B4M: Breaking Low-Rank Adapter for Making Content-Style Customization

YU XU, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of the Chinese Academy of Sciences, Beijing, China
FAN TANG and JUAN CAO, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
YUXIN ZHANG, Institute of Automation, Chinese Academy of Sciences, Beijing, China
OLIVER DEUSSEN, University of Konstanz, Konstanz, Germany
WEIMING DONG, Institute of Automation, Chinese Academy of Sciences, Beijing, China
JINTAO LI, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
TONG-YEE LEE, National Cheng Kung University, Tainan, Taiwan

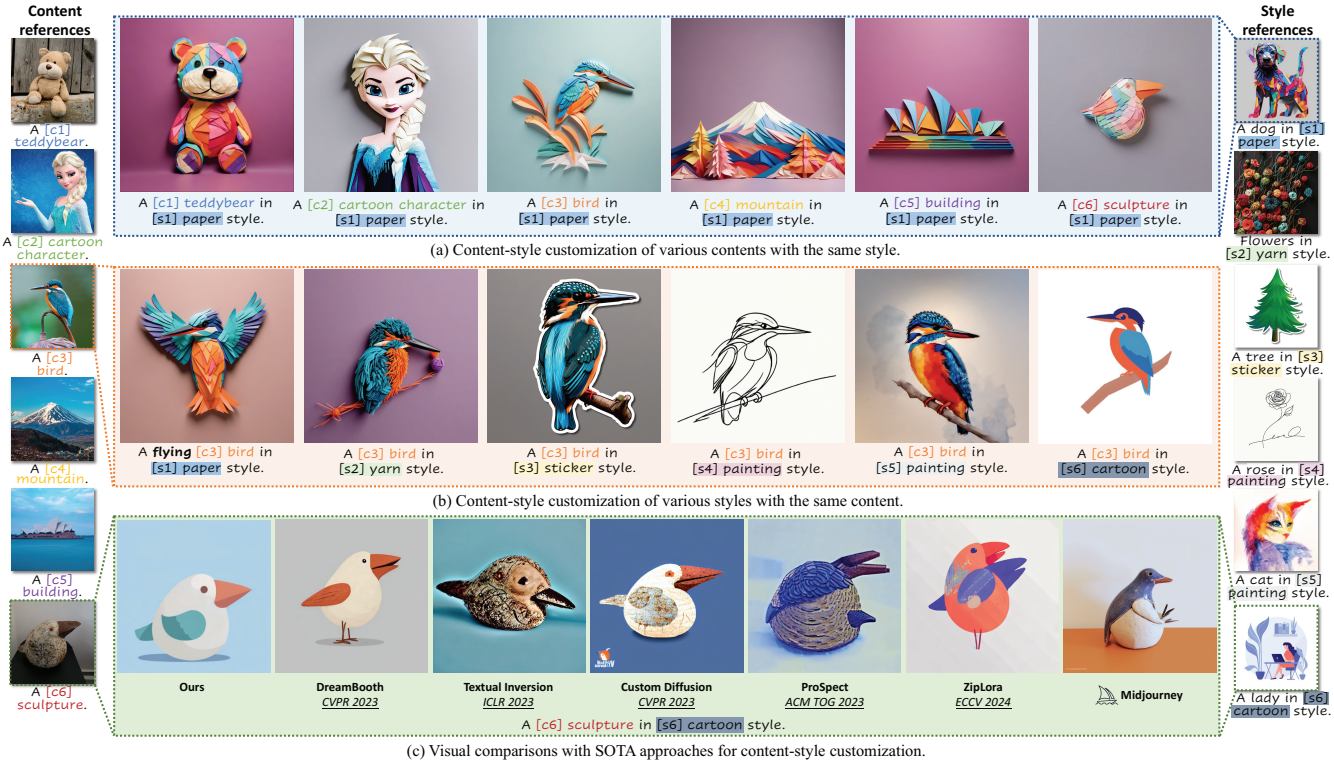

Fig. 1. By separately learning content and style in "partly learnable projection" (**PLP**), our method is able to generate images of customized content and style aligned with various prompts while successfully disentangling content and style and maintaining high fidelity. We use the "\blend" instruction in Midjourney for customized content-style generation.

This work was partly supported by the Beijing Science and Technology Plan Project under grant no. Z231100005923033, the National Science and Technology Council under grant no. 113-2221-E-006-161-MY3, Taiwan, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC 2117–422037984.

Authors' Contact Information: Yu Xu, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of the Chinese Academy of Sciences, Beijing, China; e-mail: xuyu21b@ict.ac.cn; Fan Tang (Corresponding author), Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China; e-mail: tfan.108@gmail.com; Juan Cao, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China; e-mail: caojuan@ict.ac.cn; Yuxin Zhang, Institute of Automation, Chinese Academy of Sciences, Beijing, China; e-mail: zhangyuxin2020@ia.ac.cn; Oliver Deussen, University of Konstanz, Konstanz, Baden-Württemberg, Germany; e-mail: oliver.deussen@uni-konstanz.de; Weiming Dong, Institute of Automation, Chinese Academy of Sciences, Beijing, China; e-mail: weiming.dong@ia.ac.cn; Jintao Li, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China; e-mail: jtli@ict.ac.cn; Tong-Yee Lee, National Cheng Kung University, Taiwan, Taiwan; e-mail: tonylee@mail.ncku.edu.tw.

ACM 0730-0301/2025/04-ART21
https://doi.org/10.1145/3728461

Personalized generation paradigms empower designers to customize visual intellectual property with the help of textual descriptions by adapting pre-trained text-to-image models on a few images. Recent studies focus on simultaneously customizing content and detailed visual style in images but often struggle with entangling the two. 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 learning, we propose a novel framework that separates the parameter space to facilitate individual learning of content and style by introducing "partly learnable projection" (PLP) matrices to separate the original adapters into divided sub-parameter spaces. A "**break-for-make**" customization learning pipeline based on PLP is proposed: we first **break** the original adapters into "up projection" and "down projection" for content and style concept under orthogonal prior and then **make** the entity parameter space by reconstructing the content and style PLP matrices by using Riemannian preconditioning to adaptively balance content and style learning. Experiments on various styles, including textures, materials, and artistic style, show that our method outperforms state-of-the-art single/multiple concept learning pipelines regarding content-style-prompt alignment. Code is available at https://github.com/ICTMCG/Break-for-make.

CCS Concepts: • **Computing methodologies** → **Image manipulation;**

Additional Key Words and Phrases: Customize generation, content-style fusion, text-to-image generation

ACM Reference Format:
Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Oliver Deussen, Weiming Dong, Jintao Li, and Tong-Yee Lee. 2025. **B4M**: **B**reaking Low-Rank Adapter **for M**aking Content-Style Customization. *ACM Trans. Graph.* 44, 2, Article 21 (April 2025), 17 pages. https://doi.org/10.1145/3728461

1 Introduction

Text-to-image (T2I) models based on diffusion technology [Ho et al. 2020; Ho and Salimans 2021; 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 their outstanding generalization ability, it is challenging for users to generate specific visual concepts using only textual descriptions with T2I models.

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 images (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 [Avrahami et al. 2023; Kumari et al. 2023] aims to learn and generate different contents by manipulating or constraining cross-attention mechanisms. However, the intricate nature of visual style, which is often entangled with content, poses challenges in effectively decoupling content and style concepts due to their shared parameter space and lack of disentanglement strategies employed by these methods. Recent work, such as ZipLoRA [Shah et al. 2024], 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 notable performance degradation of the merged model, leading to ineffective preservation of the distinct qualities of both content and style [Yadav et al. 2023]. B-LoRA [Frenkel et al. 2024], published concurrently with the present work, proposes a method for stylizing specific content by decomposing images into style and content representations and optimizing different LoRA layers. While this approach separates style and content in an impressive way, it poses limitations in generating customized content and style aligned with various prompts. Therefore, a method is needed that decouples the learning of content and style and recombines them in a generation process without interference.

