Balanced Image Stylization with Style Matching Score

Yuxin Jiang[1,2] Liming Jiang[3] Shuai Yang[4] Jia-Wei Liu[1] Ivor W. Tsang[2,3] Mike Zheng Shou[1†]

[1]Show Lab, National University of Singapore
[2]CFAR & IHPC, Agency for Science, Technology and Research (A*STAR)
[3]Nanyang Technological University [4]Wangxuan Institute of Computer Technology, Peking University

Abstract

The document provides supplementary information not elaborated on in our main paper due to space constraints: implementation details (Section A), derivation of the Style Matching Objective (Section B), spectrum-based style analysis (Section C), additional optimization-based comparisons (Section D), further studies (Section E), more qualitative comparisons (Section F), additional results (Section G), limitations (Section H), and broader impact (Section I).

A. Implementation Details

A.1. SMS Training Procedure

Algorithm 1 details our style matching training procedure.

A.2. Style Data

We leverage the off-the-shelf style LoRA from Civitai [1] to support diverse artistic styles. To ensure a fair comparison with baseline methods that use different style representations, we carefully make the following adaptations: 1) Text-driven (*e.g.*, FreeStyle [4], DDS [5]): Use descriptive text prompts to capture the style. 2) Exemplar-guided (*e.g.*, StyleID [3], InstantStyle-Plus [15]): Source reference images from training data. 3) Collection-based (*e.g.*, StyleLoRA): Use Exactly the same LoRA.

A.3. Training time

Table 1 reports the per-image runtime (seconds) for baselines on an NVIDIA L40 GPU under default settings. Although some baselines require only forward steps, they require additional processing steps such as DDIM [13] inversion and substantial preparation (*e.g.*, ControlNet [17] training for InstantStyle-Plus [15] and Style-LoRA). In contrast, our SMS runs without any model-specific preparation, achieving a comparable overall runtimes. Moreover, the optimization-based nature of SMS enables it to extend to more complex, parameterized representations, which is not straightforward with other methods.

Algorithm 1: SMS Training Procedure

Input: Source image x^{src}; text prompt y^{src}; editing instruction y^{edit}; number of training iterations N
Output: Trained generator G_θ
Require: Pretrained SD diffusion denoiser ϵ_{real}; style-specific LoRA integrated into ϵ_{real} yielding ϵ_{style}; trainable LoRA integrated into ϵ_{real} yielding $\epsilon^{\phi}_{\text{fake}}$; SD VAE encoder $\mathcal{E}$
Initialization: $\epsilon^{\phi}_{\text{fake}} \leftarrow \text{copyWeights}(\epsilon_{\text{real}})$

for $i = 1$ **to** N **do**

/* Generate stylized images */
$x^{\text{tgt}} \leftarrow G_\theta(x^{\text{src}})$

/* Prepare latents */
$z_0^{\text{src}} \leftarrow \mathcal{E}(x^{\text{src}})$
$z_0^{\text{tgt}} \leftarrow \mathcal{E}(x^{\text{tgt}})$
// Adaptive Narrowing Sampling
Sample $t \sim \mathcal{U}(t_{\text{min}}, t_{\text{upper}})$
Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
$z_t^{\text{src}} \leftarrow \sqrt{\bar{\alpha}_t} z_0^{\text{src}} + \sqrt{1-\bar{\alpha}_t}\epsilon$
$z_t^{\text{tgt}} \leftarrow \sqrt{\bar{\alpha}_t} z_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\epsilon$

/* Update generator */
// Semantic-Aware Gradient Refinement
$\mathcal{R}(z_t^{\text{src}}, t) = \text{Norm}(|\epsilon_{\text{real}}(z_t^{\text{src}}; y_{\text{edit}}, t) - \epsilon_{\text{real}}(z_t^{\text{src}}; y_\emptyset, t)|)$
$\mathcal{L}_{\text{style}} \leftarrow ||\mathcal{R} \odot [w_t(\epsilon_{\text{style}}(z_t^{\text{tgt}}; y_{\text{src}}, t) - \epsilon^{\phi}_{\text{fake}}(z_t^{\text{tgt}}; y_{\text{src}}, t))]||_2^2$
// Progressive Spectrum Reglarization
$\mathcal{L}_{\text{freq}} \leftarrow ||\mathcal{F}_{\text{low}}(z_0^{\text{tgt}}, t), \mathcal{F}_{\text{low}}(z_0^{\text{src}}, t)||_2^2$
$\mathcal{L}_{\text{SMS}} \leftarrow \mathcal{L}_{\text{style}} + \lambda \mathcal{L}_{\text{freq}}$
$G_\theta \leftarrow \text{update}(\theta, \nabla_\theta \mathcal{L}_{\text{SMS}})$

/* Update trainable LoRA */
Sample $t \sim \mathcal{U}(t_{\text{min}}, t_{\text{max}})$
Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
$z_t^{\text{tgt}} \leftarrow \sqrt{\bar{\alpha}_t} z_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\epsilon$
$\mathcal{L}^{\phi}_{\text{denoise}} \leftarrow ||\epsilon^{\phi}_{\text{fake}}(z_t^{\text{tgt}}, t) - \epsilon||_2^2$
$\epsilon^{\phi}_{\text{fake}} \leftarrow \text{update}(\phi, \nabla_\phi \mathcal{L}^{\phi}_{\text{denoise}})$

Table 1. Image stylization per-image runtime comparison.

