Latent Pyramid Guidance for Training-Free Diffusion Style Transfer

Dae-Young Song
Content Research Division
ETRI
Daejeon, Republic of Korea
eadyoung@etri.re.kr

Jung-Jae Yu
Content Research Division
ETRI
Daejeon, Republic of Korea
jungjae@etri.re.kr

***Abstract*—Diffusion models have attracted significant attention not only for image synthesis and editing, but also for style transfer. However, diffusion-based style transfer often produces sharp images, yet the stochasticity in sampling can perturb spatial layouts, causing structural drift across denoising steps. Additional tuning can mitigate this issue, yet diffusion models contain many parameters and both full training and test-time adaptation impose substantial computational cost. We introduce a training-free dual classifier-free guidance strategy that balances content and style objectives during sampling. We further show that pyramidal composition in latent space yields strong content preservation while maintaining low-frequency geometry. Our approach requires no parameter updates and remains compatible with off-the-shelf backbones.**

***Index Terms*—style transfer, artistic synthesis, diffusion models**

I. Introduction

Diffusion models [1], [2] have rapidly become a milestone for image synthesis and editing tasks. Their iterative denoising procedure supports fine control over image appearance and enables high perceptual fidelity for complex concepts. These strengths have renewed interest in style transfer, where the goal is to render a content image in the appearance of a target style while preserving the underlying content. Despite these advances, diffusion-based style transfer still struggles with content preservation. During the reverse process of diffusion, excessive style guidance can overwhelm the structural guidance and cause the collapse of the target content image. Therefore, achieving a balance between content preservation and style strength is crucial in diffusion-based style transfer to produce high-quality, aesthetically pleasing results.

Modern pipelines perform diffusion in latent space for computational efficiency [2], which makes explicit geometric manipulation difficult. Because direct pixel-space operations do not translate cleanly to the latent representation, therefore this choice mainly limits the practicality of applying explicit geometric controls during sampling. As a result, many methods avoid explicit geometric operations and instead inject style into the cross-attention layers of architectures [3]–[5], leverage auxiliary modules [6]–[9] to encode style or content, or resort to test-time optimization [10], [11]. Additionally, several studies have explored inversion-based [12]–[14] editing. Methods such as DDIM inversion [12] reconstruct a deterministic denoising trajectory that reproduces the input image, after which controlled edits are applied along that path to preserve high content fidelity.

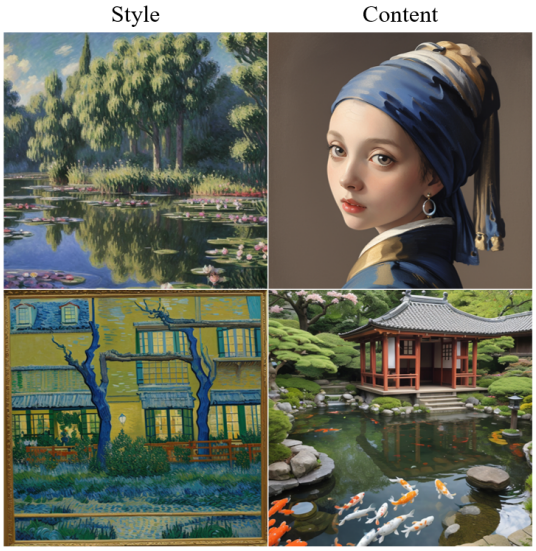

Fig. 1. (Left) Example reference image for style. (Right) Example reference image for content.

Attention injection methods manipulate cross- or self-attention to import style features from a reference, which is simple and often training free, but provides only indirect control over geometry. Auxiliary modules encode style or enforce structure using adapters or control branches, which reduces trial and error but requires additional training, annotations, and compute, and may generalize poorly across styles. Inversion-based editing reconstructs a trajectory that reproduces the input image and then applies edits along that path to preserve content fidelity, yet most formulations assume deterministic sampling and become sensitive to noise schedules when used with stochastic samplers, which calls for extra design such as noise recovery and path consistency. Test-time optimization adapts sampling to each input by solving an objective during generation, which can improve faithfulness at the cost of per-image latency.

Therefore, we introduce a training-free dual classifier-free

guidance strategy that separates content preservation from style induction and balances the two during sampling. In parallel, we perform pyramidal composition in latent space so that low-frequency geometry is preserved while high-frequency bands carry style details. Our method requires no parameter updates or auxiliary networks, and operates with off-the-shelf backbones.

