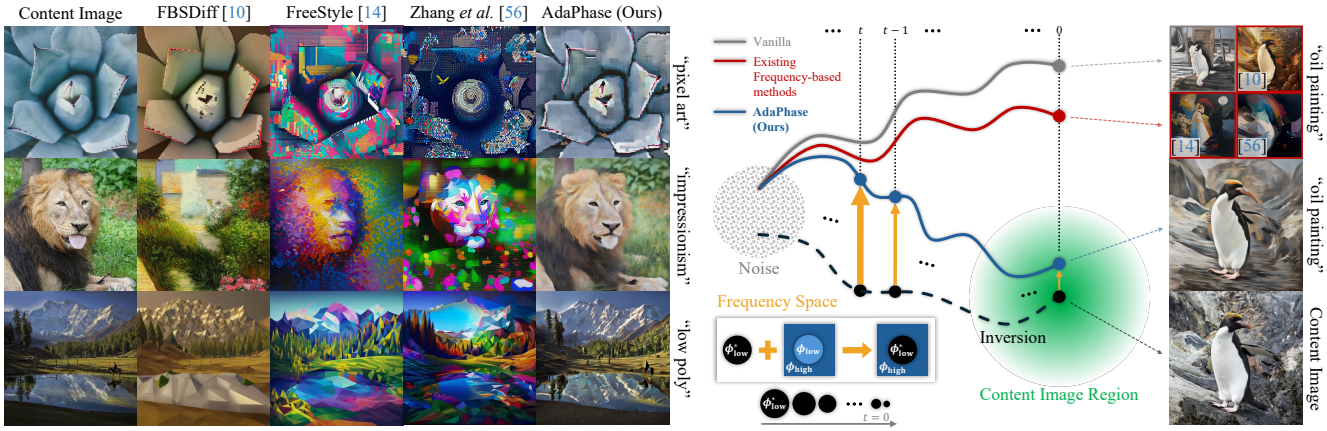
AdaPhase: Navigating the Trajectory with Adaptive Frequency Modulation for Content-preserving Stylization

Sekwang Lee[1] Ilwoo Lyu[1,2]

[1]Dept. of Computer Science and Engineering, [2]Graduate School of AI, POSTECH

{sekwang, ilwoolyu}@postech.ac.kr



(a) Failure Cases of Existing Frequency-based Methods

(b) Conceptual Illustration of Latent Denoising Trajectories

Figure 1. (a) Qualitative comparison of stylization outcomes between AdaPhase and existing frequency-based diffusion stylization models [10, 14, 56]. (b) The frequency-manipulated latent trajectory of AdaPhase maintains close proximity to the predefined inversion path. This enables content-preserving stylization by allowing the final latent to converge near the inversion endpoint (z_0^{inv}). We demonstrate this through a PCA-based projection of latent trajectories in Sec. 5.2. For (a) and (b), we use four content images from the DIV2K dataset [1].

Abstract

*We propose **AdaPhase**, a training-free stylization method for diffusion models that achieves stylization without compromising content by modulating frequency and leveraging the signal-to-noise ratio (SNR). Text-guided image stylization fundamentally struggles with the trade-off between preserving content structure and generating stylistic attributes, and existing training-based approaches incur high computational costs. Recent training-free frequency-based methods successfully disentangle content and style but often suffer from over-stylization. This results from a lack of analysis regarding the implications of frequency manipulation in latent space and a reliance on heuristic strategies. AdaPhase consists of low-frequency phase refinement derived from DDIM inversion latents to align the denoising trajectory toward the predefined inversion path. To ensure AdaPhase generates stylistic attributes flexibly without compromising the content structure, we introduce adaptive spectral alignment that scales the refinement range through an analytical assessment of the noise power at each timestep. We demonstrate that AdaPhase effectively resolves the content-style trade-off, evidenced by PCA-based projection of latent trajectories and qualitative/quantitative comparisons. The code is available at* https://github.com/Shape-Lab/AdaPhase.

1. Introduction

Text-to-image diffusion models [37] have been widely adopted beyond simple image generation. Leveraging strong priors learned from large-scale datasets [39], these models are increasingly utilized for artistic image stylization. Fundamentally, the essence of this diffusion stylization lies in how to fuse two disparate visual instructions provided via user prompts—*what* to draw (content) and *how* to draw (style)—within the diffusion model's latent space.

Prompt-based diffusion stylization can be categorized into two types: (1) using content text with a style image,

and (2) using a content image with style text. In the first setting, stylization is typically achieved by manipulating attention mechanisms during the denoising process. Recent works inject the stylistic attributes of the style image into attention layers as keys and values [4, 34, 38, 45], or modulate feature statistics in attention [16, 28]. These methods disentangle spatial structure and extract style from the image, but the content text provides high-level semantic guidance without rigorous structural constraints.

In contrast, the second setting requires the model to translate style text into stylistic attributes while preserving the structure of the content image during generation. Under such constraints, these competing objectives inevitably induce a content-style trade-off. Thus, balancing this trade-off is a key challenge in text-guided image stylization of diffusion models. Recent studies address this issue by employing external control adapters [54], fine-tuning on large-scale datasets [2, 12, 47], training additional learnable parameters for style guidance [19, 57], or using test-time optimization [50]. However, these approaches require substantial additional training or increase inference latency.

To mitigate these limitations, recent studies [10, 14, 56] propose training-free stylization methods that manipulate the frequency in latent space, inspired by signal processing theory in pixel space. These methods offer a compelling framework for disentangling content structure and stylistic attributes. Despite their promise, they directly transfer pixel-space intuitions without sufficient analysis of what frequency manipulation implies in latent space. They also lack a principled justification for how frequency manipulation should vary across denoising timesteps, relying instead on heuristic evaluation. Consequently, existing frequency-based methods often over-stylize and distort content details, as shown in Fig. 1(a). This indicates that these methods do not fully overcome the trade-off between content and style.

We introduce **AdaPhase**, a novel training-free stylization method that achieves stylization while preserving content by leveraging frequency domain properties and the diffusion model's SNR. We inject the low-frequency phase of the content image's DDIM inversion latents [41] into the denoising process. This refinement shifts latents toward the predefined inversion path and ensures the preservation of content structure, as illustrated in Fig. 1(b). PCA-based projection of latent trajectories demonstrates that our strategy effectively aligns the latents, particularly at timestep $t = 0$. Furthermore, observing that high noise energy renders structural predictions unreliable, we adaptively reduce the injection range proportional to the noise power of the diffusion process. This enables the generation of stylistic attributes without compromising the content structure.

Our main contributions are summarized as follows:

- We present a training-free framework that utilizes low-frequency phase refinement within the denoising process for content-preserving stylization, without the need for complex sampling control techniques.
- We introduce an adaptive spectral refinement that systematically modulates the frequency range based on the noise power at each timestep to ensure both effective stylization and content preservation.
- By analyzing frequency-manipulated latent trajectories via PCA, we identify that content preservation is achievable through inversion path alignment.
- Our method resolves the trade-off between content and style and outperforms existing state-of-the-art methods in both qualitative and quantitative comparisons.

2. Related Work

2.1. Text-guided Image Stylization

Conventional image stylization methods rely on a reference style image to transfer artistic features to the content image [11, 20, 23, 30]. Breaking this paradigm, CLIPstyler [26] pioneered text-guided image stylization by leveraging the text-image embedding space of a pretrained CLIP model [35]. Following the rapid advancement of diffusion models, recent studies have increasingly focused on text-guided diffusion stylization. The standard denoising process starts from random noise, which inherently disrupts the source image structure. Addressing this, existing approaches typically employ auxiliary modules with training schemes to maintain structural integrity [2, 12, 19, 47, 54, 57]. However, these methods face high training costs, and some exhibit inconsistent style adherence. ControlNet [54] incorporates spatial conditions but imposes constraints that limit the geometric deformations needed for stylization. InstructPix2Pix [2], trained on synthetic data, often prioritizes content preservation over faithful stylization.

In parallel, training-free approaches manipulating the denoising process have emerged to eliminate the need for training. Plug-and-Play [43] injects features and self-attention maps from the DDIM sampling into the U-Net [37]. DiffArtist [22] controls the U-Net's attention and skip connections based on two independent diffusion trajectories. Despite their efficiency, these methods often struggle to balance content preservation and stylization. This is largely due to the structural entanglement in diffusion models; the text conditioning via cross-attention influences both semantics and spatial layout simultaneously [15, 37].

2.2. Frequency Analysis in Diffusion Models

Diffusion models exhibit an inherent frequency bias [8, 21, 36, 51]. By integrating this insight with classical signal processing theory, recent studies leverage frequency domain analysis to improve training mechanism or generation quality [8, 18, 21, 36, 40, 51, 53]. Huang *et al*. [21] identify that standard Gaussian white noise inadequately represents

the frequency components reconstructed by the denoising network. They propose injecting blue noise characterized by predominantly high-frequency components during the training process. FreeU [40] observes that skip connections in the U-Net primarily transmit high-frequency features to the decoder. They strategically reweight the contributions sourced from the U-Net's skip connections and backbone feature maps without additional training.

These approaches have been extended to diffusion stylization by controlling the denoising process. FB-SDiff [10] extracts and transplants specific frequency bands (low/mid/high) of latents from the DDIM sampling using discrete cosine transform and binary masking. FreeStyle [14] generates separate content and style features in a dual-stream encoder and modulates their frequencies in a single-stream encoder. Zhang *et al*. [56] apply spectrum fusion using the phase and magnitude extracted from the content image, alongside semantic injection. These methods demonstrate that integrating frequency perspectives enables training-free image stylization. However, the impact of frequency manipulation on the latent trajectory remains underexplored in these methods, which often compromise content preservation. Moreover, their manipulation strategies per denoising timestep are determined heuristically.