	FreeStyle	StyleID	InstantStyle+	Style-LoRA	DDS	SMS
Style	Text	Exemplar	Exemplar	LoRA	Text	LoRA
Train	-	-	ControlNet ($\sim$ 600 h) + IPAdapter ($\sim$ 192 h)	ControlNet ($\sim$ 600 h)	-	-
DDIM Inv	-	6.553	23.688	-	-	-
Inference	28.136	2.683	18.375	2.323	31.716	87.582

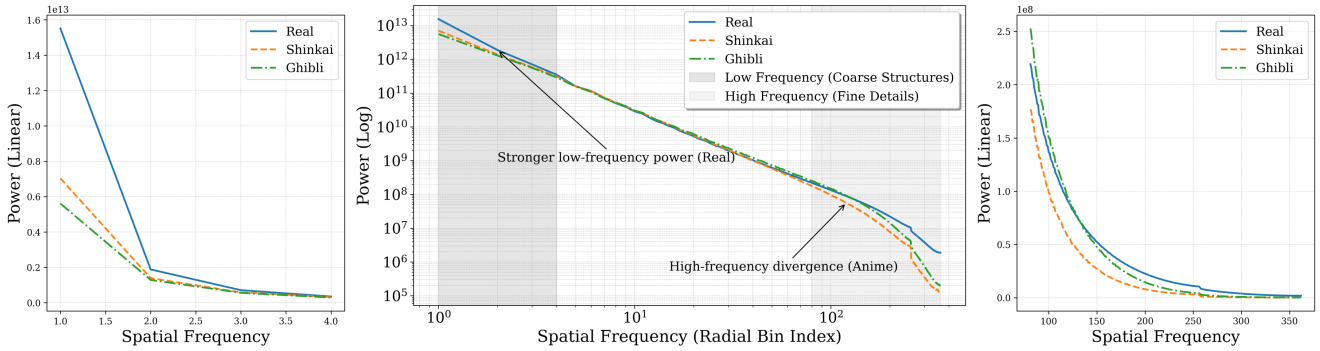

Figure 1. **Comparison of RAPSD**: Real-world images (Real) vs. Anime styles (Shinkai and Ghibli). **Left:** Zoom-in on low-frequency range (linear scale). **Middle:** Full-spectrum analysis (log-log scale). **Right:** Zoom-in on the high-frequency range (linear scale), revealing reduced high-frequency power in anime styles, which corresponds to smoother textures and stylized simplicity.

B. Derivation for Style Matching Objective

We derive the style matching objective (see Section 3.2 in the main paper) by using score functions approximated by DMs to minimize the KL divergence between the generated distribution p_{G_θ} and the target style distribution p_{style}. This derivation connects Equation (1) to Equation (2) in the main paper.

B.1. Gradient of the KL Divergence

Starting from the KL divergence:

$$D_{\text{KL}}(p_{G_\theta}||p_{\text{style}}) = \int p_{G_\theta}(x^{\text{tgt}}) \log \frac{p_{G_\theta}(x^{\text{tgt}})}{p_{\text{style}}(x^{\text{tgt}})} dx^{\text{tgt}}, \quad (1)$$

where $x^{\text{tgt}} = G_\theta(x^{\text{src}})$ and G_θ is the generator parametrized by θ. Our goal is to compute the gradient of D_{KL} with respect to θ:

$$\nabla_\theta D_{\text{KL}} = \nabla_\theta \int p_{G_\theta}(x^{\text{tgt}}) \log \frac{p_{G_\theta}(x^{\text{tgt}})}{p_{\text{style}}(x^{\text{tgt}})} dx^{\text{tgt}}. \quad (2)$$

Using the property $\nabla_\theta p_{G_\theta}(x) = p_{G_\theta}(x) \nabla_x \log p_{G_\theta}(x)$, we can express the gradient as:

$$\nabla_\theta D_{\text{KL}} = \int p_{G_\theta}(x) \nabla_\theta \log p_{G_\theta}(x) \log \frac{p_{G_\theta}(x^{\text{tgt}})}{p_{\text{style}}(x^{\text{tgt}})} dx^{\text{tgt}}. \quad (3)$$

Since p_{style} does not depend on θ, we have $\nabla_\theta \log p_{\text{style}}(x^{\text{tgt}}) = 0$. Furthermore, using the chain rule, we can compute $\nabla_\theta \log p_{G_\theta}(x^{\text{tgt}})$ as follows:

$$\begin{aligned} \nabla_\theta \log p_{G_\theta}(x^{\text{tgt}}) &= \left(\nabla_{x^{\text{tgt}}} \log p_{G_\theta}(x^{\text{tgt}})\right) \frac{\partial x^{\text{tgt}}}{\partial \theta} \\ &= s_{G_\theta}(x^{\text{tgt}}) \frac{\partial G_\theta(x^{\text{src}})}{\partial \theta}, \end{aligned} \quad (4)$$

where $s_{G_\theta}(x) := s_{\text{fake}}(x) = \nabla_x \log p_{G_\theta}(x)$ is the score function of the generated distribution. Following DMD [16], we name it the fake score. Substituting back into Equation (3):

$$\nabla_\theta D_{\text{KL}} = \int p_{G_\theta}(x^{\text{tgt}}) s_{\text{fake}}(x^{\text{tgt}}) \log \frac{p_{G_\theta}(x^{\text{tgt}})}{p_{\text{style}}(x^{\text{tgt}})} \frac{\partial G_\theta(x^{\text{src}})}{\partial \theta}. \quad (5)$$

Recognizing that the gradient of the log-density ratio is the difference of the score functions:

$$\nabla_x \log \frac{p_{G_\theta}(x)}{p_{\text{style}}(x)} = s_{\text{fake}}(x) - s_{\text{style}}(x), \quad (6)$$

where $s_{\text{style}}(x) = \nabla_x \log p_{\text{style}}(x)$ is the score function of the target style distribution. The integral can be expressed as an expectation over $x \sim p_{G_\theta}$:

$$\nabla_\theta D_{\text{KL}} = \mathop{\mathbb{E}}_{x^{\text{tgt}} \sim p G_\theta} \left[\left(s_{\text{style}}(x^{\text{tgt}}) - s_{\text{fake}}(x^{\text{tgt}})\right) \frac{\partial G_\theta(x^{\text{src}})}{\partial \theta} \right], \quad (7)$$

indicating that the gradient is pointing in the direction that moves p_{G_θ} closer to p_{style}.

B.2. Approximating Score Functions with Diffusion Models

We approximate the score functions $s_{\text{style}}(x^{\text{tgt}})$ and $s_{\text{fake}}(x^{\text{tgt}}$ using diffusion models [14, 16]. The score function of the data distribution $s(x)$ is related to the time-dependent score function $s(z_t, t)$ through the diffusion process, where z_t is obtained by adding Gaussian noise to $z_0 = \mathcal{E}(x)$.