II. Method

Classifier-free guidance (CFG) [15] for text-to-image diffusion is defined as follows:

$$\hat{\varepsilon}_\theta(x_t, c) = \varepsilon_\theta(x_t, \varnothing) + w(\varepsilon_\theta(x_t, c) - \varepsilon_\theta(x_t, \varnothing)), \quad (1)$$

where x_t is the noisy latent at timestep t, $\varepsilon_\theta(\cdot, \cdot)$ denotes the predicted noise under the given condition, c is the text condition, $\varnothing$ is the null condition for the unconditional guidance, and $w \geq 1$ is the guidance scale that controls the strength of conditioning.

A. Dual Classifier-Free Guidance

A diffusion model can produce a predicted noise ε_θ for a given noisy latent x_t under text condition c. Hence, for two text embeddings that represent content and style, we can obtain two conditional predictions $\varepsilon_{\text{con}} := \varepsilon_\theta(x_t, c_{\text{con}})$ and $\varepsilon_{\text{sty}} := \varepsilon_\theta(x_t, c_{\text{sty}})$. As a naive idea, we can treat each conditional prediction as a direction from the unconditional prediction and linearly combine them at every timestep.

$$\hat{\varepsilon} = \varepsilon_{uc} + w_{\text{con}}(\varepsilon_{\text{con}} - \varepsilon_{uc}) + w_{\text{sty}}(\varepsilon_{\text{sty}} - \varepsilon_{uc}), \quad (2)$$

where $\hat{\varepsilon}$ is the merged prediction, ε_{uc} the unconditional prediction, ε_{con} the conditional prediction for content, ε_{sty} the conditional prediction for style, w_{con} the guidance scale for content, and w_{sty} the guidance scale for style, respectively. As shown in Fig. 1, the two reference images in style transfer often have markedly different spatial layouts. In such cases, a naive interpolation of the two conditions becomes ambiguous about which geometry to follow, which makes it difficult to preserve the layout of the content reference image.

B. Latent Pyramids

Although the latent representation is not identical to pixel space, it still retains much of the coarse scene geometry. Building on this observation, we perform a pyramid decomposition of the predicted noise $\varepsilon_\theta(x_t, c)$ in latent space and carry out level-wise fusion: we anchor global content with the downsampled (low-frequency) levels, while the higher-resolution levels convey stylistic details. Given an L-level pyramid, we construct the following laplacian pyramid:

$$\varepsilon^{(k)} = \mathcal{D}^k(\varepsilon^{(0)}) - \mathcal{U}(\mathcal{D}^{k+1}(\varepsilon^{(0)})),\ k = 0, \ldots, L-1\ (L \geq 2), \quad (3)$$

where $\mathcal{D}$ is the downsampling operation by a factor of 2, $\mathcal{U}$ the upsampling, k the pyramid level, and $\varepsilon^{(0)} := \varepsilon_\theta(x_t, c)$. To ensure stable composition across K levels, we compute a gating scale for each level as:

$$g_{\text{con}}^{(k)} = \frac{k}{L-1}, \quad k = 0, \ldots, L-1 \quad (L \geq 2), \quad (4)$$

$$g_{\text{sty}}^{(k)} = 1 - g_{\text{con}}^{(k)}, \quad (5)$$

where g_{con} is the gating scale for content, and g_{sty} for style. Finally, dual CFG scales are computed as follows:

$$\hat{w}_L^{(k)} = \frac{1}{2} w_{\text{con}} g_{\text{con}}^{(k)}, \quad \hat{w}_H^{(k)} = \frac{1}{2} w_{\text{sty}} g_{\text{sty}}^{(k)}. \quad (6)$$

Then, (2) can be modified:

$$\hat{\varepsilon}^{(k)} = \varepsilon_{uc}^{(k)} + \hat{w}_L^{(k)}(\varepsilon_{\text{con}}^{(k)} - \varepsilon_{uc}^{(k)}) + \hat{w}_H^{(k)}(\varepsilon_{\text{sty}}^{(k)} - \varepsilon_{uc}^{(k)}). \quad (7)$$

Subsequently, all $\hat{\varepsilon}^{(k)}$ components are upsampled back to the original resolution and then summed.