3. Preliminaries

3.1. Deterministic Trajectories in LDM

Diffusion models comprise a diffusion process that progressively injects Gaussian random noise into data and a reverse denoising (sampling) process that reconstructs the data to generate new samples. To mitigate computational complexity, Latent Diffusion Model (LDM) [37] operates within a compressed latent space rather than in pixel space.

While the diffusion process is typically formulated as a stochastic differential equation (SDE), it can alternatively be modeled as a probability flow ordinary differential equation (PF-ODE) [42]. This deterministic process shares the same marginal probability density p_t as the SDE at any given timestep t. In the latent space $z_t \in \mathcal{Z}$, the PF-ODE is defined as:

$$dz_t = \left[f(z_t, t) - \frac{1}{2} g(t)^2 \nabla_{z_t} \log p_t(z_t) \right] dt, \qquad (1)$$

where $f(z_t, t)$ and $g(t)$ denote the drift and diffusion coefficients of the SDE, respectively, and $\nabla_{z_t} \log p_t(z_t)$ represents the score function. The PF-ODE yields a unique deterministic trajectory based on z_t and t. This guarantees an invertible mapping between the initial latent z_0 and the noise prior $z_T \sim \mathcal{N}(0, I)$, where T is the total number of diffusion timesteps. In this framework, denoising integrates from $t = T$ to 0, whereas inversion integrates from $t = 0$ to T. In practice, the score function is approximated by a neural network (*e.g*., U-Net), and the ODE is solved using discretized numerical solvers (*e.g*., DDIM and Euler method). These solvers converge to the PF-ODE solution with controllable error bounds under sufficiently small step sizes.

3.2. Frequency Analysis in Latent Space

In the frequency domain analysis of images, low-frequency components form the global structure and semantic content, whereas high-frequency components capture fine-grained details such as edges and textures. Furthermore, the structural information and identity of an image are predominantly governed by the phase [32]. Preserving the phase enables reconstructing the original structural features, even if the magnitude is discarded or replaced. Conversely, the loss of phase renders the image content unrecognizable.

Since the latent space represents a perceptually equivalent mapping of the pixel space [37], these spectral characteristics are similarly preserved. The low-frequency components in the latent space encode semantic content [18, 40, 51], and the phase plays a role in determining the structural geometry [53]. Given a latent $z \in \mathbb{R}^{C \times H \times W}$, the 2D discrete Fourier transform can be applied channel-wise. Let $z_c(h, w)$ denote the value at spatial coordinates (h, w) for the c-th channel. The forward transform $\mathcal{F}$ and its inverse $\mathcal{F}^{-1}$ are defined as:

$$\begin{aligned} Z_c(u, v) &= \mathcal{F}\left(z_c\right)(u, v) \\ &= \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} z_c(h, w) e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}, \qquad (2) \\ z_c(h, w) &= \mathcal{F}^{-1}\left(Z_c\right)(h, w) \\ &= \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} Z_c(u, v) e^{j2\pi\left(\frac{u}{H}h + \frac{v}{W}w\right)}, \qquad (3) \end{aligned}$$

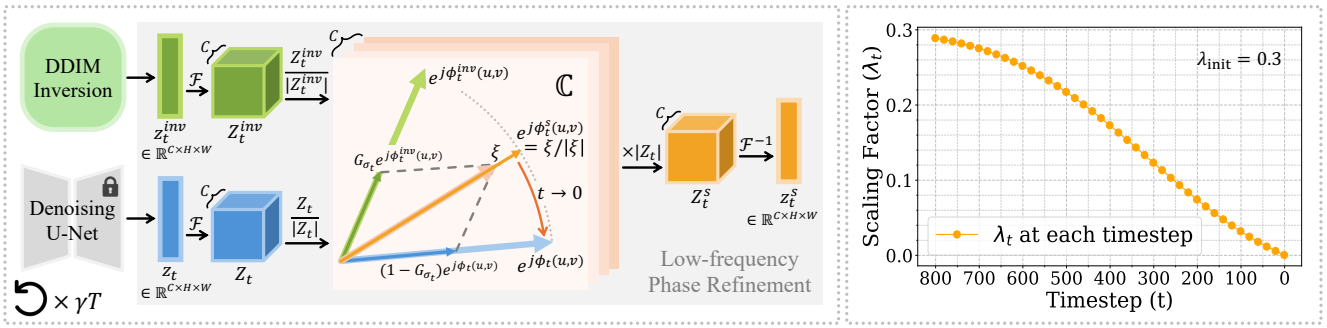
where $j^2 = -1$, and (u, v) represents the frequency coordinates. Here, Z_c denotes the frequency representation of the channel c. For notational simplicity, we omit the channel index c in the following equations, as all spectral operations are performed in a channel-wise manner.

4. Method

The overall pipeline of AdaPhase is illustrated in Fig. 2. Based on the observations in Sec. 3.2, we propose a method that injects the low-frequency phase of the source latents—extracted via the DDIM inversion process—at each timestep of the denoising process. This process effectively rectifies the latent trajectory, prevents drift, and constrains it to stay close to the inversion path to ensure content preservation.

4.1. Low-frequency Phase Refinement

Diffusion models tend to restore low-frequency components before generating high-frequency ones [8, 21, 36, 51]. Notably, the low-frequency component of the noise serves as



(a) The Overview of Framework (b) Adaptive Spectral Alignment

Figure 2. **Overall pipeline of AdaPhase.** (a) Direct linear interpolation of the phases ϕ_t^{inv} and ϕ_t can introduce phase wrapping artifacts due to periodicity. To mitigate this, we map the phases onto the complex plane $\mathbb{C}$ as unit vectors, compute their weighted sum, and project the result back onto the unit circle. This channel-wise operation preserves the unit norm of the original phase vectors. For DDIM inversion, we use empty-text conditioning. (b) The scheduling factor λ_t adaptively controls the standard deviation of G_{σ_t} based on the timestep.

a structural seed for generation [46]. Since the latent is initialized with Gaussian noise, the structural information predicted early in the denoising process often fails to align with the original image. Instead of relying on the low-frequency phase predicted by the model, we selectively replace it with the phase derived from the source DDIM inversion latents.

Given a content image x_0, we first encode it into the latent space to obtain z_0 and perform DDIM inversion to compute the reference path $\{z_t^{inv}\}_{t=0}^{\gamma T-1}$. The generative process is initialized at $z_{\gamma T}$—obtained by diffusing z_0 to timestep γT with denoising strength $\gamma \in [0, 1]$—and denoises it conditioned on a target style text. At each timestep t, we transform both the inversion latent z_t^{inv} and the current generative latent z_t into the frequency domain via a Fourier transform $\mathcal{F}$, obtaining Z_t^{inv} and Z_t, respectively. To disentangle structural information from magnitude, we extract the phase components as unit vectors, defined as:

$$e^{j\phi_t(u,v)} = \frac{Z_t(u,v)}{|Z_t(u,v)|}, \; e^{j\phi_t^{inv}(u,v)} = \frac{Z_t^{inv}(u,v)}{|Z_t^{inv}(u,v)|}. \quad (4)$$

Our primary objective is to preserve the global structure of the content image by injecting the phase information of z_t^{inv} predominantly into the low-frequency bands. To achieve this, we employ an unnormalized Gaussian low-pass filter G_{σ_t} (with a peak value of 1) as an interpolation weight, represented as:

$$G_{\sigma_t}(u,v) = e^{-\frac{u^2+v^2}{2\sigma_t{}^2}}. \quad (5)$$

This ensures a smooth spectral transition and prevents artifacts caused by abrupt frequency cutoffs. We construct the refined phase vector $e^{j\phi_t^s(u,v)}$ through a spectral soft-blending governed by G_{σ_t} as follows:

$$\xi = G_{\sigma_t}(u,v)e^{j\phi_t^{inv}(u,v)} + (1 - G_{\sigma_t}(u,v))e^{j\phi_t(u,v)}, \quad (6)$$

$$e^{j\phi_t^s(u,v)} = \xi/|\xi|. \quad (7)$$

This operation aligns the phase with the inversion path in low-frequency regions (where $G_{\sigma_t} \approx 1$), while retaining the generative phase in high-frequency regions (where $G_{\sigma_t} \approx 0$). The resulting phase vector thus bridges the source and generative components effectively.

We construct the refined frequency representation Z_t^s by combining the phase vector $e^{j\phi_t^s}$ with the original magnitude from the generative latent as

$$Z_t^s = |Z_t|\, e^{j\phi_t^s}. \quad (8)$$

An inverse Fourier transform $\mathcal{F}^{-1}$ is then applied to recover the updated spatial latent z_t^s as

$$z_t^s = \mathcal{F}^{-1}(Z_t^s). \quad (9)$$

This frequency-guided refinement is applied iteratively for all predefined $t \in \{\gamma T - 1, \ldots, 0\}$. Upon reaching the final timestep ($t = 0$), the resulting latent z_0^s is decoded to produce the final stylized image x_0^s.

4.2. Adaptive Spectral Alignment

As discussed in Sec. 4.1, noise injected during the initial timesteps can introduce deviations from the source image's underlying structure. This implies that high noise energy in these steps renders structural predictions unreliable.

