaptation parameters. LoRA's lightweight nature and ability to generate customized content/style without full model fine-tuning make it highly flexible. Various LoRA-based methods have been proposed for more effective and efficient training [Dettmers et al. 2024; Edalati et al. 2022; Hyeon-Woo et al. 2021; Valipour et al. 2022; Zhang et al. 2023a]. Po et al. [2023] design multiple LoRAs to separately train different content and generate multiple contents simultaneously in one image. By integrating adapter modules, AdapterFusion [Pfeiffer et al. 2020] allows adaptation to downstream tasks via fine-tuning only the adapter parameters. Liu et al. [2023] propose Cones, a layout guidance method for controlling multiple customized subject generation. Perfusion [Tewel et al. 2023] introduces a new mechanism locking new concepts' cross-attention Keys to their superordinate category to avoid overfitting, and a gated rank-1 approach to control a learned concept's influence during inference and combine multiple concepts. NeTI [Alaluf et al. 2023] and ProSpect [Zhang et al. 2023b] introduce an expanded text-conditioning space over diffusion time steps for fine-grained control. These concept-customized generation methods primarily focus on the quality of generated outputs, addressing general concept customization. In contrast, we focus mainly on the fusion generation of customized content and style.

2.2 Customized Content Style Fusion

The goal of the content-style customization is to generate an image that incorporates specific content and style based on reference images, while ensuring the unique characteristics of both content and style are distinctively represented and aligned with prompts. Previous works jointly train content and style on customized generation models [Gal et al. 2022; Kumari et al. 2023; Ruiz et al. 2023]. During inference, these methods generate images blending both content and style based on given prompts. However, these straightforward approaches do not optimize the learning between content and style, often resulting in their entanglement in the generated results. DreamArtist [Dong et al. 2022] employs a positive-negative prompt-tuning learning strategy for customized generation and discusses content-style image fusion in the experiments. SVDiff [Han et al. 2023] fine-tunes the singular values of weight matrices and proposes a Cut-Mix-Unmix data-augmentation technique to help multi-subject and content-style image generation. StyleDrop [Sohn et al. 2023] improves the quality of generating stylized images via iterative training with human or automated feedback. ProSpect [Zhang et al. 2023b] leverages learning word embeddings specific to content and style, incorporating them at different diffusion time steps to control customized content-style image generation. However, relying on step-wise diffusion priors limits ProSpect's generability across different content and visual styles. Recent work ZipLoRA [Shah et al. 2023]learns hybrid coefficients to optimize conflicts arising when merging two separately trained LoRAs, partially mitigating disentanglement issues. However, it concurrently modifies the distribution of learned parameters, subsequently influencing reconstruction outcomes. Compared to related approaches, our proposed "partly learnable projection" and "multi-correspondence projection learning" strategy trains content and style separately in different sub-parameter spaces within low-rank adaptations with data augmentation. This effectively disentangles content and style in generated images while maintaining high image fidelity.

3 VANILLA SOLUTIONS FOR CONTENT-STYLE CUSTOMIZATION

In this section, we first introduce the task definition of content-style customization in image generation. Then, we review existing related methods, including the basic low-rank adaptation fine-tuning method, joint training method, and merging after independent training method. Note that our primary focus is on methods based on low-rank adaptations, as these are both efficient and effective for fine-tuning large T2I models. We then investigate why these methods fail to generate images of disentangled content and style faithfully. In response to these challenges, we propose our novel solution.

The goal of content-style customization is to generate images that effectively present both user-specified content and style while ensuring their unique characteristics are distinctively represented [Shah et al. 2023; Zhang et al. 2023b]. Formally, given a content reference image I_c, a style reference image I_s, and a prompt P, we aim to generate an output image I_{out} that contains the same content as I_c, and has the same style as I_s, while aligning with the provided prompt P.

However, accurately generating specific content while effectively rendering it in a reference style without conflict is challenging. For example, the inherent style of the user-provided content image may interfere with the reference style, leading to style conflict when generating a customized content-style image. LoRA [Hu et al. 2021], a lightweight adaptation method, has been applied in customized generation, enabling effective learning of content and style within the LoRA module. This motivates us to train content and style in separate subspaces of the low-rank adaptation parameters.

Low-Rank Adaptation Fine-Tuning. LoRA [Hu et al. 2021] is an efficient adaptation strategy for fine-tuning large pre-trained models while retaining high quality. Initially proposed for fine-tuning large language models, LoRA has also proven suitable for fine-tuning vision models like diffusion models for text-to-image generation. For a per-trained weight matrix $W_o \in \mathbb{R}^{m \times n}$, each LoRA module consists of an up-projection matrix $W_{up} \in \mathbb{R}^{m \times r}$ and a down-projection matrix $W_{down} \in \mathbb{R}^{r \times n}$, where the rank $r \ll min(m, n)$. Given an input z, during training, the forwards pass is:

$$I = W_0 z + W_{up} W_{down} z, \tag{1}$$

and only W_{up} and W_{down} are updated to find a suitable adaptation $\Delta W = W_{up} W_{down}$. In this work, we incorporate LoRA modules into the cross-attention components of the diffusion model for fine-tuning [Simo 2023]. After that, we can directly merge the LoRA module with the per-trained weight matrix and obtain new weights $W = W_0 + \Delta W$, which can perform inference as usual.

Jointing Training. A straightforward method for customized content-style generation is jointly training LoRA modules with customized content images and style images. In simple terms, LoRA modules W for learning specific content and style are trained using a squared error loss function as follows:

$$L = [||\hat{W}_\theta(z_c|c_c, t) - x_c||_2^2] + [||\hat{W}_\theta(z_s|c_s, t) - x_s||_2^2], \tag{2}$$

where (z_c, c_c, x_c) and (z_s, c_s, x_s) are data-conditioning-target pairs of the specific content and style (image latent, text embeddings and target images), respectively. t is diffusion process time $t \sim ([0, 1])$, and θ represents model parameters. The training pipeline is presented in Fig. 2 (a). However, this training approach mixes the parameter spaces of content and style during the training stage, resulting in the entanglement of content and style when weights W multiplied with the input, as analyzed in Fig. 3 (a).

Merging after Independent Training. Another primary method involves independently training two LoRA modules — one dedicated to content and the other to style — in the first stage. Subsequently, in the second stage, these modules are merged with certain constraints, as shown in Fig. 2(b). Given a set of learned LoRA weights ΔW_i optimized on content and style, the merged weight is simply given by

$$W_{merged} = W_0 + \sum_i \lambda_i W_i, \tag{3}$$

where λ_i is a scalar representing the relative strength of content and style. However, directly merging two independently trained LoRAs may lead to parameter conflict. Specifically, when merging a parameter that is influential for one LoRA but redundant for the other, the influential value may be obscured by the redundant values, resulting in a decrease in overall effectiveness [Yadav et al. 2023]. This interference leads to the loss of content and style features learned during independent training stages, as analyzed in Fig. 3 (b). ZipLoRA [Shah et al. 2023] learns mixing coefficients for both content and style LoRAs to mitigate conflicts. Nevertheless, to some extent, it affects the distribution of content and style parameters learned during the training phase. Although this approach shows improved disentanglement performance, the fidelity of reconstruction is somewhat reduced. This motivates us to pursue separate training for content and style, aiming to achieve both precise reconstruction on customized content and style as well as effective disentanglement.

4 OUR METHOD

In this section, we first introduce our proposed "partly learnable projection" (**PLP**) method, a parameter separation training framework for LoRA that enables better control over the training parameters. This facilitates the generation of images that are more faithfully aligned with the specified conditions while maintaining higher fidelity. We then present "Multi-Correspondence Projection Learning" (**MCP**), a technique for training content and style representations during the customization process to mitigate overfitting between the two. By utilizing both the proposed **PLP** and **MCP** methods, we enable the generation of customized content-style images that achieve effective disentanglement of content and style, while also preserving a high degree of image fidelity.

4.1 Partly Learnable Projection

To address the aforementioned issues, we propose "Partly Learnable Projection" (**PLP**) matrices to separate the LoRA module and search for the optimal content and style parameters within distinct sub-parameter spaces. Specifically, we consider a LoRA module ΔW with input dimension n, rank r, and output dimension m. The W_{down} and W_{up} matrices of ΔW are decomposed into two submatrices along the feature dimension, respectively. The W_{up} can be formed as:

$$W_{up} = \begin{bmatrix} A & B \end{bmatrix}^{-1}, \tag{4}$$

where

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1r} \\ \vdots & \ddots & \vdots \\ A_{d1} & \cdots & A_{dr} \end{bmatrix}, B = \begin{bmatrix} B_{(m-d)1} & \cdots & B_{(m-d)r} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mr} \end{bmatrix}. \tag{5}$$