Here, we introduce a two-stage learning approach for customized content-style generation, which we call "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 a unified parameter space by combining the content and style PLP matrices, followed by fine-tuning the integrated adapter to achieve content-style customized results. To address the challenges of balancing content and style learning while mitigating concept overfitting and leakage from reference images, we introduce a novel Riemannian preconditioning approach. This method adaptively scales gradients of both content and style PLP during the fine-tuning process and balances the learning of content and style features. We present our results in Figures 1(a) and 1(b); these results demonstrate our method's capability to generate high-quality, customized images that faithfully adhere to both content and style features across diverse content references and style references.

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-Correspondence Projection (MCP) mechanism is proposed to maintain generalization and Riemannian preconditioning is proposed to balance the content and style training process.
- 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 T2I generation [Betker et al. 2023; Chang et al. 2023; Rombach et al. 2022; Saharia et al. 2022]. T2I 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. 2023; 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. 2023; Edalati et al. 2022; Hyeon-Woo et al. 2021; Valipour et al. 2023; Zhang et al. 2023a]. Po et al. [2024] design multiple LoRAs to separately train different content and generate multiple contents simultaneously in one image. By integrating adapter modules, AdapterFusion [Pfeiffer et al. 2021]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 instances of 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 content–style customization is to generate an image that incorporates specific content and style based on reference images while ensuring that 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. Instruct-Imagen [Hu et al. 2024] proposes a model that tackles content–style image generation by fine-tuning pre-trained models with retrieval-augmented training and multi-modal instruction-tuning. 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 featuring ZipLoRA [Shah et al. 2024] learns hybrid coefficients to optimize conflicts arising when merging two separately trained LoRAs, partially mitigating entanglement issues. However, it concurrently modifies the distribution of learned parameters, subsequently influencing reconstruction outcomes. Compared with related approaches, our proposed "partly learnable projection", "multi-correspondence projection learning", and "Riemannian Preconditioning" strategies train content and style separately in different sub-parameter spaces within low-rank adaptations with data augmentation to disentangle content and style information.

3 Vanilla Solutions for Content-Style Customization

In this section, we first introduce the task definition of content–style customization in image generation. Then, based on the typical customization method, low-rank adaptation fine-tuning [Hu et al. 2021], we introduce and discuss initial solutions through joint training or merging after independent training. 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.

Formulation. Content–style customization aims to generate images that effectively present user-specified content and style while ensuring that their unique characteristics are distinctively represented [Shah et al. 2024; Zhang et al. 2023b]. Formally, given one or a few content reference images I_c, style reference images I_s, and

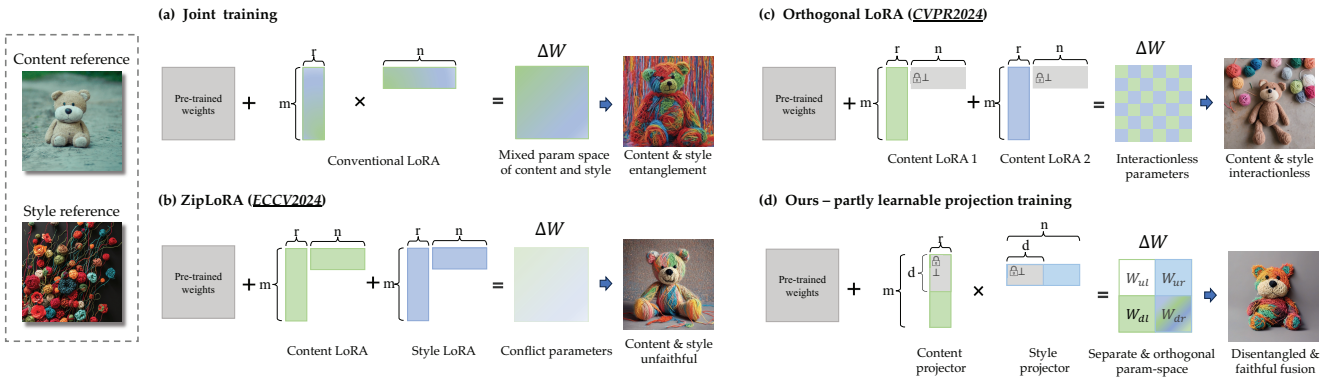

Fig. 2. Frameworks of existing approaches and ours for customized content–style image generation. Joint training LoRA will mix the parameter space of content and style, leading to the entanglement of both. ZipLoRA [Shah et al. 2024] effectively merges independently trained content and style LoRAs. However, the conflicting parameters between the content LoRA and style LoRA can lead to unfaithful reproduction of content and/or style after fusion. Orthogonal LoRA [Po et al. 2024] focuses on multi-subject customization by learning orthogonal LoRAs for each subject. These adapters are composed in a 'continual learning' manner, in which different concepts do not influence each other. However, this approach leads to a failure in content–style fusion. Our method trains content and style in separated parameter subspaces of LoRA, resulting in a disentangled and faithful fusion of content and style.

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.

Given a weight matrix $W_0 \in \mathbb{R}^{m\times n}$ of UNet [Ronneberger et al. 2015] for a pre-trained diffusion model, 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 forward pass is

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

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 training, we can directly merge the LoRA modules with the pre-trained weight matrix and obtain new weights $W = W_0 + \Delta W$, which can perform inference as usual.

Joint 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:

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

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. 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 Figure 2(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. In the second stage, these modules are merged with certain constraints. 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, merging independent LoRAs can cause parameter conflicts, in which influential values from one LoRA may be obscured by redundant values from the other, reducing overall effectiveness. ZipLoRA [Shah et al. 2024] 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, as analyzed in Figure 2(b). A recent work [Po et al. 2024] focuses on multi-content customization by learning orthogonal LoRAs for each content. The adaptors are composed in a "continual learning" manner—different concepts do not influence each other. Thus, it is effective for distinct objects/persons. However, such design exhibits conceptual fusion when overlapping elements (i.e., content and style) are presented due to a lack of feature/weight fusion/interaction. On the other hand, orthogonal LoRA controls the disentangled generation of two contents by designing two orthogonal up projections. In contrast, our approach avoids interference from redundant parameters by fixing a portion of the orthogonal initialization in both the up and down projections, as analyzed in Figure 2(c). This motivates us to pursue separate training for content and style, subsequently integrating concepts with harmonious fusion, 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 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 MCP, a technique for training content and style representations during the customization process to mitigate overfitting between the two. Finally, we introduce **"Riemannian Preconditioning" (RP)**, a technique that adaptively balances the learning of content and style PLP by taking consideration of gradient steps of each other. By utilizing the proposed PLP, MCP, and RP 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 and fusion.

4.1 Partly Learnable Projection

To address the aforementioned issues, we propose 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

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

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 that

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

$$= \begin{bmatrix} \sum_r A_{i,r} C_{r,j} & \sum_r A_{i,r} D_{r,j} \\ \sum_r B_{i,r} C_{r,j} & \sum_r B_{i,r} D_{r,j} \end{bmatrix}, \tag{9}$$

where 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, which is further discussed 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 Equations (5) and (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:

$$\sum_r A_{i,r} C_{r,j} = 0. \tag{10}$$

The upper-right part of ΔW in Equation (4.1) represents **only** the parameters of submatrix D. Similarly, the lower-left part of ΔW in Equation (4.1) represents **only** the parameters of submatrix B, and the lower-right part of ΔW in Equation (4.1) relates to B and D, allowing us to learn interactive features between them.

The forward pass during training yields

$$z_{out} = W_0 z + \begin{bmatrix} 0 & \sum_r A_{i,r} D_{r,j} \\ \sum_r B_{i,r} C_{r,j} & \sum_r B_{i,r} D_{r,j} \end{bmatrix} z, \tag{11}$$

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

Our partitioned matrices method separates content and style parameters, allowing input features to multiply with corresponding parameters during training. This distinctly represents content and style in different parameter subspaces, mitigating entanglement while maintaining high fidelity.