Equivalence of Noise and Data Prediction Before proceeding with the substitution into the gradient expression, it is beneficial to convert the data prediction models μ to noise prediction models ϵ. This conversion simplifies the derivation and aligns with practical implementations, as DMs are typically trained to predict the noise. The relationship is given by [9]:

$$\mu(z_t, t) = \frac{z_t - \sigma_t \epsilon(z_t, t)}{\alpha_t}. \quad (8)$$

Rewriting the score function in terms of the noise prediction model, we have:

$$s(z_t, t) = \nabla_{z_t} \log p(z_t) = \frac{z_t - \alpha_t \mu(z_t, t)}{\sigma_t^2} = \frac{\epsilon(z_t, t)}{\sigma_t} \tag{9}$$

Target style score. The target style distribution $p_{\text{style}}(x)$ is modeled using a pretrained DM with a style-specific LoRA ϵ_{style}. The score function is: $s_{\text{style}}(z_t, t) = \frac{\epsilon_{\text{style}}(z_t, t)}{\sigma_t}$.

Generated fake score. Similarly, we model the generated distribution p_{G_θ} using a DM with trainable LoRA $\epsilon^{\phi}_{\text{fake}}$. The score function is: $s_{\text{fake}}(z_t, t) = \frac{\epsilon^{\phi}_{\text{fake}}(z_t, t)}{\sigma_t}$. We train $\epsilon^{\phi}_{\text{fake}}$ to model the distribution of the generated images $z_0^{\text{tgt}} = \mathcal{E}(G_\theta(x^{\text{src}}))$ by minimizing the standard denoising objective [6]:

$$\mathcal{L}^{\phi}_{\text{denoise}} = ||\epsilon^{\phi}_{\text{fake}}(z_t, t) - \epsilon||_2^2, \tag{10}$$

Substituting the approximations into Equation (7), we obtain:

$$\nabla_\theta D_{\text{KL}} \simeq \mathbb{E}_{t,\epsilon}\left[w_t \left(\epsilon_{\text{style}}(z_t, t) - \epsilon^{\phi}_{\text{fake}}(z_t, t) \right) \frac{\partial G_\theta(x^{\text{src}})}{\partial \theta} \right]. \tag{11}$$

C. Spectrum-Based Style Analysis

To identify and quantify the gap between the real and style domains, we analyze their spectral differences, focusing on two representative anime styles: Shinkai and Ghibli. Using $5,958$ Shinkai images [7], 714 Ghibli images and $90,000$ real-world images [12], we calculate the Radially Averaged Power Spectral Density (RAPSD) for each domain.

Figure 1 shows that real images have consistently higher power at both low and high frequencies. In contrast, anime styles demonstrate reduced high-frequency power, suggesting smoother textures and a uniform representation of details. This aligns with its artistic choices in anime, where sharp transitions and clean edges are emphasized while avoiding natural noise and irregularities in real-world images. Inspired by this gap, we propose a progressive spectrum regularization term (see Section 3.3) that aligns the spectral properties of generated images with the target style domain, allowing faithful stylization while maintaining structural fidelity.

D. Additional Optimization-based Method Comparisons

In the main paper, we select DDS [5] as the representative optimization-based method for clarity. Although other score distillation methods such as SDS [11] and PDS [10] are technically relevant, our experiments show that these methods fail in global style transfer, resulting in poorer performance (see Figure 2(Row 1,3)). Furthermore, when we apply the same style LoRA priors to these text-guided optimization methods, the results (see Figure 2(Row 2,4)) indicate that they do not fully leverage the style LoRA for capturing style information.

Figure 2. Comparison of optimization-based methods (SDS [11], DDS [5] and PDS [10]) with and without style LoRA priors on Ghibli and oil painting styles.

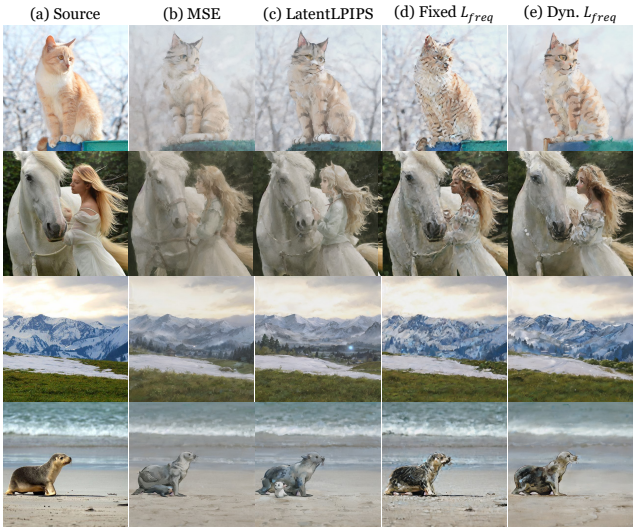

Figure 3. **Comparison of identity loss variants.**

E. Further Studies

E.1. Identity Loss Variant Study

In Section 3.3 of the main paper, we introduce a novel progressive spectrum regularization in the frequency domain, instead of traditional spatial domain identity preservation losses. While we have already ablated its effectiveness

