A Training-Free Style-Personalization via SVD-Based Feature Decomposition

Supplementary Material

A. Comprehensive analysis of our method

A.1. Role of step 3 in global feature formation

The choice of the 3rd step can be supported not only empirically but also from a structural perspective based on its spatial resolution (4×4). This resolution corresponds to the minimum scale at which global image appearance can be spatially expressed, as evidenced by prior image generation frameworks [5, 6] , which adopt this resolution as the initiating stage of the generation process.

Accordingly, the 3rd step can be interpreted as the earliest stage where globally expressive representations emerge, while later steps mainly refine local structure conditioned on this global representation. Variations at this step propagate through subsequent steps, inducing large-scale changes in the final image, consistent with our Step-wise analysis in Sec. 4.

A.2. Additional results for Key step feature analysis

Although the 3rd step feature F_3 encodes both content and style, its coarse spatial resolution inherently biases it toward low-frequency statistics. Prior works [8, 9] suggest that early stages are critical in shaping image style. Consistent with this perspective, our Key step feature analysis in Sec. 4 reveals that stylistic variations at F_3 are dominated by the leading principal component.

Building on this observation, we further analyze the structure of F_3 using singular value decomposition (SVD). In Sec. 4-(2), we showed that replacing only the largest singular component of F_3 primarily alters stylistic attributes while largely preserving content. To further validate this observation, we extend the SVD-guided manipulation experiment by varying the number of preserved singular values.

We use the same prompt setup and intervention protocol as in the main paper: we construct 100 mixed prompt pairs $(T, \hat{T})$, each differing in both object category and color (e.g., "`A photo of a red truck`" vs. "`A photo of a purple cat`"). For each prompt, we perform singular value decomposition $F_3 = U\Sigma V^\top$, and reconstruct truncated variants that retain only the top-k components:

$$F_3^{(k)} = U\Sigma^{(k)}V^\top, \tag{1}$$

where $\Sigma^{(k)}$ is constructed by preserving only the largest k singular values while zeroing out the remaining entries. We evaluate $k \in \{1, 2, 4, 8, 16, 32\}$, and for each k, we generate SVD-guided outputs by replacing the corresponding portion of $\hat{F}_3$:

$$\hat{F}_3 \leftarrow F_3^{(k)} + \hat{F}_3^{\text{res}(k)}, \tag{2}$$

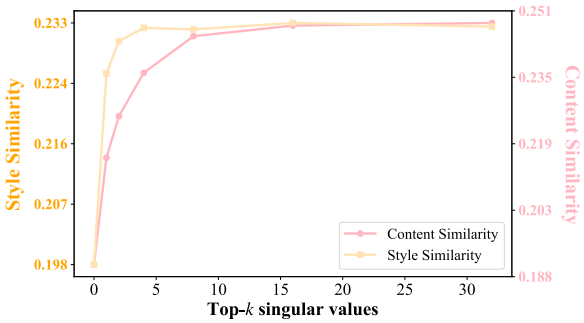


Figure 7. Qualitative results of SVD-guided feature replacement with varying top-k singular values. From left to right: the baseline output generated from $\hat{T}$, SVD-guided outputs with $k \in \{1, 2, 4, 8, 16, 32\}$, and the baseline output generated from T.

where $\hat{F}_3^{\text{res}(k)} = \hat{F}_3 - \hat{F}_3^{(k)}$ preserves the remaining feature components.

We measure object-related and color-related CLIP similarity following the same evaluation protocol used for the $k = 1$ experiment in the main paper. As shown in Fig. 7, color-related similarity sharply increases at $k = 1$ and saturates thereafter, demonstrating that the dominant singular direction primarily captures style. In contrast, object-related similarity increases gradually as k grows, indicating that higher-rank components encode structural information. CLIP similarity values are not directly comparable to S_{img} in Tab. 1, as we measure text-image similarity in our analysis, while Tab. 1 reports image-image similarity, which typically yields higher scores due to the shared visual domain.

Qualitative examples in Fig. 9 show a similar trend: the $k = 1$ output transfers texture and color while preserving object shape, whereas larger k values begin to alter geometry and object identity. These results further support our main finding that the first principal component of F_3 predominantly encodes style, also justifying our exponential reweighting design in the main method.

A.3. Extended evaluation

We extend the evaluation with 10 additional style sets to assess generalization across a wide range of artistic expressions, from impressionism to pixel art. This extended benchmark further facilitates the assessment of performance on highly abstract or non-textural styles, such as cubism and minimalism art. We compare against the top four models from our quantitative evaluation–StyleAligned [4], IP-Adapter [17], DreamBooth-LoRA (DB-LoRA) [10], and B-LoRA [2]–shown in Tab. 3-(a-d), as well as three recent

