FREESTYLE: An Anchor-Free Mechanism for Training-Free Style-Aligned Image Generation

Supplementary Material

A. Generalizability of our method

A.1. Incorporating ours into other scale-wise autoregressive models

To verify the generalization capability of our framework, we integrate it into Switti [5], another Text-to-Image model that adopts the scale-wise autoregressive generation paradigm. Our method is applied to each residual feature $\mathbf{R}_s$ of Switti without modifying any pretrained parameters or training procedures.

As shown in Fig. 1, the model equipped with our method produces image sets that exhibit more consistent style alignment across diverse object prompts while faithfully preserving the semantic content described by the text. Compared to the vanilla Switti, our integration enhances cross-sample style coherence, demonstrating that the proposed mechanisms are architecture-agnostic and can be seamlessly incorporated into other scale-wise autoregressive T2I models to improve stylistic consistency without retraining.

Figure 1. Qualitative results on Switti with *FREESTYLE*. The row (a) uses prompt: {A friendly robot, A woman walking a dog, Cherryblossom} in flat cartoon illustration style. The row (b) uses prompt: {Flowers, Golden gate bridge, Trees} in water color painting style.

A.2. Incorporating ours into diffusion-based model

While our method is developed within the scale-wise autoregressive paradigm, we further demonstrate its generality by adapting *FREESTYLE* to diffusion-based generation. Specifically, we apply our approach to SDXL [3], a representative diffusion T2I model. Unlike autoregressive models, where *FREESTYLE* operates on residual features at each scale, we inject style information at the noise residual of each denoising step without modifying model weights or architecture.

As shown in Fig. 2, applying *FREESTYLE* to SDXL enforces consistent style across images while preserving the semantic content defined by the prompts. This result highlights that our method is not restricted to scale-wise autoregressive models and can be seamlessly integrated into diffusion-based pipelines in a training-free, anchor-free manner.

Figure 2. Qualitative results on SDXL with *FREESTYLE*. The row (a) uses prompt: {A friendly robot, A woman walking a dog, Cherryblossom} in flat cartoon illustration style. The row (b) uses prompt: {Flowers, Golden gate bridge, Trees} in water color painting style.

B. Style alignment from user-provided reference

We further extend our framework to support user-specified style control by combining scale-wise autoregressive inversion with our method. Following [2], we first invert a reference image into its scale-wise feature trajectory. During generation, the recovered style representations are reintroduced at each scale as guidance signals, while our method enforces batch-wise style consistency without any additional training or parameter updates.

As shown in Fig. 3, our approach successfully transfers the reference style across diverse object prompts while preserving textual semantics. These results demonstrate that our framework not only yields style-consistent outputs in model-inherent styles but also faithfully adapts to externally

provided reference style images in a controlled, prompt-aligned manner.

Figure 3. Results of reference image-based style-aligned image generation.

C. Additional analysis

C.1. Impact of cross-attention maps

To further analyze the effect of coarse and noisy cross-attention maps, we provide additional results examining how the quality of cross-attention masks relates to the final generation outcomes. Specifically, we use SAM3 [1] to obtain object segmentation masks and measure the MIoU between these masks and the cross-attention masks across all backbone scales. Fig. 4 compares the cases with high and low MIoU samples and shows that the style alignment remains stable. This suggests that cross-attention maps provide sufficient localization of prompt-critical regions for generation, even when they are relatively noisy, and our method remains robust with these key areas.

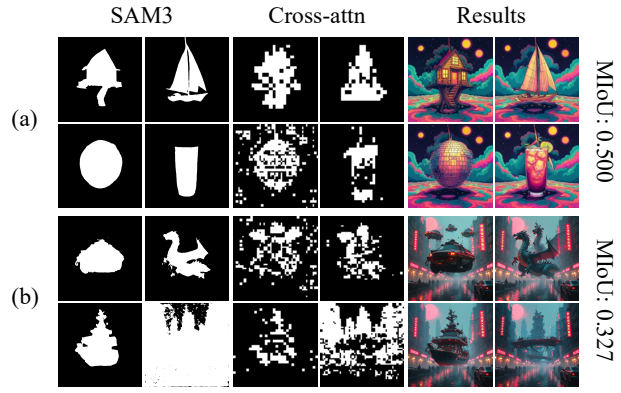

Figure 4. Visualization of the impact of cross-attention map accuracy.

C.2. Complex multi-object prompts

We provide additional qualitative results on complex multi-object prompts. These examples are intended to examine whether our method remains effective in more challenging settings involving multiple entities and attributes. While the models generally follow the prompts, some objects are occasionally omitted. As shown in Fig. 5, the baseline already fails to render the "fluffy cat," suggesting that such omissions mainly stem from limitations of the underlying backbone in the training-free regime. Nevertheless, our method yields more stylistically coherent results while largely maintaining the requested object composition.