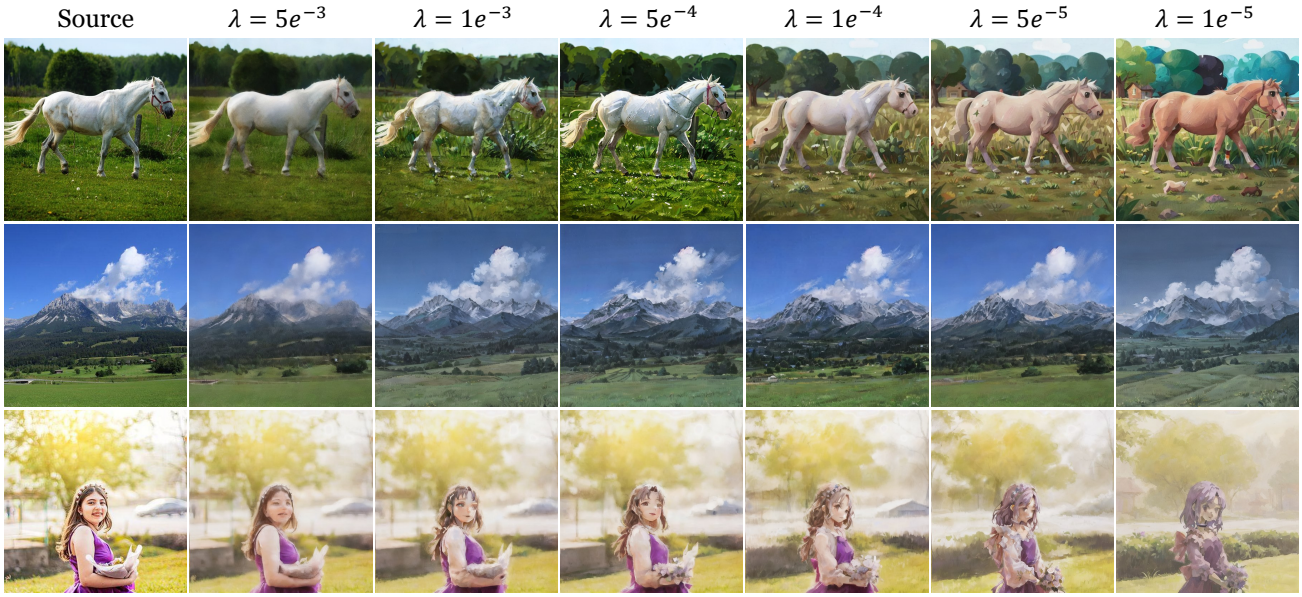

Figure 4. **Effects of the loss weight λ.** The first row shows results in kids illustration style, and the last two rows show results in Ghibli style.

in Section 5.3, we further verify its utility by comparing it against other latent space identity loss variants: spatial Mean Square Error (MSE) [10] and E-LatentLPIPS [8]. Additionally, we test a fixed frequency threshold $\text{thld}(t) = 0.3$, retaining the top 30% of low-frequency components, as opposed to our timestep-aware progressive approach.

The qualitative results, presented in Figure 3, illustrate the limitations of these alternatives. The MSE loss applied uniform regularization across all pixels, leading to blurriness and an inability to balance content fidelity with style adaptation (see Figure 3(b)). The LatentLPIPS loss, despite focusing on high-level feature alignment, struggles to maintain sufficient identity while incorporating style details. Adopting a fixed frequency cutoff results in oversharpened artifacts, underscoring the necessity of timestep-aware frequency regularization. In contrast, our method successfully translates intricate high-frequency style textures, such as hairs details (see Figure 3(e), Row 2) and seal skin (Row 4) while preserving low-frequency structure fidelity, like the mountain ridgeline (Row 3). Our progressive spectrum regularization strikes a balance between high-frequency style transfer fidelity and low-frequency content preservation.

E.2. Effects of Loss Weight λ Study

The strength of the explicit identity regularization term is determined by the loss weight λ. As shown in Figure 4, increasing λ enhances content fidelity, while reducing it allows for stronger stylization, demonstrating a clear trade-off between style and content. It provides a user-controllable knob for adjusting the stylization strength.

F. More Qualitative Comparisons

We present additional qualitative comparisons with five state-of-the-art methods. As shown in Figure 6, SMS achieves superior content preservation, maintaining structural integrity and ensuring a harmonious color balance, all while delivering comparable stylization results.

G. Additional Results

We provide additional examples of images generated by SMS on the DIV2K dataset [2] to showcase its superior high-quality balanced stylization ability across different styles. Figures 7, 8, 9, 10, 11, 12 displays stylizations in watercolor, oil painting, Ghibli, Ukiyo-e, kids illustration and sketch styles, respectively.

H. Limitations

Despite the promising results, our method has certain limitations. SMS relies on style-specific LoRAs, and if a LoRA lacks sufficient content diversity, especially for specific object categories, distortions may occur. For example, using an oil painting style LoRA that trained with few or no images of jellyfish can result in stylized outputs where jellyfish are inaccurately transformed into other objects, such as human figure (see Figure 5(a)). This issue arises because the LoRA has not learned appropriate representations for those unseen or underrepresented content types. Increasing the content preservation parameter λ may mitigate this problem, albeit at the cost of reduced stylization strength.

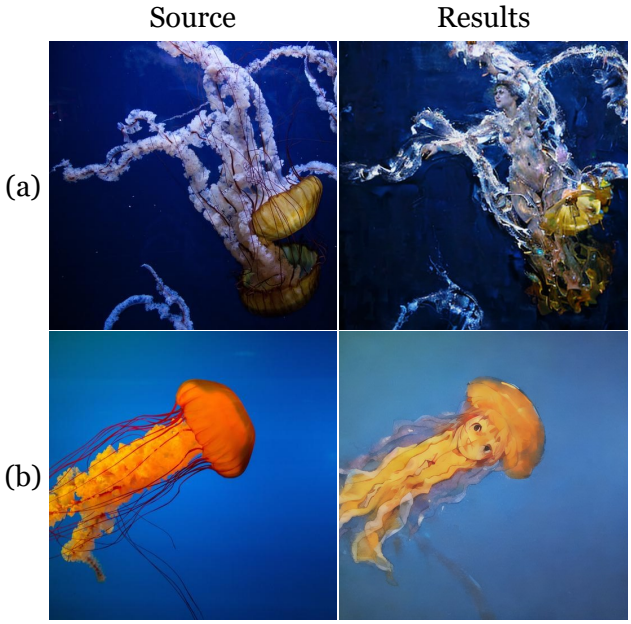

Figure 5. **Certain failure cases.** (a) oil painting style with text prompt y^{src}: *a group of jellyfish floating on top of a body of water.* (b) watercolor style with text prompt y^{src}: *a jellyfish swimming in the ocean.*

I. Broader Impact

Our stylization framework has significant societal impacts. Positively, it can enhance creativity in graphic design, animation, and digital art, offering powerful tools for high-quality style transfer. It also holds promise for personalized education and immersive entertainment experiences.

However, we must be mindful of potential negative consequences. Biases present in training datasets can propagate through generative models, potentially amplifying societal inequities. Furthermore, the ability to train a style-LoRA with limited artistic works and use SMS to transform other images into an artist's style raises concerns regarding intellectual property rights and copyright protection. Careful ethical considerations and adherence to copyright laws are crucial to mitigate these risks.