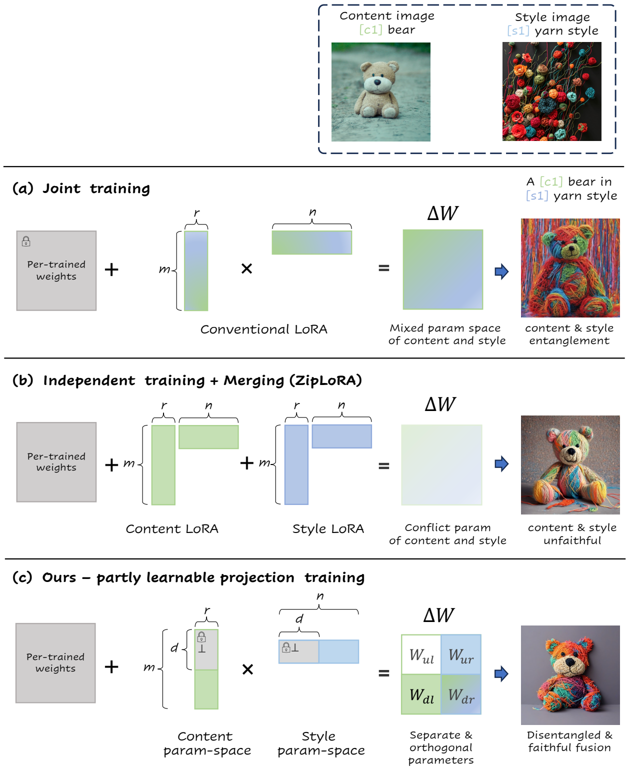

Content image
[c1] bear
Style image
[s1] yarn style
(a) Joint training
A [c1] bear in
[s1] yarn style
Per-trained weights
r
m
n
Conventional LoRA
ΔW
Mixed param space of content and style
content & style entanglement
(b) Independent training + Merging (ZipLoRA)
Per-trained weights
Content LoRA
Style LoRA
Conflict param of content and style
content & style unfaithful
(c) Ours – partly learnable projection training
Per-trained weights
d
Content param-space
Style param-space
W_{ul} W_{ur} W_{dl} W_{dr}
Separate & orthogonal parameters
Disentangled & faithful fusion

Figure 3: Analysis of the two main methods and ours using LoRA for customized content-style image generation. Joint training LoRA will mix the parameter space of content and style, leads to entanglement of content and style. Merging LoRAs after independent training has problem of conflict parameters from content LoRA and style LoRA, leads to content and/or style unfaithful after fusion. Our proposed method trains the content and style in separate parameter subspaces of the LoRA modules, with orthogonal fixed parameter spaces, resulting in disentangled and faithful fusion of content and style.

Similarly, the W_{down} matrix can be formed as:

$$W_{down} = \begin{bmatrix} C & D \end{bmatrix}, \tag{6}$$

where

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1d} \\ \vdots & \ddots & \vdots \\ C_{r1} & \cdots & C_{rd} \end{bmatrix}, D = \begin{bmatrix} D_{1(n-d)} & \cdots & D_{1n} \\ \vdots & \ddots & \vdots \\ D_{r(n-d)} & \cdots & D_{rn} \end{bmatrix}. \tag{7}$$

According to the rules of partitioned matrix multiplication, we have

$$\Delta W = W_{up} W_{down} \tag{8}$$

$$= \begin{bmatrix} W_{ul} & W_{ur} \\ W_{dl} & W_{dr} \end{bmatrix}, \tag{9}$$

where

$$W_{i,j}^{ul} = \sum_r A_{i,r} C_{r,j}, \tag{10}$$

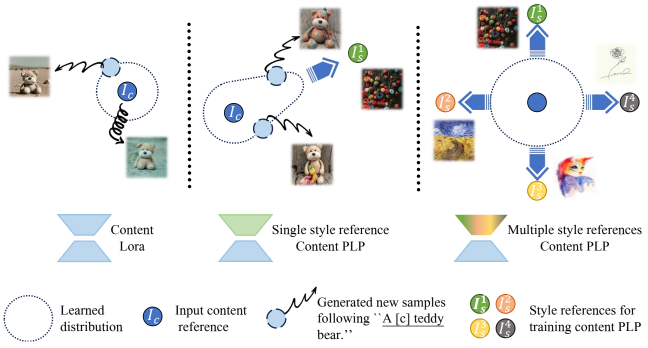


Figure 4: Illustration of the multi-correspondence projection. We present the learned content distribution on the left of the top row. When training specific content and style in a one-to-one manner, the content will tend to overfit to the specific style, as illustrated on the middle of the top row. By leveraging our proposed multi-correspondence projection, we learn multiple styles with the content in PLP, enhance the generalization of the learned content.

$$W_{i,j}^{ur} = \sum_r A_{i,r} D_{r,j}, \tag{11}$$

$$W_{i,j}^{dl} = \sum_r B_{i,r} C_{r,j}, \tag{12}$$

$$W_{i,j}^{dr} = \sum_r B_{i,r} D_{r,j}. \tag{13}$$

Here, d represents the feature dimension of the fixed parameters. Adjusting the size of d implies modifying the ratio of frozen to trainable parameters within the matrix. We will discuss it in Section 5.7. After multiplication, we obtain a partitioned matrix, which can be visualized as the original matrix decomposed into a set of horizontal and vertical submatrices.

We propose PLP with orthogonal parameters for better disentanglement of content and style during training. Specifically, the matrices A and C in Eq. (5) and Eq. (7) are kept frozen during the training process. We initialize A and C as approximately orthogonal to reduce redundant parameters and achieve better disentanglement of content and style:

$$W_{i,j}^{ul} = \sum_r A_{i,r} C_{r,j}, \tag{14}$$

$$= 0. \tag{15}$$

We can notice that, the up-right part of ΔW in Eq. (11) represents **only** the parameters of submatrix D only. Similarly, the down-left part of ΔW in Eq. (12) represents **only** the parameters of submatrix B only, $W_{i,j}^{dr}$ in Eq. (13) relates to B and D, allowing us to learn the interactive features between them.

The forward pass during training yields:

$$I = W_0 z + \begin{bmatrix} 0 & \sum_r A_{i,r} D_{r,j} \\ \sum_r B_{i,r} C_{r,j} & \sum_r B_{i,r} D_{r,j} \end{bmatrix} z, \tag{16}$$

where $A_{i,r}$ and $C_{r,j}$ are frozen during training.

So far, we have demonstrated that our proposed method based on partitioned matrices can effectively separate the parameters associated with content and style. This enables the multiplication of input features with the corresponding parameters during training, thereby distinctly representing the acquired content and style in different subspaces of the parameter space. Consequently, this mitigates the entanglement between content and style during image generation while preserving a high degree of fidelity.

Our method is illustrated in Fig. 3 (c), which indicates that after separating the LoRA module into two parts and performing forward matrix multiplication, the resulting partitioned matrices exhibit the following advantageous characteristics: the top-left part consists of zeros due to the multiplication of orthogonal vectors, the top-right part represents the style submatrix for learning style image feature parameters, the bottom-left part represents the content submatrix for learning content image feature parameters, and the bottom-right part is utilized for learning interactive feature parameters between content and style. The four distinct parts in the partitioned matrix demonstrate that our method successfully separates content and style for training in different LoRA parameter subspaces. This circumvents the parameter conflict issues introduced by merging methods and allows us to obtain disentangled content and style feature representations. Meanwhile, the interactive parameters between content and style enable the generation of more naturalistic fusion images with high visual quality.

4.2 Multi-Correspondence Projection Learning

When training specific content and style in a one-to-one manner, overfitting issues may arise, resulting in suboptimal performance when reconstructing the content-style modules in the second stage for image generation. To mitigate this problem between content and style during training, we introduce a multi-correspondence projection ("**MCP**") learning method involving diversified content-style training data pairs. Specifically, when training for a particular content, we update the parameters of B in Eq. (5) with the particular content image and update the parameters of D in Eq. (7) with various style images, vice versa. In simple terms, a LoRA model W for learning specific content is trained using a squared error loss function as follows:

$$L = [\|\hat{W}_\theta(z_c|c_c, t) - x_c\|_2^2] + \frac{1}{n} \cdot \sum_{i=1}^{n} [\|\hat{W}_\theta(z_s|c_s, t) - x_s\|_2^2], \quad (17)$$

where (z_c, c_c, x_c) and (z_s, c_s, x_s) are data-conditioning-target triplets of the specific content and diverse styles (image latents, text embeddings, and target images), respectively. n represents the number of different styles. t is the diffusion process time $t \sim ([0, 1])$, and θ represents the model parameters. The loss function for training the style LoRA model is similar to Eq. (17). This training approach prevents overfitting issues that arise when learning specific content-style pairs (see the ablation study in Fig. 14), simultaneously enhancing the method's generalization ability and improving the effectiveness of diverse content-style combinations.

Inference. After training, we obtain $LoRA_c$ which contains the learned parameters of the specific content, and $LoRA_s$ which contains the learned parameters of the specific style. We then combine the up-projection part of $LoRA_c$ with the down-projection part of $LoRA_s$ to reconstruct $LoRA_f$ as the fusion adapters. With a few dozen fine-tuning steps of $LoRA_f$ on the given content and style images, we can effectively obtain the final adapters capable of generating content-style disentangled images with high fidelity. For single content or style generation, we can directly perform inference using the learned $LoRA_c$ or $LoRA_s$, respectively.

5 EXPERIMENTS

In this section, we conduct qualitative comparisons and quantitative evaluations to demonstrate our method outperforms state-of-the-art customized content-style fusion baselines. We also conduct ablation studies to analyze the impact of certain crucial modules in our approach on the generation results.

Datasets. For fair and unbiased evaluation, we use concept images and style images from related works [Gal et al. 2022; Ruiz et al. 2023; Shah et al. 2023; Zhang et al. 2023b] together with diverse images from the Internet. For training content images, we collect three to five images of the same content and five different styles, each style consisting of one image. For training style images, we collect one to three images of the same style and five different contents, each content consisting of one image.