Prompt : {a dreamy woman and a calm bulldog and a fluffy cat and a white swan, ... } in impressionist painting style.

Figure 5. Qualitative results with complex multi-object prompts.

C.3. Additional quantitative analysis

We provide additional quantitative comparisons with recent SOTA editing baselines using expanded evaluation metrics, including CSD [4] and $S_{\text{arithmetic}}$. Because image editing models operate in a reference-guided setting, a style reference image is required for evaluation. To this end, we first generate one image from the text prompt alone and then use it as the reference image for generating the remaining samples. Here, S_{CSD} measures style consistency and serves as an additional style-sensitive metric complementary to S_{DINO} and S_{CLIP}.

As shown in Tab. 1, our method achieves the best overall performance across most of the reported metrics, while remaining substantially faster than the other baselines. Qwen-Image-Edit achieves a high S_{DINO} score, but its relatively low S_{obj} and S_{whole} suggest less reliable preservation of the requested object composition and prompt semantics. In contrast, FLUX.2 obtains comparatively stronger object-related scores, but its lower style-related metrics suggest less faithful style alignment. Both S_{harmonic} and $S_{\text{arithmetic}}$ are included in the table for completeness.

C.4. AdaIN operation in Majority Voting (MV)

Majority Voting (MV) extracts a batch-wise representative residual style feature R_s^{maj} by selecting dominant residuals across samples. However, because this feature is taken

Table 1. Quantitative results with additional models and metrics.

Method	$S_{\text{harmonic}} \uparrow$	$S_{\text{arithmetic}} \uparrow$	$S_{\text{obj}} \uparrow$	$S_{\text{CSD}} \uparrow$	$S_{\text{DINO}} \uparrow$	$S_{\text{whole}} \uparrow$	$S_{\text{CLIP}} \uparrow$	Time (s) $\downarrow$
StyleAR	<u>0.455</u>	<u>0.541</u>	<u>0.281</u>	<u>0.768</u>	0.559	0.330	0.772	335.23
StyleAligned	0.447	0.525	<u>0.281</u>	0.723	0.530	<u>0.331</u>	0.762	11.25
IP-Adapter	0.442	0.522	0.278	0.708	0.529	0.324	0.772	<u>10.14</u>
Qwen-Image-Edit	0.431	0.526	0.257	0.663	**0.625**	0.301	<u>0.784</u>	154.80
FLUX.2	0.445	0.529	0.271	0.702	0.561	0.327	0.782	146.93
Ours	**0.463**	**0.553**	**0.284**	**0.772**	<u>0.589</u>	**0.332**	**0.791**	**1.98**

from real generated samples, it may still retain object-specific signals that interfere with purely style-driven propagation. To further isolate stylistic information, during majority voting, we apply AdaIN to R_s^{maj}, normalizing object-associated regions using statistics from style-dominant regions, thereby suppressing content cues while preserving style characteristics.

Tab. 2 reports quantitative results with and without the AdaIN refinement. Applying AdaIN yields consistent improvements in style consistency—both DINO-based (S_{DINO}) and CLIP-based metrics (S_{CLIP})—while also providing slight gains in prompt-related metrics (S_{obj}, S_{whole}). These results confirm that AdaIN effectively removes residual content signals from R_s^{maj}, producing cleaner style representations that lead to more robust and faithful style alignment.

Table 2. Quantitative ablation study on AdaIN operation in Majority Voting (MV).

Method	$S_{\text{harmonic}} \uparrow$	$S_{\text{obj}} \uparrow$	$S_{\text{DINO}} \uparrow$	$S_{\text{whole}} \uparrow$	$S_{\text{CLIP}} \uparrow$
W/o AdaIN	0.086	0.283	0.584	0.331	0.787
Ours	**0.088**	**0.284**	**0.589**	**0.332**	**0.791**

D. Limitation and future work

Our anchor-free method is training-free and aggregates batch-wise style information from generated samples to enforce a representative, shared style across the batch. While this approach effectively enhances style consistency when the model has a well-formed prior for the target style, it also inherits limitations from the pretrained backbone. When the base model has limited exposure to the target style, individual samples may express it inconsistently, resulting in noisy batch statistics and making it difficult to extract a stable representative style. In such cases, the final output may fail to fully reflect the intended style, not because the alignment mechanism fails, but because the underlying model lacks a reliable style manifold to align toward. These limitations suggest future extensions that expand style priors or incorporate external style memory for more robust alignment under unfamiliar styles.