	Method	$S_{\text{txt}} \uparrow$	$S_{\text{img}} \uparrow$	$S_{\text{harmonic}} \uparrow$	$\text{VQA}_{\text{txt}} \uparrow$	$\text{VQA}_{\text{leak}} \downarrow$	Time (s) $\downarrow$
(a)	StyleAligned [4]	0.314	<u>0.737</u>	**0.440**	0.682	0.529	64.58
(b)	IP-Adapter [17]	0.303	0.775	0.435	0.600	0.603	<u>10.13</u>
(c)	DB-LoRA [10]	<u>0.330</u>	0.599	0.425	0.750	<u>0.248</u>	342.01
(d)	B-LoRA [2]	<u>0.330</u>	0.568	0.417	0.734	**0.209**	630.42
(e)	Qwen-Image-Edit [13]	0.328	0.634	0.432	<u>0.776</u>	0.394	246.27
(f)	Flux.1 Kontext [7]	0.316	0.631	0.421	0.676	0.304	58.05
(g)	USO [14]	0.286	**0.804**	0.421	0.588	0.641	36.04
(i)	Ours	**0.333**	0.640	<u>0.438</u>	**0.808**	0.258	**3.58**

Table 3. Additional quantitative comparison on an extended style set, including top-performing models from the main evaluation and recent training-free methods, evaluated on a total of 2000 images. The symbol ↑ indicates that higher is better. The best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

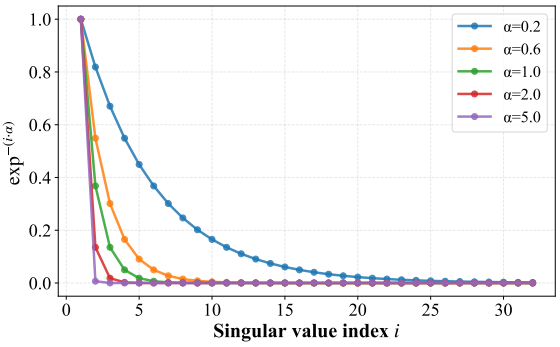


Figure 8. Visualization of exponential decay rates α with respect to the singular value index $i \in \{0, 1, \ldots, r-1\}$.

training-free models–Qwen-Image-Edit [13], Flux.1 Kontext [7], and USO [14]–shown in Tab. 3-(e-g).

Furthermore, we introduce two additional VQA-based metrics to supplement S_{img} and S_{txt}: VQA_{leak}, which detects whether objects from the reference style image appear in the generated image (content leakage), and VQA_{txt}, which evaluates text alignment. For VQA_{leak}, we use the prompt "`Does the image contain the following object: {}? Please answer yes or no`", where {} is populated with objects extracted from the style reference, while VQA_{txt} uses the input text prompt. Under this evaluation setup, our method achieves comparable or superior performance with significantly faster inference, demonstrating strong generalization across diverse style domains. In particular, it shows substantially lower content leakage (VQA_{leak}) than other models with higher S_{img}, while maintaining strong VQA_{txt} and S_{txt}.

A.4. Hyperparameter analysis

We conduct additional ablation studies on hyperparameters in our method, using an extended style set of 20 images, as used in Tab. 3.

Effect of the exponential decay rate α. We first investigate the impact of the exponential decay rate α in Principal Feature Blending (PFB). As shown in Tab. 4, our method remains robust across different values of α, exhibiting only a minor trade-off between style fidelity and prompt fidelity. Fig. 8 provides a visualization of how varying α controls the exponential decay of weights across singular values.

Table 4. Additional ablation study on exponential decay rate (α) in Principal Feature Blending (PFB). The symbol ↑ indicates that higher is better. The best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

alpha (α)	$S_{\text{txt}} \uparrow$	$S_{\text{img}} \uparrow$	$S_{\text{harmonic}} \uparrow$
0.2	0.321	**0.651**	0.430
0.6	<u>0.329</u>	<u>0.643</u>	0.435
1.0 (Ours)	**0.333**	0.640	**0.438**
2.0	**0.333**	0.636	<u>0.437</u>
5.0	**0.333**	0.634	<u>0.437</u>

Table 5. Additional ablation study on Structural Attention Correction (SAC). The symbol ↑ indicates that higher is better. The best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

Method	$S_{\text{txt}} \uparrow$	$S_{\text{img}} \uparrow$	$S_{\text{harmonic}} \uparrow$
w/o SAC	0.320	<u>0.644</u>	0.427
Step 3-7	<u>0.331</u>	0.636	<u>0.435</u>
Step 8-12	0.316	**0.649**	0.425
Step 3-12 (Ours)	**0.333**	0.640	**0.438**

Decreasing α, which increases the influence of the higher-rank singular components, naturally elevates the risk of content leakage during style injection, reducing prompt fidelity. This behavior is consistent with our hypothesis that the dominant singular value predominantly encodes style-related information over the remaining components. We set $\alpha = 1.0$ as it provides the most balanced performance.