From this perspective, we propose adaptive spectral alignment. This strategy adaptively modulates the frequency range in proportion to the noise power at each timestep. The core idea is to expand the low-frequency phase refinement range during high noise energy intervals, which are characterized by high uncertainty, to enforce structural consistency. Conversely, as the process transitions to low noise energy intervals where noise is reduced and reliability improves, we gradually narrow the range to allow for greater generative freedom.

To quantify this, we utilize the signal-to-noise ratio (SNR). Given the noise schedule of the diffusion process,

the SNR at timestep t is derived as [5, 13]:

$$\text{SNR}(t) = \frac{(\sqrt{\bar{\alpha}_t})^2}{(\sqrt{1-\bar{\alpha}_t})^2} = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}, \tag{10}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ denotes the cumulative product of the noise schedule terms [17]. We design the standard deviation σ_t of the Gaussian low-pass filter G_{σ_t} to be proportional to the ratio of noise power to the total signal power. Specifically, this imposes stronger structural constraints in regimes where the noise variance is dominant. Since the added noise has zero mean, this variance is equivalent to its expected power. Hence, the scheduling is formulated as:

$$\lambda_t = \frac{\lambda_{\text{init}}}{1+\text{SNR}(t)} = \lambda_{\text{init}}(1-\bar{\alpha}_t), \tag{11}$$

$$\sigma_t = \lambda_t r_{\text{max}}, \tag{12}$$

where λ_{init} is an initial injection strength hyperparameter, and λ_t denotes the time-dependent scaling factor determined by the noise level. r_{max} represents the maximum radial distance on the discrete Fourier transform grid, *i.e.*, the Euclidean distance from the center to the farthest corner.

This strategy can be justified by Parseval's theorem: the total energy of a signal is conserved across both spatial and frequency domains. In the diffusion process, the noise variance at timestep t is $1-\bar{\alpha}_t$. Since the added noise is Gaussian white noise, its power spectrum is uniformly distributed across all frequencies. Thus, $1-\bar{\alpha}_t$ can serve as a proxy for the total spectral noise power. As illustrated in Fig. 2(b), λ_t exhibits a non-linear, monotonically decreasing trend following the schedule of $\bar{\alpha}_t$. With σ_t converging to 0 as denoising progresses, the model is able to generate detailed stylistic attributes flexibly.

5. Experiments

5.1. Implementation Details

We conduct all experiments using the pretrained Stable Diffusion XL 1.0 (SDXL) [33] image-to-image pipeline. We use the Euler discrete sampler with 50 inference steps, the model's default sampling algorithm. We set the denoising strength to $\gamma = 0.8$, corresponding to 40 steps starting from a partially noised latent. To regulate the initial low-frequency band, we choose $\lambda_{\text{init}} = 0.3$. Unless otherwise specified, we adhere to the default hyperparameters of SDXL. All experiments are conducted on NVIDIA RTX 6000 Ada Generation GPUs.

We utilize four distinct datasets covering diverse domains: LAION-5B (web-scale images) [39], FFHQ (human faces) [25], WikiArt (artistic paintings) [24], and DIV2K (high-resolution images) [1]. For Sec. 5.2, we construct a subset of 3,000 images by randomly sampling 1,000 images from each of LAION-5B, FFHQ, and WikiArt. For qualitative analysis, we select 10 images from DIV2K spanning content types including human faces, architecture, nature, animals, and inanimate objects. For quantitative analysis, we build an evaluation set of 2,000 stylized images by pairing 100 source images (50 randomly selected from LAION-5B and 50 from FFHQ) with 20 curated style texts. The full list of style texts is provided in the Supplementary Material.

Table 1. **Mean directional similarity and displacement error of refined trajectories relative to the inversion path.** To analyze the trajectory evolution, the refinement strategy is applied during the initial 40% of the process, and metrics are computed over the subsequent 60% where refinement is deactivated. We exclude adaptive spectral alignment for all strategies and fix $\lambda_t = 0.3$ for all low-frequency (phase) strategies to evaluate pure reconstruction capability. **Bold values** indicate the best results.

Refinement Source	Refinement Type	CosSim $(\Delta\tilde{z}_t^{inv}, \Delta\tilde{z}_t^s)\uparrow$	$\|\tilde{z}_t^{inv} - \tilde{z}_t^s\|_2\downarrow$	$\|\tilde{z}_0^{inv} - \tilde{z}_0^s\|_2\downarrow$
$\{z_t^{inv}\}_{t=0}^{T-1}$	Low-freq. Phase (Ours)	**0.89±0.11**	**26.8±7.7**	**15.7±11.4**
	Entire Phase	0.87±0.25	58.3±12.0	30.8±7.13
	Low-freq.	0.86±0.19	59.0±17.7	32.7±16.1
z_0^{inv}	Low-freq. Phase	0.82±0.33	70.8±14.1	36.2±12.2
	Entire Phase	0.82±0.35	65.2±16.9	33.3±10.2
	Low-freq.	0.76±0.31	78.5±13.2	36.2±14.3
-	No Refinement	0.79±0.23	87.6±16.1	34.3±14.3

5.2. PCA-based Trajectory Projection

Based on Sec. 3.1, we analyze the geometric properties of latent trajectories. We employ principal component analysis (PCA) to effectively map the high-dimensional DDIM inversion path without significant information loss. As established in [44], the diffusion trajectory predominantly lies within a 2D subspace spanned by the latent z_T and z_0.

We define a space of the first two principal components based on the inversion path. Since stylization has no definitive ground truth, we project the refined trajectories under empty-text reconstruction onto the inversion path to show close alignment. We evaluate the geometric proximity to the inversion path, as the DDIM inversion path reveals the original image reconstruction path in reverse [41]. We note that even if two trajectories are close in latent space at different timesteps, their images can differ as the reconstruction process is conditioned on t. We focus on two metrics: trajectory displacement error and directional similarity. These are quantified by the mean L2 distance between the projected inversion latent $\tilde{z}_t^{inv}$ and the projected refined latent $\tilde{z}_t^s$, and the mean cosine similarity of $\Delta\tilde{z}_t = \tilde{z}_t - \tilde{z}_{t+1}$, respectively. For a comprehensive analysis, we consider multiple refinement strategies, including the entire phase and the low-frequency (phase + magnitude). Additionally, we vary the refinement source by utilizing both the timestep-wise inversion latents and the static initial latent z_0^{inv}.

As shown in Tab. 1, the trajectory obtained by refining the low-frequency phase of the inversion latents significantly outperforms other methods. These results are qual-

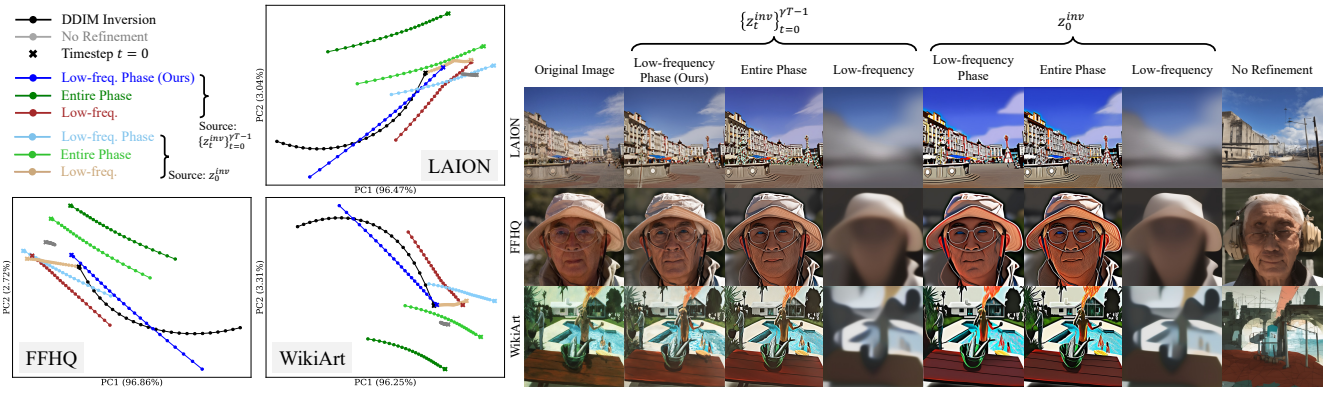


Figure 3. *Left:* PCA plots for randomly selected samples from the LAION, FFHQ, and WikiArt datasets, as reported in Tab. 1. *Right:* Visualization of decoded images of z_0^s for the corresponding samples.

Table 2. **PCA endpoint alignment and content preservation.** For low-frequency phase refinement using the inversion path $\{z_t^{inv}\}_{t=0}^{\gamma T-1}$, we select 100 samples around the first and third quartiles of the PCA endpoint-displacement distribution. The lower-displacement group achieves higher DINO and 1−LPIPS scores.

Group	$\|\tilde{z}_0^{inv} - \tilde{z}_0^s\|_2 \downarrow$	DINO [3] ↑	1−LPIPS [55] ↑
Lower displacement (Q1)	7.660±0.179	0.811±0.082	0.589±0.055
Higher displacement (Q3)	20.543±0.422	0.763±0.143	0.558±0.087

itatively supported by Fig. 3. While our method closely reconstructs original images, others tend to produce artifacts like overly emphasized contrast or blurring. As discussed in Sec. 3.2, high-frequency components capture fine-grained details such as edges and textures. Consequently, preserving the entire phase spectrum, including high-frequency components, may reduce the spatial flexibility available to the denoising process (emphasized contrast in Fig. 3). Interestingly, using the initial latent z_0^{inv} directly as a refinement source leads to a substantial increase in displacement error. This suggests that naively injecting the original latent anchors the trajectory rigidly, disrupting the restoration of the underlying structure. Again, we do not seek to replicate the inversion path. Although the paths may vary due to the style text, the final latent converges near the inversion endpoint; this serves as a proxy for the preservation of the content structure, and we verify its image-level relevance in Tab. 2.