References

[1] Civit AI, Inc. https://civitai.com/. 1

[2] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR workshops*, 2017. 4

[3] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, 2024. 1

[4] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. FreeStyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, 2024. 1

[5] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta Denoising Score. In *ICCV*, 2023. 1, 3

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3

[7] Yuxin Jiang, Liming Jiang, Shuai Yang, and Chen Change Loy. Scenimefy: learning to craft anime scene via semi-supervised image-to-image translation. In *ICCV*, 2023. 3

[8] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling Diffusion Models into Conditional GANs. In *ECCV*, 2024. 4

[9] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022. 2

[10] Juil Koo, Chanho Park, and Minhyuk Sung. Posterior Distillation Sampling. In *CVPR*, 2024. 3, 4

[11] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3

[12] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *ICCV*, 2021. 3

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[14] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[15] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. InstantStyle-Plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024. 1

[16] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 2

[17] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1

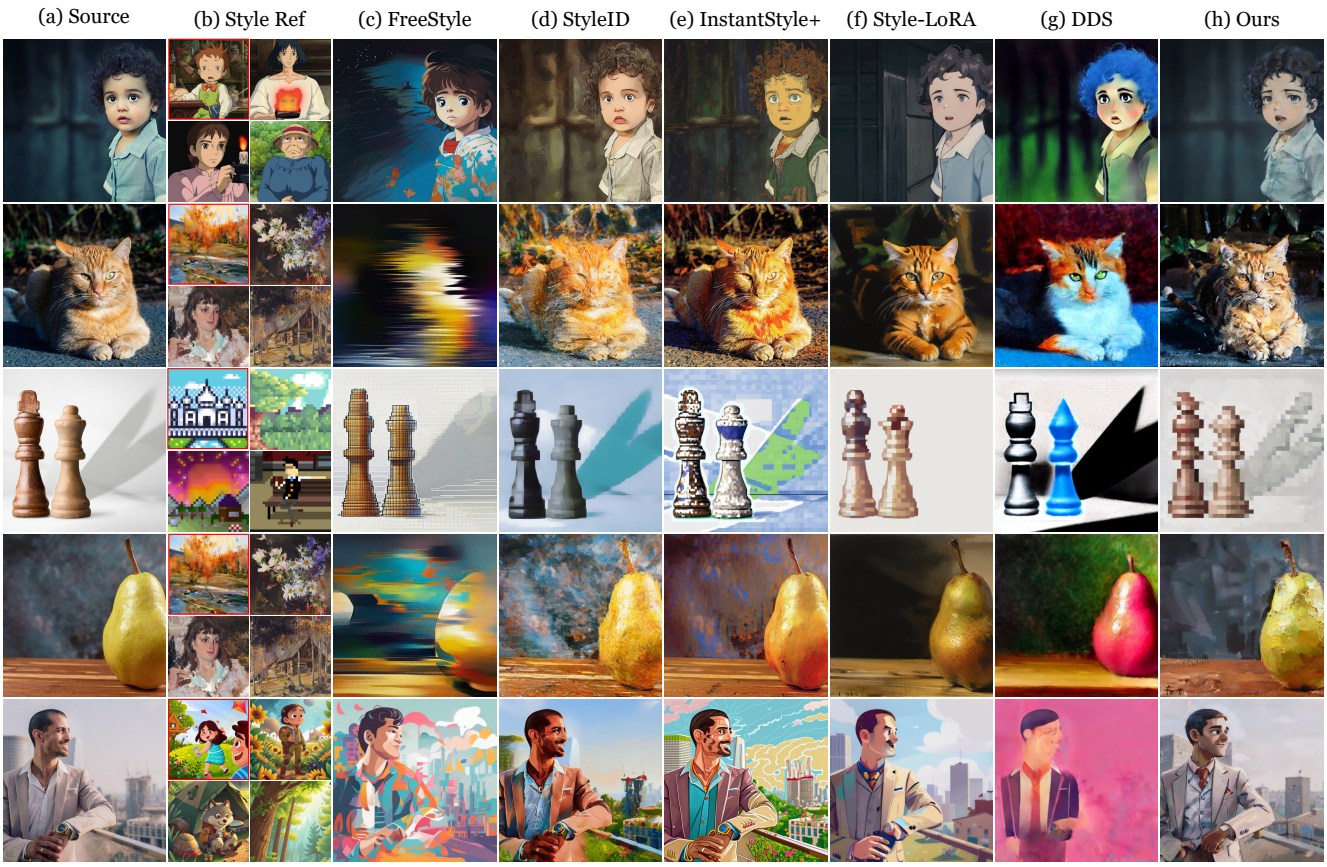

Figure 6. **Additional qualitative comparison** between SMS (Ours) and five representative methods.