As shown in Figure 2(d), after separating the LoRA module and performing forward matrix multiplication, the resulting partitioned matrices exhibit a zero top-left part due to orthogonal vector multiplication, a top-right style submatrix, a bottom-left content submatrix, and a bottom-right part for learning interactive feature parameters. This approach avoids parameter conflicts from merging methods and achieves disentangled content and style representations. The interactive parameters enable the generation of naturalistic fusion images with high visual quality. Additional visualization analyses of the LoRA parameter spaces, with and without the application of our proposed method, are provided in the supplementary material.

4.2 Multi-Correspondence Projection Learning

When training a one-to-one mapping between a specific content and style, the content and style distribution tends to drift away from the desired representation, leading to potential overfitting issues and 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 an MCP learning method involving diversified content–style training data pairs. Specifically, when training for a particular content, we update the parameters of B in Equation (5) with the particular content image and update the parameters of D in Equation (7) with various style images and vice versa. In simple terms, a LoRA model W for learning specific content is trained using a squared error loss function as follows:

$$L = \left[\left\| \hat{W}_\theta(z_c | c_c, t) - x_c \right\|_2^2 \right] + \frac{1}{n} \cdot \sum_{i=1}^{n} \left[\left\| \hat{W}_\phi(z_s | c_s, t) - x_s \right\|_2^2 \right], \tag{12}$$

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 Equation (12). This training approach prevents overfitting issues that arise when learning specific content–style pairs, simultaneously enhancing the method's generalization ability and improving the effectiveness of diverse content–style combinations, as illustrated in Figure 3. After the first stage training, we obtain $LoRA_c$ and $LoRA_s$, containing learned parameters for specific content and style, respectively. We then reconstruct $LoRA_f$ as fusion adapters by combining the up-projection part of $LoRA_c$ with the down-projection part of $LoRA_s$.

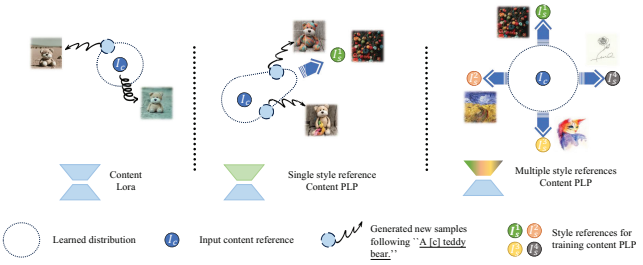

Fig. 3. Illustration of the multi-correspondence projection. We present the learned content distribution on the left of the top row. When training specific content and style in a one-to-one manner, the content will tend to overfit to the specific style, as illustrated in the middle of the top row. By leveraging our proposed multi-correspondence projection, we learn multiple styles with the content in PLP and enhance the generalization of the learned content.

4.3 Gradient Scaling with Riemannian Preconditioning

We fine-tune $LoRA_f$ in the second stage with a few dozen steps for better fusion of content and style. During fine-tuning, we observe a disparity in convergence rates between content and style PLPs. This discrepancy often leads to concept overfitting and concept leakage of either content or style. Overfitting can also result in learned content or style representations that fail to generate images well aligned with diverse prompts. Drawing from the Riemannian metric, which simultaneously incorporates the objective function and constraints within the matrix optimization [Mishra and Sepulchre 2016; Zhang and Pilanci 2024], we introduce a dynamic preconditioner in each gradient step of PLP to adaptively balance the learning of content and style features. Specifically, we scale the gradients of content PLP and style PLP by employing an $r \times r$ preconditioner for including the information of both; the parameter update is formulated as

$$M_{t+1} = M_t - \alpha \left(N_t^T N_t\right)^{-1} \left(\nabla_{M_t} \mathcal{L}\right), \tag{13}$$

$$N_{t+1} = N_t - \alpha \left(\nabla_{N_t} \mathcal{L}\right) \left(M_t M_t^T\right)^{-1}, \tag{14}$$

where M and N represent the learnable part of style PLP and content PLP, respectively. $(M_t M_t^T)^{-1}$ and $(N_t^T N_t)^{-1}$ are our introduced preconditioners, α is the learning rate, $\mathcal{L}$ is the training loss, and t is the gradient step.

Compared with the standard AdamW optimizer [Loshchilov and Hutter 2019], Riemannian preconditioning facilitates beneficial information exchange between the up and down projections during training. This exchange ensures that each partition is not isolated but rather is informed by the other, leading to a more holistic and effective training process, thereby improving the overall quality and generalizability of the customized model, as illustrated in Figure 14. Our method not only enhances the stability of LoRA training but also yields more coherent and diverse generations that better capture the intended content and style features.

5 Experiments

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. 2024; Zhang et al. 2023b] together with diverse images from the Internet. Our datasets include 30 content types and 20 style types. For training content PLP, we collect three to five images of the same content and five different styles, each style consisting of one image. For training style PLP, we collect one to three images of the same style and five different contents, each content consisting of one image. The influence of the number of references is discussed in the supplementary material.

Compared Methods. We compare our method against four state-of-the-art customization approaches: **textual inversion (TI)** [Gal et al. 2022], **ProSpect** [Zhang et al. 2023b], **custom diffusion (CD)** [Kumari et al. 2023] and **ZipLoRA** [Shah et al. 2024]. TI and ProSpect are based on prompt tuning for frozen T2I modes, meaning they can directly merge different concepts by operating prompts. CD extends DreamBooth to learning multiple concepts. ZipLoRA is the representative work for merging after independent training. As official ZipLoRA codes have not yet been released, we adopt a popular implementation [mkshing 2023], which initially trains the content and style models separately and performs LoRA merging. Furthermore, we implement the **joint training (JTtrain)** fashion following [Simo 2023], where both DreamBooth and LoRA are adopted for learning content and style concepts into one model together.

Metrics. For quantitative comparisons, we mainly assess three metrics: **content alignment** and **style alignment** between the generated images and reference images, as well as **prompt alignment** between the generated images and the corresponding prompts. Following quantitative experiment settings of ProSpect [Zhang et al. 2023b] and ZipLoRA [Shah et al. 2024], we compare cosine similarities between CLIP [Ilharco et al. 2021] features for calculating style and prompt alignment and DINOv2 [Oquab et al. 2023] features for content.

Implementation Details. In our experiments, we utilize 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 of the same input and output feature dimensions, which means that m in Equation (5) equals n in Equation (7). 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. In the supplementary material, we showcase more of our results with diverse content and styles.

Qualitative Comparison. We first present results of generating the same content image with multiple style images in Figure 4. Then, we present the same style image with multiple content images in Figure 5. Results indicate that our methods successfully disentangle content and style in one image while maintaining a high level of fidelity. JTtrain usually generates images of unnatural content style fusion (the result of "mountain" with "yarn style" and "oil painting style" in Figure 4) and images of the mixed style ("vase" and "teapot" with "glass style" in Figure 5). The observed entanglement phenomenon aligns with the analysis presented in Section 3.