5.3. Comparison of Stylization Performance

We validate AdaPhase by comparing it with seven state-of-the-art models in two categories: text-guided stylization (CLIPstyler [26], FBSDiff [10], FreeStyle [14], DiffArtist [22], and Zhang *et al*. [56]) and text-guided image editing (ControlNet [54] and InstructPix2Pix [2]), all of which have a style text as input. Excluding ControlNet, InstructPix2Pix, and CLIPstyler, the remaining methods are training-free. Among these, FBSDiff, FreeStyle, Zhang *et al*., and our method leverage the frequency domain. We use the Canny2Image model for ControlNet and configure FBSDiff to modulate high-frequency components primarily used for stylization, yielding the best quality. All methods are evaluated using their optimal hyperparameters.

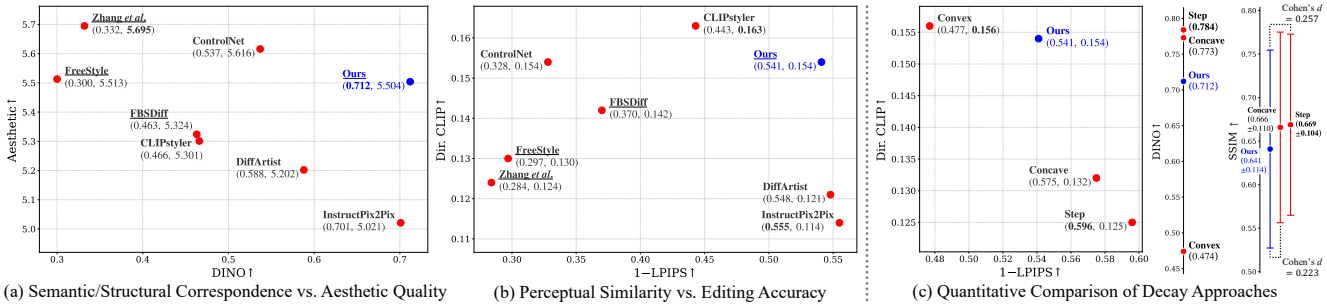
Qualitative Comparison. As demonstrated in Fig. 4, our method achieves superior structural preservation while successfully stylizing all content-style pairs. In contrast, competing methods exhibit a trade-off between content preservation and stylization. While FreeStyle and Zhang *et al*. generate distinct styles, they suffer from the degradation of structural details. FBSDiff fails to retain content in most samples. InstructPix2Pix and DiffArtist often fail to manifest the target style when the content structure is preserved (*e.g*., columns 3, 4, and 6). Notably, InstructPix2Pix exhibits structural collapse in some cases where the style is applied (*e.g*., columns 7 and 10). Although ControlNet and CLIPstyler generally maintain the content structure, they frequently introduce unnatural color shifts and struggle to capture stylistic attributes precisely (*e.g*., columns 5 and 7).

Quantitative Comparison. To verify that our method overcomes the trade-off between content and style, we employ four metrics. LPIPS [55] measures perceptual content preservation, while DINO score [3] assesses semantic and structural consistency. LAION Aesthetic score [39] evaluates human-perceived visual quality, and directional CLIP similarity (dir. CLIP) score [9] quantifies how accurately the edit direction reflects the text-driven change.

As shown in Fig. 5(a), higher DINO scores tend to correlate with lower Aesthetic scores in the baseline methods. In contrast, our method achieves the highest DINO score while maintaining a competitive Aesthetic score. Although ControlNet, FreeStyle, and Zhang *et al*. exhibit higher Aesthetic scores, Fig. 4 reveals that their outputs are often highly saturated; the Aesthetic score has a known bias toward high saturation [29]. Notably, these methods yield lower DINO

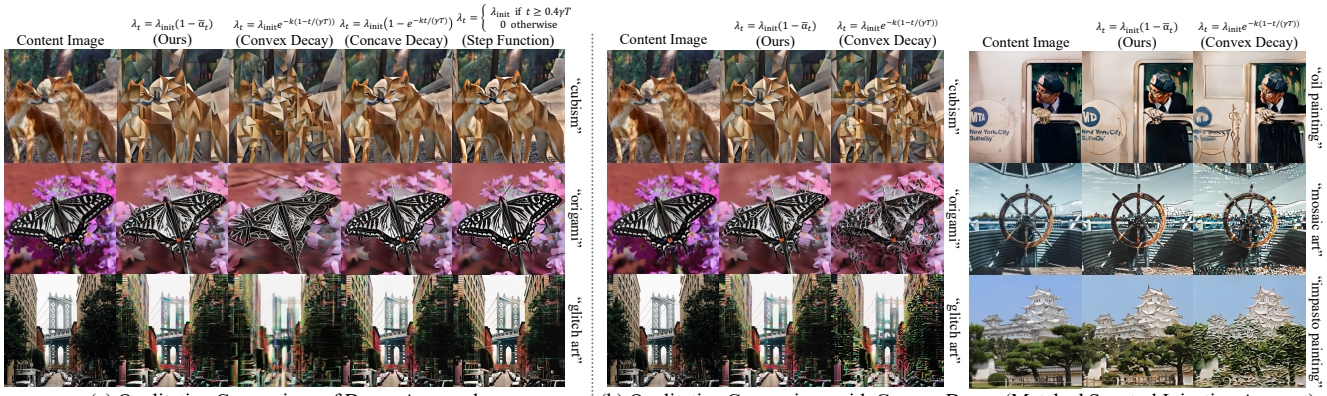


Figure 4. **Qualitative comparison of stylization outcomes with existing methods.** Underlined models utilize the frequency domain. The stylistic details of certain images (*e.g.*, "mosaic" and "cartoon") are visible upon zooming in.



(a) Semantic/Structural Correspondence vs. Aesthetic Quality (b) Perceptual Similarity vs. Editing Accuracy (c) Quantitative Comparison of Decay Approaches

Figure 5. **(a, b) Quantitative comparison with existing methods and (c) quantitative ablation study of different decay approaches.** We plot 1 − LPIPS so that higher values indicate better perceptual similarity. For all dir. CLIP evaluations, the source text is set to "a photo". Underlined models utilize the frequency domain; **bold values** indicate the best results.



(a) Qualitative Comparison of Decay Approaches

(b) Qualitative Comparison with Convex Decay (Matched Spectral Injection Amount)

Figure 6. **Qualitative ablation study of different decay approaches.** (a) We set $k = 3$ to accentuate convex/concave profiles for clearer distinction from our method. (b) To equalize injection, we set $\lambda_{\text{init}} = 0.4403$ for the convex decay. This corresponds to the $\sum_t \sigma_t^2$ of our method ($\lambda_{\text{init}} = 0.3$) and ensures the cumulative variance of G_{σ_t} over all timesteps matches that of our method.

scores and $1 - \text{LPIPS}$, which implies insufficient preservation of high-level semantic and structural consistency.

In Fig. 5(b), we find that FreeStyle and Zhang *et al.* exhibit low dir. CLIP scores despite their high Aesthetic scores. This suggests that these methods prioritize aesthetic quality over faithful adherence to the style text. Conversely, our method shows a high dir. CLIP score comparable to ControlNet, while outperforming CLIPstyler in $1 - \text{LPIPS}$. Consequently, our model is positioned in the upper-right corner of the plot, representing an optimal balance between style editing accuracy and content preservation.

5.4. Analysis of Spectral Injection Strategy

We validate Sec. 4.2 by comparing it with three alternative scheduling strategies. These methods exhibit decay patterns misaligned with the injection strength λ_t from the SNR schedule. First, convex decay diminishes the injection strength rapidly during the early denoising stage. Second, concave decay sustains high injection strength until the mid-to-late stages before dropping sharply. Finally, the step function maintains a constant injection strength for the first 60% of the timesteps and then stops completely. We additionally employ SSIM [48] to measure pixel-level structural similarity for a granular assessment of content fidelity.

Fig. 6(a) demonstrates the impact of the decay pattern on the generation results, and our method shows content-style-balanced images. Convex decay results in structural collapse, while concave decay and the step function lead to weak stylization. Fig. 5(c) supports these observations; our method achieves a similar dir. CLIP score to that of convex decay without compromising the DINO score and $1 - \text{LPIPS}$. The decrease in structural metrics relative to the concave decay and step function is marginal, with a small SSIM effect size (Cohen's $d < 0.26$) [6]. Crucially, the significantly low dir. CLIP scores of the concave decay and step function indicate limited stylization. Convex decay suffers from lower DINO scores and $1 - \text{LPIPS}$. We further equalize the total spectral injection of convex decay with our method to ensure the collapse is not due to insufficient injection. In Fig. 6(b), convex decay fails to preserve content structure even with larger λ_{init}. This implies the injection schedule is critical for content-preserving stylization.

Table 3. Computational overhead compared with vanilla SDXL.

Methods	Peak Allocated VRAM (GiB)	Inversion Storage (MiB)	Time (sec)
Vanilla SDXL	16.171	-	6.165
AdaPhase (Ours)	16.177	6.375	10.023

5.5. Computational Overhead

As shown in Tab. 3, AdaPhase introduces negligible GPU memory overhead, increasing the peak allocated VRAM by 6 MiB (0.037%). The increase in inference time is primarily due to the additional DDIM inversion required to extract the reference path. Storing the path requires only 6.375 MiB.

