Supplemental Materials for "**B4M**: **B**reaking Low-Rank Adapter **for** **M**aking Content-Style Customization"

A CONCEPT AND STYLE RECONSTRUCTION

The proposed two-stage training paradigm employs a multiple correspondence projection 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. 1 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 with Riemannian precondition 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. 1. These results verify that content and style are disentangled in the first stage and have better effects after undergoing the fine-tuning process.

B 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 with Riemannian precondition to enable the model to generate images with customized content and style. The results of ablating the fine-tuning step are presented in Fig. 1. 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. 1. These results illustrate that our proposed method successfully disentangles content and style while retaining the capability to generate individual content or style faithfully.

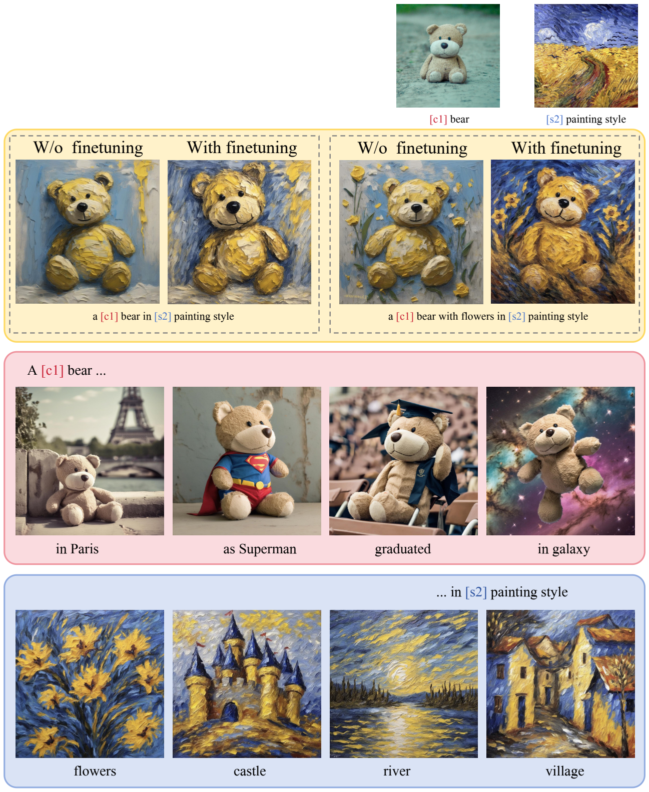

Fig. 1. **Individual content/style generation of our method.** Our method can generate diverse content/style images individually with a high level of fidelity and disentanglement. Fine-tuning enhances the final effect.

C PARAMETER DISTRIBUTION VISUALIZATION

In this section, we employ t-SNE (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 present three different content-style customized samples and utilize t-SNE to reduce the dimensionality of the low-rank adapter parameters to two dimensions. Additionally, we present the average results of these samples to obtain general findings. We compare our method with a joint training baseline. For a fair comparison, we also present results of joint training with zeroing out the orthogonal parameter components to align with the parameter formulation of our method. The first row in Fig. 2 depicts the parameter distribution of the low-rank adapters from our proposed method after applying t-SNE for dimensionality reduction and visualization. The two distinct clusters observed indicate the presence of two separate sub-distributions within the LoRA parameters. The second and third rows in Fig. 2 show the parameter distribution of the joint training baseline with and without zeroing out the orthogonal parameter components, respectively. The results demonstrate that our method successfully separates the parameter space of the low-rank adapters from a uniform distribution, in contrast to the joint training approach.

D VARIOUS REFERENCE OF THE SAME STYLE

During training, each style class contains 1-3 different images of the same style, which has an influence on the outcome. For example, in Fig. 4, the background color of the image in previous results column may influenced by the last style reference image. We have now included all the style images in Figure 4 in main text for further clarification. To address the concern about this, we incorporate Riemannian preconditioning method to mitigating overfitting in the revised version. The updated results, shown on the right side of Fig. 4, demonstrate a clearer and more consistent alignment with the references compared to the previous version. We also update Figure 5 and other results to showcase these improvements, providing a

Author's address:

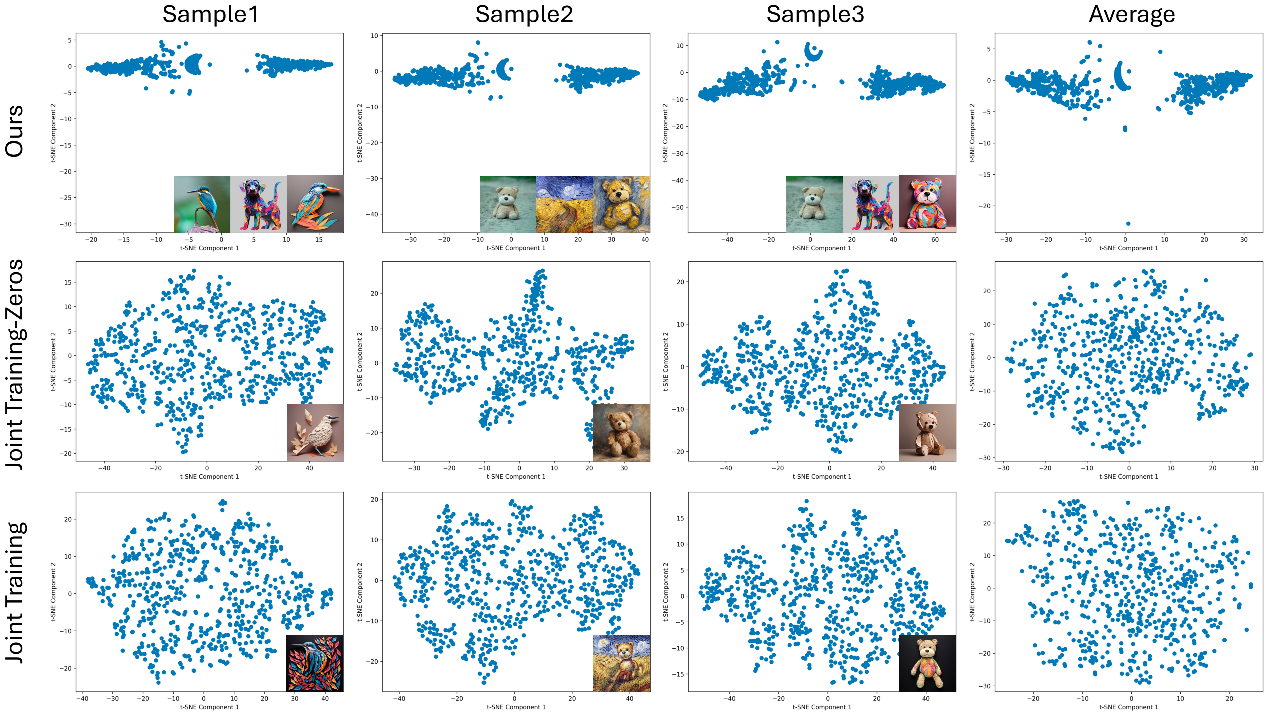

Fig. 2. **Visualizing Low-Rank Adapter Parameter Distributions via t-SNE.** Compared with the joint training baseline, the two clusters of our method observed in the t-SNE visualization indicate the presence of two distinct sub-distributions within the LoRA.

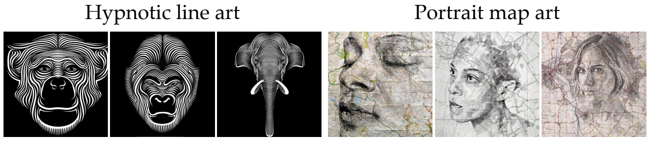

Fig. 3. Style references used for training Hypnotic line art and Portrait map art.

more representative and compelling presentation of our technique's effectiveness.

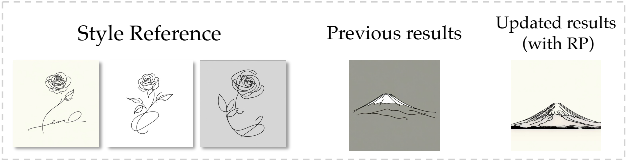

Fig. 4. **Reference and results.** Previous results tend to influenced by single reference, with Riemannian preconditioning, results are more align with style.

E STYLE REFERENCES IN APPLICATION-III

We showcase content-style customization results of two modern art Hypnotic line art and Portrait map art in Application-III of main text, here we present style references used for training in Fig. 3. For each style, we collect 3 reference images.

F TRAINING EXAMPLES DIVERSITY FOR MCP

As shown in Table 1, more images in MCP will increase the alignment. However, the improvement from 5 to 10 is insignificant. Using images of the same content/style for MCP decreases alignment.

Table 1. **Training examples diversity for MCP.** Our method adopts five different classes of images in MCP, taking into account both the generation effect and computational efficiency.