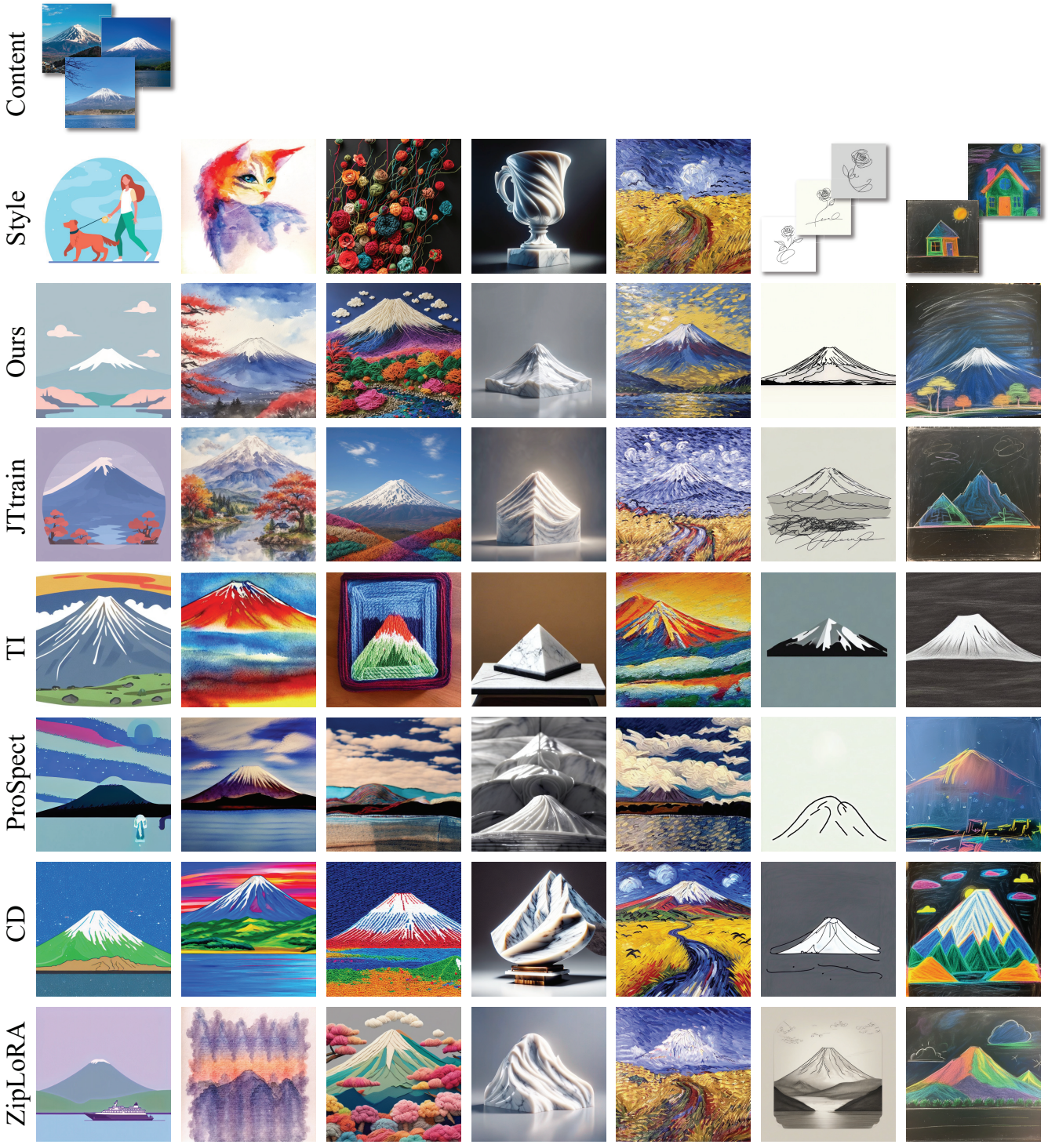

Fig. 4. Qualitative evaluation and comparison of DB+LoRA, TI, ProSpect, CD, ZipLoRA, and our method in diverse styles. We present the results of customized generation of the same content and different styles. Results indicate that our method generates harmonious fusion images of the content and style while preserving the disentanglement of content and style, as well as maintaining high-level fidelity.

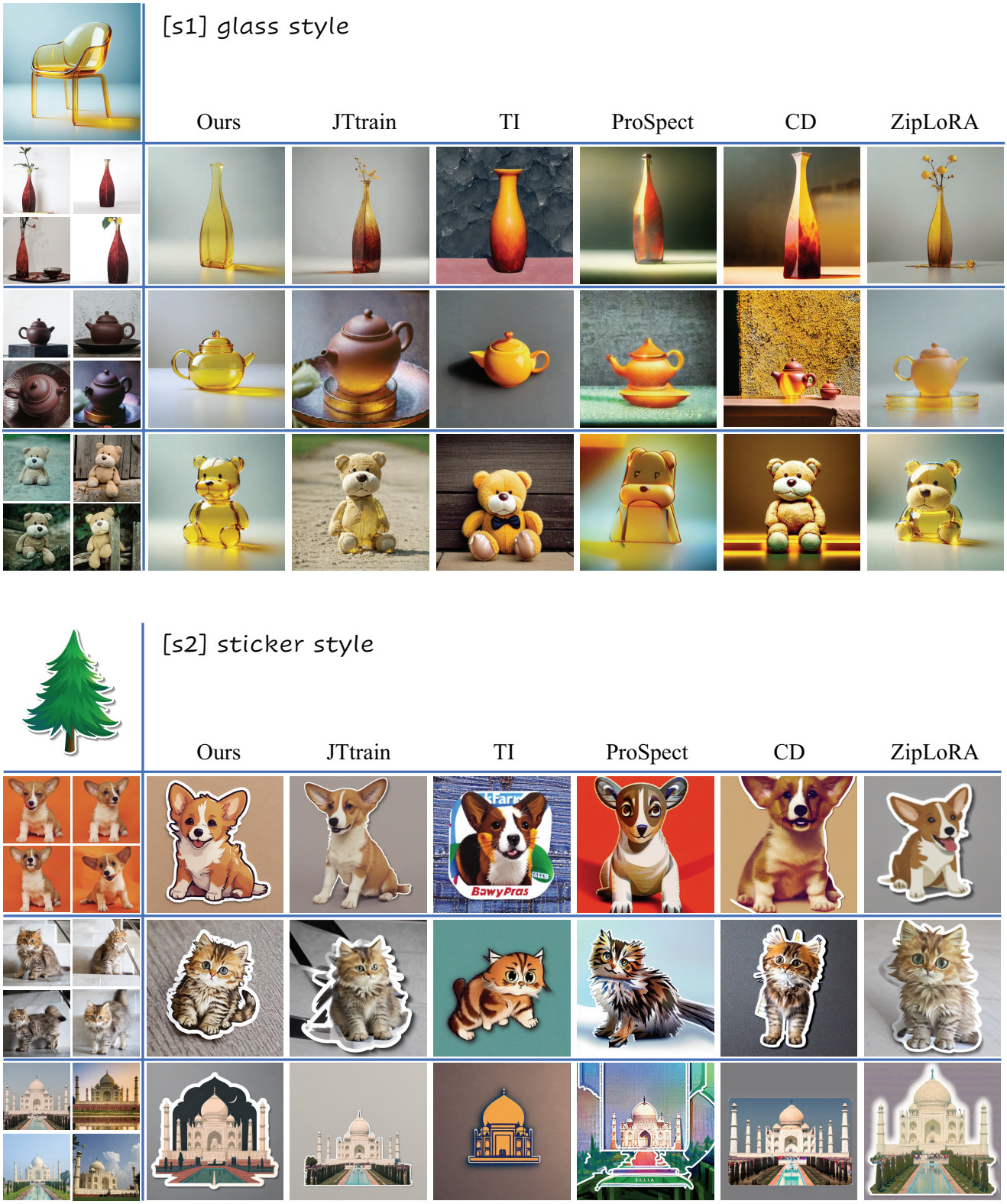

Fig. 5. Qualitative evaluation and comparison of DB+LoRA, TI, ProSpect, CD, ZipLoRA, and our method in diverse contents. The results indicate that our method generates harmonious content–style fusion images with diverse contents while preserving the disentanglement of content and style as well as maintaining high-level fidelity.

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

Methods	JTtrain	TI	ProSpect	CD	ZipLoRA	w/o MCP	w/o Orth	w/o RP	Ours
Content alignment (↑)	0.5221	0.4942	0.4816	0.5181	**0.5319**	0.5175	0.5265	0.5119	0.5288
Style-alignment (↑)	0.5438	0.6092	0.6165	0.6345	0.6403	0.5835	0.6297	0.6615	**0.6754**
Prompt-alignment (↑)	0.3038	0.2836	0.3156	0.2778	0.3319	0.3971	0.4046	0.3908	**0.4107**
Average (↑)	0.4566	0.4623	0.4712	0.4768	0.5014	0.4994	0.5203	0.5214	**0.5383**

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.