6. Conclusion

We introduced AdaPhase, leveraging frequency domain properties and the diffusion model's SNR for training-free diffusion stylization. AdaPhase injects low-frequency phase derived from DDIM inversion latents to maintain the structure of the content image. This capability is supported by our analysis of PCA-based trajectory projection. By setting the injection range proportional to the noise power at each denoising timestep, AdaPhase can generate stylistic attributes flexibly. Our method outperforms existing methods, including frequency-based approaches, in both content preservation and stylization. We believe that AdaPhase provides insights into frequency manipulation in latent space and contributes to the framework of advanced content-preserving stylization.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant RS-2024-00333931 and Grant RS-2025-02216257; in part by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant RS-2019-II191906 (AIGS Program) and Grant RS-2026-25616313.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 1, 5, 12, 15

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 2, 6, 13

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[4] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8619–8628, 2024. 2

[5] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11472–11481, 2022. 5

[6] Jacob Cohen. A power primer. *Psychological Bulletin*, 112 (1):155–159, 1992. 8

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 14

[8] Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A fourier space perspective on diffusion models. *arXiv preprint arXiv:2505.11278*, 2025. 2, 3

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 6

[10] Xiang Gao and Jiaying Liu. Fbsdiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4101–4109, 2024. 1, 2, 3, 6

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[12] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1968–1978, 2025. 2, 13

[13] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7441–7451, 2023. 5

[14] Feihong He, Gang Li, Fuhui Sun, Mengyuan Zhang, Lingyu Si, Xiaoyan Wang, and Li Shen. Freestyle: Free lunch for text-guided style transfer using diffusion models. *Pattern Recognition*, page 113093, 2026. 1, 2, 3, 6, 13

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[16] 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. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[18] Chao Huang, Susan Liang, Yunlong Tang, Jing Bi, Li Ma, Yapeng Tian, and Chenliang Xu. Fresca: Scaling in frequency space enhances diffusion models. *arXiv preprint arXiv:2504.02154*, 2025. 2, 3

[19] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3370–3383, 2024. 2

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[21] Xingchang Huang, Corentin Salaun, Cristina Vasconcelos, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. Blue noise for diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 2, 3

[22] Ruixiang Jiang and Chang Wen Chen. Diffartist: Towards structure and appearance controllable image stylization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9598–9607, 2025. 2, 6, 13

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[24] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger

Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013. 5

[25] 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. 5

[26] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18062–18071, 2022. 2, 6, 13

[27] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 13, 14

[28] Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, and Chi Zhang. Stylestudio: Text-driven style transfer with selective control of style elements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23443–23452, 2025. 2

[29] Shufan Li, Harkanwar Singh, and Aditya Grover. Instructany2pix: Image editing with multi-modal prompts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 594–619, 2025. 6

[30] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 2

[31] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 12

[32] A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 3

[33] 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. 5

[34] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8693–8702, 2024. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[36] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. *arXiv preprint arXiv:2206.13397*, 2022. 2, 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 14

[38] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free stylization using reference-based modulation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1, 5, 6

[40] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2024. 2, 3

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali 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*, pages 1921–1930, 2023. 2

[44] Binxu Wang and John J Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023. 5

[45] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2

[46] Qingsong Wang, Zhengchao Wan, Mikhail Belkin, and Yusu Wang. Seeds of structure: Patch pca reveals universal compositional cues in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 4

[47] Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7847–7856, 2025. 2, 13

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[49] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 13, 14

[50] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22873–22882, 2023. 2

[51] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023. 2, 3

[52] 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. 14

[53] Yu Zeng, Charles Ochoa, Mingyuan Zhou, Vishal M Patel, Vitor Guizilini, and Rowan McAllister. Neuralremaster: Phase-preserving diffusion for structure-aligned generation. *arXiv preprint arXiv:2512.05106*, 2025. 2, 3

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 6, 13

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[56] Siyuan Zhang, Wei Ma, Libin Liu, Zheng Li, and Hongbin Zha. Training-free fourier phase diffusion for style transfer. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2386–2394, 2025. 1, 2, 3, 6, 13

[57] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7396–7404, 2024. 2

Table A1. Categorization of style texts.

Category	Style Text
Traditional Arts (8)	"oil painting", "watercolor", "impressionism", "ink wash painting", "impasto painting", "sketching", "pointillism", "ukiyo-e"
Fragmented Forms (4)	"pixel art", "mosaic art", "cubism", "stained glass art"
3D & Spatial Forms (4)	"voxel art", "claymation", "low poly", "origami"
Urban & Digital Arts (4)	"cartoon", "cyberpunk", "glitch art", "graffiti painting"

Supplementary Material

A. Full List of Style Texts

Tab. A1 presents 20 style texts for the quantitative comparison of the main paper. To evaluate style generation capabilities across visual attributes such as painterly textures and geometric patterns, we curate the texts to cover four distinct categories.

B. Additional Experiments

B.1. Impact of Initial Injection Strength

Fig. A1 demonstrates that $\lambda_{\text{init}} \in \{0.1, 0.2\}$ is insufficient to mitigate the structural distortion caused by the initial noise. Conversely, $\lambda_{\text{init}} \in \{0.4, 0.5\}$ leads to an excessive accumulation of the low-frequency phase across all timesteps. Furthermore, the Gaussian filter preserves more mid/high-frequency phase as λ_{init} rises, limiting stylization flexibility. We adopt $\lambda_{\text{init}} = 0.3$ to strike a balance between these conflicting objectives.

We further quantify the phase injection amount based on the cumulative spectral content of each frequency band that passes through the Gaussian low-pass filter. Since there is no universal standard for frequency partitioning in the 2D Fourier transform, we define the frequency bands based on multi-resolution analysis [31]. We apply the hierarchical decomposition ratios of the wavelet transform to a circular partitioning scheme, reflecting the isotropy of the 2D Fourier spectrum in contrast to the conventional rectangular grid. Specifically, the spectrum is divided into low- (r = 0–25%), mid- (r = 25–50%), and high- (r = 50–100%) frequency bands.

As shown in Tabs. A2 and A3, $\lambda_{\text{init}} \in \{0.1, 0.2\}$ results in insufficient low-frequency phase injection compared to our method, whereas $\lambda_{\text{init}} \in \{0.4, 0.5\}$ leads to a significant increase in mid/high-frequency phase injection. Notably, for $\lambda_{\text{init}} \in \{0.4, 0.5\}$, the phase injection in the mid/high-frequency bands becomes comparable to or exceeds that in the low-frequency band, as indicated in Tab. A4.

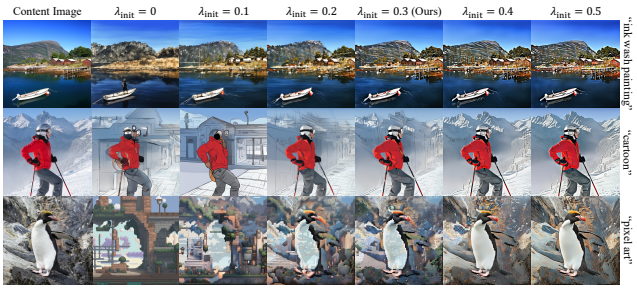


Figure A1. Qualitative results with varying initial injection strengths. $\lambda_{\text{init}} = 0$ corresponds to no phase refinement. We use three content images from the DIV2K dataset [1].

Table A2. Total low-frequency phase injection accumulated across all timesteps, normalized to our method ($\lambda_{\text{init}} = 0.3$).

Frequency Band	$\lambda_{\text{init}} = 0.1$	$\lambda_{\text{init}} = 0.2$
Low-freq. Band (r = 0–25%)	24.2%	69.7%

Table A3. Total mid- and high-frequency phase injection accumulated across all timesteps, normalized to our method ($\lambda_{\text{init}} = 0.3$).

Frequency Band	$\lambda_{\text{init}} = 0.4$	$\lambda_{\text{init}} = 0.5$
Mid-freq. Band (r = 25–50%)	172.2%	231.9%
High-freq. Band (r = 50–100%)	388.0%	828.2%

Table A4. Ratios of cumulative mid- and high-frequency phase injection to the low-frequency phase injection.

Frequency Ratio	$\lambda_{\text{init}} = 0.3$ (Ours)	$\lambda_{\text{init}} = 0.4$	$\lambda_{\text{init}} = 0.5$
Mid-freq. Band / Low-freq. Band	92.9%	134.6%	163.8%
High-freq. Band / Low-freq. Band	28.0%	91.3%	176.1%

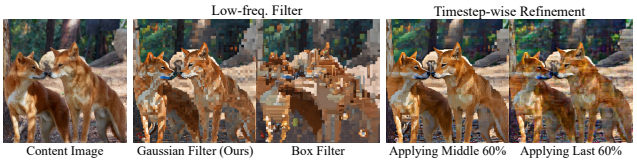


Figure A2. Ablation study of filter and timestep-wise refinement strategies. We use the style text "pixel art" for all images.

B.2. Additional Ablation Study

We conduct an additional ablation study to verify the Gaussian low-pass filter's soft-blending effect and the effectiveness of alternative refinement strategies that deviate from the SNR trend. As shown in Fig. A2, replacing the Gaussian filter with a box filter creates an abrupt frequency cutoff instead of a smooth spectral transition, which disrupts the content structure. Furthermore, applying low-frequency phase refinement during the middle or last 60% of the timesteps leads to stylization failures or visual artifacts.

