CLoSeR: Continual Learning in VQ-GAN for Test-time Style Refinement

Anonymous authors
Paper under double-blind review

Abstract

While existing artistic style transfer methods enable cross-domain image synthesis, they often struggle to strike a balance among stylistic realism, inference efficiency, and geometric consistency. To address this limitation, we propose a *test-time refinement* (TTR) framework that universally enhances stylistic fidelity through a self-supervised VQ-GAN, without requiring any gradient updates to the pre-trained generator. Our primary contribution is a continual learning framework for VQ-GAN, which combines *Low-Rank Adaptation* (LoRA) with incremental codebook expansion. This design enables efficient adaptation to diverse artistic styles while preserving previously learned knowledge, significantly reducing the computational and memory overhead of deploying models across multiple domains. Notably, our approach reduces the number of trainable parameters by up to 94% compared to full-model fine-tuning, offering a highly parameter-efficient solution for test-time refinement. Furthermore, we introduce positional embeddings into the latent embedding space, which strengthens the model's geometry awareness and improves structural coherence in the generated results. We name our approach CLoSeR (***C**ontinual **L**earning in VQ-GAN fo**r** **S**tyl**e** **R**efinement*), and evaluate it across multiple style transfer benchmarks under a test-time adaptation setting. Experimental results show that CLoSeR improves style fidelity and structural consistency, achieving a maximum relative reduction of 44% in *Fréchet Inception Distance* (FID), demonstrating significant gains in generation quality. The code will be released.

1 Introduction

Artistic style transfer (AST) has witnessed rapid progress through a variety of approaches, most notably neural style transfer (NST) (Gatys et al., 2016; Huang & Belongie, 2017; Liu et al., 2021; Hong et al., 2023) and generative adversarial networks (GANs) (He et al., 2018; Lee et al., 2020; Huang et al., 2024). These methods typically rely on one or a few reference style images to guide the stylization process. More recently, diffusion models (Zhang et al., 2023; Chung et al., 2024; Wang et al., 2024; Zhou et al., 2025), autoregressive (AR) approaches (Li et al., 2024), and flow-based generative methods (Lipman et al., 2022; Geng et al., 2025) have demonstrated impressive capabilities in producing high-quality and diverse stylizations, often supporting multimodal inputs. These advances highlight the growing importance of transferable representations that capture both content and stylistic priors, enabling more flexible and controllable AST.

However, existing methods struggle to achieve an optimal balance between content consistency, stylistic realism, and inference efficiency. NST and GAN-based methods (Gatys et al., 2017; Selim et al., 2016; Zhu et al., 2017) enable fast inference and preserve geometric structure well, but often fail to learn sufficiently rich representations of artistic textures. Diffusion models (Zhang et al., 2023; Wang et al., 2024; Ye et al., 2025) generate high-quality results with nuanced style patterns, yet suffer from hallucinated content, weak content–style correspondence, and the high computational cost due to iterative sampling. Reducing inference steps typically degrades image quality significantly. Moreover, both diffusion and AR models often yield over-smoothed textures, suggesting that their learned representations do not fully align with the expressive nature of real-world artistic styles. Few-shot or training-free adaptation methods (Chung et al., 2024; Farhadzadeh et al., 2025) further face challenges in building robust representations for unseen domains. Thus, learning domain-aligned and structurally consistent representations remains an open challenge for AST.

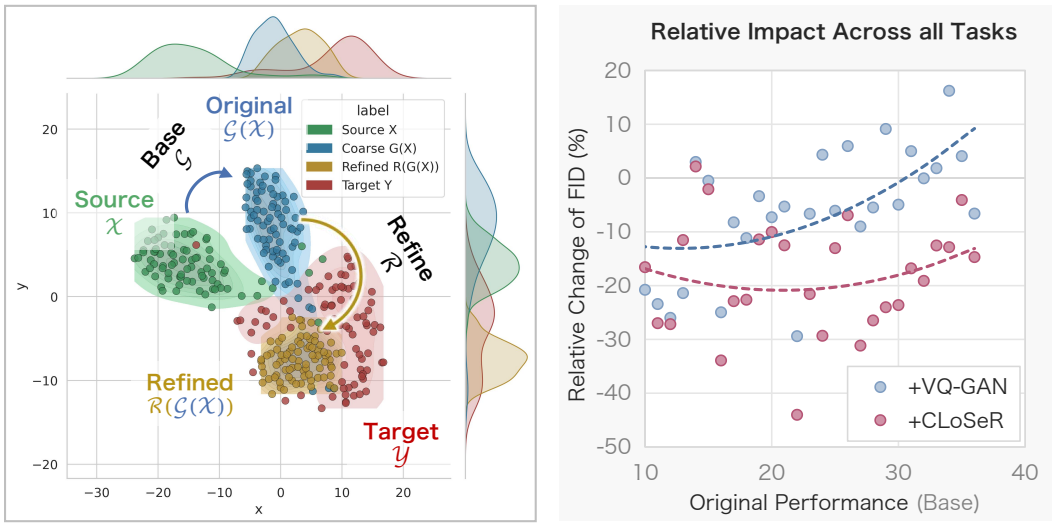

Figure 1: Motivation of CLoSeR. **Left:** illustration of the distribution shift from the source domain (CelebAMask-HQ (Lee et al., 2020)) to the target domain (MetFace (Karras et al., 2020)). StyleID (Chung et al., 2024) serves as the base model to generate coarse outputs, while our CLoSeR produces refined results that align more closely with the target domain. Features are extracted with VGG-19 (Simonyan & Zisserman, 2014) and visualized via t-SNE.) **Right:** scatter plot of refined performance versus original performance across diverse base models and artistic styles. Lower FID values indicate better style fidelity.

As illustrated in Figure 1, the motivation for our approach stems from the persistent distributional gap between stylized outputs and the target domain. While existing image translation models—such as GAN-, attention-, and diffusion-based methods (Huang & Belongie, 2017; Liu et al., 2021; Yi et al., 2019; Chung et al., 2024; Zhou et al., 2025) —can roughly map source content into the target style space, their outputs often exhibit significant deviations from the authentic target distribution, particularly in terms of stylistic fidelity and geometric consistency (left panel). These gaps indicate a representation mismatch between the generated outputs and the target artistic domain.

To address this, we explore an alternative perspective: rather than retraining or modifying the generator, we refine its outputs at test time through reconstruction in the embedding space. Inspired by the ability of VQ-GAN (Esser et al., 2021) to learn a compact, self-supervised representation of the target domain, we propose a *test-time refinement* (TTR) framework that leverages VQ-GAN as a domain anchor. In other words, VQ-GAN refines coarse stylized images by aligning their features with a pre-learned target domain representation in its latent codebook, eliminating the need for generator updates.