C. Gate Scheduling

During sampling, diffusion models tend to establish coarse global content at early timesteps and progressively refine local details at later ones. Motivated by this observation, we modulate the contribution of each pyramid level to preserve global content. Let T denote the total number of sampling steps and let $t \in \{0, \ldots, T-1\}$ be the current timestep, scheduling factors are computed as:

$$\lambda_L = (1 - \frac{t}{T-1})^\gamma, \quad \lambda_H = 1 - \lambda_L, \quad (8)$$

where γ is a damping factor. we set γ to 1.5. Then (6) can be modified as follows:

$$\tilde{w}_L^{(k)} = \lambda_L w_{\text{con}} g_{\text{con}}^{(k)}, \quad \tilde{w}_H^{(k)} = \lambda_H w_{\text{sty}} g_{\text{sty}}^{(k)}. \quad (9)$$

III. Experimental Results

Given a content prompt and a style prompt, we first generate reference images I_{con} and I_{sty} by applying standard CFG to each prompt separately. To this end, we used ChatGPT [16] to generate 36 style prompts and 39 content prompts. By pairing each of the 36 style prompts with each of the 39 content prompts, we obtain $36 \times 39 = 1{,}404$ prompt pairs and we generate one image per pair conditioned on both prompts for evaluation. We synthesize $1{,}404$ images conditioned on both prompts utilizing three methods: dual CFG (DCFG), pyramidal dual CFG (PDCFG), and scheduled pyramidal dual CFG (SPDCFG). For evaluation, we use the style reference I_{sty} to compute Art-FID [17] as a measure of stylistic fidelity, and the content reference I_{con} to compute LPIPS as a measure of structural consistency. We employed DreamShaper 8 [18] for this experiments. Qualitative and quantitative results are presented in Fig. 2 and TABLE I, respectively. For DCFG, although style is captured most faithfully, global structure is not explicitly preserved. Because frequency separation is not considered in the latent space, the style prompt exerts a strong influence on the overall layout. In our experiments we observe that pyramid guidance can maintain global structure as shown in Fig. 2. Especially, SPDCFG preserved the original layout best. This trend is also evident in TABLE I.