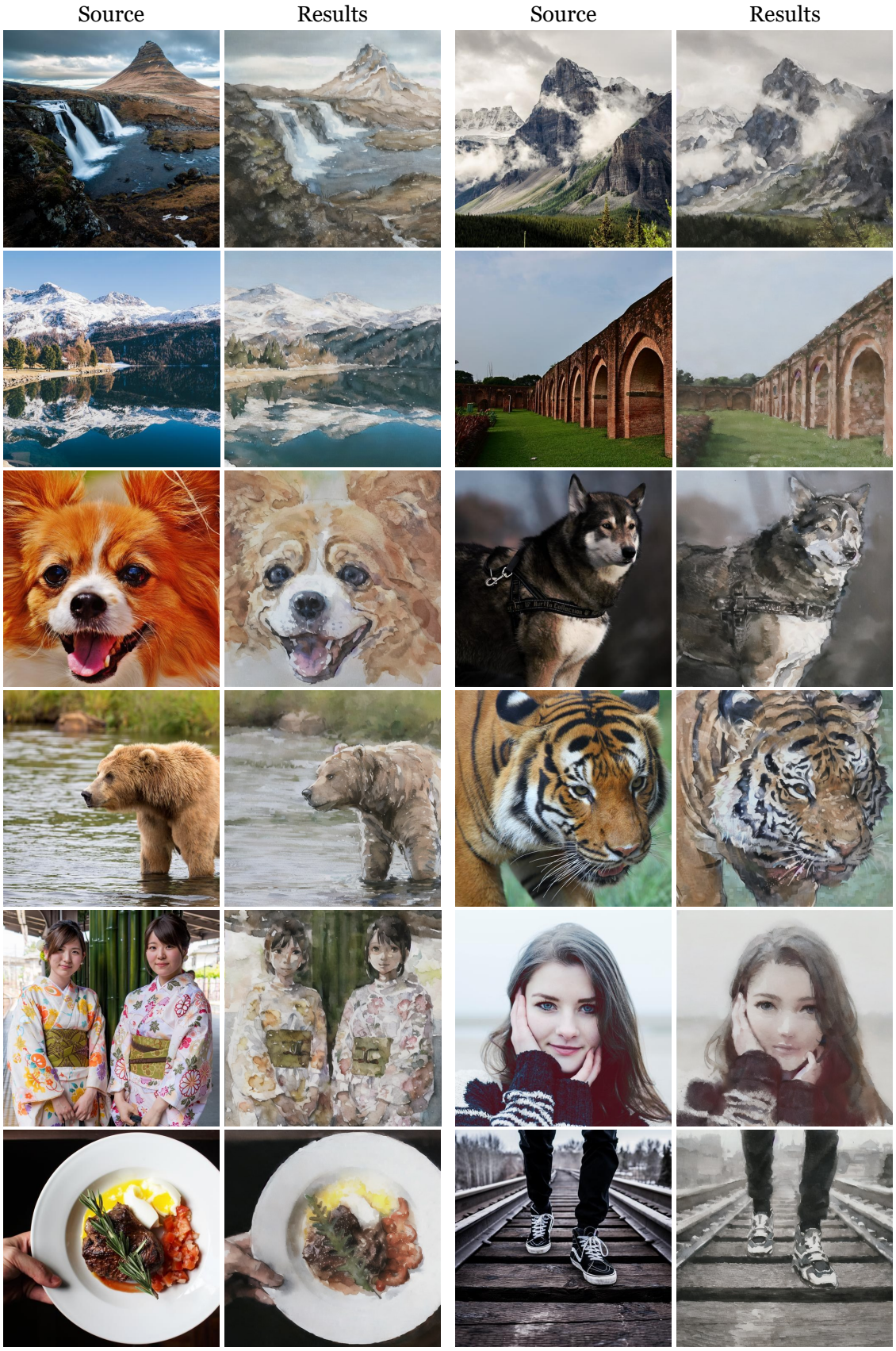

Figure 7. **Watercolor style.** The results capture the fluid and translucent qualities typical of watercolor paintings, with gentle color gradients and soft edges.

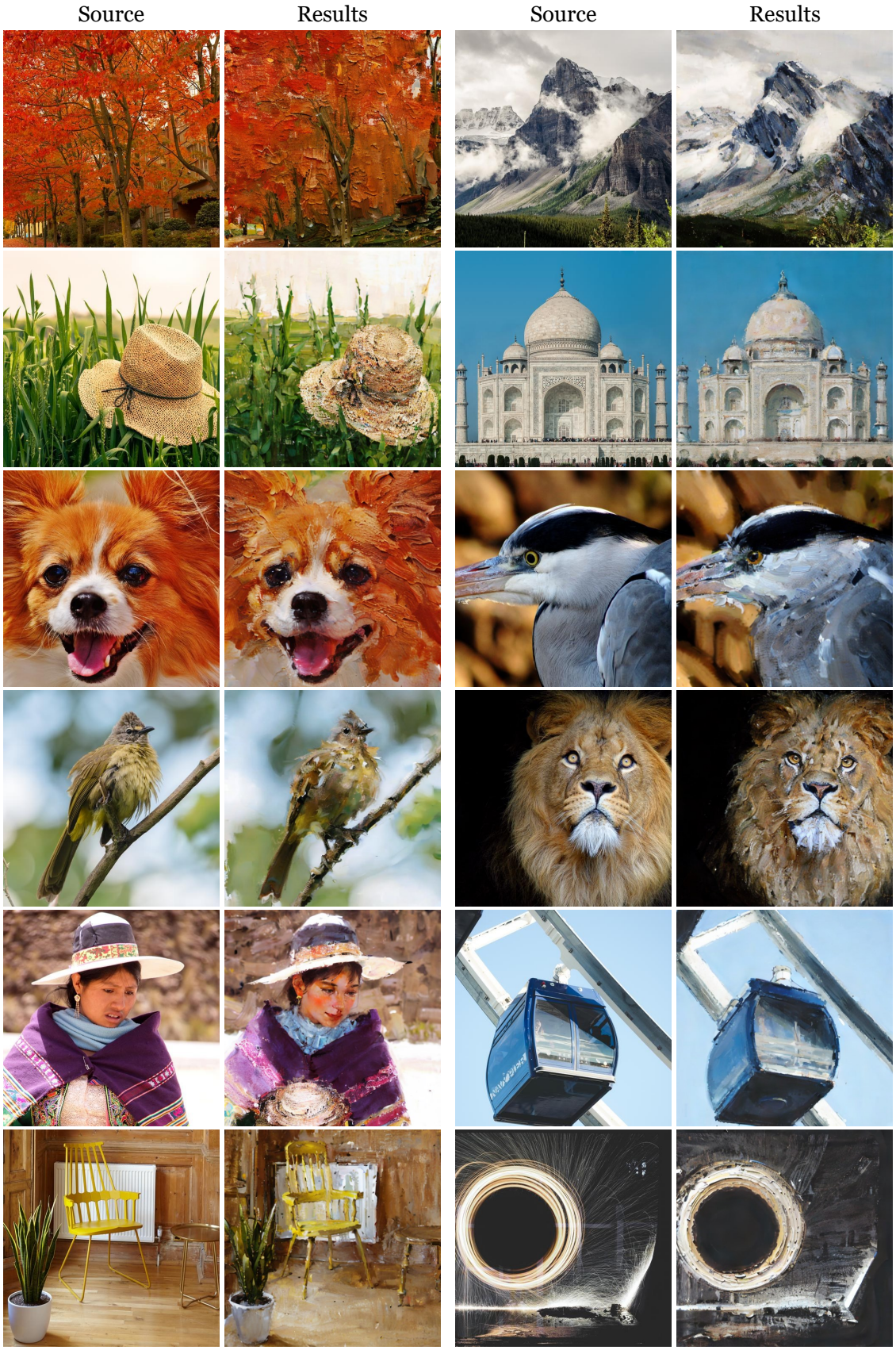

Figure 8. **Oil painting style.** The results reflect the rich textures and bold brushstrokes associated with oil paintings, emphasizing depth and vibrancy.