Compared Methods. We compare our method against state-of-the-art baselines on the task of content-style customization.

- Dreambooth+LoRA (DB+LoRA) [Simo 2023] leverages LoRA for customized content and style generation. We jointly train the LoRA with content and style images and their corresponding prompts. In each epoch, we first update parameters based on content loss for the same model, followed by updating parameters based on style loss.
- Textual inversion (TI) [Gal et al. 2022] inverts a customized image back to text embeddings and bands it with a token. Then, the token can be composed into a prompt to generate related content or style in the output image. We learn content and style on two separate tokens and simultaneously incorporate these two tokens into the prompt for content-style customization.
- ProSpect [Zhang et al. 2023b] proposes a novel approach that adds different conditions to the diffusion model in different generation steps to achieve fine-grained and controllable generation. In our experiment, we add customized content and style as conditions in different steps of generation to achieve content-style fusion, according to their paper.
- Custom Diffusion (CD) [Kumari et al. 2023] proposes an efficient fine-tuning method for simultaneously generating multiple customized content or content with style. We followed the official open-source code, conducting joint training for content and style.
- ZipLoRA [Shah et al. 2023] is a recently released method that provides a novel approach to merge trained content and style LoRAs by learning mixing coefficients for LoRAs. As official codes of ZipLoRA have not been released yet, we evaluate this method with a popular implementation in the GitHub community [mkshing 2023]. We initially train the content and style models separately and then perform LoRA merging based on the parameters specified in the paper.

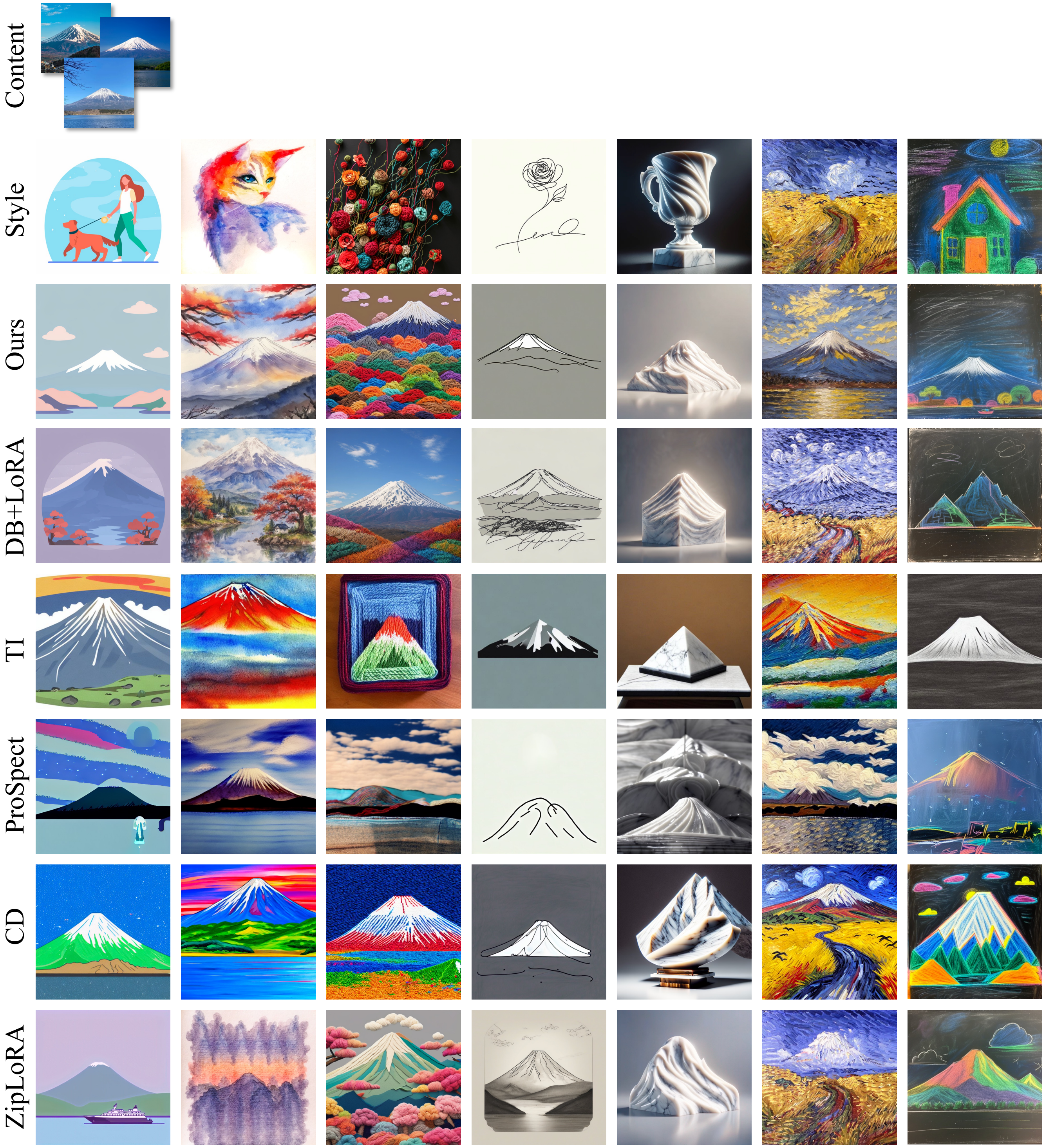


Figure 5: Qualitative evaluation and comparison of DB+LoRA, TI, ProSpect, CD, ZipLoRA and our method in diverse styles. We present the results of customized generation of the same content and different styles. Results indicate that our method generates harmonious fusion images of the content and the style, while preserving the disentanglement of content and style, as well as maintaining high-level fidelity of them.

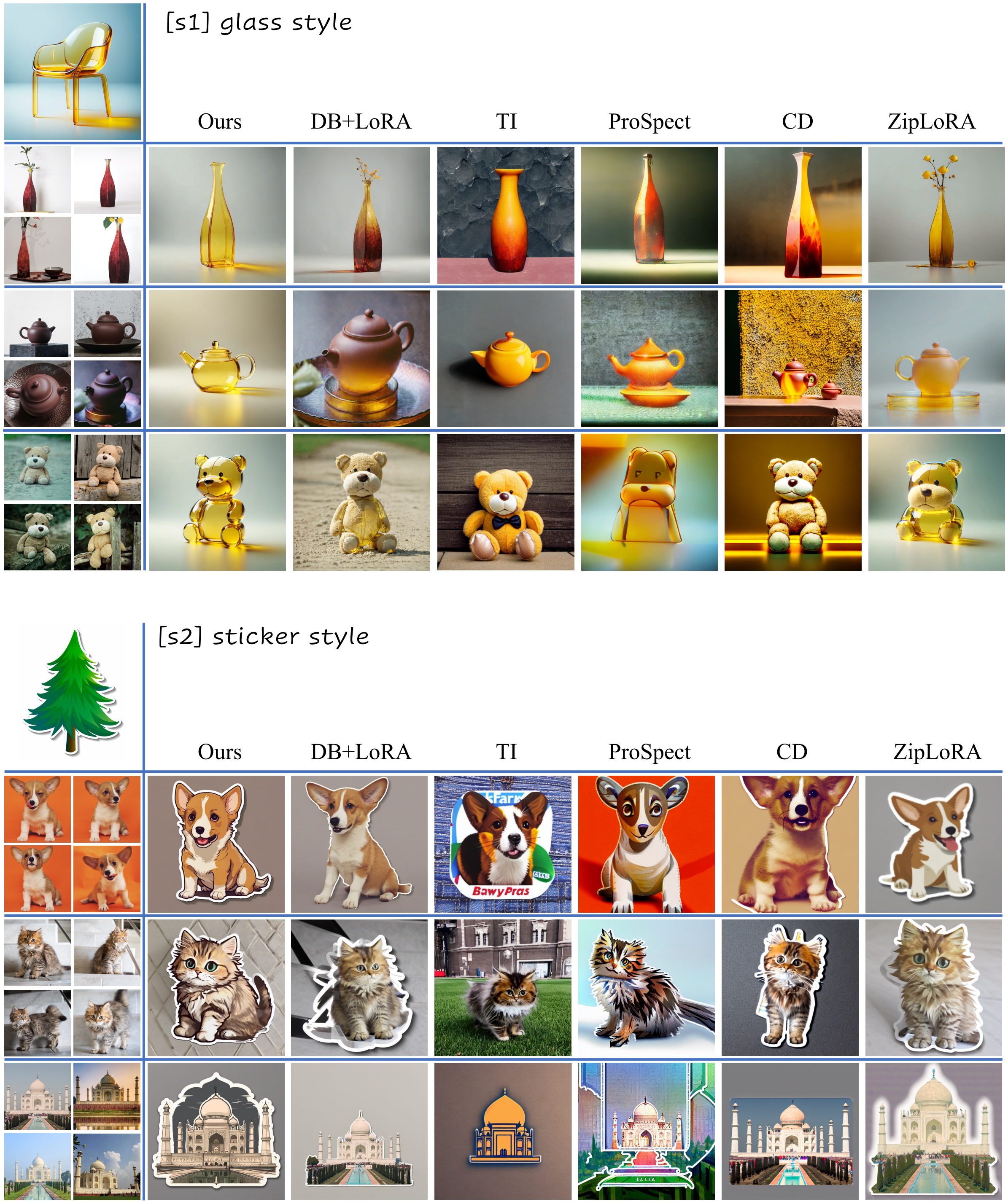


Figure 6: Qualitative evaluation and comparison of DB+LoRA, TI, ProSpect, CD, ZipLoRA, and our method in diverse contents. The results indicate that our method generates harmonious content-style fusion images with diverse contents while preserving the disentanglement of content and style, as well as maintaining high-level fidelity.