However, directly fine-tuning VQ-GAN for each new style remains computationally expensive and lacks scalability. To overcome these limitations, we propose a TTR framework dubbed **CLoSeR**, *i.e.* ***C**ontinual **L**earning in VQ-GAN for **S**tyle **R**efinement*. CLoSeR enables efficient continual adaptation by incrementally enriching the learned representation space through *Low-Rank Adaptation* (LoRA) (Hu et al., 2022) and codebook expansion. This design drastically reduces the number of trainable parameters—by over 94% compared to full fine-tuning—while preserving previously acquired representations of earlier styles. Furthermore, to mitigate structural distortions caused by the lack of spatial awareness in vanilla VQ-GAN (Esser et al., 2021), we incorporate 2D sine-cosine positional embeddings (Vaswani et al., 2017; Carion et al., 2020) into the latent representation space, endowing the codebook and decoder with explicit spatial priors. Together, these components enable CLoSeR to refine generation quality through representation learning, achieving both high-fidelity stylization and geometric consistency across diverse artistic domains.

We conduct extensive experiments to evaluate the effectiveness and generality of our approach. The results demonstrate that CLoSeR consistently improves generation quality across diverse style transfer pipelines—including GAN- (Yi et al., 2019; Zhang et al., 2022), attention- (Liu et al., 2021;

Hong et al., 2023), and diffusion-based (Kwon & Ye, 2022; Chung et al., 2024; Zhou et al., 2025) models—under both single-style and continual learning settings. The framework enhances stylistic realism and structural consistency, while also learning transferable representations. As shown in the right panel of Figure 1, a scatter plot of FID improvement reveals that both the baseline VQ-GAN and CLoSeR reduce stylization errors compared to the original outputs, but CLoSeR achieves significantly greater FID reduction, particularly in challenging cases with higher baseline errors. This confirms its superior refinement capability and scalability in real-world deployment scenarios.

2 Related Works

Artistic Style Transfer. Early approaches leveraged CNNs to decouple style and content representations, enabling stylized image synthesis (Gatys et al., 2016; Johnson et al., 2016; Jing et al., 2019). Subsequent methods aimed to enhance style diversity and generalization by introducing adaptive normalization and attention-based mechanisms (Huang & Belongie, 2017; Park & Lee, 2019; Hong et al., 2023). More recently, diffusion-based approaches have emerged as powerful alternatives for style and domain transfer (Ho et al., 2020; Kwon & Ye, 2022; Gu et al., 2022). These methods have been applied to stylization, latent space disentanglement, and domain adaptation by exploiting denoising priors and structured noise injection (Kwon & Ye, 2022; Su et al., 2022; Parmar et al., 2024; Zhou et al., 2025). In addition, training-free paradigms have been explored to achieve lightweight and interpretable transfer (Chung et al., 2024). Despite these advances, both CNN-based and diffusion-based pipelines often struggle with preserving structure and maintaining style fidelity in complex artistic domains.

Vector Quantization. Vector Quantization (VQ) has emerged as a powerful technique for learning discrete representations. VQ-VAE (Van Den Oord et al., 2017) pioneered vector quantization in generative modeling, and VQ-GAN (Esser et al., 2021) further advanced this direction. Building on the success of VQ-GAN, a variety of works have emerged, such as VQ-Diffusion (Gu et al., 2022) for text-to-image generation and QuantArt (Huang et al., 2023) for artistic style transfer. Reconstruction and generation using VQ have also been widely studied (Zhu et al., 2024; Yu et al., 2024; Yao et al., 2025). In the autoregressive paradigm, Li et al. (2024) propose eliminating discrete quantization entirely by modeling per-token distributions, while MergeVQ (Li et al., 2025) unifies representation learning and generation through token merging and a lookup-free quantization strategy.

Continual Learning. Continual learning has been extensively studied, but its application to artistic domains remains relatively underexplored. Traditional style transfer methods often require retraining for each new style (Gatys et al., 2016; Johnson et al., 2016), making them inefficient and vulnerable to catastrophic forgetting. To address these limitations, modular and parameter-efficient approaches have been proposed (Liang & Li, 2024; Zhu et al., 2025; He et al., 2025; Roy et al., 2023). More recently, continual generative learning has incorporated strategies such as replay (Caccia et al., 2020; Jeon et al., 2023), distillation (Lesort et al., 2019; Zhao et al., 2020), and modularization (Yoon et al., 2018). LoRA-based adapters (Hu et al., 2022; Farhadzadeh et al., 2025) have proven particularly effective, enabling lightweight, style-specific modules to be integrated into frozen backbones for scalable, efficient, and largely forget-free adaptation. However, they still suffer from increasing knowledge degradation as the number of tasks grows (Liang & Li, 2024).

3 Method

3.1 Overview

We propose CLoSeR (***C**ontinual **L**earning in VQ-GAN fo**r** **S**tyl**e** **R**efinement*), a *test-time refinement* (TTR) framework that enhances both stylistic realism and geometric consistency in artistic style transfer. The pipeline of CLoSeR is shown in Figure 2. Building upon VQ-GAN (Esser et al., 2021), we integrate parameter-efficient adaptation through *Low-Rank Adaptation* (LoRA) and incremental codebook expansion, supporting continual adaptation to new styles with minimal overhead. For each new style, only a lightweight LoRA module and a style-specific discriminator are trained, while the shared VQ-GAN backbone remains frozen. This strategy enables scalable deployment in dynamic and long-tail style scenarios. In addition, our approach introduces *geometry-aware*

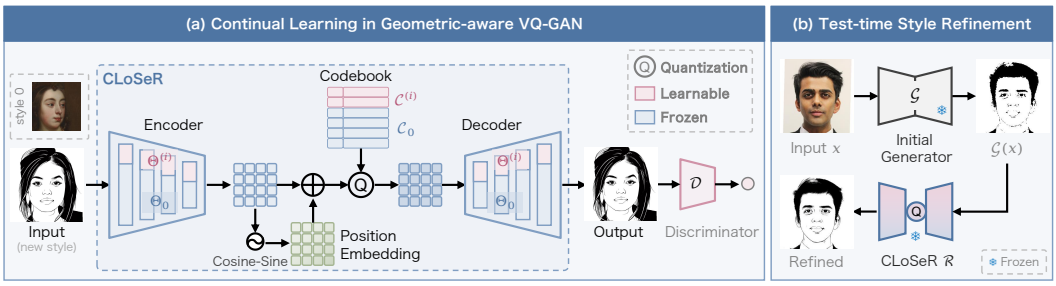

Figure 2: Overview of **CLoSeR**, *i.e.*, ***C**ontinual **L**earning in VQ-GAN **f**or test-time **S**tyl**e** **R**efinement*. (a) New styles are integrated by expanding the codebook ($\mathcal{C}^{(i)}$) while retaining the base style representation $\mathcal{C}_0$ (style 0). The encoder features are enriched with cosine-sine positional embeddings and reconstructed by the decoder with LoRA-based adaptation. (b) Given an initial coarse stylized output $\mathcal{G}(x)$ from any generator, CLoSeR reconstructs it through the learned codebook, aligning the result with the target style domain.

vector quantization by embedding positional encodings into the latent space, allowing the model to incorporate explicit spatial priors during reconstruction and thereby correcting geometric distortions and local artifacts commonly present in coarse stylized outputs. Finally, CLoSeR operates in a *plug-and-play* manner and can be applied to enhance outputs from arbitrary generative models.