Effect of Structural Attention Correction (SAC). We further analyze the effect of SAC by applying it at early steps (3–7) and later steps (8–12) after PFB. As shown in Tab. 5, applying SAC at early steps effectively corrects structural distortions introduced by PFB, whereas applying it at later

Algorithm 1 Dual-path style-personalized image generation

Input: Style reference image I^{sty}, text prompt T
Output: Stylized image I^{gen}

$\{F_s^{\text{sty}}\}_{s=1}^{S} \leftarrow \mathcal{E}_I(I^{\text{sty}})$ # Multi-scale style features
Initialize $F_0^{\text{con}}, F_0^{\text{gen}}$ # Same initial condition, same prompt T
for $s = 1$ to S **do**
  # (1) Dual-stream iterative update (Eq. (1), (2))
  $F_s^{\text{con}} \leftarrow \mathcal{M}(F_{s-1}^{\text{con}}, \mathcal{E}_T(T))$
  $F_s^{\text{gen}} \leftarrow \mathcal{M}(F_{s-1}^{\text{gen}}, \mathcal{E}_T(T))$
  if $s = 3$ **then**
    # (2) Principal Feature Blending (PFB)
    $F_3^{\text{gen}} \leftarrow \Phi(F_3^{\text{sty}}) + (F_3^{\text{gen}} - \Phi(F_3^{\text{gen}}))$
  end if
  if $s \in \mathbf{S}_{\text{fine}}$ **then**
    # (3) Structural Attention Correction (SAC)
    $Q_s^{\text{gen}} \leftarrow Q_s^{\text{con}} = W_Q F_s^{\text{con}}$
    $K_s^{\text{gen}} \leftarrow K_s^{\text{con}} = W_K F_s^{\text{con}}$
  end if
end for
$I^{\text{gen}} \leftarrow \text{Decoder}(F_S^{\text{gen}})$
return I^{gen}

steps has limited impact, as the global structure is already established during the autoregressive generation process. Accordingly, we apply SAC from early steps through all subsequent steps to provide consistent guidance and achieve the best performance.

B. Details of the dual-stream generation mechanism

We provide a detailed description of our dual-stream generation process in Algorithm 1. Both the *content path* and *generation path* are conditioned on the same text prompt T ("`<content> in <style>`") and are executed jointly within a single inference batch. Using identical conditioning prevents semantic mismatch between the two streams and ensures that both evolve under the same textual supervision.

The content path follows the original inference process of the pre-trained model without modification, producing a sequence of features $\{F_s^{\text{con}}\}_{s=1}^{S}$, which serve as a structural reference. Meanwhile, the generation path produces its own feature sequence $\{F_s^{\text{gen}}\}_{s=1}^{S}$, which is selectively modulated by our proposed mechanisms (PFB, SAC). Throughout inference, the content path provides structural guidance to the generation path, enabling it to preserve spatial consistency while integrating style information from the reference style image.

C. Implementation details

C.1. Implementation setup of comparison models

We conduct extensive comparisons against existing style-personalized image generation methods. To ensure fair and reproducible evaluation, all baseline models are run using publicly released implementations and their default hyperparameters, without additional tuning or prompt engineering unless explicitly required. We categorize baselines into two groups: (1) tuning-based approaches, which require style-specific fine-tuning before inference, and (2) training-free or pre-trained approaches, which operate directly without per-style optimization. For each method, we follow the official configurations provided in their respective repositories unless otherwise stated.

Tuning-based approaches These methods require fine-tuning a model for each reference style image. For each style reference, we performed style-specific fine-tuning following the official instructions of each repository, and report the **total runtime** consisting of both (1) training time per style and (2) inference time per image in the main paper.

- **B-LoRA** [2]: Official implementation: https://github.com/yardenfren1996/B-LoRA
- **DB-LoRA** [10]: Official implementation: https://github.com/huggingface/diffusers/tree/main/examples/dreambooth
- **DreamStyler** [1]: Official implementation: https://github.com/webtoon/dreamstyler
- **StyleDrop** [11]: Unofficial PyTorch reproduction: https://github.com/zideliu/StyleDrop-PyTorch

Training-free or pre-trained approaches These methods do not require additional fine-tuning per style. Instead, they operate either using a pre-trained style adapter or through direct inference-time conditioning. We evaluate all methods using their **official inference settings** and do not perform retraining or additional dataset-specific tuning.

- **IP-Adapter** [17]: Official implementation: https://github.com/tencent-ailab/IP-Adapter
- **StyleAligned** [4]: Official implementation: https://github.com/google/style-aligned
- **CSGO** [16]: Official implementation: https://github.com/instantX-research/CSGO
- **StyleAR** [15]: Official implementation: https://github.com/wuyi2020/StyleAR
- **Qwen-Image-Edit** [13]: Official implementation: https://huggingface.co/Qwen/Qwen-Image-Edit
- **Flux.1 Kontext** [7]: Official implementation: https://huggingface.co/black-forest-labs/FLUX.1-Kontext-dev

- **USO** [14]: Official implementation: https://github.com/bytedance/USO

All models are evaluated under a unified hardware environment using a single NVIDIA A6000 GPU with PyTorch.

C.2. Styles and prompts for generation

Fig. 10 presents the style prompts for the reference images used in the paper. Images marked with * indicate those used for the quantitative evaluation, for which we use the same prompts as Finestyle [18]. The style prompts serve as a high-level guide during the generation process, allowing the model to better align visual features with the target style. This provides a lightweight, training-free alternative to methods that require additional training. Note that our method does not rely on detailed prompts; simple, high-level style categories (e.g., "oil painting," "3d rendering," etc.) are sufficient.