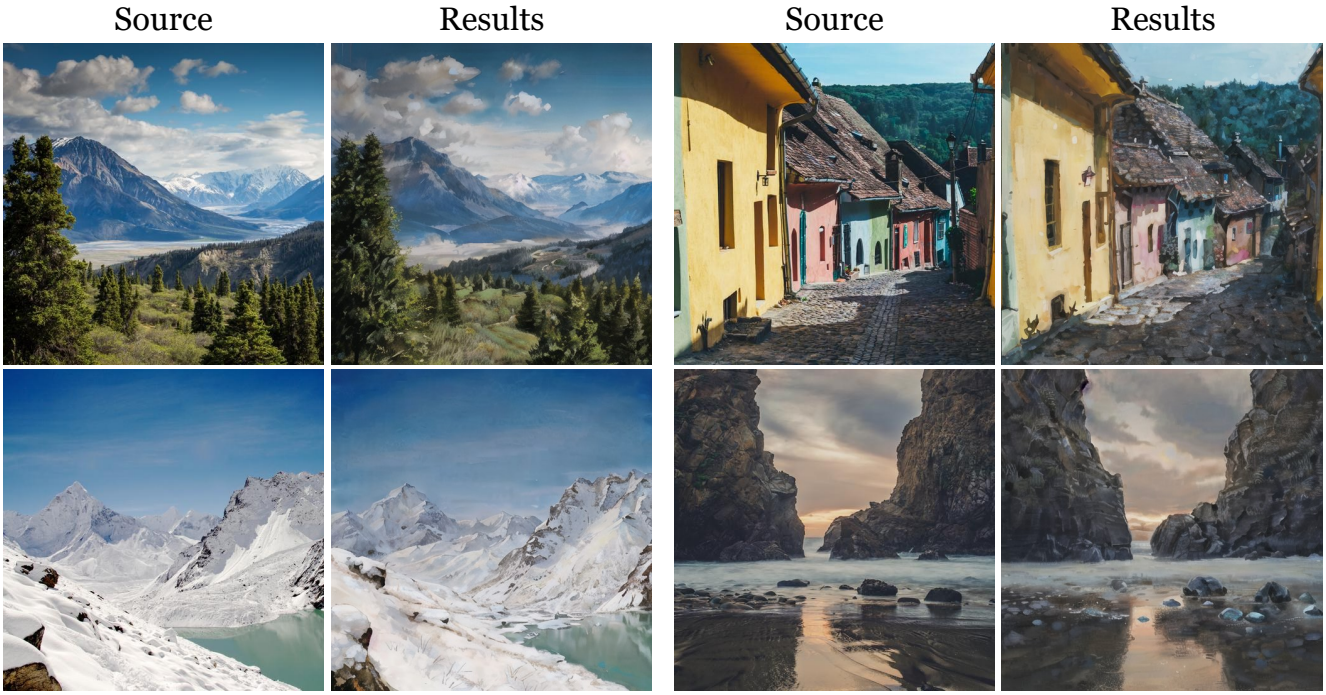

Figure 9. **Ghibli style.** The results create a harmonious blend of realism and painterly artistry characteristic of Studio Ghibli, combining intricate pre-designed brush-like strokes in the scenes.

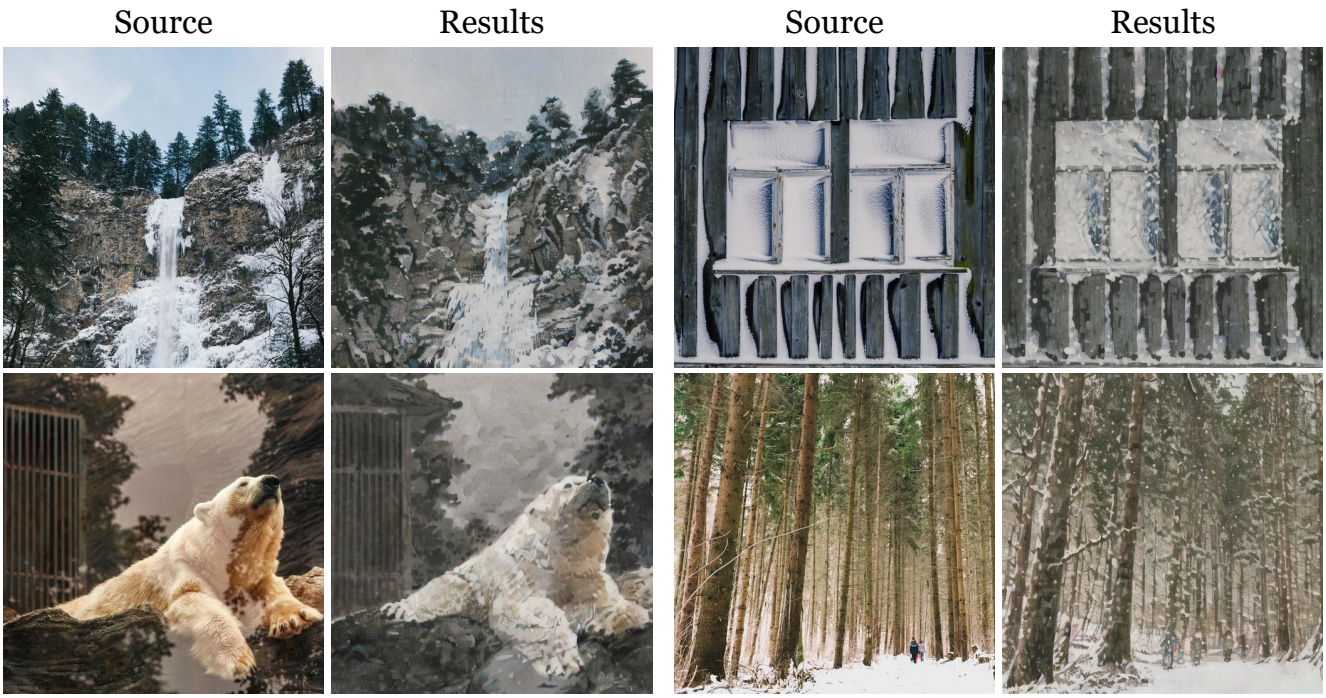

Figure 10. **Ukiyo-e style.** The results reflect the essence of traditional Japanese ukiyo-e woodblock prints.