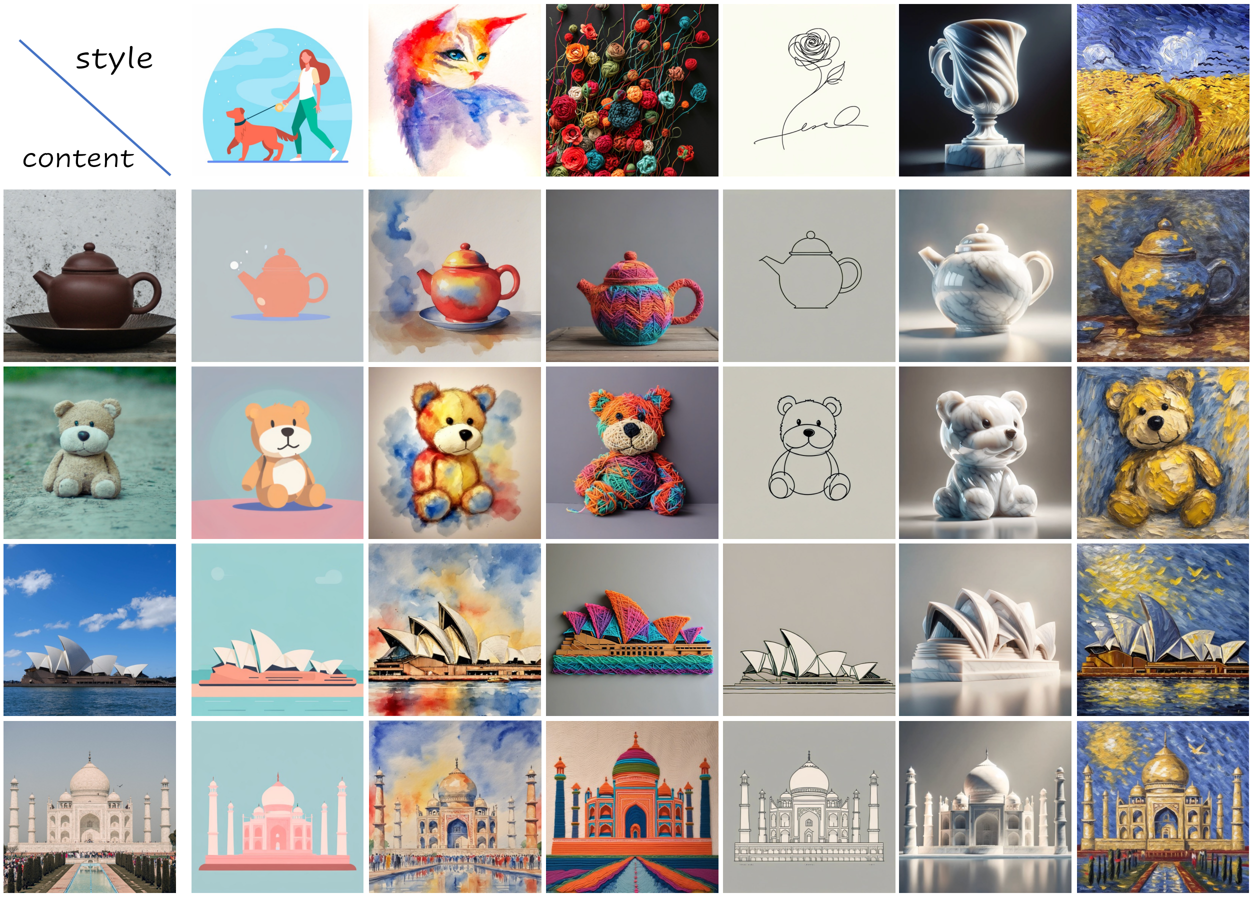


Figure 7: More results of diverse content and style generated by our method.

Metrics. We primarily conduct qualitative and quantitative comparisons between our method and baseline methods. For qualitative comparisons, we primarily present and compare the visual quality of the generated images. For quantitative comparisons, we mainly assess three metrics: content alignment and style alignment between the generated images and reference images, as well as text alignment between the generated images and the corresponding prompts. Following quantitative experiment settings of ProSpect [Zhang et al. 2023b] and ZipLoRA [Shah et al. 2023], we compare cosine similarities between CLIP [Ilharco et al. 2021] (for style and prompt) and DINOv2 [Oquab et al. 2023] features (for content) of the generated images and reference contents, styles and prompts respectively.

Implementation Details. In our experiments, we utilized Stable Diffusion XL v1.0 [Podell et al. 2023] with default hyperparameters and set a base learning rate of 0.0001. During training, we set the batch size to 1, text encoders of SDXL are kept frozen, and the refiner of SDXL is not utilized. Based on the orthogonal fixed parameters we proposed, we train LoRA modules with the same size as the input and output feature dimensions. The rank of LoRA is set to 64.

5.1 Main Results

In this section, we present Qualitative and Quantitative Comparisons between our method and baseline approaches. Additionally, we showcase more of our results with diverse contents and styles in Fig. 7.

Qualitative Comparison. We compare our method with five content-style customization methods: DreamBooth+LoRA (DB+LoRA), Textual Inversion (TI), ProSpect, Custom Diffusion (CD) and ZipLoRA. We first present the results of generating the same content image with multiple style images in Fig. 5, then we present the same style image with multiple content images in Fig. 6. Results indicate that our methods successfully disentangle content and style in one image while maintaining a high level of fidelity. The DB+LoRA method usually generates images of unnatural content style fusion (the result of “mountain” with “yarn style” and “oil painting style” in Fig. 5) and images of the mixed style (“vase” and “teapot” with “glass style” in Fig. 6), the observed entanglement phenomenon aligns with the analysis presented in Section 3. The TI method only updates parameters in the text embedding space, thus having a relatively weaker learning capability. At times, it struggles to

Table 1: Comparison of cosine similarity between CLIP(for style and prompt) and DINO features(for content) of the generated images and reference style, content, and prompt, respectively. Our method has the best average score, indicating that our approach successfully customizes the generation of the target content and style while aligning with the prompt.

Methods	DB+LoRA	TI	ProSpect	CD	ZipLoRA	w/o MCP	w/o Orth	Ours
Content-alignment (↑)	0.7982	0.7292	0.6165	0.6845	0.7103	0.6242	**0.6874**	0.6615
Style-alignment (↑)	0.4974	0.3942	0.4816	0.4381	0.5414	0.5331	0.5626	**0.6219**
Prompt-alignment (↑)	0.3894	0.2836	0.3156	0.2778	0.3319	0.3867	**0.4035**	0.3908
Average (↑)	0.5617	0.4690	0.4712	0.4668	0.5279	0.5147	0.5512	**0.5581**

accurately learn content/style features, leading to a decrease in the fidelity of generated images (“mountain” with “Minimalism painting style” and “marble style” in Fig. 5, the loss of feature “transparent glass” in “glass style” in Fig. 6). ProSpect learns the reference image with a specific token and trains the embedding of this token, then adds it as a condition in different steps during inference. This method has achieved effective control over content and style to some extent, as seen in examples such as “vase” and “teapot” in Fig. 6, shape and material are presented in generated images. However, it is constrained by its learning capability, which leads to low-quality content-style customization results (the result of “mountain” with “watercolor painting style” and “yarn style” in Fig. 5). The CD also encounters entangling issues between content and style. In cases of “glass style” with “vase” and “teapot” in Fig. 6, the reference images of content influence the style of the generated images. In the case of ZipLoRA, the generated results may not accurately present the reference content or style. For example, in instances like “vase” and “teapot” in Fig. 6, the outputs of ZipLoRA lack the texture style of “transparent glass” in the reference set. In instances of generating “mountain” with “oil painting” style and “blackborad painting” style, the mountain cannot be generated faithfully as the reference. This also reflects the manifestation of fidelity degradation due to parameter conflicts. Compared with the above methods, our method maintains a high level of fidelity and harmonious content-style interaction when generating various styles for the same content. This also demonstrates the strong generalization capability of our approach. One more interesting thing is that the instance of the “sticker style” images includes a dual style, encompassing both sticker and cartoon styles. When evaluating it as the reference, our method successfully generates images in the sticker style. It simultaneously transfers the content into a cartoon style, while the results of other methods are kept in a realistic style.

Quantitative Comparison. We present quantitative comparison results in Table 1, evaluating the style-alignment, prompt-alignment (using CLIP feature extraction), and content-alignment (using DINO feature extraction) metrics. Additionally, we report the average of these three metrics, where higher values indicate better performance. Our method achieves the highest average score among all baselines, suggesting it generates customized content-style images that align well with the content and style references while corresponding to the given prompt. Note that in the content-alignment metric, our score is not the highest because other methods tend to generate images that retain more features from the content reference images. However, this could compromise accurate expression of style and adherence to the prompt in the generated images, affecting the effectiveness of style transfer and prompt alignment, as indicated by the lower style-alignment and prompt-alignment metrics for other methods in Table 1. Additionally, the comparative display in Fig.5 and Fig. 6 supports this observation.