TI struggles to accurately learn content/style features, leading to a decrease in the fidelity of generated images ("mountain" with "Minimalism painting style" and "marble style" in Figure 4, the loss of feature "transparent glass" in "glass style" in Figure 5). ProSpect has achieved effective control over content and style to some extent; as seen in examples such as "vase" and "teapot" in Figure 5, 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 Figure 4). CD also encounters entangling issues between content and style. In cases of "glass style" with "vase" and "teapot" in Figure 5, 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 such as "vase" and "teapot" in Figure 5, 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 "blackboard 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. Note that the instance of "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 content-alignment, style-alignment, and prompt-alignment metrics. For each unique content–style pair, we generate 10 images with random seeds. A total of 6,000 images are used for quantitative comparisons. Additionally, we report the average of these three metrics, in which higher values indicate better performance. Our method achieves the highest average score among all baselines, suggesting that 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 ZipLoRA tends to generate images that retain more features from the content reference images. However, this could compromise the 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 Figures 4 and 5 supports this observation.

5.2 Editability Evaluation

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 use of ZipLoRA are obtained from their original paper. As illustrated in Figure 6, 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. We show more generation results from diverse prompts in the bottom two rows of Figure 6. The results show 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 with 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 Comparisons with Two-Stage 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 [Brooks et al. 2023; Hertz et al. 2022; 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 state-of-the-art real-image editing method to edit style in the second stage. The results are presented in Figure 7. We observe that 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.

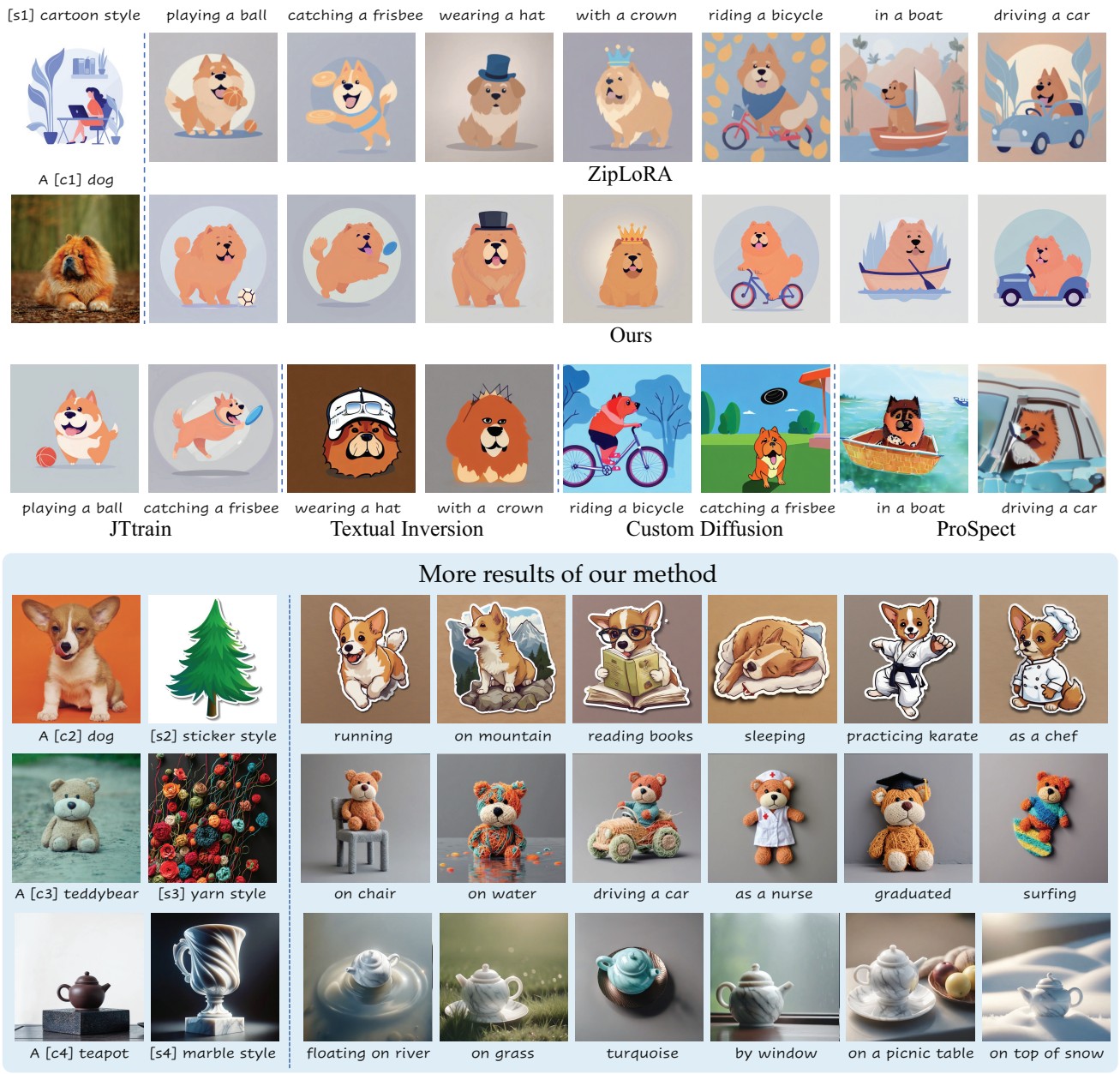

Fig. 6. Results of generating diverse customized content–style images. This indicates that our method exhibits excellent editing capabilities as well as generalization capabilities to both content and style.

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 with the two-stage training and editing methods.

5.4 User Study

We conduct a user study to assess the images generated by our method and other baseline methods. A total of 60 participants (32 female and 28 male, between 14 and 65 years old) took part in the survey, including 25 researchers in computer graphics or computer vision. Each participant took part in the following three settings; in total, we got 3, 600 votes.

— *User Study I: alignment preference.* In this setting, participants were shown the content and style inputs as well as

Fig. 7. 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**.

the results by all methods. They were asked to select the generated images that most closely aligned with the given content/style/both reference images. Statistics are shown in Figure 8(a). The results indicate a strong preference for our method's outputs across all methods. This user study validates our method's effectiveness in learning disentangled yet cohesive content–style representations from references. It highlights our framework's capabilities in accurately interpreting and presenting targeted features from references while seamlessly combining them per user intent expressed through prompts.

— *User Study II: success rate.* Diffusion models are known to be highly sensitive to the initial seed noise, which can substantially influence eventual results. To assess the robustness and reliability of our framework across varying initializations, we conduct A/B testing, analyzing the impact of different seed images on the generated outputs. We randomly generate nine seeds for each question and randomly select one baseline method. The selected and our methods generate nine images according to the nine seeds. Participants were asked to judge which set of nine images contained more content–style customized generated images. Statistics are shown in Figure 8(b). Evaluation results indicate that our method generates more content–style customized images than other methods, suggesting a higher success rate in satisfying image generation.

Fig. 8. User study results.

— *User Study III: consistency.* Besides the reliability of given input references, consistency among generated outputs is also crucial for real-world applications. We evaluate the generation stability across multiple samples produced by our method and baseline methods in the same setting as User Study II. Participants are required to assess the coherence and style congruency within each set of nine outputs. Statistics are shown in Figure 8(c). Our results demonstrate a remarkably high degree of consistency, with the vast majority of samples exhibiting faithful adherence to the specified content semantics and style attributes across the entire set. In contrast, baseline methods often suffer from greater sample variance, producing results that appear substantially more divergent from the intended prompt, both in terms of content and style preservation.