C.3. User study details

To complement our quantitative evaluation, we conduct a user study involving 30 participants (ages 20s–50s). Participants compare results across two criteria: **prompt fidelity** (semantic alignment with text) and **style fidelity** (visual similarity to the reference style). Each comparison presents participants with a reference style image, a target text prompt, and outputs from multiple models.

We select comparison models based on their quantitative performance: StyleAligned [4] and IP-Adapter [17], which achieved the highest S_{img} (style fidelity), and DB-LoRA [10] and B-LoRA [2], which achieved the highest S_{txt} (prompt fidelity). This selection ensures that the user study compares the strongest-performing baselines under each metric.

As shown in Tab. 6, our method achieves the highest preference in **prompt fidelity** (35.3%) while maintaining competitive **style fidelity** (32.0%). Notably, prompt-tuned baselines (DB-LoRA, B-LoRA) exhibit strong semantic alignment but fail to preserve style, while style-focused baselines (StyleAligned, IP-Adapter) preserve style but lack semantic consistency. An example of the interface used in the study is shown in Fig. 11.

Table 6. User study preference results (percentage).

Model	Prompt Fidelity ↑	Style Fidelity ↑
StyleAligned [4]	4.3%	30.7%
IP-Adapter [17]	5.0%	23.3%
DB-LoRA [10]	26.7%	8.3%
B-LoRA [2]	28.7%	5.7%
Ours	**35.3%**	**32.0%**

D. Generalization across scale-wise autoregressive models

Unlike diffusion-based models that operate at a fixed spatial resolution, scale-wise autoregressive models generate images progressively across multiple scales, resulting in scale-dependent feature dynamics. This structural difference makes naive adaptation of prior attention-based methods ineffective in this setting, as directly applying such methods to our scale-wise autoregressive backbone, Infinity, without modification fails to yield meaningful results.

Our method is designed to be model-agnostic within the family of scale-wise autoregressive generative models, as it operates directly without modifying model weights or requiring retraining. To validate its generalization ability, we apply our method to two additional models beyond our primary backbone, Infinity-2B [3].

We first implement our method on Infinity-8B, a larger variant of our baseline model with increased capacity and parameter count. As shown in Fig. 12-(Top), our method produces consistent and stable stylization effects across this stronger model configuration, demonstrating robustness to architectural scaling without additional tuning or adaptation. We further apply our method to Switti [12], a distinct scale-wise autoregressive text-to-image model that differs structurally from Infinity. Despite architectural differences, our plug-and-play modules function reliably without modification, producing coherent, style-personalized generations, as shown in Fig. 12-(Bottom). This result supports our approach to generalizing across models that share the scale-wise autoregressive generation paradigm.

E. Future work and limitations

Our work presents a training-free style-personalized image generation framework grounded in a comprehensive analysis of a scale-wise autoregressive model. By identifying a key step that significantly influences the output image and demonstrating that dominant singular components of its feature space effectively capture style information, we establish a principled mechanism for style extraction and injection. We believe that this analysis opens up several promising future directions, enabling more precise and flexible control over style, content, and other visual attributes in personalized image generation systems.

Despite these strengths, our method faces limitations when the style reference image contains heterogeneous or conflicting stylistic attributes (e.g., mixed artistic media or multiple visual motifs), as it lacks an explicit mechanism to disentangle and selectively transfer specific sub-styles. Since our style extraction relies on dominant singular components, the injected style may reflect a blended representation of multiple styles rather than a feature representing a single, isolated style. Future research could incorporate

localized style decomposition, spatially variant basis representations, or user-guided selection to enable more fine-grained style control.

F. Additional qualitative results

F.1. Additional results

Fig. 13 presents additional qualitative results demonstrating that our method faithfully transfers style-specific information from the reference image while suppressing irrelevant details, effectively avoiding content leakage or mode collapse. This enables expressive and robust style personalization that generalizes well across diverse scenes and artistic styles.

F.2. Style-aligned image generation

Furthermore, we demonstrate that our model can perform style-aligned image generation using only a style prompt, without requiring a reference style image, by including a dedicated style pathway in the same batch derived from the style text prompt and leveraging its third feature as the style representation. As shown in Fig. 14, our model shows competitive performance compared to representative style-aligned image generation models [4, 19], indicating its capability in style-aligned image generation. These results validate that our method can operate effectively in both image-guided and text-guided style-related generation scenarios in a unified and training-free manner.