E. Additional qualitative results

We present additional qualitative samples in Fig. 6 and Fig. 7, demonstrating our method across diverse styles and prompt variations. The results show that our approach consistently preserves content semantics while enforcing coherent style characteristics across samples. Unlike anchor-based propagation, our method does not depend on any specific sample, thereby avoiding artifact transfer and maintaining stable performance even under large variations in object prompts and stylistic conditions. These findings further validate the robustness and generality of our framework in practical, style-aligned generation scenarios.

F. Details of User Study

We conducted a user study with 40 participants (ages 20–50) using the interface shown in Fig. 8. The study is divided into two evaluation tasks, each presented as a multi-choice selection from four options (Options 1–4). The four options correspond to our method and three top-performing baselines selected based on quantitative scores, and their order was randomized for every trial to avoid positional bias.

Part 1: Text Relevance. Each trial displayed a content prompt along with four candidate image sets. Participants were asked to select the option that best matched the semantic meaning of the prompt.

Part 2: Style Consistency. Each trial presented different objects paired with the same style prompt, again with four candidate image sets. Participants were instructed to choose the set that most consistently reflects a unified visual style across samples.

Responses for both tasks were collected as single categorical choices per trial, allowing separate aggregation for text relevance and style consistency.

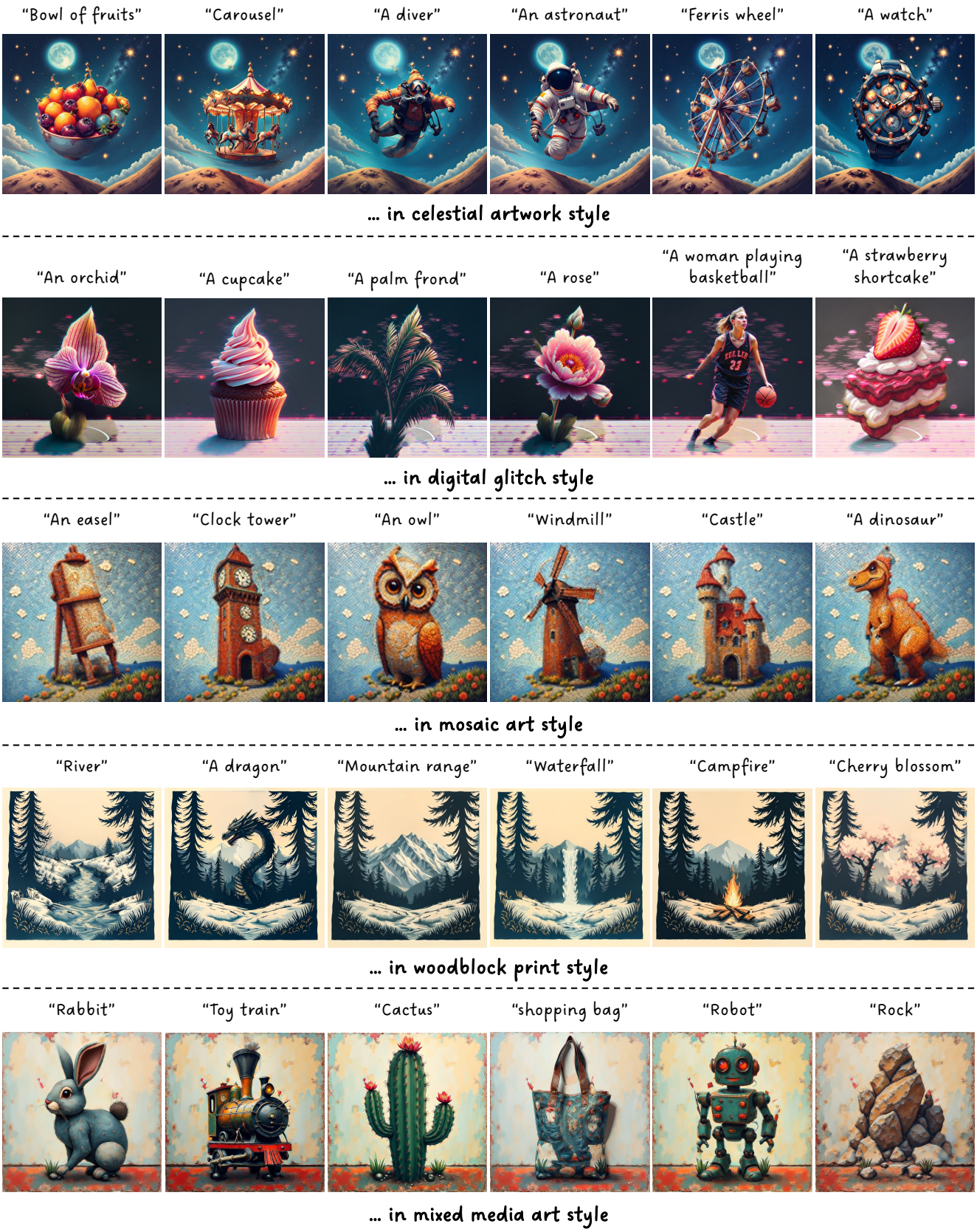

Figure 6. Additional qualitative results of *FREESTYLE* under diverse style-aligned image generation settings.

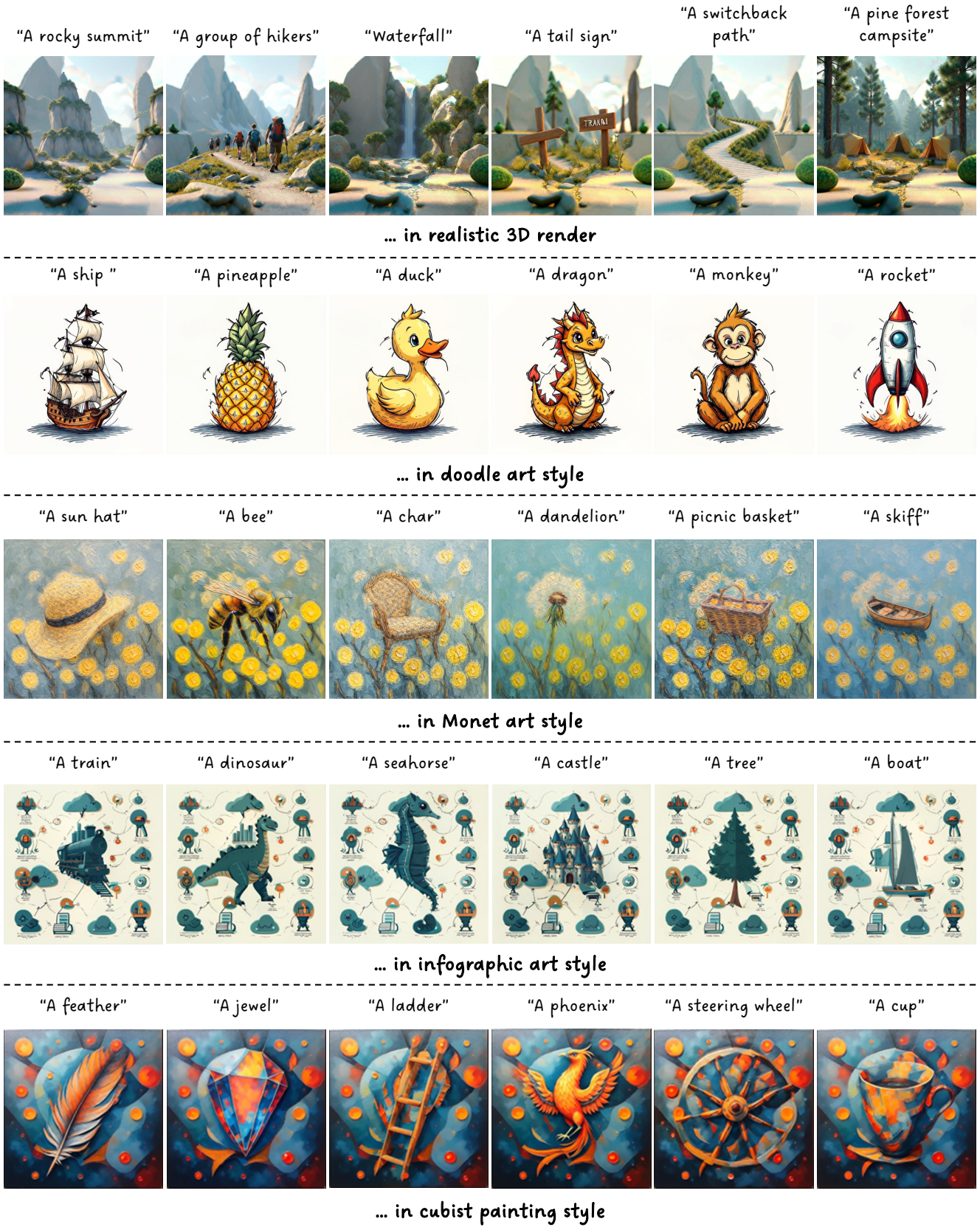

Figure 7. Additional qualitative results of *FREESTYLE* under diverse style-aligned image generation settings.