5.5 Comparison with Concurrent Work

B-LoRA [Frenkel et al. 2024], published concurrent to the present work, proposes a method for image stylization by decomposing images into style and content representations and optimizing different LoRA layers. However, our approach differs from B-LoRA in several significant aspects. Firstly, while B-LoRA primarily focuses on image stylization, our work aims to enable customized content-style generation, allowing for the synthesis of images with diverse poses, scenes, views, and other content variations. Secondly, whereas B-LoRA decomposes images into style and content representations and optimizes separate LoRA layers for each, our method trains LoRA in content and style sub-parameters using our proposed PLP and MCP techniques. This integrated approach facilitates a more seamless fusion of content and style components while maintaining high fidelity of reference content and style. We present a comparison of our method and B-LoRA in Figure 9. B-LoRA exhibits limitations in generating customized content and style aligned with various prompts. For example, B-LoRA's results with prompts such as "driving a car" and "catching a frisbee" show

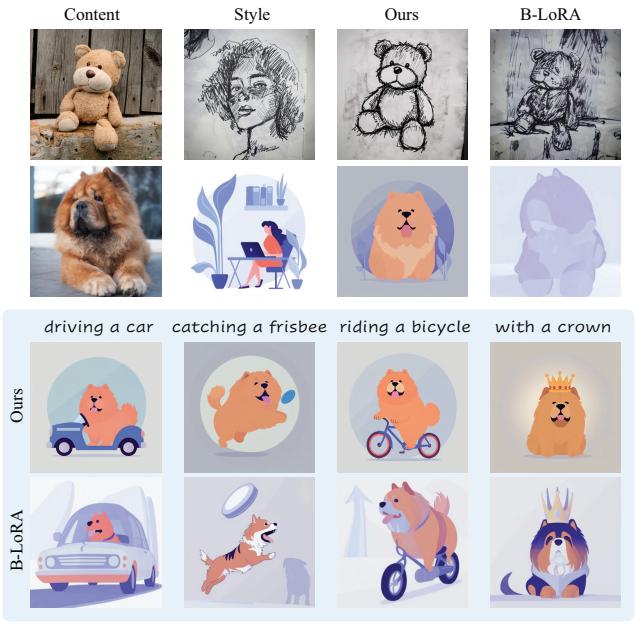

Fig. 9. Comparison with concurrent work. Our method can generate customized images with high fidelity to the content reference, style reference, and diverse prompts. Moreover, our results demonstrate stronger consistency in preserving both content and style. The results of B-LoRA are from [Frenkel et al. 2024].

less fidelity to the content reference, whereas the "with crown" prompt displays lower fidelity to the style reference. In contrast, our method learns content and style using the proposed PLP and MCP techniques in separate parameter spaces of LoRA, enabling the generation of disentangled content and style with high fidelity and alignment to various prompts. Moreover, our results demonstrate stronger consistency in preserving content and style when generating images conditioned on different prompts.

5.6 Comparison with Joint Training for Accommodating Multiple Concepts

As the proposed MCP leverages multiple content and style references for training to mitigate overfitting, for fair comparison, we also accommodate multiple style and content concepts for JTtrain with two settings. In the first setting, we leverage multiple style and content concepts for training in the first stage, and fine-tuning with the specific content-style pair in the second stage. In the second setting, the first stage involves training a specific style in the down projection while using multiple content in the up projection, and vice versa. In the second stage, the trained down projection (style) is combined with the up projection (content) for fine-tuning. Both settings leverage Riemannian Preconditioning for gradient scaling. The results are shown as JTtrain-II and JTtrain-III in Figure 10 and quantitative results in Table 2.

From Figure 10 we can see that, for JTtrain-II, training a variety of style and content concepts jointly within a shared parameter space in the first stage complicates the accurate learning of reference features, resulting in an inability to effectively represent the reference features in the second stage (e.g., the shape of the

Fig. 10. Comparison of results from our method with JTtrain methods accommodated with multiple style and content concepts. JTtrain leverages the specific content–style pair for training. JTtrain-II leverages multiple style and contents for training in the first stage, and fine-tuning with the specific content–style pair in the second stage. JTtrain-III trains specific style in down projection with multiple contents in up projection and vice versa, then combines the trained down projection (style) with up projection (content) for fine-tuning.

Table 2. Comparison with Joint Training for Accommodating Multiple Concepts

Methods	Content (↑)	Style (↑)	Prompt (↑)	**Average** (↑)
JTtrain	0.5221	0.5438	0.3038	0.4566
JTtrain-I	0.5095	0.5317	0.3027	0.4479
JTtrain-II	0.5237	0.5572	0.3102	0.4637
Ours	**0.5288**	**0.6754**	**0.4107**	**0.5383**

"teapot" and the material of its handle, the identity of the "teddy bear", and the identity and flat cartoon style of the "dog"). Results of JTtrain-II have even worse alignment than JTtrain, since JTtrain only uses the specific content–style pair. For JTtrain-III, although specific style and content are learned within the LoRA down and up projections in the first stage, the parameter space remains mixed after projection multiplication, resulting in outputs that cannot maintain high fidelity to the references after combining and fine-tuning in the second stage (e.g., the fine texture of the "teapot", the yarn art style of the "teddy bear" and the identity of the "dog"). As shown in Table 2, compared with our method, the alignments of JTtrain-II significantly decrease, indicating low fidelity to both the image references and the prompts. JTtrain-III achieves slightly better results than JTtrain-II owing to improved concept learning in the first stage. However, the mixed parameter space still leads to suboptimal performance.

5.7 Ablation Study

The Optimal Dimension d for the Fixed Parameters. The hyperparameter d, as the row dimension of the fixed parameters, represents the proportion of fixed parameters in the parameter subspace. Without loss of generality, we set d as 0, 1/8, 1/4, 3/8, 1/2, 5/8, 3/4,

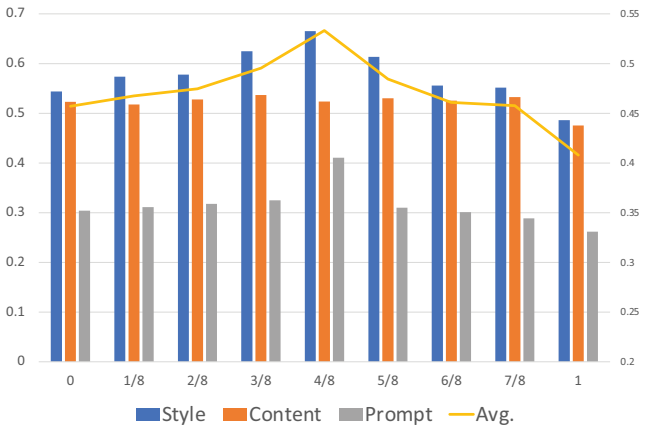

Fig. 11. Cosine similarity between features of output and reference style, content and prompt of different ratios for d. When $d = 0.5m$, the average cosine similarity of the features reaches its maximum, indicating optimal alignment between the generated results and the reference content, style, and prompt. The vertical axis on the right side of the chart represents the line labeled "Avg.".

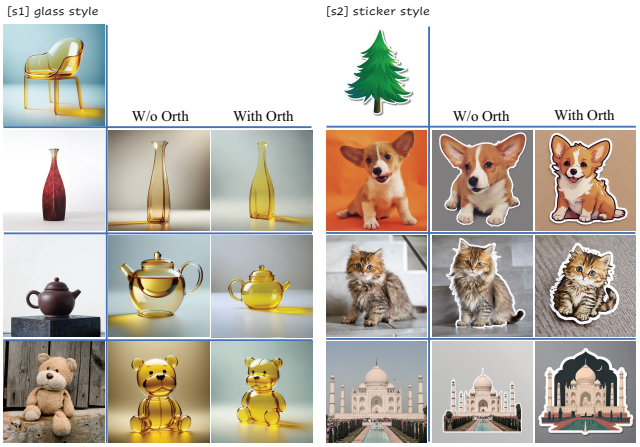

Fig. 12. Ablation study evaluating the impact of the proposed orthogonal fixed parameters. The **w/o Orth** shows results without orthogonal fixed parameters, while the **with Orth** demonstrates the improved image quality achieved by our full method incorporating orthogonal fixed parameters. The visual comparison highlights the effectiveness of orthogonal fixed parameters in enhancing the content and style fidelity of generated images.