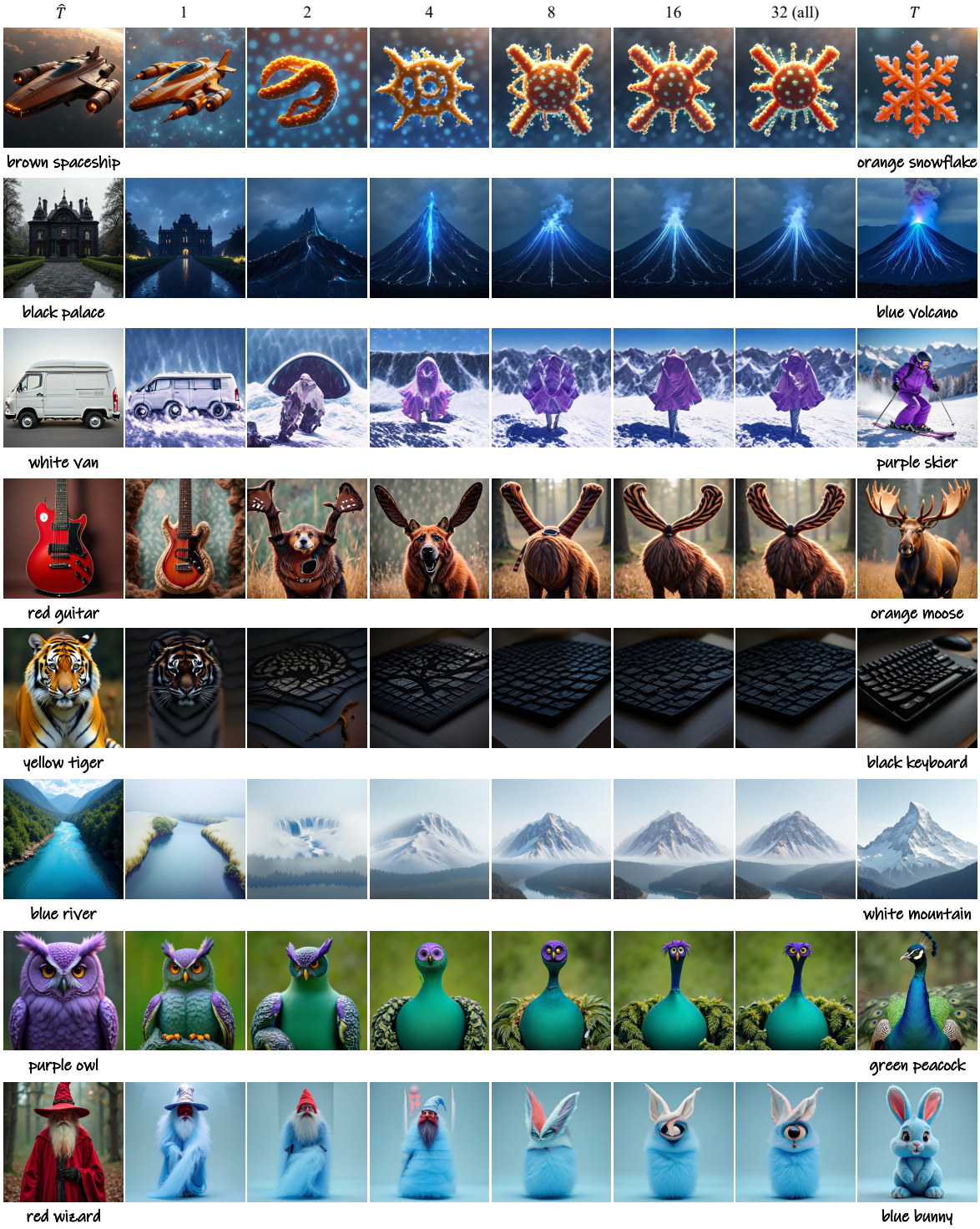


Figure 9. Qualitative results of SVD-guided feature replacement with varying k. From left to right: the baseline output generated from $\hat{T}$, SVD-guided outputs with $k \in \{1, 2, 4, 8, 16, 32\}$, and the baseline output generated from T.