Images	2	5(ours)	10	5(same style)
Content-alignment (↑)	0.5213	0.5288	0.5291	0.5221
Style-alignment (↑)	0.6142	0.6754	0.6752	0.6373
Prompt-alignment (↑)	0.4001	0.4107	0.4109	0.4051

G DCO AND PARETO CURVE

DCO [Lee et al. 2024] aims to improve customization consistency for a single concept. Direct merging models by DCO is sub-optimal without the disentanglement of content and style concepts. However, the proposed comprehensive caption (comp.) method can benefit B4M, as shown in Table 2.

Table 2. **Comprehensive caption from DCO for B4M.**

Images	B4M	DCO	B4M+comp.
Content-alignment (↑)	0.5288	0.5031	0.5292
Style-alignment (↑)	0.6754	0.6157	0.6771
Prompt-alignment (↑)	0.4107	0.3685	0.4122
Average (↑)	0.5383	0.4109	0.5395

H MORE RESULTS

In this section, we show more results for different content-style pairs in Fig. 6. Fig. 8 shows the results for the cross-mixing setting of target and style references. The results indicate that our methods successfully disentangle content and style within a single image while maintaining a high level of fidelity to the respective content and style references.

I MORE RESULTS LEVERAGING THE PLP STRATEGY WITH RIEMANNIAN PRECONDITIONING

To further investigate the effectiveness of our PLP strategy, we conduct additional experiments that ablate RP and MCP of our method, as shown in Fig. 5. The results demonstrate that even without the benefits of RP and MCP, our PLP approach achieves content and style fidelity to references. While some degree of concepts leakage or overfitting is observed in our method without RP or MCP, the outputs still maintain good structural consistency and stylistic adherence to the references. This improvement can be attributed to PLP's fundamental design of separating content and style into distinct parameter spaces, allowing for more precise representation over each aspect.

REFERENCES

Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. 2024. Direct consistency optimization for compositional text-to-image personalization. arXiv preprint arXiv:2402.12004 (2024).

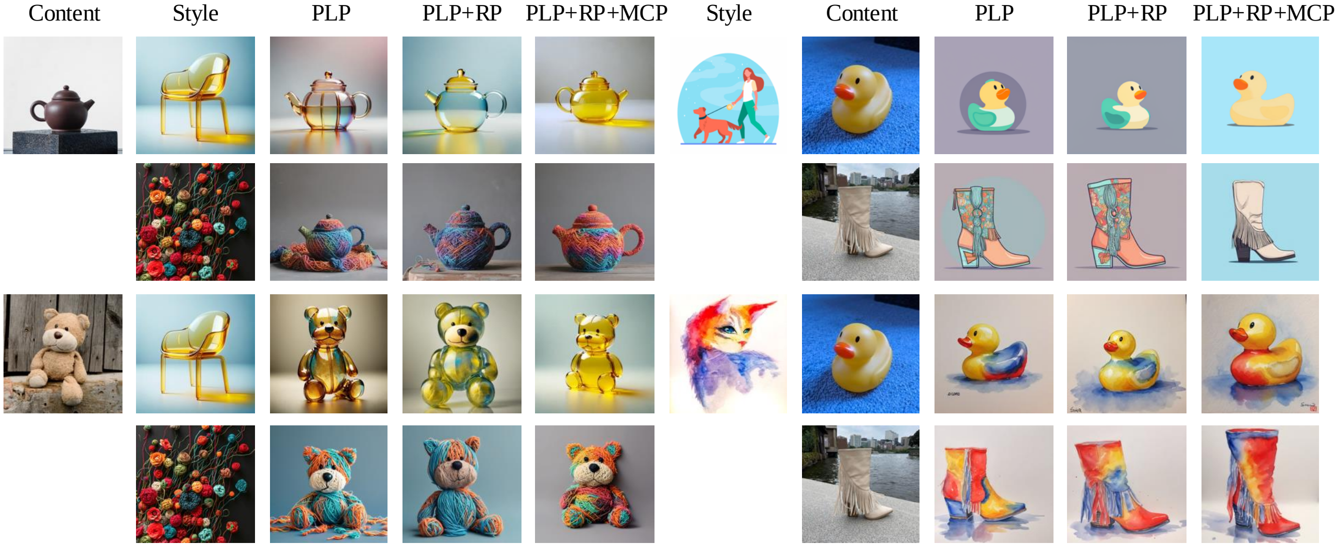

Fig. 5. More comparison results of PLP method and PLP+RP method.