7/8, and all of the dimensions for the corresponding pre-trained weights (n or m). We report the content-alignment, style-alignment, and prompt-alignment metrics for various values of d in Figure 11. From the histogram, we observe that as the value of d increases and the style alignment and prompt alignment gradually rise, reaching their peaks when $d = 0.5m$, then gradually declining. The average of the three alignments reaches its maximum at a ratio of 0.5, indicating that the optimal alignment occurs at a ratio of 0.5 with 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" and "style parameter subspace" 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.

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 to remove the orthogonal fixed parameters and replace them with randomly fixed parameters. We present the results in Figure 12 for comparison. Without adding the orthogonality prior to 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 emphasize describing the image content.

Multi-Correspondence Projection Learning. We conduct two ablation studies with different experimental settings to evaluate the impact of the proposed MCP. In the first setting, we train the target content PLP (e.g., "vase") with the style references (e.g., "glass style") in a one-to-one manner. In the second setting, we train the model on the target content with a non-target style (e.g., "yarn style" or "painting style") and then combine the content PLP with the trained style (e.g., "glass style") PLP. The results are shown as **w/o MCP-I** and **w/o MCP-II** in Figure 13. We can observe that in the first setting, the generated images exhibit a degree of overfitting to the reference images (e.g., "teapot" with "chair legs" from the style reference image), resulting in a decrease in fidelity to the content or style, thereby reducing the quality of the outputs. In the second setting, the results exhibit some features (e.g., the color from "yarn style" or "painting style") of the style trained in the first stage. As the final model does not incorporate this style, this is mainly due to the fact that, without MCP, the "yarn style" and "painting style" influence the parameter space of the content during the training stage. We present the results of our method 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. We also present quantitative comparisons in Table 1. Without MCP, the content, style, and prompt alignment decrease significantly, indicating that our method with MCP can faithfully preserve the content and style of the reference images while achieving a high degree of alignment with provided prompts.

Gradient Scaling with Riemannian Preconditioning. In our ablation study, we evaluated the impact of Riemannian preconditioning by comparing it with the conventional AdamW optimizer [Loshchilov and Hutter 2019]. Figure 14 illustrates the comparative results. Using the conventional AdamW optimizer, we observe significant overfitting to the patterns of reference images (e.g., the poses of "teapot" and "teddy bear", and have no

Fig. 13. 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 direct inference after training. Results are presented in the **w/o MCP-I** column. We train specific content (e.g., "vase") and style (e.g., "yarn" or "painting") in a one-to-one manner and then combine the content (e.g., "vase") adapter with style (e.g., "glass") adapter in the second stage, then inference with the combined adapters. We present results trained with "[s2] yarn style" in the **w/o MCP-II-[s2]** column and trained with "[s3] painting style" in the **w/o MCP-II-[s3]** 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.

interaction with "water" or "skateboard"), leading to suboptimal alignment with different prompts and limiting the model's ability to generalize across different prompts. We can also observe subject leakage from the reference images in the bottom row of the figure (e.g., a plant from the style reference unexpectedly appears in the result of "cartoon dog"). In contrast, our method demonstrates markedly improved performance through balanced training content and style in up and down projections with Riemannian preconditioning. The results show better alignment with input prompts while preserving the intended content and style features. Quantitatively, as shown in Table 1, the conventional AdamW optimizer results in decreased content-alignment, style-alignment, and prompt-alignment scores compared with our approach, thus validating the crucial role of Riemannian preconditioning over conventional optimization techniques in mitigating overfitting and enhancing the model's generalization capabilities.

Concepts Learning Ablation. We aim to evaluate the learning effect of the desired content or style compared to the baseline stable diffusion model. To achieve this, we employ pseudo words for training and inference of specific content and style. For comparison, we describe content and style using prompts for generation. The results presented in Figure 15—solely relying on prompts to describe the desired content or style without learning these representations—fail to capture detailed features from reference images, leading to unfaithful generation of the customized content and style.

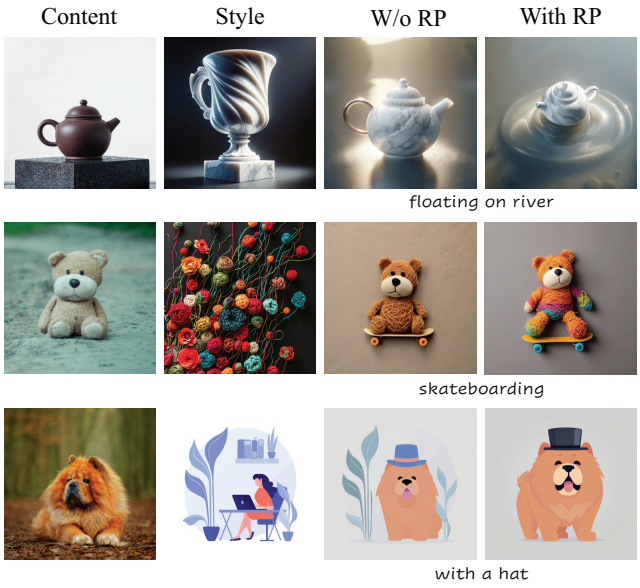

Fig. 14. Ablation study results demonstrating the effectiveness of Riemannian preconditioning. **RP** indicates the Riemannian preconditioning, results with our proposed Riemannian preconditioning method exhibiting improved prompt alignment and better content-style fusion.

Fig. 15. Comparison of results with and without learning concepts. We present output images generated with and without learning reference content or style in orange (by our method) and green (by DreamBooth method) boxes, respectively. We also show images directly generated by the basic Stable Diffusion-XL model in the blue box. The prompts used for inference are displayed at the top. Without learning content or style in pseudo words, models that rely solely on prompts cannot generate desired content or styles faithfully.

5.8 Computational Time Costs

We evaluate computational time costs by calculating the average time consumption on a single NVIDIA A100 GPU. We design two application scenarios to evaluate computational time costs with baseline methods. In the first scenario, a new style or content concept is given and needs to compose with a trained content or style concept. In the second scenario, both style and content concepts are given new. Results are presented in Tables 3 and 4, respectively.

In the first scenario, except for JTtrain and CD, the first-stage training can be done separately for content and style. Thus, in our method, given a new style, the style PLP can be independently trained (with the specific style reference and multiple content

Table 3. Comparison of Time Cost when Given a New Style or Content Concept

Method	Time Cost	Method	Time Cost
JTtrain	28 min 41 s	CD	18 min 41 s
TI	27 min 4 s	ZipLoRA	21 min 39 s + 8 min 46 s
Prospect	28 min 47 s	B4M	21 min 58 s + 2 min 25 s

Table 4. Comparison of Time Cost When Given Both New Style and Content Concepts

Method	Time Cost	Method	Time Cost
JTtrain	28 min 41 s	CD	18 min 41 s
TI	27 min 4 s ×2	ZipLoRA	21 min 39 s ×2 + 8 min 46 s
Prospect	28 min 47 s ×2	B4M	21 min 58 s ×2 + 2 min 25 s

Our method achieves the most balance between required time and generation quality. We have greatly saved time, especially compared with the two-stage method (e.g., ZipLoRA).

references) for an average of 21 min 58 s in the first stage and then combined with a pre-trained content PLP for a fine-tuning average of 2 min 25 s. The total average time cost is 24 min 23 s, which is more efficient than ZipLoRA's 30 min 25 s (21 min 39 s for style training plus 8 min 46 s for fine-tuning). Conversely, the process works similarly for a new content concept. Results are presented in Table 3.