3.2 Continual Learning in VQ-GAN via LoRA

Adapting to new artistic styles while preserving previously learned knowledge remains challenging due to catastrophic forgetting and the large parameter overhead of full fine-tuning. To enable efficient and scalable continual learning, we integrate LoRA (Hu et al., 2022) and incremental codebook expansion into the VQ-GAN framework, allowing CLoSeR to adapt to new styles with minimal trainable parameters while keeping the shared backbone frozen.

LoRA-based Encoder–Decoder Adaptation. We apply LoRA to all convolutional layers of the encoder and decoder, injecting trainable low-rank matrices to modulate features in a style-specific manner. Specifically, each pre-trained weight $W_0 \in \mathbb{R}^{d\times k}$ is updated as:

$$W = W_0 + \frac{\alpha}{r} AB, \quad A \in \mathbb{R}^{d\times r},\ B \in \mathbb{R}^{r\times k}, \tag{1}$$

where A and B are the low-rank adaptation matrices. A is initialized with zeros, B with a standard normal distribution, α is a scaling factor, and r is the rank (set to 8 in our experiments). The original weights W_0 remain frozen and are shared across all styles.

Incremental Codebook Expansion. For each new style s_i, we expand the codebook with $\Delta K = 1024$ additional entries:

$$\mathcal{C}^{(i)} = \mathcal{C}0 \cup eK_0 + 1, \ldots, e_{K_0+\Delta K}, \tag{2}$$

where $\mathcal{C}_0$ denotes the initial codebook. This strategy enables the model to encode style-specific visual primitives while preserving previously learned representations.

Training. During training on style s_i, only three components are updated: the LoRA parameters $\Theta^{(i)}_{\text{LoRA}}$, the newly added codebook entries $\mathcal{C}^{(i)} \setminus \mathcal{C}_0$, and a lightweight style-specific discriminator $\mathcal{D}^{(i)}$. All other parameters—including the encoder, decoder, and the base codebook—remain frozen.

Inference. At inference, given an initial stylized result $\mathcal{G}(x)$ from any pre-trained generator, the refined output for style s_i is computed as:

$$\hat{y} = \mathcal{R}\big(\mathcal{G}(x); \Theta^{(i)}\big), \quad \text{with} \quad \Theta^{(i)} = \Theta^{(i)}_{\text{LoRA}}, \mathcal{D}^{(i)}. \tag{3}$$

This modular design enables plug-and-play refinement: users select the target style, and the system loads the corresponding lightweight parameters, thereby avoiding redundant computation and supporting efficient deployment in dynamic or long-tail scenarios.

3.3 GEOMETRY-AWARE VQ-GAN

To improve spatial structure preservation in artistic style reconstruction, we enhance the VQ-GAN framework (Esser et al., 2021) with 2D sine–cosine positional embeddings injected into the latent representation space. Unlike standard VQ-GAN, which processes latent features without explicit spatial inductive bias, our method embeds positional priors prior to quantization—thereby enabling geometry-aware refinement without introducing any additional learnable parameters.

Similar to Transformer (Vaswani et al., 2017), for each spatial position $(m, n) \in \{1, \ldots, h\} \times \{1, \ldots, w\}$ of the continuous latent feature map $f_s \in \mathbb{R}^{h \times w \times d}$, we generate a corresponding 2D positional embedding $P_{m,n} \in \mathbb{R}^d$ using an extended sine-cosine scheme:

$$\begin{aligned} P_{m,2i} &= \sin\left(\frac{m}{10000^{\frac{2i}{d}}}\right), \quad P_{m,2i+1} = \cos\left(\frac{m}{10000^{\frac{2i}{d}}}\right), \\ P_{n,2i} &= \sin\left(\frac{n}{10000^{\frac{2i}{d}}}\right), \quad P_{n,2i+1} = \cos\left(\frac{n}{10000^{\frac{2i}{d}}}\right), \end{aligned} \tag{4}$$

where m and n denote the row and column indices, i is the dimension index, and d is the embedding dimension. The positional embedding $P_{m,n}$ is then added element-wise to the latent feature f_s:

$$f_{pe} = f_s + P_{m,n}, \tag{5}$$

forming spatially enriched features that preserve both content semantics and explicit spatial structure.

The enhanced features f_{pe} are then passed to the codebook for quantization:

$$Q_{\mathcal{C}}(f_{pe}) := \arg\min_{\mathbf{c}_i \in \mathcal{C}} \|f_{pe} - \mathbf{c}_i\|, \tag{6}$$

where $\mathbf{c}_i$ denotes the i-th code vector in the codebook $\mathcal{C}$. By integrating explicit spatial priors into the vector quantization pipeline, our approach effectively improves geometric consistency in the reconstructed outputs, particularly in structure-sensitive artistic domains.

3.4 LOSS FUNCTIONS

To balance pixel-level fidelity, perceptual quality, quantization alignment, and adversarial realism, we adopt a composite loss composed of multiple complementary objectives.

Reconstruction Objective. The reconstruction objective combines an L1 pixel-wise loss and a perceptual loss in deep feature space. Given the input image x_s and its reconstruction y_s, the pixel-level reconstruction loss is defined as $\mathcal{L}_{\text{rec}} = \|y_s - x_s\|_1$. To capture higher-level semantic consistency, we further employ the LPIPS metric (Zhang et al., 2018) as a perceptual loss:

$$\mathcal{L}_{\text{perc}} = \text{LPIPS}(x_s, y_s). \tag{7}$$

The total reconstruction loss is then given by:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{perc}} \cdot \mathcal{L}_{\text{perc}}, \tag{8}$$

where λ_{perc} controls the relative weight of perceptual similarity.