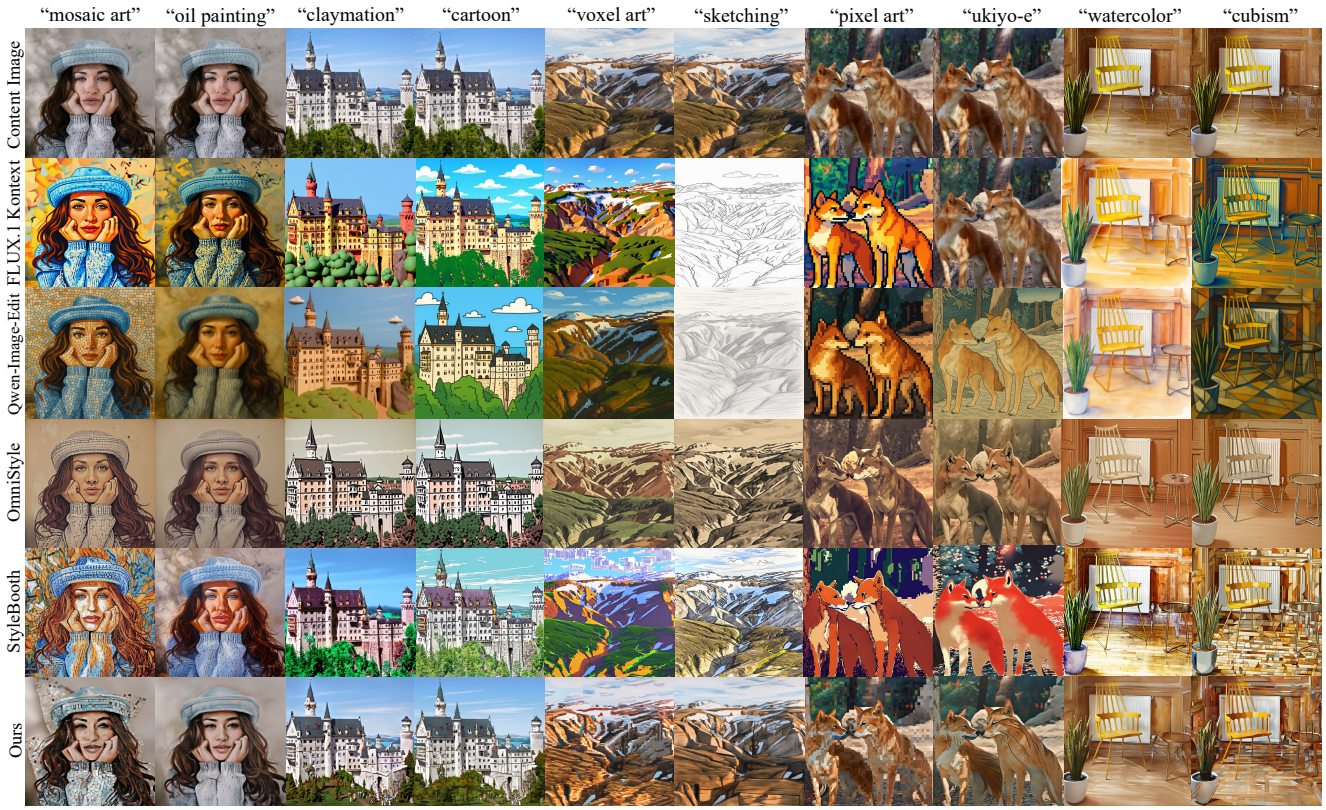


Figure A3. Qualitative comparison of stylization outcomes with recent works.

B.3. User Study

We conduct a user study with 70 participants to evaluate the overall stylization quality in Fig. 4 of the main paper. To mitigate potential bias, we randomize the order of the methods in Fig. 4 and mask the method names. Participants are provided with the entire figure and asked to select the model whose outputs best preserve the content while faithfully reflecting the style. In total, 40.0% of the participants prefer AdaPhase, the highest among all methods. FreeStyle [14], Zhang *et al.* [56], CLIPstyler [26], InstructPix2Pix [2], ControlNet [54], and DiffArtist [22] receive 20.0%, 14.3%, 10.0%, 7.1%, 4.3%, and 4.3%, respectively.

B.4. Comparison with other SOTA models

We further validate AdaPhase by comparing it with four recent state-of-the-art models in two categories: general-purpose text-to-image models (FLUX.1 Kontext [27] and Qwen-Image-Edit [49]) and text-guided stylization models (OmniStyle [47] and StyleBooth [12]). OmniStyle and StyleBooth require additional training, and are evaluated using their optimal hyperparameters.

Qualitative Comparison. As demonstrated in Fig. A3, while most methods generate plausible stylized images, our method consistently achieves content-preserving stylization across all images. FLUX.1 Kontext and Qwen-Image-Edit generally preserve semantic content during stylization. However, FLUX.1 Kontext fails to stylize certain images (*e.g.*, columns 1, 8 and 10), and Qwen-Image-Edit occasionally exhibits weaker content preservation compared to the other methods (*e.g.*, columns 1 and 7). OmniStyle and StyleBooth exhibit weak stylization on several samples (*e.g.*, columns 5, 7, and 8).

Table A5. Quantitative comparison with recent works. Experiment settings are identical to those in Fig. 5 of the main paper. **Bold values** and underlined values indicate the best and second best results, respectively.

Methods	FLUX.1 Kontext	Qwen-Image-Edit	OmniStyle	StyleBooth	Ours
1−LPIPS ↑	0.378	0.243	0.454	<u>0.467</u>	**0.541**
DINO ↑	<u>0.671</u>	0.495	**0.712**	0.621	**0.712**
Aesthetic ↑	<u>5.564</u>	**6.166**	5.547	5.430	5.504
Dir. CLIP ↑	0.112	**0.166**	0.114	0.153	<u>0.154</u>

Quantitative Comparison. As shown in Tab. A5, our method demonstrates superior performance in content preservation. Furthermore, our competitive Dir. CLIP score indicates AdaPhase's precise style reflection capability. We

Table A6. Quantitative comparison including standard deviations.

Methods	DINO	1−LPIPS	Aesthetic	Dir. CLIP
ControlNet	0.537±0.140	0.328±0.076	5.616±0.577	0.154±0.058
InstructPix2Pix	0.701±0.269	0.555±0.197	5.021±0.586	0.114±0.078
CLIPstyler	0.466±0.175	0.443±0.110	5.301±0.695	0.163±0.053
FBSDiff	0.463±0.188	0.370±0.086	5.324±0.654	0.142±0.062
FreeStyle	0.300±0.166	0.297±0.120	5.513±0.691	0.130±0.058
DiffArtist	0.588±0.215	0.548±0.146	5.202±0.721	0.121±0.078
Zhang *et al.*	0.332±0.166	0.284±0.109	5.695±0.738	0.124±0.062
FLUX.1 Kontext	0.671±0.208	0.378±0.219	5.564±0.726	0.112±0.073
Qwen-Image-Edit	0.495±0.189	0.243±0.091	6.166±0.567	0.166±0.049
OmniStyle	0.712±0.153	0.454±0.104	5.547±0.746	0.114±0.053
StyleBooth	0.621±0.223	0.467±0.132	5.430±0.664	0.153±0.061
AdaPhase (Ours)	0.712±0.142	0.541±0.098	5.504±0.816	0.154±0.070

Table A7. Comparison of inference time with recent works.

Methods	ControlNet	InstructPix2Pix	CLIPstyler	FBSDiff	DiffArtist	FreeStyle
Time (sec)	6.180	8.017	19.861	69.581	10.912	42.729
Methods	**Zhang *et al.***	**FLUX.1 Kontext**	**Qwen-Image-Edit**	**OmniStyle**	**StyleBooth**	**Ours**
Time (sec)	12.215	65.573	191.939	50.743	59.410	10.023



Figure A4. Robustness across different denoising strengths and numbers of inference steps. We use the style text "pixel art" for all images and set the denoising strength to $\gamma = 0.8$ when varying the number of inference steps.



Figure A5. Robustness across different samplers and model backbones. We use the style text "pixel art" for all images.

additionally report standard deviations of quantitative results in Tab. A6.

AdaPhase is training-free and does not require complex sampling control techniques. Accordingly, it achieves a competitively shorter inference time compared to other methods, as shown in Tab. A7.

B.5. Generalization of AdaPhase

AdaPhase demonstrates robust content-preserving stylization across varying denoising strengths and numbers of inference steps, as shown in Fig. A4.

PCA-based trajectory projection (Sec. 5.2 of the main paper) is theoretically grounded in the deterministic trajectory of latent diffusion models (Sec. 3.1 of the main paper). Consequently, AdaPhase maintains its robustness across sampling methods and model backbones when the underlying generative dynamics can be formulated as deterministic ODEs, as demonstrated in Fig. A5. Since the spectral characteristics (discussed in Sec. 3.2 of the main paper) are preserved in latent space, AdaPhase can be extended to latent generative models, including both diffusion [37] and flow-matching (with Transformer-based backbones) formulations [7, 27, 49]. For Stable Diffusion 2.1 [37], Stable Diffusion 3 [7], FLUX.1 Kontext, and Qwen-Image-Edit, we construct inversion paths using the dynamics of each model and adapt the low-frequency phase bandwidth to each model's noise schedule.