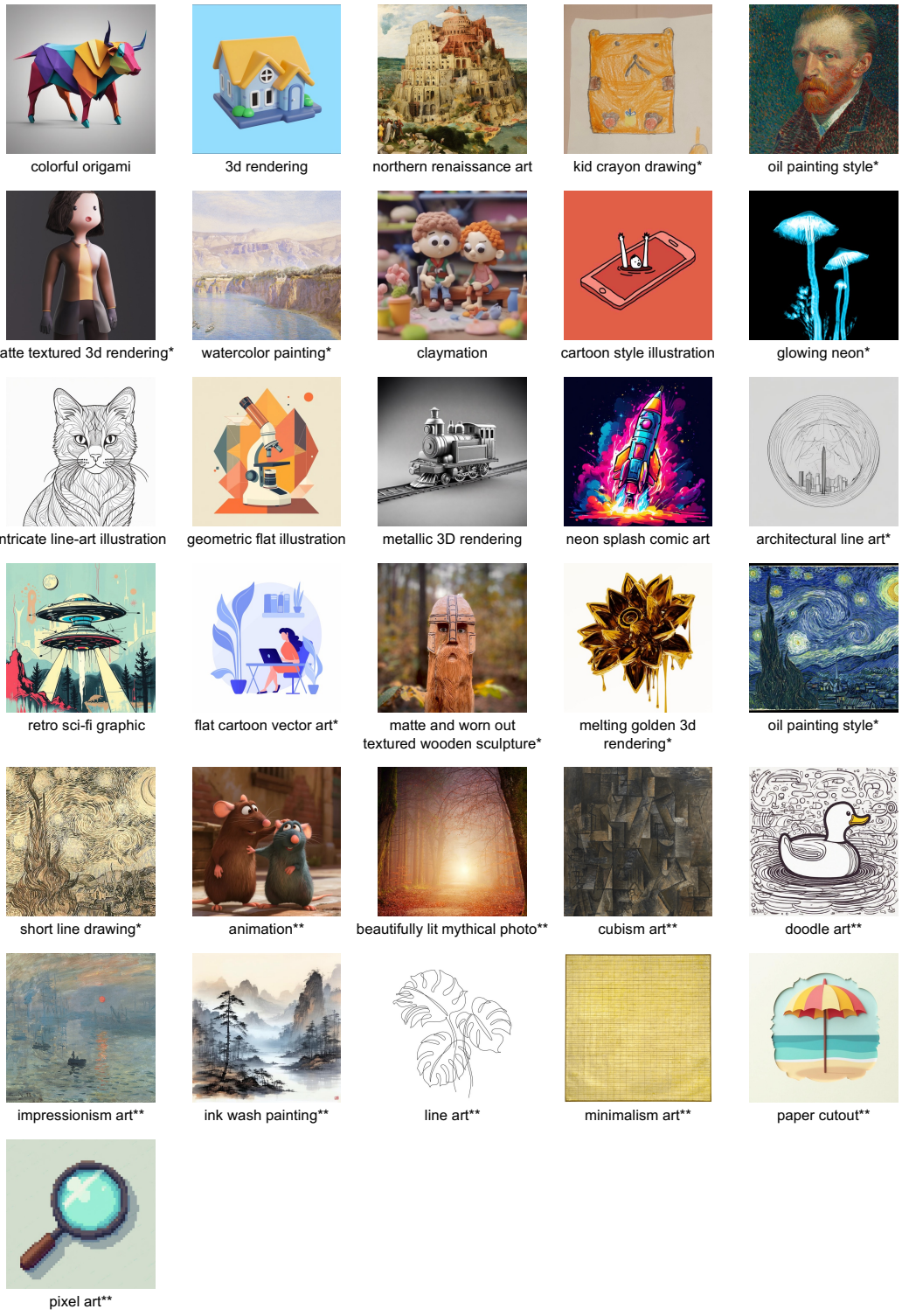
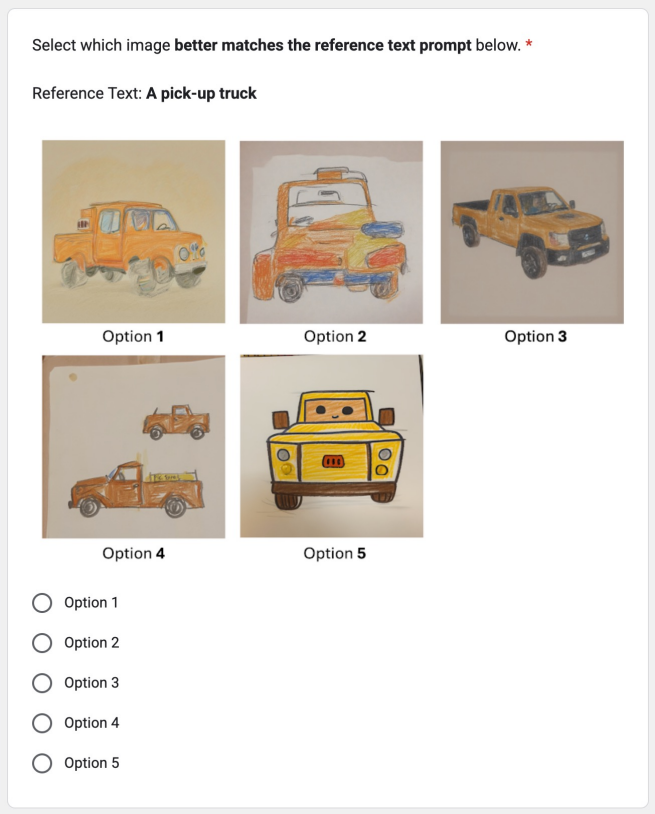


Figure 10. Style images and their corresponding prompts. The symbol * indicates those used for quantitative evaluation in the main paper, and ** indicates those used for extended evaluation in this supplementary material.

[Prompt Fidelity]



[Style Fidelity]

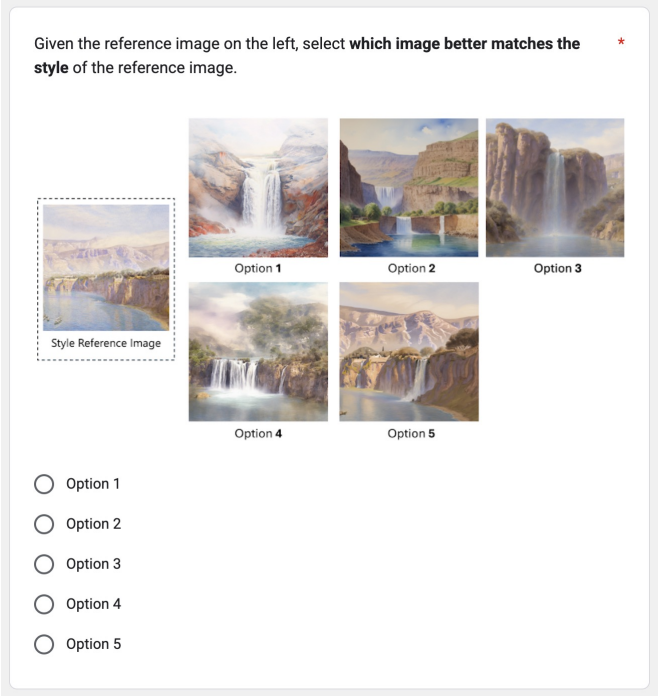


Figure 11. Example interface used in the user study. Participants selected the best-performing method among five candidates for each evaluation criterion.