VQ Loss. Following standard practice in vector quantized models (Esser et al., 2021), we incorporate a vector quantization (VQ) loss to align the latent space with the codebook. Let $f_s \in \mathbb{R}^{B \times C \times H \times W}$ denote the continuous latent features from the encoder. We enrich these features with 2D sine-cosine positional encoding (see Section 3.3) to obtain f_{pe}, which is then flattened and mapped to the nearest entries in a learnable codebook $\mathcal{C} \in \mathbb{R}^{K \times D}$, where K is the number of codebook vectors and D is the embedding dimension. The quantized output z_{q} replaces each feature in f_{pe} with its closest codebook entry under the Euclidean distance. To jointly optimize the codebook and encoder, we use the following VQ loss:

$$\mathcal{L}_{\text{VQ}} = \|\mathbf{sg}[z_{\text{q}}] - f_{pe}\|_2^2 + \beta \|\mathbf{sg}[f_{pe}] - z_{\text{q}}\|_2^2, \tag{9}$$

where $\mathbf{sg}[\cdot]$ denotes the stop-gradient operator and β is a hyperparameter controlling the codebook update strength.

Adversarial Loss. For adversarial training, we adopt the standard cross-entropy objective as in VQ-GAN (Esser et al., 2021). The discriminator $\mathcal{D}^{(i)}$ for style s_i is optimized as:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}[\log \mathcal{D}^{(i)}(y)] - \mathbb{E}[\log(1 - \mathcal{D}^{(i)}(y_s))], \tag{10}$$

where y and y_s denote real and reconstructed images.

Total Loss. The overall training objective is a weighted combination of all components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \mathcal{L}_{\text{adv}}, \tag{11}$$

where λ_{VQ} is set to 0.1 by default. This multi-objective formulation ensures high-fidelity, geometrically coherent, and stylistically realistic reconstructions.

4 Experiments

4.1 Settings

Datasets & Metrics. For the **Artistic Portrait** domain, we use MetFace (Karras et al., 2020), APDrawing (Yi et al., 2019), and FS2K (Fan et al., 2022) as style datasets, with facial photos from CelebAMask-HQ (Lee et al., 2020) and FS2K serving as content images. For the **Natural Scene** domain, we collect data from Flickr and WikiArt. We adopt standard metrics—ArtFID (Wright & Ommer, 2022), FID (Heusel et al., 2017), and KID (Bińkowski et al., 2018)—to quantitatively evaluate our results. All images are resized to 256×256 before training and evaluation.

Implementation Details. Following the architecture design of QuantArt (Huang et al., 2023), the encoder and decoder each consist of four blocks, with two ResBlocks (He et al., 2016) and a downsampling/upsampling layer. The quantized feature map has a spatial resolution of 16×16 and an embedding dimension of 256. The codebook contains $N = 1024$ entries, each of dimension $d = 256$. For training, we set the batch size to 8 and the momentum queue length to 1024. For each newly added style, the codebook is expanded by 1024 tokens. We use the Adam optimizer (Adam et al., 2014) with a learning rate of 4.5×10^{-6}. Our CLoSeR framework is implemented in PyTorch (Paszke et al., 2019), and all experiments are conducted on a single NVIDIA RTX 4090 GPU.

Baseline Models. We evaluate our method against a set of state-of-the-art methods, including neural style transfer (QuantArt (Huang et al., 2023), AesPA-Net (Hong et al., 2023), CAST (Zhang et al., 2022), AdaAttN (Liu et al., 2021)), and diffusion-based stylized image generation (DiffuseIT (Kwon & Ye, 2022), InST (Zhang et al., 2023), StyleID (Chung et al., 2024) and AttenDistill (Zhou et al., 2025)). For fair comparison, we use publicly available implementations with their recommended configurations. As shown in Figure 4, our method outperforms all base models in both stylization fidelity and semantic consistency. Note that APDrawingGAN (Yi et al., 2019) is specialized for pen drawings, thus we evaluate it only in its intended settings to ensure fairness.

4.2 Performance Evaluation

4.2.1 Natural Scene Style Transfer

Unlike the standard style transfer task, we train our model to reconstruct the input and use this to refine the results of artistic style transfer results. The model is first trained on the Monet dataset and then continually extended to Van Gogh and Ukiyo-e, enabling progressive refinement across multiple styles. Experimental results demonstrate the effectiveness of our approach.

Quantitative Analysis. As illustrated in Figure 3, for both Monet and Van Gogh, the average values of all three evaluation metrics consistently decrease after the initial refinement with VQ-GAN and are further reduced when applying our proposed CLoSeR. Notably, across all baselines, our method achieves substantial improvements: FID is reduced by approximately 25% on Monet and Van Gogh, KID drops by more than 30%, and ArtFID decreases by over 20%.

Qualitative Analysis. As shown in Figure 4, CLoSeR consistently enhances base models by recovering structural details and enriching textures. Without refinement, AdaAttN and AesPA-Net tend to produce over-smoothed outputs, while vanilla VQ-GAN introduces texture but often causes distortions. In contrast, CLoSeR yields more faithful style expression—Monet's color gradients appear

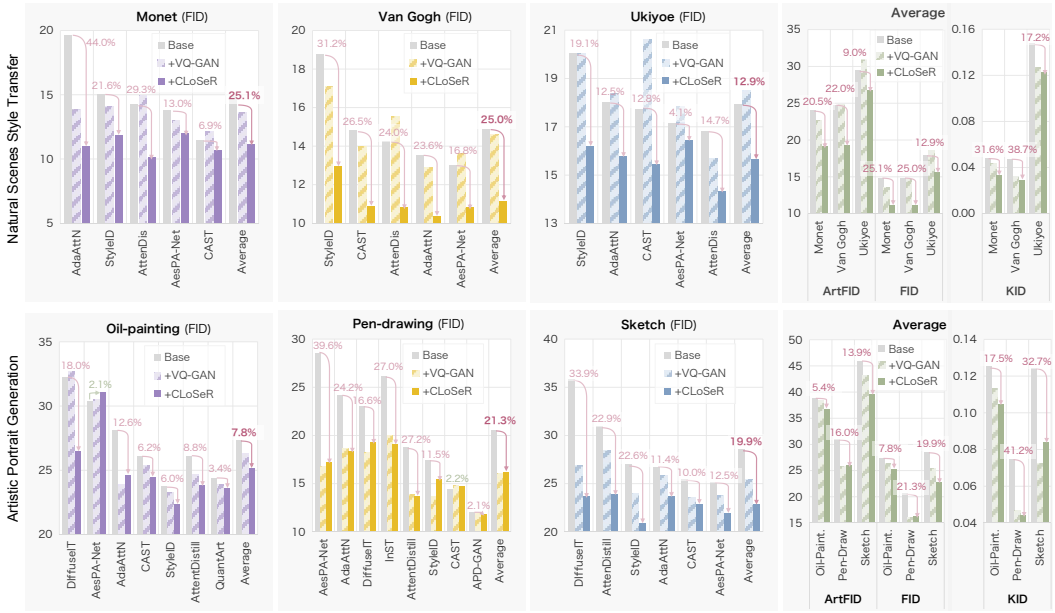

Figure 3: Quantitative performance on artistic style transfer for natural scenes and facial portraits.