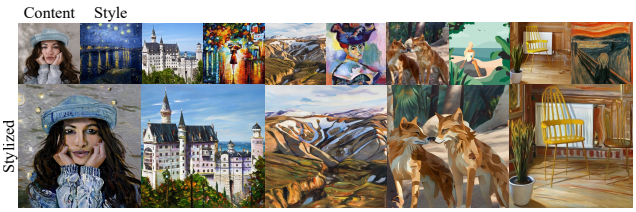


Figure A6. Application of AdaPhase combined with (SDXL-based) IP-Adapter.



Figure A7. Low-frequency component transfer. While our default setting provides a robust baseline for content preservation, increasing the denoising strength can further enhance low-frequency stylization.

B.6. Application of AdaPhase

We additionally integrate our method with IP-Adapter [52] and demonstrate its applicability to tasks that take a style image as input. As shown in Fig. A6, AdaPhase successfully reflects the reference style while achieving content preservation. As noted in Sec. B.5, AdaPhase generalizes across diverse generative models.

B.7. Transferring styles involving low-frequency components

While AdaPhase preserves content structure by injecting the low-frequency phase of the inversion latent, it still allows low-frequency attributes (*e.g.*, object contours and colors) to be stylized, as shown in Fig. A7. The denoising network continues to iteratively refine the latent under text guidance, allowing low-frequency stylization to emerge throughout the generation process.

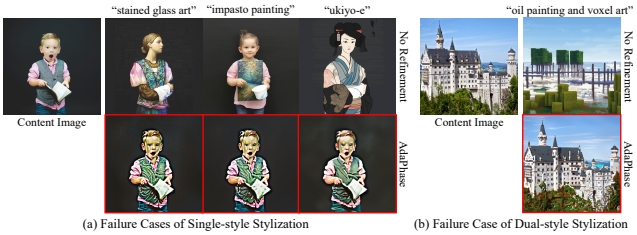


Figure A8. Failure cases of AdaPhase. *No Refinement* denotes the standard SDXL pipeline. (a) For "stained glass art" and "impasto painting", SDXL shows incomplete stylization, with style appearing only locally. For "ukiyo-e", the style is inaccurate in the SDXL output. (b) For "oil painting and voxel art", the SDXL result reflects only the "voxel art" style and ignores the "oil painting" style.

B.8. Additional Qualitative Results

To verify the performance of AdaPhase across the 20 style texts described in Sec. A, we conduct qualitative experiments with 19 content images from the DIV2K dataset [1]. As demonstrated in Figs. A9 to A12, AdaPhase consistently preserves content structure while achieving high-quality stylization. This capability is supported by the close alignment of the trajectory with the inversion path, as illustrated in Fig. A12. For Figs. A9 to A12, see the last four pages.

C. Limitations and Future Work

AdaPhase enables content-preserving stylization for various style texts by leveraging spectral modulation and SNR-based adaptive scheduling. However, as shown in Fig. A8(a), instances of incomplete stylization remain. This limitation arises because AdaPhase's stylization relies on the SDXL 1.0 image-to-image pipeline as its backbone. If this backbone fails to adequately generate the target style for a specific content image, AdaPhase inevitably inherits this failure, resulting in unfaithful stylization. This dependency is further highlighted in Fig. A8(b). When the backbone pipeline fails to handle text containing two distinct styles, AdaPhase likewise fails to capture both styles in the final output. Building upon the alignment of the denoising trajectory with the inversion path, future research on AdaPhase should expand toward a theoretical analysis of how the denoising process incorporates specific style texts.

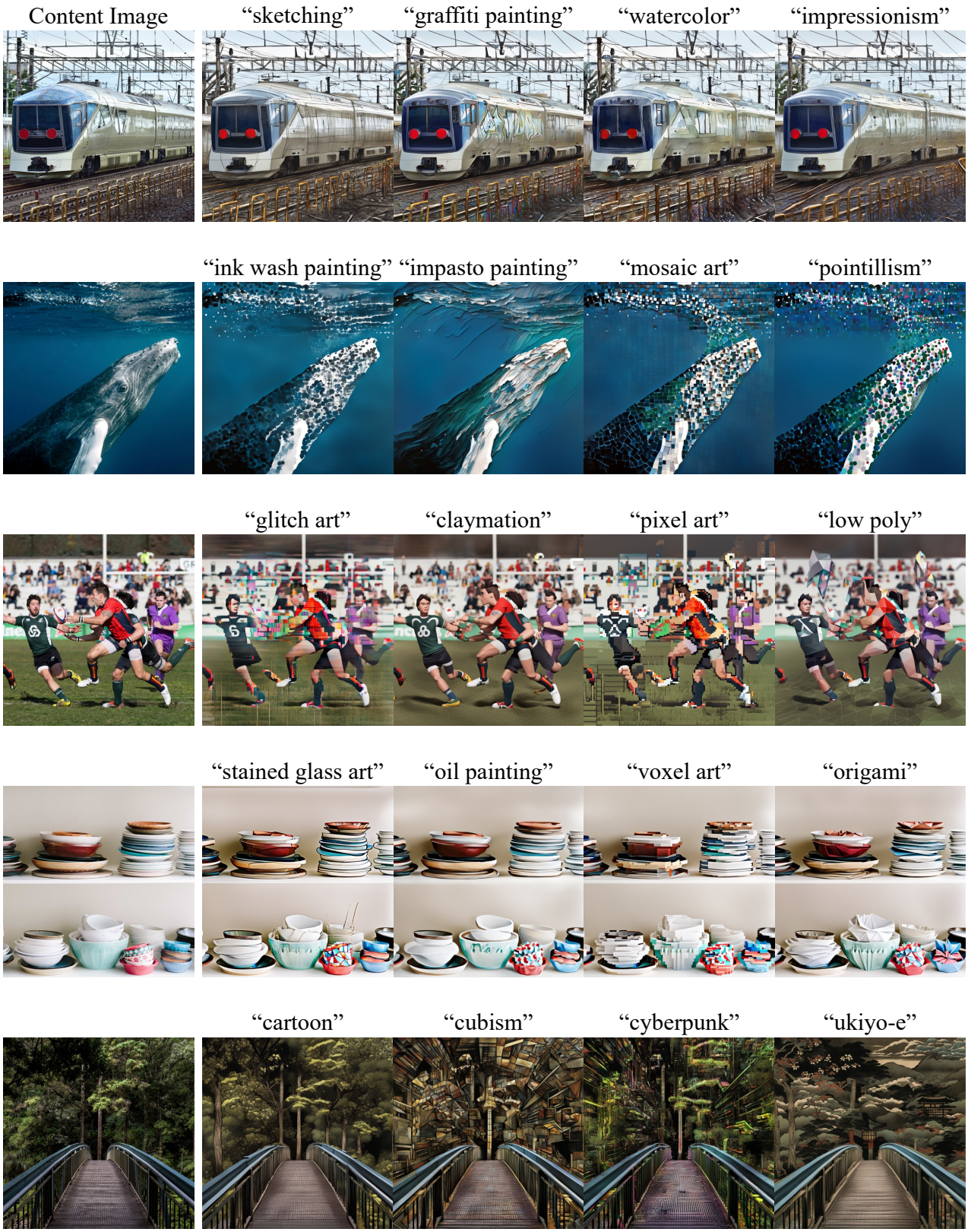


Figure A9. Additional qualitative results of AdaPhase using 20 style texts (Part 1).