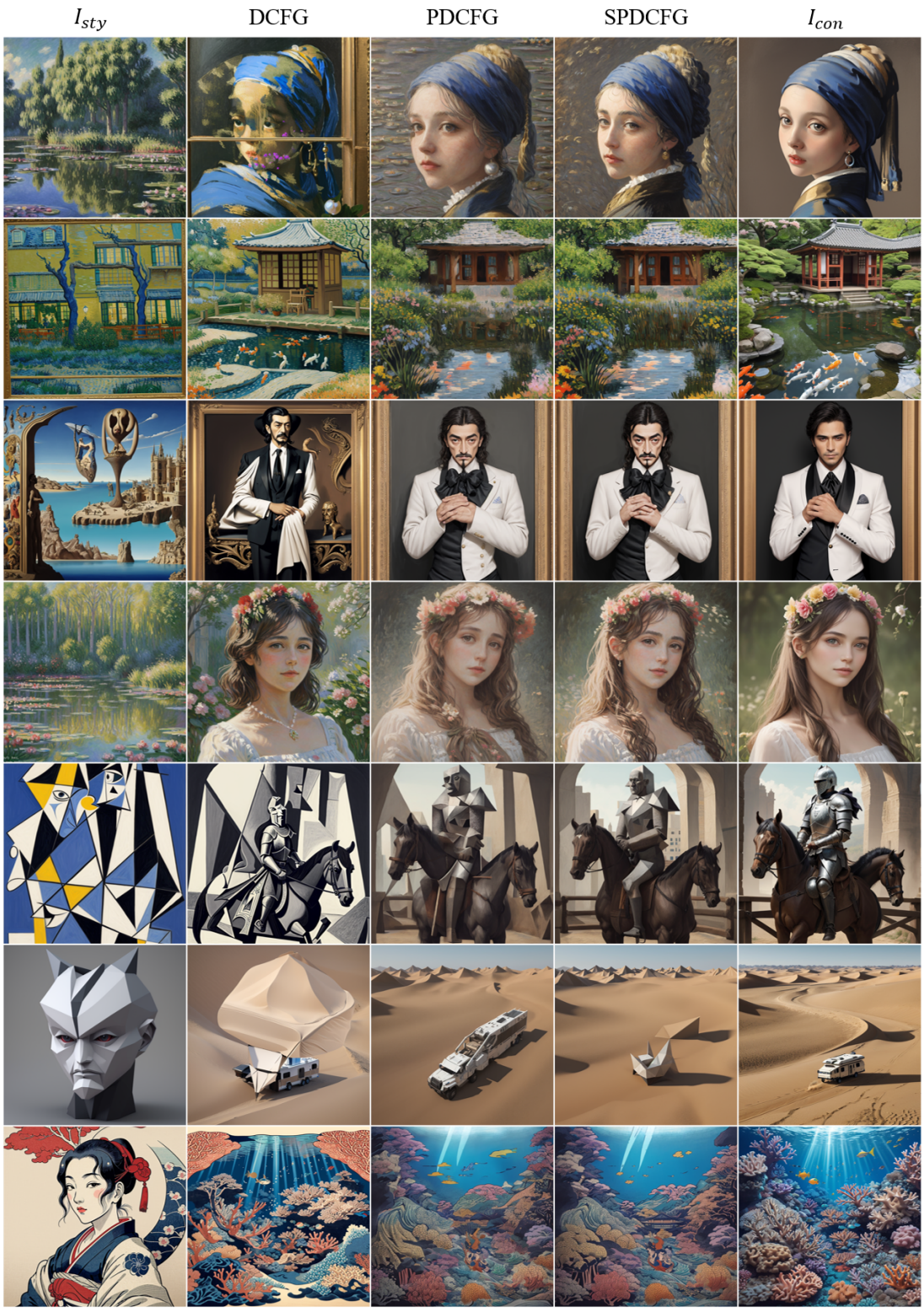

Fig. 2. Qualitative results on 1,404 prompt pairs.

IV. Conclusion

In this work we addressed training-free stylization with a single pre-trained diffusion model. We introduced a latent pyramid that decomposes the predicted noise and schedules the contribution of each level during sampling. This design preserves global structure while injecting high-frequency style,

TABLE I
GENERATION RESULTS WITH DUAL-PROMPT CLASSIFIER-FREE GUIDANCE. **BOLD**: BEST.

Method	Art-FID ↓	LPIPS ↓
DCFG (II-A)	**5.5678**	0.6487
PDCFG (II-B)	7.8932	0.5353
SPDCFG (II-C)	7.2813	**0.4812**

and it integrates with off-the-shelf backbones without parameter updates. Qualitative and quantitative results indicate improved structural fidelity at competitive visual quality, and the proposed SPDCFG preserves the original layout most effectively. However, these approaches are limited by the coverage of the pre-trained model, since the model should observe the target style and objects. In addition, When the style and content are extremely mismatched or when the style is highly abstract relative to photorealistic content, stylization becomes unreliable. In this work, we used hand-crafted manipulation approaches for diffusion guidance. We expect future works to address these constraints with more adaptive schemes.

ACKNOWLEDGMENT

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025. (Project Name : Development of AI Agent Technology Based on Artists Unique Characteristics for Interactive Culture Creation, Project Number : RS-2025-02312732)

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[3] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[4] J. Chung, S. Hyun, and J.-P. Heo, "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8795–8805.

[5] Y. Deng, X. He, F. Tang, and W. Dong, "z*: Zero-shot style transfer via attention rearrangement," *arXiv preprint arXiv:2311.16491*, 2023.

[6] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 3836–3847.

[7] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 5, 2024, pp. 4296–4304.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *International Conference on Learning Representations (ICLR)*, 2022.

[9] S. Li, "Diffstyler: Diffusion-based localized image style transfer," *arXiv preprint arXiv:2403.18461*, 2024.

[10] Y.-Y. Tsai, F.-C. Chen, A. Y. Chen, J. Yang, C.-C. Su, M. Sun, and C.-H. Kuo, "Gda: Generalized diffusion for robust test-time adaptation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 23 242–23 251.

[11] S. Kim, Y. Min, Y. Jung, and S. Kim, "Controllable style transfer via test-time training of implicit neural representation," *Pattern Recognition (PR)*, vol. 146, p. 109988, 2024.

[12] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6038–6047.

[13] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1921–1930.

[14] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6007–6017.

[15] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[16] OpenAI, "ChatGPT," https://chatgpt.com, 2025, accessed: 2025-08-18.

[17] M. Wright and B. Ommer, "Artfid: Quantitative evaluation of neural style transfer," in *DAGM German Conference on Pattern Recognition (GCPR)*. Springer, 2022, pp. 560–576.

[18] Lykon, "DreamShaper 8," https://huggingface.co/Lykon/dreamshaper-8, 2023, accessed: 2025-08-18.