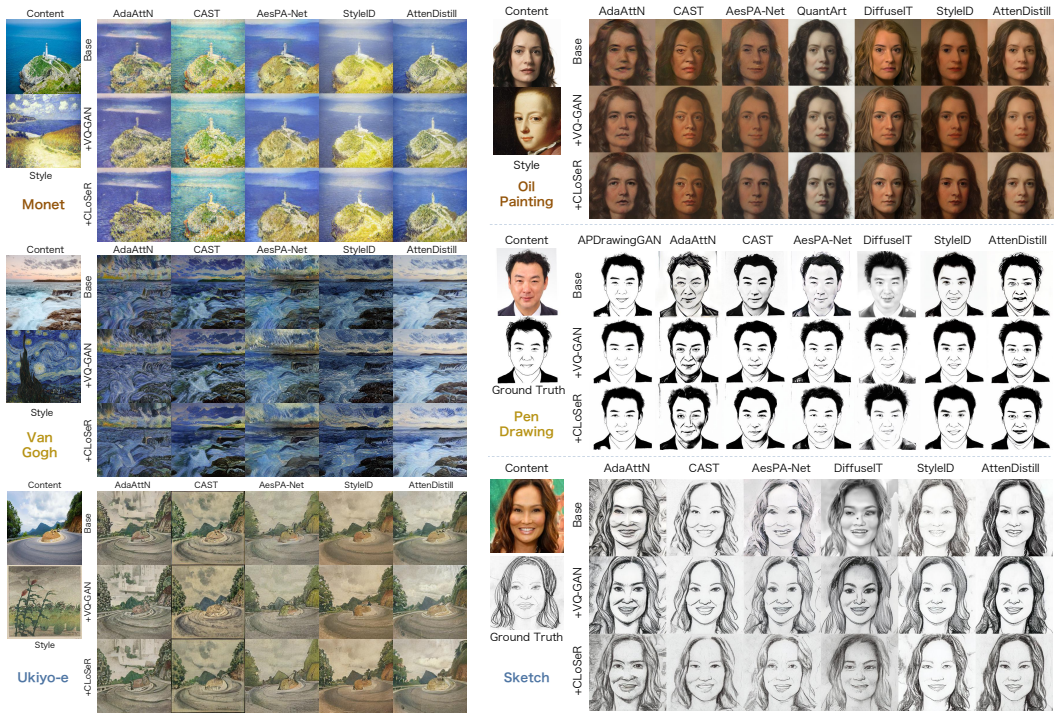

Figure 4: Generated results of different artistic styles for natural scenes and facial portraits. Please zoom in for details.

smoother, Van Gogh's bold strokes are better preserved, and Ukiyo-e's flat shading and outlines remain more coherent—demonstrating improved style fidelity and content stability across diverse models.

4.2.2 ARTISTIC PORTRAIT GENERATION

We first pre-train CLoSeR on the MetFace (Karras et al., 2020) dataset to learn robust facial representations. Building on this foundation, we extend the model to support continual refinement across two additional styles: APDrawing (Yi et al., 2019) and FS2K (Fan et al., 2022), resulting in a three-style refinement setup.

Quantitative Analysis. As illustrated in Figure 3, our CLoSeR significantly improves performance across all artistic domains and metrics. Compared to the base models, CLoSeR reduces average FID by 7.8%, 21.3%, and 19.9% on oil painting, pen drawing, and sketch styles, respectively, consistently outperforming the intermediate VQ-GAN refinement. Across tasks, CLoSeR achieves an average 5.4% ArtFID reduction and a 17.5% KID decrease on the oil painting domains. Moreover, ArtFID decreases by 16.0% on pen drawing, indicating stronger stylistic consistency. These results confirm that our method not only restores structural fidelity but also enhances stylistic realism across diverse datasets and models.

Qualitative Analysis. For Oil Paintings, AdaAttN and AesPA-Net produce over-smoothed or distorted faces, while VQ-GAN reduces artifacts but suffers from leakage and color shifts. CLoSeR better preserves identity (sharper jawlines, clearer eyes) and renders textures closer to the target style. For Pen Drawings, DiffuseIT and AesPA-Net often yield blurry or off-domain results; VQ-GAN adds stroke effects but loses detail and symmetry. CLoSeR restores crisp contours and accurate strokes, resembling ground truth. For Sketches, base models distort proportions (e.g., bloated or muddy textures), whereas CLoSeR enhances contour sharpness and line stability. These improvements highlight its ability to recover fine-grained structure while embedding faithful stylistic cues.

4.2.3 MODEL EFFICIENCY

As shown in Table 1, CLoSeR is highly efficient, requiring only 4.74 MB trainable parameters, 2.42 GB memory, and 0.0545 s inference—substantially lower than most baselines. Its lightweight test-time adaptation, without modifying the generator, offers an excellent trade-off between performance and resource cost, making it practical for low-resource applications.

Table 1: Comparison of model efficiency.

Methods	Params. (MB)	Memory (GB)	Time (s)
AdaAttN	13.63	10.80	0.066
CAST	10.52	10.01	0.056
AesPA-Net	14.11	3.39	0.148
StyleID	–	12.87	5.848
AttenDistill	49.49	3.61	57.560
CLoSeR	**4.74**	**2.42**	**0.055**

4.3 MODEL ANALYSIS

Ablation Study of CLoSeR. Figure 5 illustrates the progressive effect of each component. Base models (APDrawingGAN, AttenDistill) often produce blurry features and artifacts. Refinement with vanilla VQ-GAN improves textures but still struggles with structure and style consistency. Adding positional encoding further enhances spatial fidelity, while our final CLoSeR achieves the clearest geometry, reduced artifacts, and more natural textures.

Catastrophic Forgetting Evaluation of Continual Learning. We assess continual learning by incrementally adding new tasks on both natural scene and portrait drawing datasets. Specifically, we

APDrawingGAN	ArtFID ↓	FID ↓	KID ↓
BASE	19.57	12.03	0.0267
+ VQ-GAN	19.54	11.96	0.0171
+ VQ-GAN w/ PE	19.30	11.77	0.0170
+ CLoSeR (Ours)	**18.70**	**11.35**	**0.0073**
AttenDistill	ArtFID ↓	FID ↓	KID ↓
BASE	28.16	18.74	0.0506
+ VQ-GAN	22.32	13.87	0.0271
+ VQ-GAN w/ PE	22.24	13.74	0.0285
+ CLoSeR (Ours)	**21.95**	**13.65**	**0.0255**

Table 2: Ablation study of CLoSeR.