5.2 Editability Evaluation and Comparison

We evaluate and compare the editability of our method against other baselines by generating customized content-style fusion images using a diverse set of prompts. For a fair comparison, the prompts and results of the ZipLoRA are obtained from their original paper. As illustrated in Fig. 8, ZipLoRA is generally effective in generating customized content-style images that align well with the provided prompts. However, in some details, ZipLoRA tends to lose certain characteristics of the reference image, such as the ears and mouth in the “wearing a hat” example, and the overall appearance in the “in a boat” and “driving a car” examples. In contrast, our method maintains better consistency with the reference image in these generated results. Additionally, we showcased more generation results from diverse prompts in the bottom two rows of Fig. 8. These results demonstrate high alignment with the prompts while maintaining a high level of disentanglement between content and style, as well as preserving the fidelity of content and style representations. Overall, our method exhibits superior editability compared to existing baselines, enabling the generation of customized content-style images that faithfully integrate the provided prompts while retaining the desired characteristics of the reference content and style.

5.3 Visualizing and Comparing Parameter Distributions for Our Method and Baseline Methods

We employ t-SNE [Van der Maaten and Hinton 2008] (t-Distributed Stochastic Neighbor Embedding) to visualize the high-dimensional parameter distributions of the low-rank adapters from our method and the joint training baseline. Specifically, we use t-SNE to reduce the column dimension of the low-rank adapters parameters to 2 dimensions. Fig. 9(a) depicts the parameter distribution of the low-rank adapters from our proposed method after applying t-SNE for dimensionality reduction and visualization. For a fair comparison, we set the orthogonal part of the joint training baseline’s parameters to zero to align with our methods’ parameter formulation. Fig. 9(b) shows the t-SNE visualization of the resulting joint training baseline’s low-rank adapter parameter distribution. Additionally, Fig. 9(c) presents the parameter distribution of the joint training baseline after t-SNE visualization, but without zeroing out the

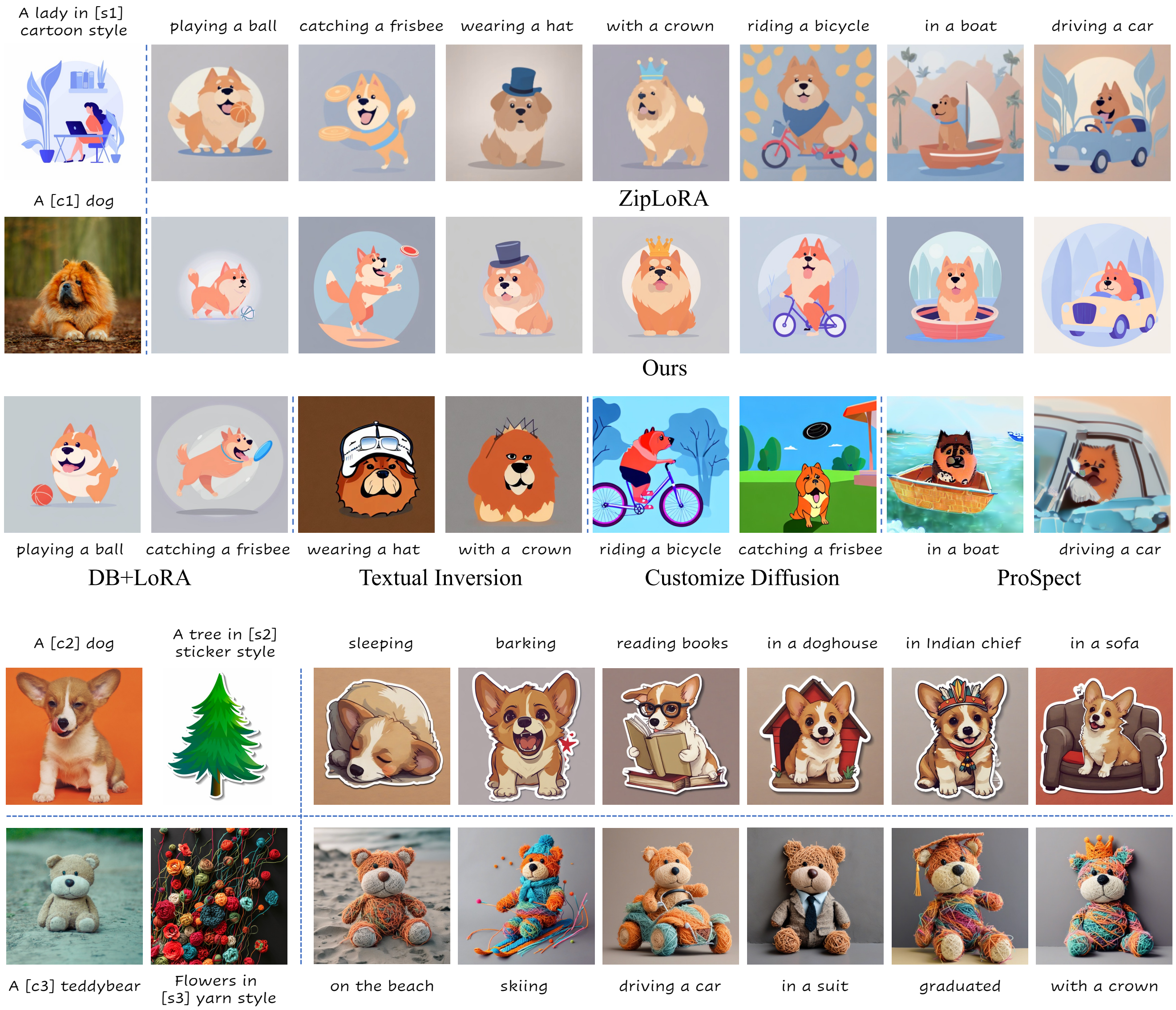
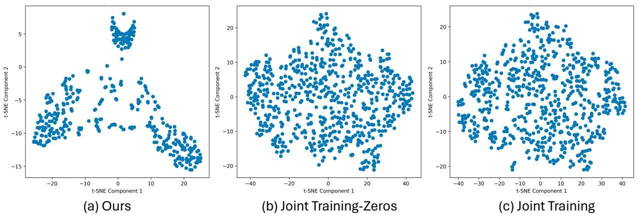


Figure 8: Results of generating diverse customized content-style images. This indicates that our method exhibits excellent editing capabilities as well as generalization capabilities to both content and style.



(a) Ours (b) Joint Training-Zeros (c) Joint Training

Figure 9: Visualizing Low-Rank Adapter Parameter Distributions via t-SNE.

orthogonal parameter components. The results indicate that, after training, compared with the joint training baseline method, our method successfully separates the parameter space of the low-rank adapters from a uniform distribution.

5.4 Concept and Style Reconstruction after First Stage Training

The proposed two-stage training paradigm employs a multi-correspondence projection learning strategy in the first stage to accurately learn the specific content and style features. Consequently, in the second stage, only a few dozen iterations of fine-tuning are required to generate customized content-style fusion images. Fig. 10

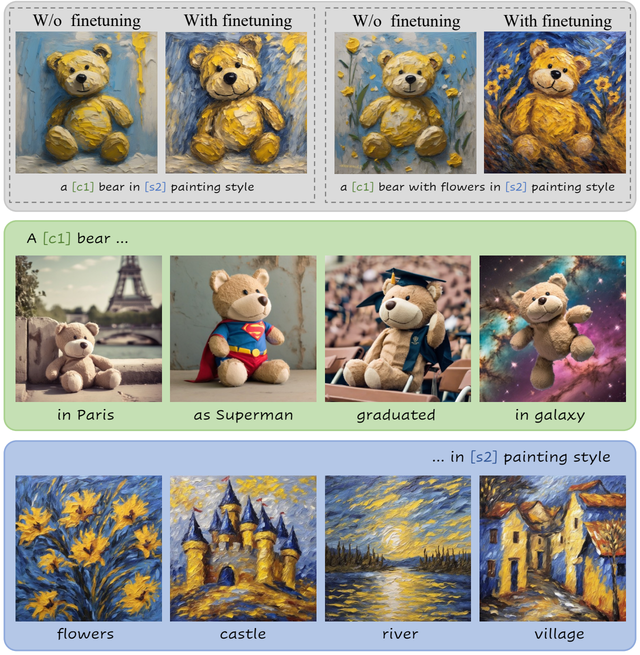


Figure 10: Individual content/style generation of our method. Our method can generate diverse content/style images individually with a high level of fidelity and disentanglement. Fine-tuning enhances the final effect.

illustrates that our approach has accurately learned the features of content or style in the first stage of training and maintained a high level of fidelity for individual content and style after the second stage of fine-tuning. Moreover, this multi-correspondence projection learning strategy prevents overfitting between content and style, thereby enabling the generation of more diverse results based on prompts. We also present images generated from directly combined adapters without fine-tuning in Fig. 10, these results verify that content and style are disentangled in the first stage and have better effects after undergoing the fine-tuning process.