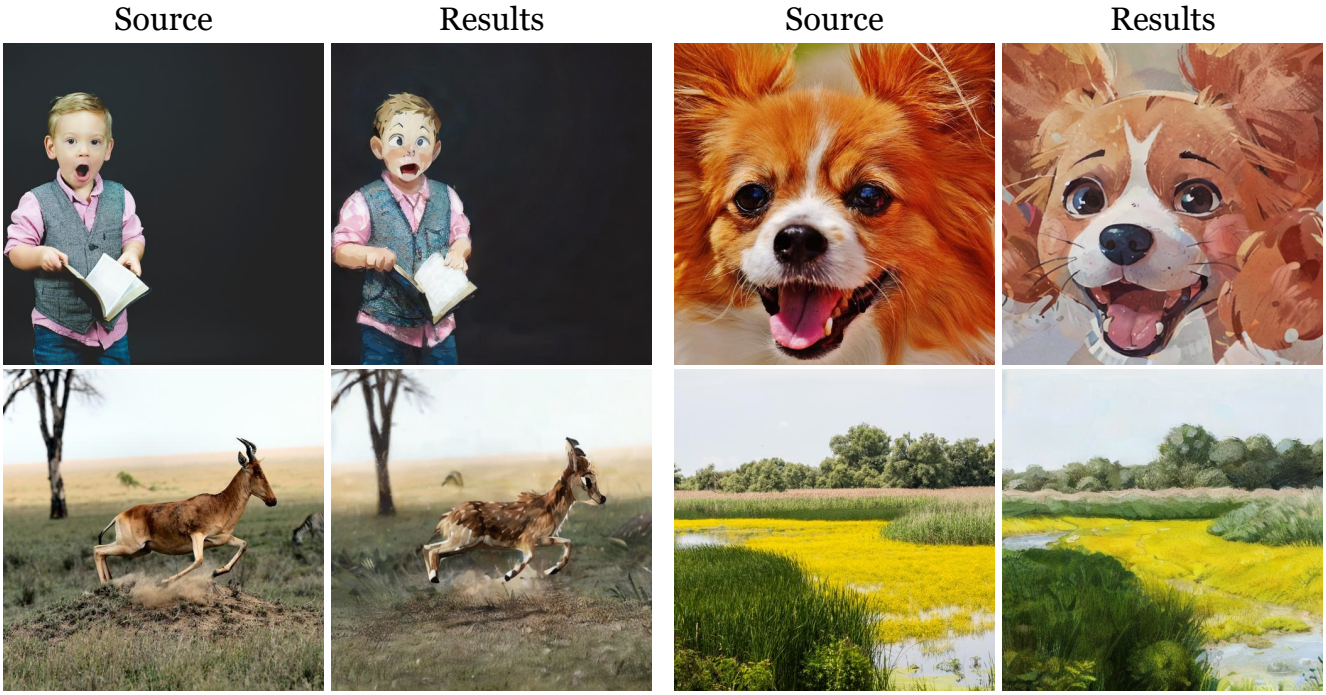

Figure 11. **Kids illustration style.** The results have playful and vibrant qualities typical of children's illustrations, featuring simplified shapes and bold outlines.

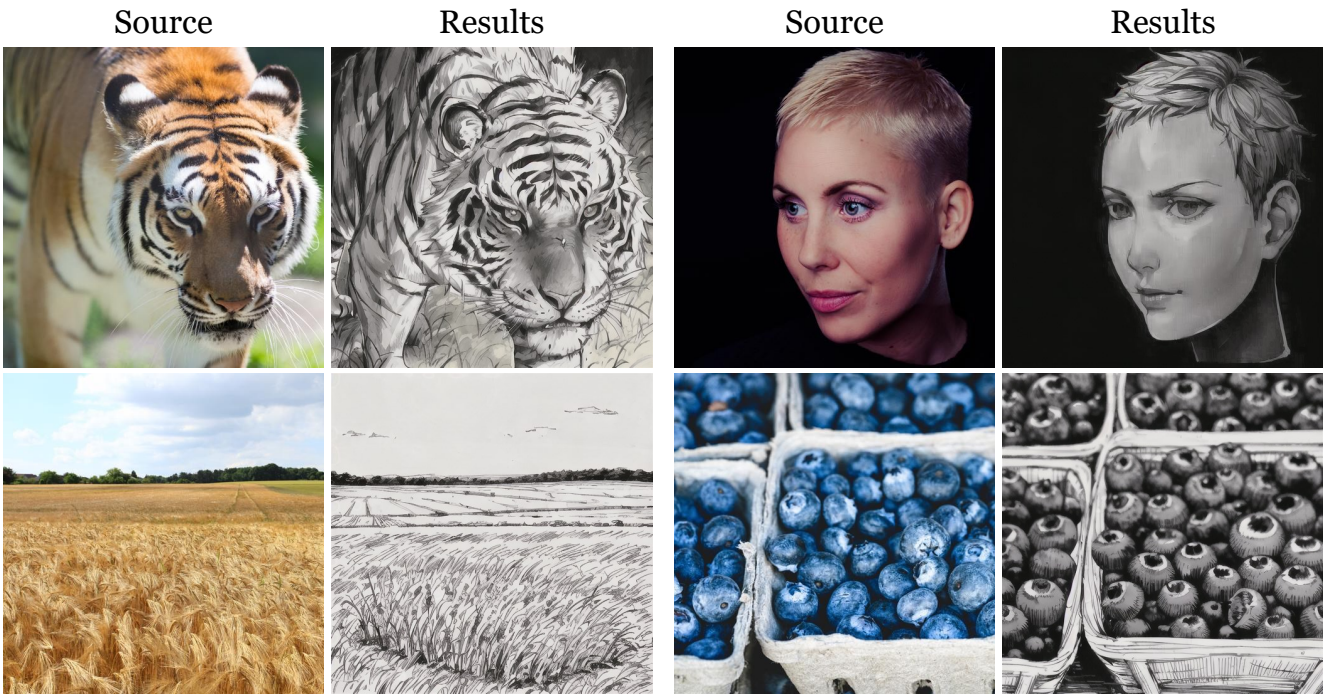

Figure 12. **Sketch style.** The results resemble hand-drawn sketches, featuring monochromatic tones and emphasized contours.