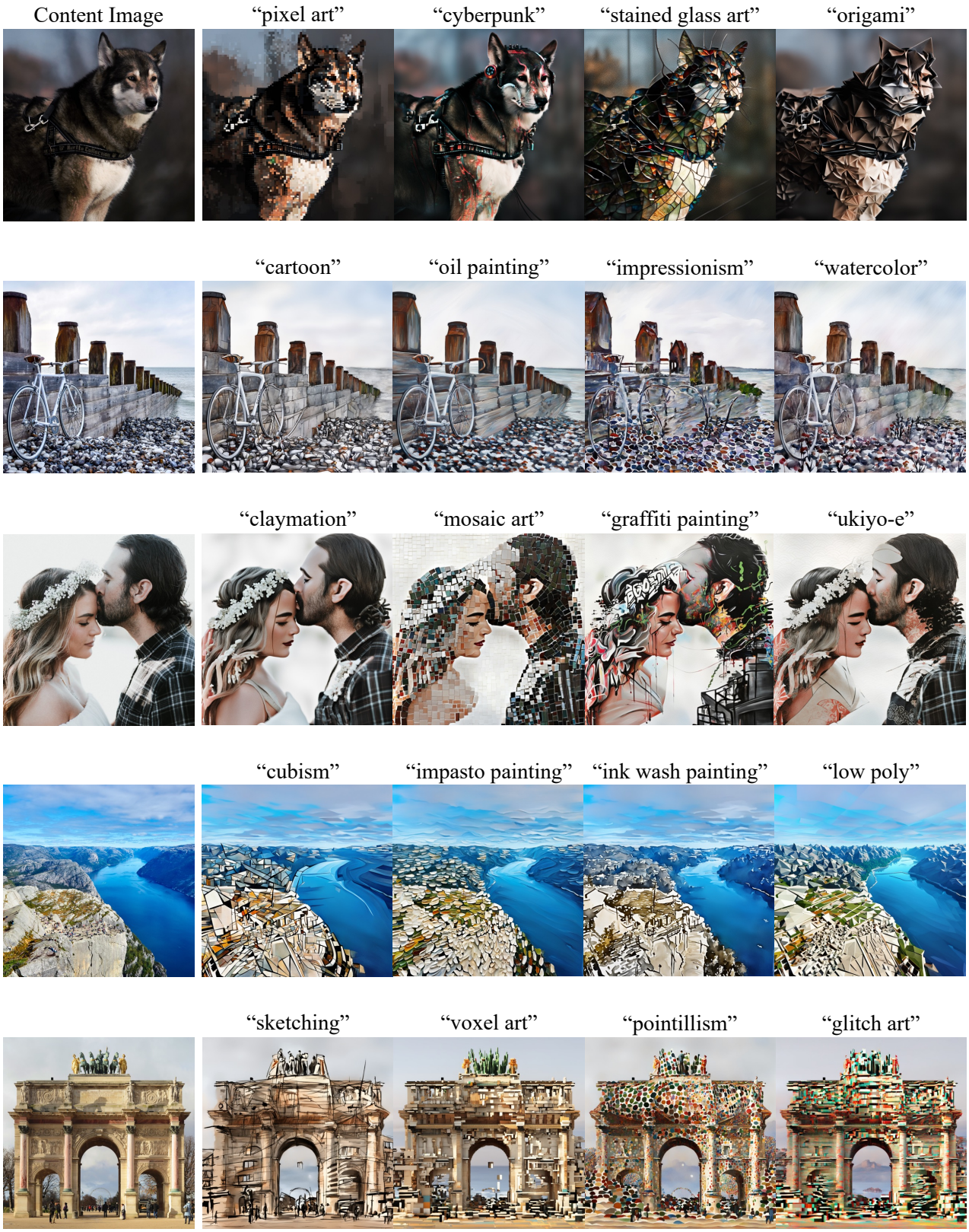


Figure A10. Additional qualitative results of AdaPhase using 20 style texts (Part 2).

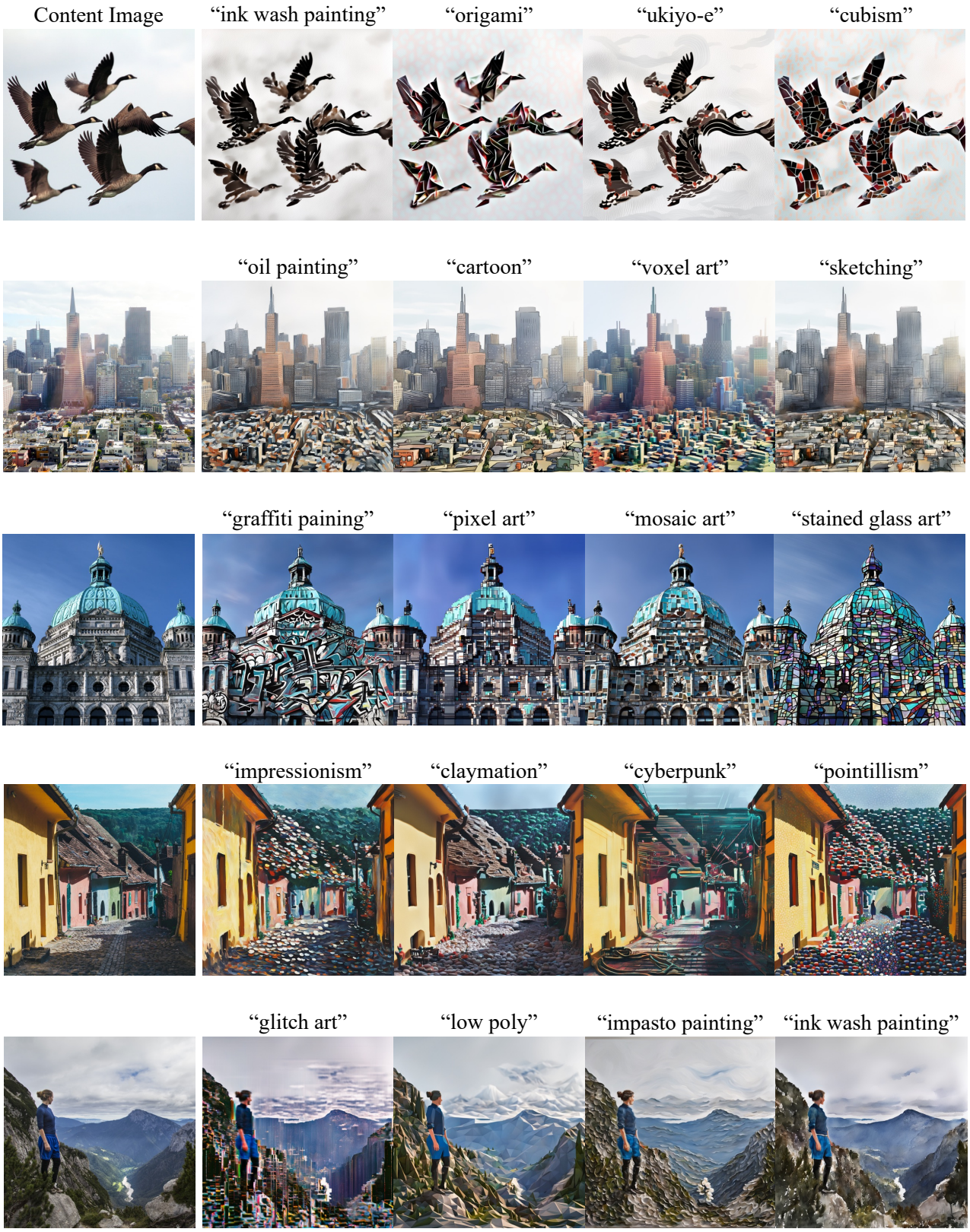


Figure A11. Additional qualitative results of AdaPhase using 20 style texts (Part 3).

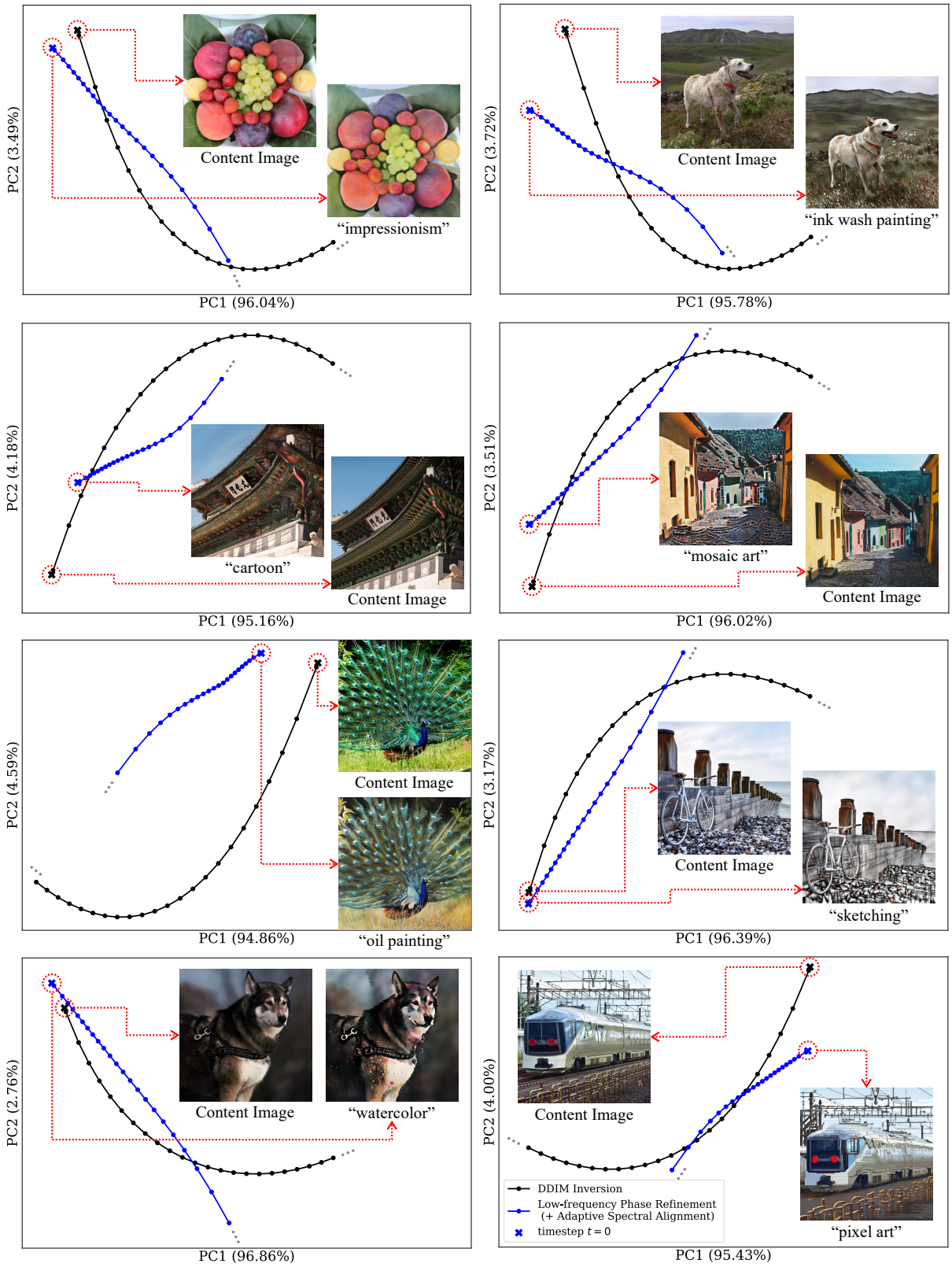


Figure A12. Additional qualitative results of AdaPhase and latent trajectories in the DDIM inversion-based PCA space.