5.5 Comparison with Other Two-Stage Content-Style Customization Paradigms

For the task of customized content-style image generation, we also evaluate other two-stage approaches that involve learning specific content/style in the first stage and subsequently learning or editing style/content [Avrahami et al. 2023b,c; Balaji et al. 2022; Bau et al. 2019; Brooks et al. 2023; Hertz et al. 2022; Kawar et al. 2023; Mokady et al. 2023; Parmar et al. 2023] based on the previous results in the second stage. In our experiments, we learn the content of reference images in the first stage and learn or edit style in the second stage. We leverage NULL-text Inversion [Mokady et al. 2023], a SOTA real image editing method to edit style in the second stage. The results are presented in Fig. 11. We observe that both the two-stage training and editing methods share similar drawbacks, primarily the entanglement between content and style features. For instance, when generating the "glass" style, the "teddy bear" retains plush features, and the "vase" and "teapot" retain opaque material from the content reference. In the case of the "sticker" style, these two methods only generate the contours as the "sticker" style, while the content of the sticker still reflects the realistic style depicted in the content reference image. Furthermore, the editing-based approach often necessitates complex prompts to accurately describe the features of the reference image, thereby increasing the difficulty of precisely customizing content-style generation. In contrast, our method effectively disentangles the content and style of the reference image, blending them together to generate high-quality customized content-style images without the need for complex prompts. Our approach demonstrates superior performance in achieving faithful content-style fusion compared to both the two-stage training and editing methods.

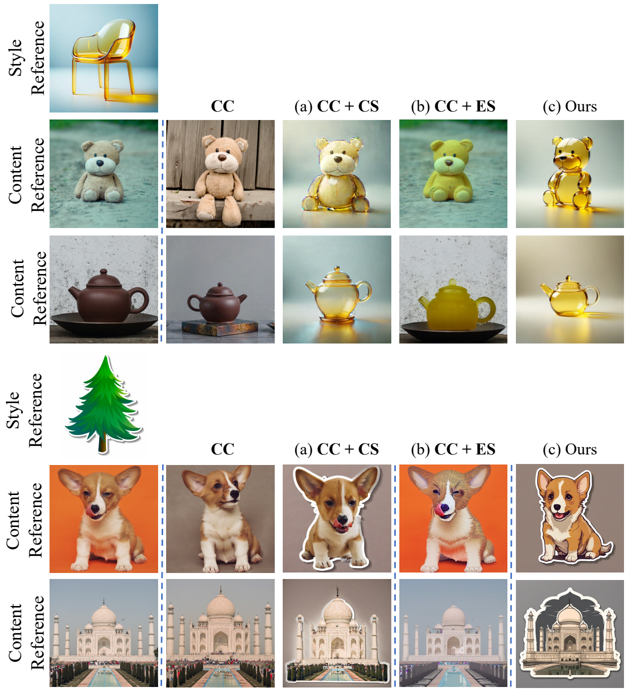
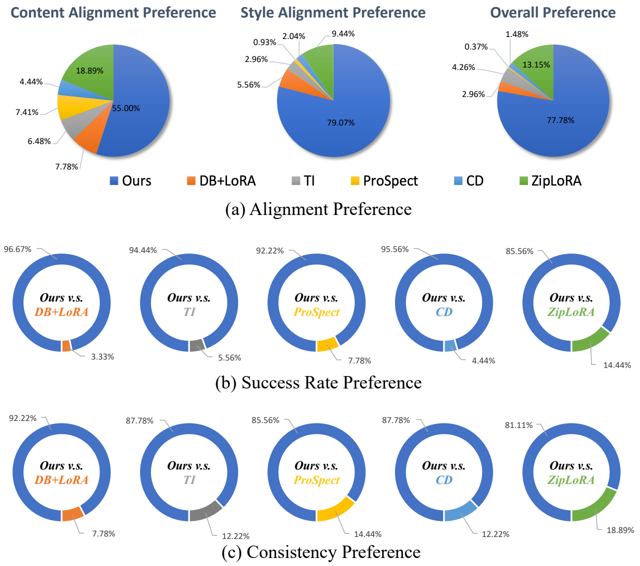


Figure 11: Comparison with other two-stage content-style customization paradigms. CC indicates custom content in the first stage, CC+CS indicates custom style in the second stage based on CC. CC+ES indicates editing style based on CC.

5.6 User Study

We conduct a user study to assess the images generated by our method and other baseline methods, employing five comparative methods: DB+LoRA, TI, ProSpect, Custom Diffusion (CD), and ZipLoRA. A total of 45 participants took part in the survey, including 20 researchers in computer graphics or computer vision. Among the participants, five are aged between 10 and 19 years old, 36 are aged between 20 and 39 years old, and four are aged between 40 and 60 years old. Additionally, there are 23 female participants and 22 male participants. The evaluation primarily



(a) Alignment Preference

(b) Success Rate Preference

(c) Consistency Preference

Figure 12: User study results.

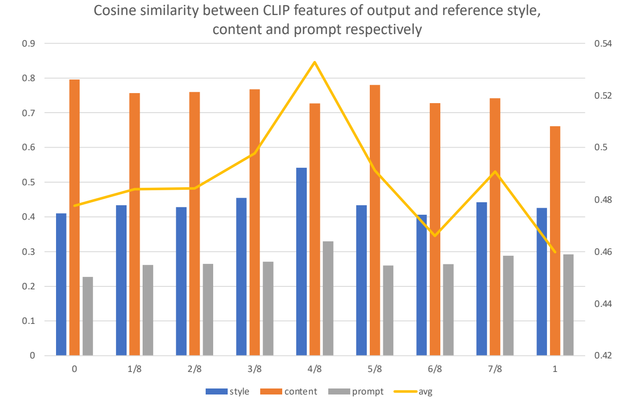


Figure 13: Cosine similarity between features of output and reference style, content and prompt of different dimension d. When d=0.5, the average cosine similarity of the features reaches its maximum, indicating optimal alignment between the generated results and the reference content, style, and prompt.

includes: I. Alignment of content between generated images and reference images; II. Alignment of style between generated images and reference images; III. Overall alignment of content and style between generated and reference images; IV. The success rate of generating images with custom content and style; IV. Stability of the generated results.

- *User Study I.* Alignment of content between generated images and reference images. One of the objectives of the content-style customization task is to ensure the alignment of content between the generated image and the reference image. In this user study, participants were tasked with selecting the image that most closely aligned with the given content reference image from six images generated using different methods (including our proposed method and baseline methods). Our method received 55.00%'s preference while DB+LoRA, TI, ProSpect, CD, ZipLoRA received 7.78%, 6.48%, 7.41%, 4.44%, 18.89%, respectively. Results are presented on the left of Fig. 12(a), indicating that our generated images have a higher level of content alignment with the reference images compared to baseline methods.
- *User Study II.* Alignment of style between generated images and reference images. We also need to evaluate the alignment of style between the generated image and the reference image. In this user study, participants were tasked with selecting the image that most closely aligned with the given style reference image from six images generated using different methods (including our proposed method and baseline methods). Our method received 79.07%'s preference while DB+LoRA, TI, ProSpect, CD, ZipLoRA received 5.56%, 2.96%, 0.93%, 2.04%, 9.44%, respectively. Results are presented in the middle of Fig. 12(a), indicating that our generated images have a higher level of style alignment with the reference images compared to baseline methods.
- *User Study III.* Overall alignment of content and style between generated images and reference images. Participants were asked to provide an overall assessment of the alignment between the generated images and both the content reference images and the style reference images, selecting the most fitting results. Our method received 77.78%'s preference while DB+LoRA, TI, ProSpect, CD, ZipLoRA received 2.96%, 4.26%, 0.37%, 1.48%, 13.15%, respectively. Results are presented in the right of Fig. 12 (a), indicating that, overall, our method aligns with both the given content and style reference images simultaneously.
- *User Study IV.* We conduct A/B testing to evaluate the success rate of generating content-style customized images between our method and other baseline methods. During the test, one of the five baseline methods is randomly selected for comparison with our method. Both methods generate nine images with different seeds. Participants were asked to select which set of nine images contained more content-style customized generated images. Our method received 96.67%'s preference while compared with DB+LoRA, 94.44%'s preference while compared with TI, 92.22%'s preference while compared with ProSpect, 95.56%'s preference while compared with CD, and 85.56%'s preference while compared with ZipLoRA, respectively. Results are shown in Fig. 12 (b) The results indicate that our method generates a greater number of content-style customized images compared to other methods, suggesting a higher success rate in satisfied image generation.
- *User Study V.* Similar to the experiment settings in User Study IV, participants were tasked with selecting which set of nine images exhibited stronger consistency among them. Our method received 92.22%'s preference while compared with DB+LoRA, 87.78%'s preference while compared with TI, 85.56%'s preference while compared with ProSpect, 87.78%'s preference while compared with CD, and 81.11%'s