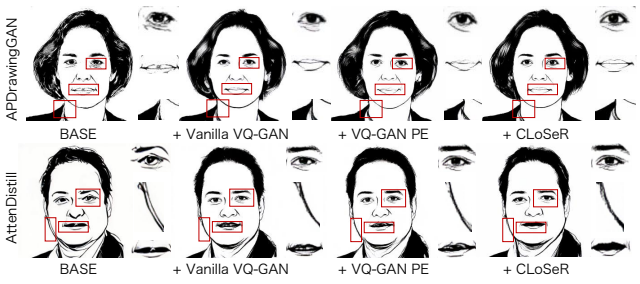

Figure 5: Qualitative results of the ablation study.

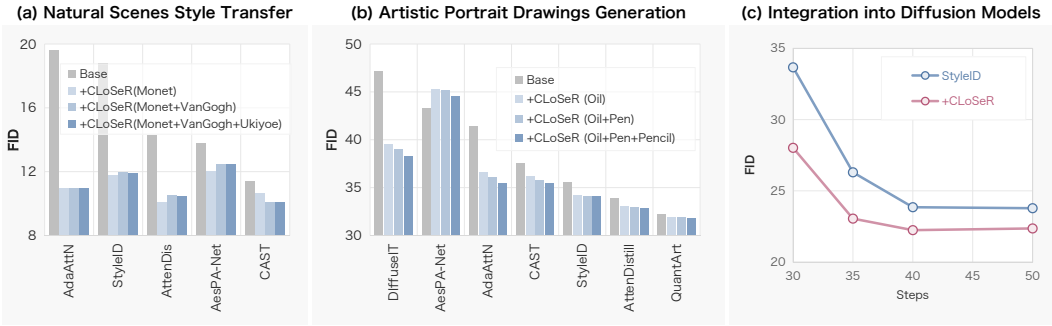

Figure 6: Catastrophic forgetting evaluation and integration into diffusion models. (a) Natural scenes style transfer with Monet as the target domain. (b) Artistic portrait drawings generation using MetFace (Oil), APDrawing (Pen), and FS2K (Pencil). (c) Integration into StyleID (Chung et al., 2024) under varying sampling steps, where CLoSeR consistently reduces FID compared to the baseline.

Table 3: Validation of *Positional Encoding* (PE) with NME (↓).

Method	AdaAttN	AesPA-Net	CAST	DiffuseIT	StyleID	AttnDistill	Average
+VQ-GAN	0.0357	0.0328	0.0348	0.0393	**0.0338**	0.0275	0.0340
+VQ-GAN w/ PE	**0.0348**	**0.0314**	**0.0325**	**0.0377**	0.0341	**0.0274**	**0.0330**

adopt MetFace (Karras et al., 2020) as the style domain for faces (denoted as *Oil*), APDrawing (Yi et al., 2019) for pen drawings (*Pen*), and FS2K (Fan et al., 2022) for pencil sketches (*Pencil*), and Monet is used for natural scenes. As shown in Figure 6, the refined models are evaluated on outputs from various base generators. The results demonstrate that performance on earlier styles remains largely stable even after introducing multiple new domains. These findings confirm that CLoSeR effectively mitigates catastrophic forgetting, retaining prior knowledge while adapting to new styles.

Validation of Positional Encoding. To evaluate the role of positional encoding (PE) in geometric consistency, we adopt YOLOv5-face (Qi et al., 2022) as the evaluation backbone and test on stylized results from the MetFace dataset (Karras et al., 2020). We report *Normalized Mean Error* (NME) as the main metric. As shown in Table 3, PE consistently reduces NME across models, confirming its benefit in preserving geometric structure. Results with *Percentage of Correct Keypoints* (PCK) under different thresholds are provided in the appendix A.

Integration into Diffusion Models. We integrate CLoSeR into the StyleID (Chung et al., 2024) diffusion framework under varying sampling steps. As shown in Figure 6(c), CLoSeR consistently reduces FID relative to the baseline, with improvements persisting across all iterations. This indicates that CLoSeR enhances domain alignment and stabilizes generation quality, even under fewer sampling steps. Additional qualitative results are provided in the appendix A.

5 Conclusions

We introduce CLoSeR, a lightweight test-time refinement framework that improves style fidelity and geometric consistency in artistic style transfer. Through LoRA-based continual adaptation, codebook expansion, and positional encoding, CLoSeR delivers parameter-efficient refinement while preserving prior knowledge across multiple domains. Extensive experiments demonstrate consistent improvements over GAN-, attention-, and diffusion-based baselines, with strong robustness against catastrophic forgetting. Future directions include extending CLoSeR to few-shot adaptation and cross-modal applications. Additionally, our current approach focuses primarily on spatial consistency and may underexplore finer temporal or semantic dynamics, particularly in video or multimodal tasks, which could be addressed in future work.

6 Ethics Statement

This work does not involve human subjects, personally identifiable information, or sensitive data. All datasets used (e.g., MetFace, FS2K, APDrawing, Monet, VanGogh, Ukiyo-e) are publicly available and widely adopted in the literature. Our research focuses purely on artistic style transfer and does not raise foreseeable ethical or societal concerns such as bias, fairness, or privacy.

7 Reproducibility Statement

We have made every effort to ensure reproducibility. All model architectures, training strategies, and evaluation metrics (FID, KID, ArtFID, NME, PCK) are described in detail in the main paper and appendix. Additional implementation details, hyperparameters, and evaluation protocols are provided in the appendix A. We will release the source code upon publication to facilitate full reproducibility of our results.

References

Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Lucas Caccia, Oleksiy Ostapenko, Massimo Trentin, Emiliano Calabrese, Eugene Brevdo, and Alexandre Lacoste. Online learned continual compression with adaptive quantization modules. In *ICLR*, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8795–8805, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. Facial-sketch synthesis: A new challenge. *Machine Intelligence Research*, 19(4):257–287, 2022.

Farzad Farhadzadeh, Debasmit Das, Shubhankar Borse, and Fatih Porikli. Zero-shot adaptation of parameter-efficient fine-tuning in diffusion models. *arXiv preprint arXiv:2506.04244*, 2025.

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*, pp. 2414–2423, 2016.

Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3985–3993, 2017.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Bin He, Feng Gao, Daiqian Ma, Boxin Shi, and Ling-Yu Duan. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1172–1180, 2018.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30534–30544, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22758–22767, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The gan is dead; long live the gan! a modern gan baseline. *Advances in Neural Information Processing Systems*, 37: 44177–44215, 2024.

Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, and Hanspeter Pfister. Quantart: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5947–5956, 2023.

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*, pp. 1501–1510, 2017.

Myeongho Jeon, Hyoje Lee, Yedarm Seong, and Myungjoo Kang. Learning without prejudices: Continual unbiased learning via benign and malignant forgetting. In *The Eleventh International Conference on Learning Representations*, 2023.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11): 3365–3385, 2019.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5549–5558, 2020.

Timothée Lesort, Alexander Gepperth, Andrei Stoian, and David Filliat. Generative models from the perspective of continual learning. *arXiv preprint arXiv:1812.09111*, 2019.

Siyuan Li, Luyuan Zhang, Zedong Wang, Juanxi Tian, Cheng Tan, Zicheng Liu, Chang Yu, Qingsong Xie, Haonan Lu, Haoqian Wang, et al. Mergevq: A unified framework for visual generation and representation with disentangled token merging and quantization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19713–19723, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6649–6658, 2021.

Dae Young Park and Kyoung Mu Lee. Arbitrary style transfer with style-attentional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5880–5888, 2019.

Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. Yolo5face: Why reinventing a face detector. In *European Conference on Computer Vision*, pp. 228–244. Springer, 2022.

Anurag Roy, Vinay K Verma, Sravan Voonna, Kripabandhu Ghosh, Saptarshi Ghosh, and Abir Das. Exemplar-free continual transformer with convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5897–5907, 2023.

Ahmed Selim, Mohamed A Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Trans. Graph.*, 35(4):129–1, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pp. 560–576. Springer, 2022.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2630–2640, 2025.

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10743–10752, 2019.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *ICLR*, 2018.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

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*, pp. 586–595, 2018.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–8, 2022.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10146–10156, 2023.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13208–13217, 2020.

Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18270–18280, 2025.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25613–25622, 2025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. *Advances in Neural Information Processing Systems*, 37:12612–12635, 2024.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

In this section, we will provide a comprehensive overview of our experimental setup, detailing all aspects of the implementation to ensure transparency and reproducibility.

A.1.1 CONTINUAL LEARNING BASED ON LORA

To enable scalable and memory-efficient continual learning in multi-style domains, we introduce Low-Rank Adaptation (LoRA) (Hu et al., 2022) into the VQ-GAN (Esser et al., 2021) framework. This is achieved by injecting LoRA modules into specific convolutional layers (conv1, conv2) of both the encoder and decoder. Each LoRA module performs a low-rank decomposition of the convolutional kernel updates, significantly reducing the number of trainable parameters during test-time refinement.

Training Phase. During training, the LoRA modules are initialized with a low-rank pair of trainable matrices $A \in \mathcal{R}^{r \times d_{in}}$, with a scaling factor α/r. These modules are only activated for target style-specific adapters, each associated with a unique *style_id*. We implement a style-wise code isolation strategy by naming and registering all LoRA parameters under their respective *style_id*.In the continual learning scenario, only LoRA parameters and newly appended codebook embeddings are optimized, while all other original weights in the encoder, decoder, and quantizer are frozen. To accommodate novel style tokens without disrupting previously learned knowledge, we expand the codebook by appending new embeddings, and apply selective gradient masking to freeze the original indices. This ensures forward compatibility and avoids catastrophic forgetting.

Inference Phase. At test time, the framework dynamically selects and activates the appropriate LoRA module based on the input *style_id*. The inference pipeline searches for the latest LoRA checkpoint corresponding to the style domain, loads its parameters, and activates only the relevant LoRA paths while disabling others. This design ensures geometric consistency and stylistic specificity across diverse domains under a single model instance. Overall, the proposed LoRA-based continual adaptation mechanism provides a lightweight, modular, and effective solution to multi-style artistic synthesis, enabling test-time refinement with up to 94% fewer trainable parameters.

A.1.2 DATASET DETAILS AND TRAINING CONFIGURATION

In this work, we employ three distinct datasets to train specialized codebooks for different artistic styles within our CLoSeR framework. Each dataset is carefully selected to represent a unique visual domain, enabling the learning of style-specific discrete representations.

Artistic Portrait Generation. We choose a pre-trained model (vqgan_metfaces_f16_1024.ckpt) from QuantArt (Huang et al., 2023) to finetune VQ-GAN (Esser et al., 2021) to achieve style-specific reconstruction. **MetFace** (Karras et al., 2020) is used to train the general facial appearance codebook. This dataset contains a total of 1336 face images, partitioned into 1,200 training samples and 136 test samples. **APDrawing** (Yi et al., 2019) datasets consist of pen-drawing portrait drawings. The dataset is divided into 420 training images and 70 test images. We initialize the VQ-GAN from the model pre-trained on the MetFace dataset (covering photorealistic facial appearances) and introduce Low-Rank Adaptation (LoRA) modules into the 'conv1' and 'conv2' of encoder and decoder. **FS2K** (Fan et al., 2022) includes 2,104 face sketches across three distinct artistic styles. We initialize the VQ-GAN from the model pre-trained on the APDrawing. We combine all three styles into a single training set to encourage the model to learn a more generalized sketch representation. The training split contains 2,004 images, with the remaining 100 reserved for testing.

Scene Oil Paingting. To further evaluate the generalization capability of our continual learning framework, we extend our experiments to three additional classical art styles:Monet,Van Gogh, and Ukiyo-e, all datasets are from WikiArt, follow the work from (Zhu et al., 2017). And we choose a pre-trained model (vqgan_wikiart_f16_1024.ckpt) from QuantArt (Huang et al., 2023) to finetune VQ-GAN to achieve style-specific scene oil painting reconstruction. **Monet** dataset comprises 1,072 training and 121 test images, capturing soft brushwork and natural light effects. **Van Gogh** dataset