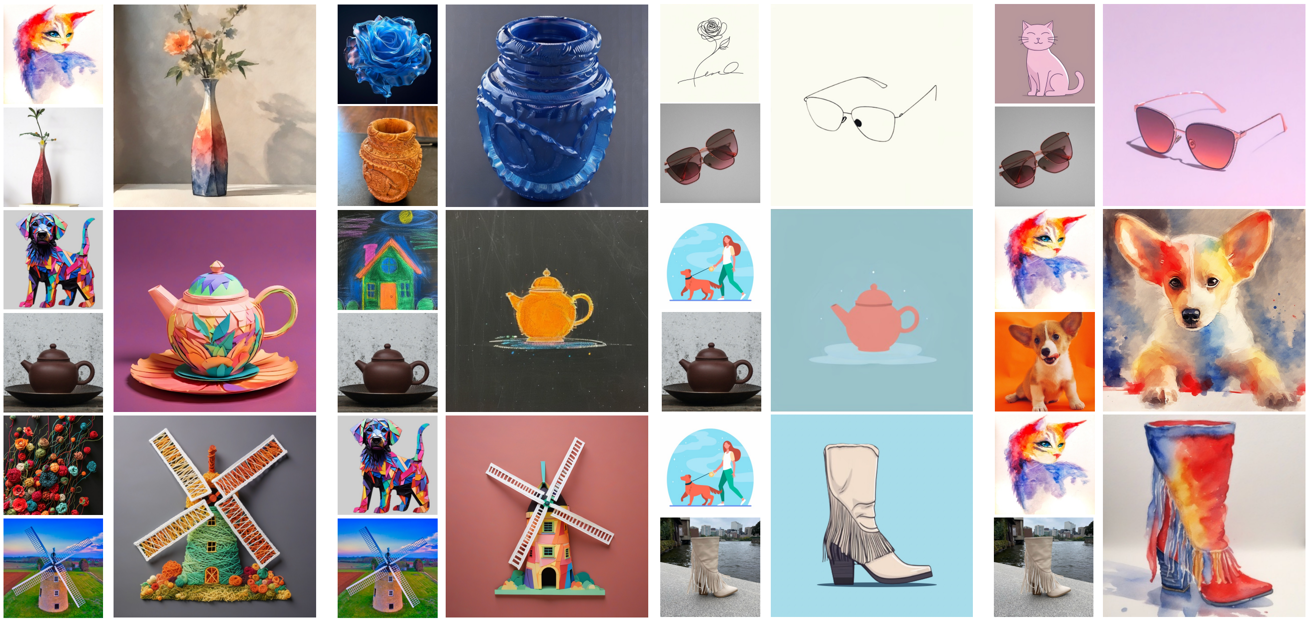

Fig. 6. More results of our method for making different content and style customizations.

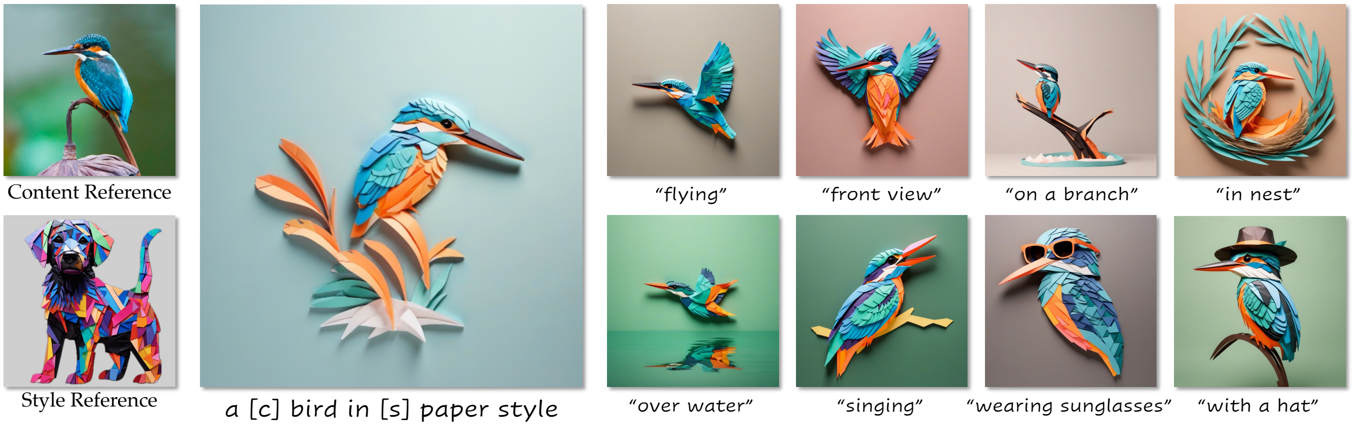

Fig. 7. More results generated by our method using different prompts.