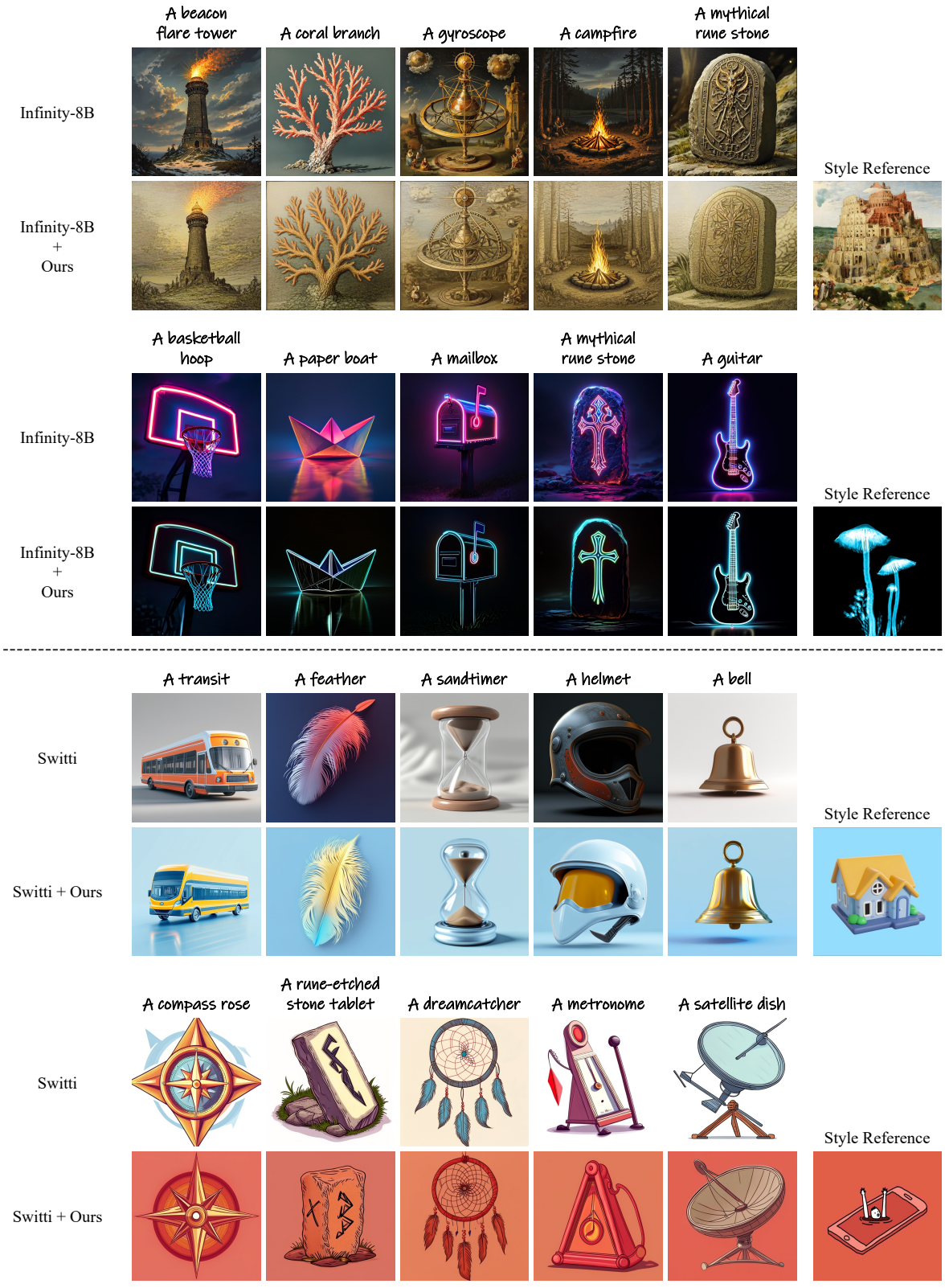


Figure 12. Qualitative results of applying our method to other scale-wise autoregressive models.