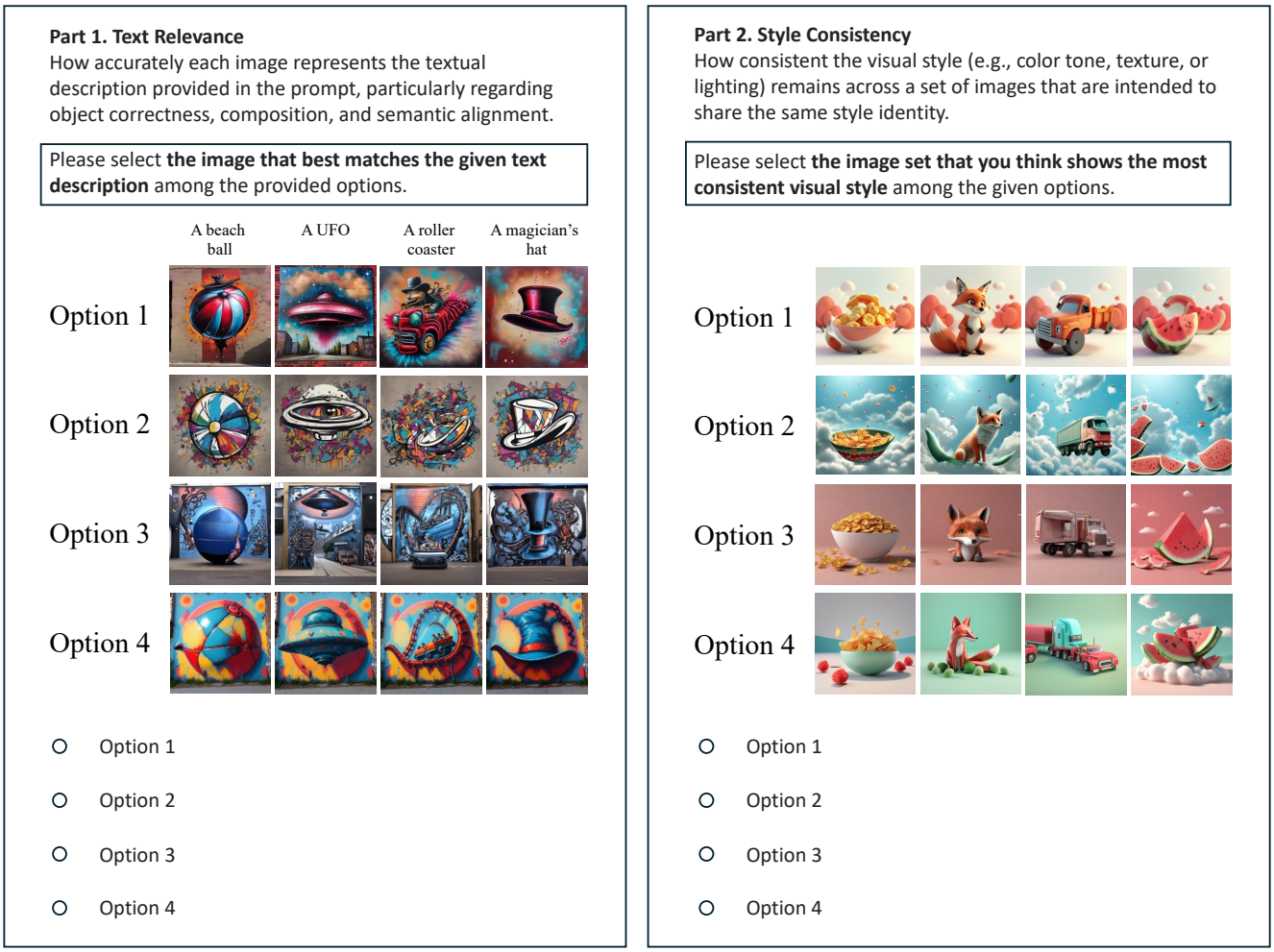

Figure 8. User study interface for text relevance (left) and style consistency (right). Among four randomized options, participants select the image that best matches the prompt or the set with the most consistent visual style.

References

[1] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 2

[2] Quan Dao, Xiaoxiao He, Ligong Han, Ngan Hoai Nguyen, Amin Heyrani Nobar, Faez Ahmed, Han Zhang, Viet Anh Nguyen, and Dimitris Metaxas. Discrete noise inversion for next-scale autoregressive text-based image editing. *arXiv preprint arXiv:2509.01984*, 2025. 1

[3] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[4] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 2

[5] 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. 1