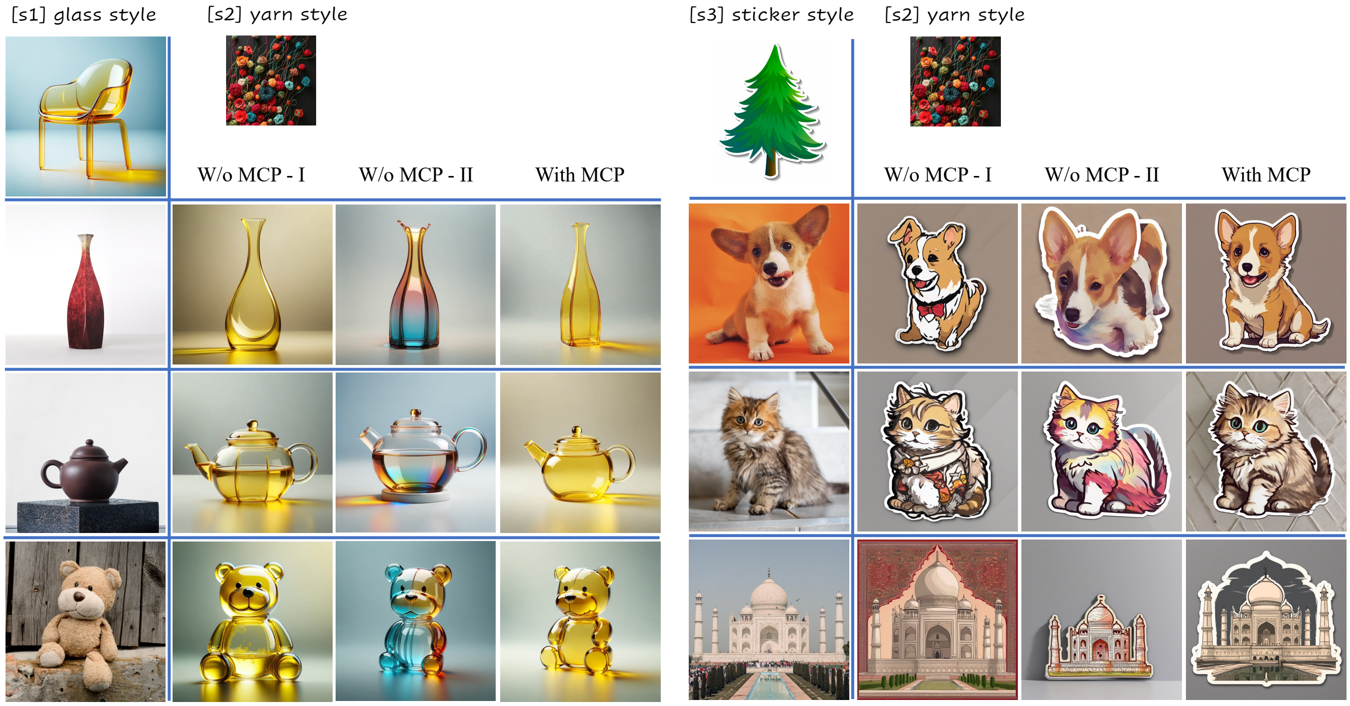


Figure 14: Ablation study evaluating the impact of the proposed Multi-Correspondence Projection ("MCP"). We train specific content (e.g., "vase") and style (e.g., "glass") in a one-to-one manner and directly inference after training. Results are presented in W/o MCP - I column. We train specific content (e.g., "vase") and style (e.g., "yarn") in a one-to-one manner in the first stage, and combine the content (e.g., "vase") adapters with other style (e.g., "glass") adapters in the second stage, then inference with the combined adapters. Results are presented in W/o MCP - II column. The visual comparison highlights the effectiveness of MCP in enhancing the details while preserving the disentanglement of content and style, as well as maintaining high-level fidelity of them.

preference while compared with ZipLoRA, respectively. The results indicate that our method generates a more significant number of content-style customized images compared to the other methods, suggesting a higher success rate in satisfied image generation. Results are shown in Fig. 12 (c) The results demonstrate that the images generated by our method exhibit stronger consistency among them, indicating that our method, comparatively, has the highest level of stability.

5.7 Ablation Study

The Optimal Dimension *d* for the Fixed Parameters. In Section 4, we introduce the hyperparameter d as the row dimension of the fixed parameters, representing the proportion of fixed parameters in the parameter subspace. We conduct experiments on eight different fixed parameter ratios, $1/8$, $1/4$, $3/8$, $1/2$, $5/8$, $3/4$, $7/8$, and 1, corresponding to the proportion of fixed parameters relative to the total parameters. We quantitatively evaluate the text alignment, content alignment, and style alignment metrics for various values of d, and the results are presented in Fig. 13. From the histogram, we observe that as the ratio of fixed parameters increases, both the values of "Style Alignment" and "Text Alignment" gradually rise, reaching their peaks at a ratio of 0.5, and then gradually decline. The "Average Alignment" reaches its maximum at a ratio of 0.5. This indicates that the optimal alignment occurs at a ratio of 0.5, resulting in better customized content-style images. This finding aligns with our theoretical framework introduced in Section 4, where a 1:1 ratio between fixed and trainable parameters results in the "content parameter subspace" in Eq. (12) and "style parameter subspace" in Eq. 11) having the maximum number of trainable parameters, thus reaching the maximum learning capacity and achieving the best generation effect. It is noteworthy that at a ratio of 0.5, the "Content Alignment" is not maximal. This is because the results of other ratios present a weaker learned style (as indicated by lower "Style Alignment" in the histogram) and are entangled with the content to some extent.

Multi-Correspondence Projection Learning. To prevent overfitting between specific content and style during the training stage, we introduce a **Multi-Correspondence Projection** ("**MCP**") learning within our work. We conduct two ablation studies with different experimental settings to evaluate the impact of the proposed MCP. Specifically, in the first study, we train the specific content (e.g., "vase") with a particular style (e.g., "glass style") in a one-to-one manner and leverage the trained model for inference. The results are shown as **W/o MCP-I** in Fig. 14. In the second study, we first train the model on a specific content (e.g., "vase") with a different

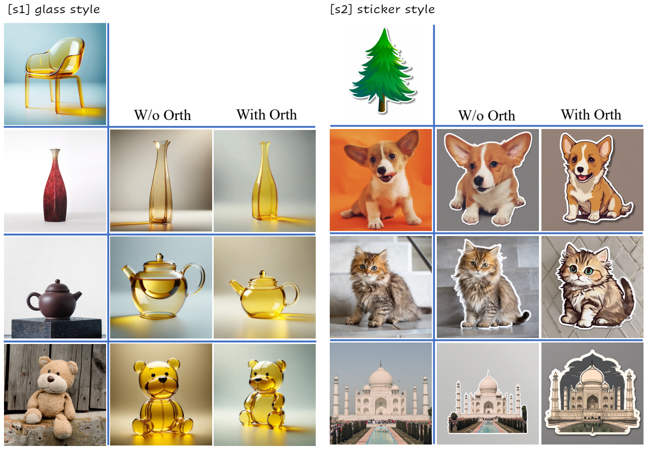


Figure 15: Ablation study evaluating the impact of the proposed orthogonal fixed parameters. The W/o Orth shows results without orthogonal fixed parameters, while the With Orth demonstrates the improved image quality achieved by our full method incorporating orthogonal fixed parameters. The visual comparison highlights the effectiveness of orthogonal fixed parameters in enhancing content and style fidelity of generated images.

style (e.g., “yarn style”) in a one-to-one manner. Subsequently, in the second stage, we combine the content adapters with the trained style (e.g., “glass style”) adapters and utilized the combined model for inference. The results are shown as **W/o MCP-II** in Fig. 14. We can observe that in the first study, the generated images exhibit a degree of overfitting to the reference images(e.g., “teapot” with “chair legs” from the style reference image, “sticker” style with realistic style from the content reference image), resulting in a decrease in fidelity to the content or style, thereby reducing the quality of the outputs. In the second study, we can observe that the results exhibit some features (e.g., the color from “yarn style”) of the style trained in the first stage. As the final model does not incorporate this style, this is mainly because due to the fact that without MCP, the “yarn style” influences the parameter space of the content during the training stage, as analyzed in Fig. 4. We present the results of our methods in **With MCP**; by comparing, we can observe that with MCP, we can effectively avoid overfitting and generate images with more disentangled content and style.

Orthogonal Fixed Parameters. To demonstrate the effectiveness of the orthogonal fixed parameters designed to enhance the content and style fidelity of generated images, we conduct an experiment where we remove the orthogonal fixed parameters and replace them with randomly fixed parameters. We present the results in Fig.15 for comparison. Without the orthogonality of the fixed parameters, it leads to decreased fidelity for the generated images. For instance, in the case of “vase”, “teapot”, and “teddy bear”, the generated images no longer preserve the original content details, and the style has also changed. In the case of the “sticker style”, the generated images lose the cartoonish style of the contents present in the reference. We also present quantitative results in Table. 1. After ablating fixed parameter orthogonalization, although the content alignment slightly increases, the style alignment decreases significantly, and the average alignment decreases as well. Note that the slight increase in text alignment is due to the increase in content alignment, as the prompts’ emphasis on describing the image content.

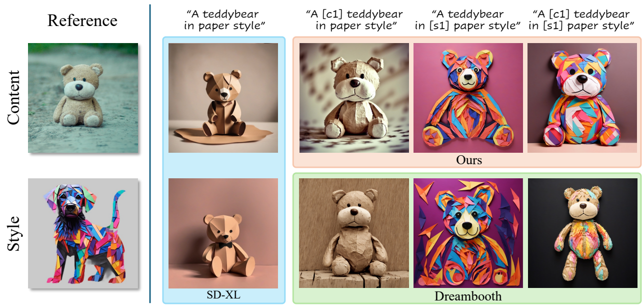

Reference
Content
Style
“A teddybear in paper style”
“A [c1] teddybear in paper style”
“A teddybear in [s1] paper style”
“A [c1] teddybear in [s1] paper style”
SD-XL
Ours
Dreambooth

Figure 16: Comparison of results with and without learning concepts. We present output images generated with and without learning reference content or style in in the orange (by our method) and green (by DreamBooth method) boxes. We also show images directly generated by basic Stable Diffusion-XL model in blue box. Prompts for inference are shown on top. Without learning content or style in pseudo words, models that rely solely on prompts cannot generate desired content or styles faithfully.