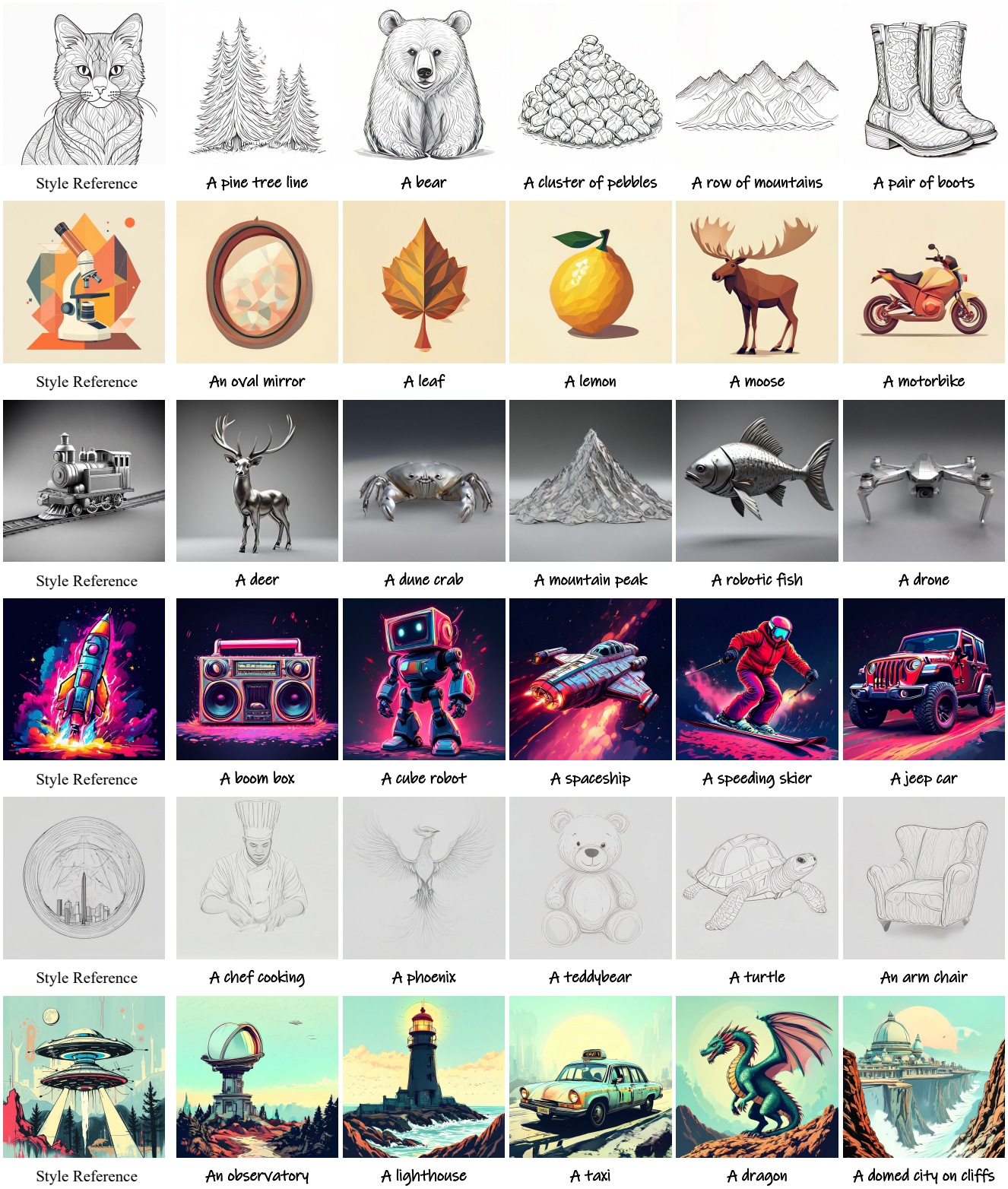


Figure 13. Various style-personalized results of our model.

Figure 14. Style-aligned image generation results with text-only style descriptions. Each row represents a different content prompt, and each column applies a distinct style, as described in the text.

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 674–681, 2024. 3

[2] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024. 1, 2, 3, 4

[3] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025. 4

[4] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 1, 2, 3, 4, 5

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[7] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 3

[8] Quang-Binh Nguyen, Minh Luu, Quang Nguyen, Anh Tran, and Khoi Nguyen. Csd-var: Content-style decomposition in visual autoregressive models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17013–17023, 2025. 1

[9] Jihun Park, Jongmin Gim, Kyoungmin Lee, Minseok Oh, Minwoo Choi, Jaeyeul Kim, Woo Chool Park, and Sunghoon Im. A training-free style-aligned image generation with scale-wise autoregressive model. *arXiv preprint arXiv:2504.06144*, 2025. 1

[10] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2022. *URL https://github.com/cloneofsimo/lora*, 10:19, 2022. 1, 2, 3, 4

[11] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 3

[12] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024. 4

[13] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 2, 3

[14] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025. 2, 4

[15] Yi Wu, Lingting Zhu, Shengju Qian, Lei Liu, Wandi Qiao, Lequan Yu, and Bin Li. Stylear: Customizing multimodal autoregressive model for style-aligned text-to-image generation. *arXiv preprint arXiv:2505.19874*, 2025. 3

[16] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 3

[17] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 3, 4

[18] Gong Zhang, Kihyuk Sohn, Meera Hahn, Humphrey Shi, and Irfan Essa. Finestyle: Fine-grained controllable style personalization for text-to-image models. *Advances in Neural Information Processing Systems*, 37:52937–52961, 2024. 4

[19] Jiexuan Zhang, Yiheng Du, Qian Wang, Weiqi Li, Yu Gu, and Jian Zhang. Alignedgen: Aligning style across generated images. *arXiv preprint arXiv:2509.17088*, 2025. 5