In the second scenario, when both the given content and style concepts are new, the style PLP and content PLP need to be independently trained in the first stage. As a result, the training time in the first stage is 21 min 58 s on average for each, totaling 43 min 56 s. The second stage remains the same as the first scenario and the total average time cost is 46 min 21 s on average, which is more efficient than ZipLoRA's 52 min 4 s (21 min 39 s for training each style LoRA and content LoRA plus 8 min 46 s for fine-tuning). Results are presented in Table 4.

Compared with ZipLoRA, B4M significantly reduces the fine-tuning time in the second stage while improving the generation quality.

6 Applications and Discussions

6.1 Applications

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

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 Figure 16. 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 with traditional rendering pipelines requiring extensive modeling and material setup, our data-driven approach significantly streamlines this process.

Fig. 16. Application I. Content-style customization of various textures, including knit, burlap, denim, and fabric texture.

Fig. 17. Application II. Content-style customization of portraits.

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 such as virtual production and artificial intelligence–assisted artwork creation. For digital artists and creative professionals, our framework efficiently synthesizes portrait imagery in various artistic styles with fine user control. Figure 17 illustrates examples in which we tasked artists to create stylized portraits using our approach in styles such as sticker, watercolor painting, and flat cartoons. Compared with 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.

Application III: Content-Style Customization of Modern Arts. Our proposed method has significant applications in the realm of computational arts and content customization, as presented in Figure 18. One compelling application is the customization of hypnotic line art. Hypnotic line art is an emerging art form that uses intricate interwoven lines and patterns to create mesmerizing visual effects [Yeh et al. 2020]. Our method could allow artists to customize this content (e.g., a dog) based on user-specified input

Fig. 18. Application III. Content-style customization of hypnotic line art [Yeh et al. 2020] and portrait map art [Zhang et al. 2023d].

Fig. 19. Bad cases. Sometimes our method struggles with cases in which overlapping opaque elements should be visible through the transparent regions.

images and prompts. This opens up possibilities for generative art pieces, animated line art visualizations driven by data inputs, and unique branding/design elements. Another application domain is portrait map art [Zhang et al. 2023d], in which stylized geographic maps are used to compose portraits or other imagery. This kind of modern art form is created by British portrait artist Ed Fairburn. Our technique could be employed to customize these map portraits through the road networks, terrain features, and map elements to emphasize specific visual qualities of the input portrait. This could produce visually striking artwork while preserving recognizable characteristics of the depicted portrait. Map portraits have applications in data visualization, creative cartography, and modern pop art.

6.2 Bad Cases and Limitation

Bad Cases. While our proposed method demonstrates considerable promise in addressing customized content and style fusion, it is essential to acknowledge instances in which certain artistic styles involve transparency or see-through elements, such as a "glass style", as shown in Figure 19. While our method can learn and apply transparent attributes to the overall content, it may struggle with cases in which overlapping opaque elements should be visible through the transparent regions. For instance, if a user prompts "with flowers" in a glass vase rendering, our current approach may fail to depict the flower stems properly inside the transparent vase body. This is because, during training, the model does not learn to decompose the scene into explicit occluded and visible components based on transparency. As a result, when opaque elements such as flower stems would normally be perceivable through the transparent glass regions, they may be omitted or masked out incorrectly. Handling such cases would require explicit modeling of occluded scene components and line-of-sight visibility, which is not currently captured in our framework. This limitation highlights challenges in photorealistic compositing for styles involving transparent surfaces or volumes.

Limitations. While our method performs well on content–style customization, generating images with complex or rare content/style solely by using textual prompts remains challenging. 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, "[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.

7 Conclusion

We introduce *Break-for-make* (B4M), a novel separated LoRA training framework that enables composable content–style customization. Our approach disentangles content and style representations, allowing independent recombination of content and style LoRA projections. When evaluated against state-of-the-art methods, our method demonstrated superior disentanglement capabilities, simultaneously preserving high fidelity to the content and style reference and alignment to diverse prompts. This decoupled yet faithful representation facilitates seamless customization across an extensive range of content–style combinations. However, our work has limitations when customizing transparent subjects, as it struggles with cases in which overlapping opaque elements should be visible through the transparent regions. Looking ahead, our method proposes a simple yet effective framework for exploring compositional generative models that can flexibly combine and remix visual elements on demand. With strong disentanglement and robust customization capabilities, we hope our method catalyzes new creative paradigms and practical applications across digital arts, media, and design domains.

Acknowledgments

We thank the reviewers for their insightful comments and suggestions, as well as Ziyao Huang and You Wu for their help with the experiments.

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. 2023. Break-A-Scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers (SA'23)*. ACM, New York, NY, Article 96, 12 pages. https://doi.org/10.1145/3610548.3618154

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, et al. 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)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.), Vol. 202. PMLR, 4055–4075.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. 2023. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems* 36 (2023), 30286–30305.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems* 36 (2023), 10088–10115.

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).

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. 2024. Implicit style-content separation using B-LoRA. *arXiv preprint arXiv:2403.14572* (2024). 181–198.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The 11th International Conference on Learning Representations*.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7323–7334.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross-attention control. In *The 11th International Conference on Learning Representations*.

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. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Edward J. Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hexiang Hu, Kelvin C. K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. 2024. Instruct-imagen: Image generation with multi-modal instruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4754–4763. *arXiv preprint arXiv:2401.01952* (2024).

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. In *International Conference on Learning Representations*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. 2021. OpenCLIP. (July 2021). Retrieved from https://doi.org/10.5281/zenodo.5143773

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, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023. Cones: Concept neurons in diffusion models for customized generation. In *Proceedings of the 40th International Conference on Machine Learning*. 21548–21566.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Midjourney. 2023. Retrieved 2023 from Midjourney. https://www.midjourney.com/. (2023).

Bamdev Mishra and Rodolphe Sepulchre. 2016. Riemannian preconditioning. *SIAM Journal on Optimization* 26, 1 (2016), 635–660.

mkshing. 2023. ZiploRA. Retrieved 2023 from https://github.com/mkshing/ziplora-pytorch. (2023).

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, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2023).

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems* 36 (2023), 66727–66754.

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. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL'21)*. Association for Computational Linguistics (ACL), 487–503.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. 2024. Orthogonal adaptation for modular customization of diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7964–7973. *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. In *The 12th International Conference on Learning Representations*.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 234–241.

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. 2024. ZipLoRA: Any subject in any style by effectively merging LoRAs. In *European Conference on Computer Vision*. 422–438.

Ryu Simo. 2023. LoRA. Retrieved from https://github.com/cloneofsimo/lora. (2023).

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. In *37th Conference on Neural Information Processing Systems (NeurIPS'23)*. Neural Information Processing Systems Foundation, 36 (2023), 66860–66889.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

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. 2023. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 3274–3287.

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. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV'23)*. IEEE, 15897–15907.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *37th Conference on Neural Information Processing Systems*. 36 (2023), 7093–7115.

Chih-Kuo Yeh, Zhanping Liu, I-Hsuan Lin, Eugene Zhang, and Tong-Yee Lee. 2020. WYSIWYG design of hypnotic line art. *IEEE Transactions on Visualization and Computer Graphics* 28, 6 (2020), 2517–2529.

Fangzhao Zhang and Mert Pilanci. 2024. Riemannian preconditioned LoRA for fine-tuning foundation models. In *41st International Conference on Machine Learning*. 59641–59669.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. Adaptive budget allocation for parameter-efficient fine-tuning. In *The 11th International Conference on Learning Representations*.

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023b. ProSpect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)* 42, 6, Article 244 (Dec 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.

Yuxin Zhang, Fan Tang, Weiming Dong, Thi-Ngoc-Hanh Le, Changsheng Xu, and Tong-Yee Lee. 2023d. Portrait map art generation by asymmetric image-to-image translation. *Leonardo* 56, 1 (2023), 28–36.

Received 13 September 2024; revised 4 March 2025; accepted 25 March 2025