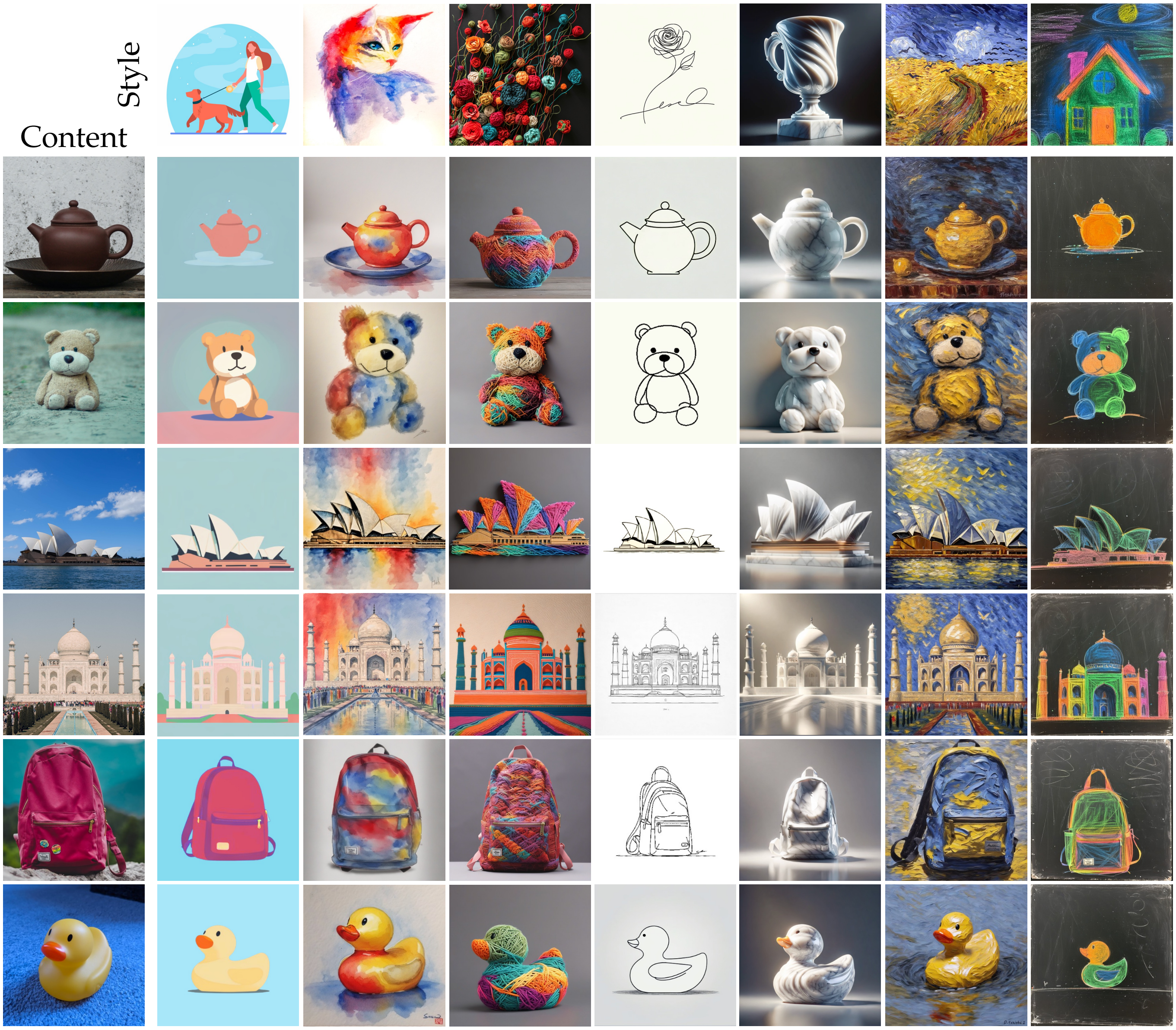

Fig. 8. More results of diverse content and style generated by our method.