Fine-Tuning of the Combined LoRA Modules. In the second stage of our pipeline, we reconstruct the entity parameter space by combining the content and style PLP matrices. Subsequently, we fine-tune the combined LoRA modules for a few dozen steps to enable the model to generate images with customized content with style. The results of ablating the fine-tuning step are presented in Fig. 10. The results demonstrate that after undergoing a few dozen fine-tuning steps, our proposed method achieves optimal visual performance. We also present individual content or style generation results in the middle and bottom rows of Fig. 10. These results illustrate that our proposed method successfully disentangles content and style while retaining the capability to faithfully generate individual content or style.

Concepts learning ablation. We aim to evaluate the learning effect of the desired content or style in comparison with the baseline stable diffusion model. To achieve this, we employ pseudo words for training and inference of specific content and style. For the purpose of comparison, we describe content and style using prompts for generation. As the results presented in Fig. 16, solely relying on prompts to describe the desired content or style, without learning these representations, fails to capture detailed features from reference images, leads to unfaithful generation of the customized content and style.

6 APPLICATIONS AND DISCUSSIONS

We demonstrate the effectiveness and versatility of our technique across various applications, including content-style customization of diverse textures and portraits.

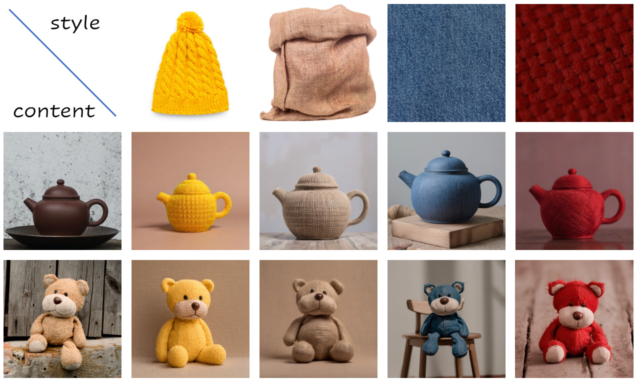


Figure 17: Application I. Content-style customization of variety texture, including knit, burlap, denim and fabric texture.

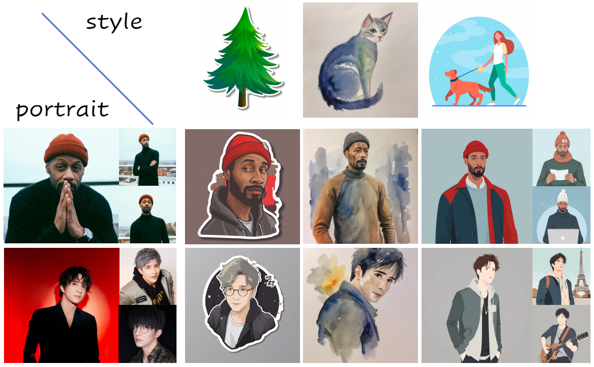


Figure 18: Application II. Content-style customization of portraits. Image credits:@Philip Martin [Philip 2023](up)

Application-I: Content-Style Customization of Various Textures. Our technique enables the synthesis of high-quality content with a wide variety of user-controlled textures and materials, which can be leveraged for customized product visualization, digital content creation, or material design applications. We present results for different textures (knit texture, burlap texture, denim texture, and fabric texture) in Fig. 17. The visualized results indicate that our method is capable of customizing generation for a diverse range of textures while maintaining content consistency with the reference images. With our approach, designers can easily showcase their products with custom material and textile options tailored to customer preferences. Compared to traditional rendering pipelines requiring extensive modeling and material setup, our data-driven approach significantly streamlines this process.

Application-II: Content-Style Customization of Portraits. Another compelling application of our technique is enabling users to generate stylized portraits adhering to diverse artistic styles and visual domains. This capability opens up new creative avenues for digital artists, as well as opportunities in areas like virtual production and AI-assisted artwork creation. For digital artists and creative professionals, our framework efficiently synthesizes portrait imagery in various artistic styles with fine user control. Fig. 18 illustrates examples where we tasked artists to create stylized portraits using our approach in styles like sticker, watercolor painting, and flat cartoons. Compared to manual digital painting, our approach dramatically accelerates this creative process while still allowing users to guide stylistic aspects and maintain consistent facial identities. A key advantage of our approach is its ability to generalize stylized portrait synthesis across numerous visual domains while still allowing users to control diverse scenes, poses, etc.

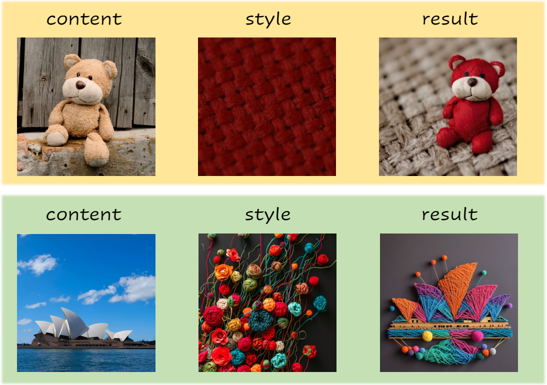


Figure 19: Bad cases of our method. Sometimes style reference can induce undesirable influences on the background generation in the output results.

Bad Cases. While our proposed method demonstrates considerable promise in addressing customized content and style fusion, it is essential to acknowledge instances where the model occasionally exhibits an influence from the style reference on the background regions of the generated images, as presented in Fig. 19. This observed influence on the background regions in the generated results can be attributed to the limited diversity among the style reference images. To mitigate this issue, future iterations of our method will incorporate rigorous regularization techniques and enhanced data preprocessing methodologies. Furthermore, the integration of cross-validation procedures and model simplification strategies will be explored to promote improved generalization performance.

Limitations. While our method performs well on content-style customization, generating images with complex or rare content/style solely by using textual prompts remains a challenging task. Specifically, our method leverages the class priors in the T2I model when learning the content or style of given images (*e.g.*, “a [c1] dog” leverages “dog” as a class prior, “a [s1] yarn style” leverages “yarn” as a class prior) [Ruiz et al. 2023]. When the customized content or style images are highly complex or rare, obtaining accurate priors through simple prompts becomes challenging, leading to a decrease in the fidelity of the generated images.

REFERENCES

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–10.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023a. Break-A-Scene: Extracting Multiple Concepts from a Single Image. In *SIGGRAPH Asia 2023 Conference Papers* (, Sydney, NSW, Australia,) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 96, 12 pages. https://doi.org/10.1145/3610548.3618154

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023b. Blended latent diffusion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–11.

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023c. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18370–18380. https://doi.org/10.1109/CVPR52729.2023.01762

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

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–11.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2 (2023), 3.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 4055–4075.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. 2024. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems* 36 (2024).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems* 36 (2024).

Ziyi Dong, Pengxu Wei, and Liang Lin. 2022. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337* (2022).

Ali Edalati, Marzieh Tahaei, Ivan Kobyzev, Vahid Partovi Nia, James J Clark, and Mehdi Rezagholizadeh. 2022. Krona: Parameter efficient tuning with kronecker adapter. *arXiv preprint arXiv:2212.10650* (2022).

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

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

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305* (2023).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

Nisha Huang, Weiming Dong, Yuxin Zhang, Fan Tang, Ronghui Li, Chongyang Ma, Xiu Li, and Changsheng Xu. 2024. CreativeSynth: Creative Blending and Synthesis of Visual Arts based on Multimodal Diffusion. *arXiv preprint arXiv:2401.14066* (2024).

Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. 2021. Fedpara: Low-rank hadamard product for communication-efficient federated learning. *arXiv preprint arXiv:2108.06098* (2021).

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023. Cones 2: Customizable Image Synthesis with Multiple Subjects. *arXiv preprint arXiv:2305.19327* (2023).

Midjourney. 2023. Midjourney. https://www.midjourney.com/.

mkshing. 2023. ZiploRA. https://github.com/mkshing/ziplora-pytorch.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. (May 2023).

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterFusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247* (2020).

Martin Philip. 2023. unsplash. https://unsplash.com/@phlmrtn.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. 2023. Orthogonal adaptation for modular customization of diffusion models. *arXiv preprint arXiv:2312.02432* (2023).

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 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. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. 2023. ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs. *arXiv preprint arXiv:2311.13600* (2023).

Ryu Simo. 2023. LoRA. https://github.com/cloneofsimo/lora.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. 2023. StyleDrop: Text-to-Image Generation in Any Style. *arXiv preprint arXiv:2306.00983* (2023).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

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*. 1–11.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2022. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558* (2022).

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

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).

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848* (2023).

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Resolving Interference When Merging Models. *arXiv preprint arXiv:2306.01708* (2023).

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512* (2023).

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023b. ProSpect: Expanded Conditioning for the Personalization of Attribute-aware Image Generation. *arXiv preprint arXiv:2305.16225* (2023).

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023c. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10156.