Table 4: Impact of CLoSeR on *Natural Scenes Style Transfer.*

Method	Monet			Vangogh			Ukiyo-e		
	ArtFID ↓	FID ↓	KID ↓	ArtFID ↓	FID ↓	KID ↓	ArtFID ↓	FID ↓	KID ↓
AdaAttN (CVPR'21)	32.34	19.63	0.0602	23.60	13.55	0.0397	30.14	18.03	0.1380
+ VQGAN	23.64 ↓26.9%	13.86 ↓29.4%	0.0681 ↓13.1%	22.58 ↓4.3%	12.88 ↓4.9%	0.0259 ↓34.8%	31.03 ↑3.0%	18.36 ↑1.8%	0.1144 ↓17.1%
+ CLoSeR (ours)	19.35 ↓40.2%	10.99 ↓44.0%	0.0467 ↓22.4%	18.46 ↓21.8%	10.35 ↓23.6%	0.0277 ↓30.2%	26.94 ↓10.6%	15.77 ↓12.5%	0.1193 ↓13.6%
CAST (SIGGRAPH'22)	19.53	11.43	0.0159	24.90	14.81	0.0410	29.06	17.75	0.0888
+ VQGAN	21.04 ↑7.7%	12.11 ↑5.9%	0.0240 ↑50.9%	24.16 ↓3.0%	14.00 ↓5.5%	0.0297 ↓27.6%	34.51 ↑18.8%	20.63 ↑16.2%	0.0710 ↓20.0%
+ CLoSeR (ours)	18.87 ↓3.3%	10.64 ↓6.9%	0.0155 ↓2.5%	19.11 ↓23.3%	10.89 ↓26.5%	0.0250 ↓39.0%	26.46 ↓8.9%	15.47 ↓12.8%	0.0814 ↓8.3%
AesPA-Net (ICCV'23)	23.58	13.82	0.0808	22.7	12.99	0.0602	29.48	17.15	0.1673
+ VQGAN	22.66 ↓3.9%	12.98 ↓6.1%	0.0631 ↓21.9%	23.75 ↑4.6%	13.64 ↓5.0%	0.0309 ↓48.7%	30.77 ↑4.4%	17.85 ↑4.1%	0.1479 ↓11.6%
+ CLoSeR (ours)	21.31 ↓9.6%	12.02 ↓13.0%	0.0639 ↓20.9%	19.17 ↓15.6%	10.81 ↓16.8%	0.0330 ↓45.2%	28.57 ↓3.1%	16.45 ↓4.1%	0.1458 ↓12.9%
StyleID (CVPR'24)	23.81	15.07	0.0370	30.63	18.78	0.0532	32.39	20.04	0.1733
+ VQGAN (ours)	22.94 ↓3.7%	14.07 ↓6.6%	0.0308 ↓16.8%	28.74 ↓6.2%	17.09 ↓9.0%	0.0400 ↓24.8%	32.95 ↑1.7%	20.03 ↓0.0%	0.1641 ↓5.3%
+ CLoSeR (ours)	19.85 ↓16.6%	11.82 ↓21.6%	0.0184 ↓50.1%	22.11 ↓27.8%	12.93 ↓31.2%	0.0353 ↓33.6%	27.13 ↓16.2%	16.21 ↓19.1%	0.1394 ↓19.6%
AtteneDist (CVPR'25)	21.22	14.29	0.0489	22.13	14.24	0.0431	26.31	16.81	0.1718
+ VQGAN	22.91 ↑8.0%	14.91 ↑4.3%	0.0320 ↓34.6%	24.82 ↑12.2%	15.54 ↑9.1%	0.0353 ↓18.1%	25.25 ↓4.0%	15.7 ↓6.6%	0.1388 ↓19.2%
+ CLoSeR (ours)	16.35 ↓23.0%	10.1 ↓29.3%	0.0216 ↓55.8%	17.9 ↓19.1%	10.82 ↓24.0%	0.0245 ↓43.2%	24.99 ↓5.0%	14.34 ↓14.7%	0.1259 ↓26.7%

includes 700 training and 100 test images, emphasizing expressive and vivid color contrasts. **Ukiyo-e** dataset contains 562 training and 263 test images, featuring flat color regions, strong outlines, and stylized compositions typical of traditional Japanese art.

All datasets are preprocessed to a consistent resolution of 256×256 with center cropping and normalized to the range $[-1, 1]$. During training, we preserve the LoRA parameters together with the corresponding discriminator for each style, enabling modular switching at inference time. This plug-and-play design supports flexible and memory-efficient multi-style generation within a single unified architecture.

A.1.3 More Metrics Details of the Tasks

We evaluate our model by ArtFID (Wright & Ommer, 2022), FID (Heusel et al., 2017), and KID (Bińkowski et al., 2018). The specific numerical metrics of Scene Oil Paintings are shown in the Table4, Face Portrait Drawings are shown in the Table 5. From the quantitative metrics, we can see that our algorithm has shown excellent performance under each base method.

A.2 More Results of CLoSeR

A.2.1 Ablation Study

Catastrophic Forgetting Evaluation of Continual Learning. Due to space constraints, we report the detailed quantitative results of continual learning in the appendix. As shown in Table 7 and Table 8, the refined models are evaluated on outputs from various base generators. The results show that performance on earlier styles remains largely stable even after introducing multiple new domains. These findings confirm that **CLoSeR** effectively mitigates catastrophic forgetting, retaining prior knowledge while adapting to new styles.

Validation of Positional Encoding. To evaluate the role of positional encoding (PE) in geometric consistency, we conduct landmark detection on stylized outputs with CelebAMask-HQ (Lee et al., 2020) as the content domain and MetFace (Karras et al., 2020) as the style domain. For each algorithm, we generate 80 stylized results, where both the vanilla VQ-GAN and VQ-GAN w/PE are trained on MetFace for 48 epochs. The qualitative comparisons of different detection algorithms are provided in Figure 8. Due to space constraints, additional *Percentage of Correct Keypoints* (PCK) results under 5%, 7%, and 10% thresholds are reported in the Appendix, as shown in Table 6.

Integration into Diffusion Models. We integrate CLoSeR into the StyleID (Chung et al., 2024) diffusion framework and evaluate under different sampling steps. As shown in Figure 7, we assess refinement on MetFace-based generations at 30, 35, 40, and 50 steps. The qualitative results clearly

Table 5: Impact of CLoSeR on *Artistic Portrait Generation.*

Method	Oil Painting			Pen Drawing			Sketch		
	ArtFID ↓	FID ↓	KID ↓	ArtFID ↓	FID ↓	KID ↓	ArtFID ↓	FID ↓	KID ↓
APDrawingGAN (CVPR'19)	–	–	–	19.56	12.02	0.0267	–	–	–
+ VQGAN	–	–	–	19.58 ↑0.1%	11.96 ↓0.5%	0.0171 ↓40.0%	–	–	–
+ CLoSeR (ours)	–	–	–	19.30 ↓1.3%	11.77 ↓2.1%	0.0073 ↓71.8%	–	–	–
AdaAttN (CVPR'21)	41.39	28.14	0.1281	37.34	24.21	0.1293	44.81	26.74	0.0905
+ VQGAN	35.26 ↓17.4%	23.90 ↓17.7%	0.1071 ↓19.6%	29.57 ↓20.8%	18.68 ↓22.8%	0.0771 ↓40.4%	43.87 ↓2.1%	25.84 ↓3.4%	0.0847 ↓6.4%
+ CLoSeR (ours)	36.62 ↓11.5%	24.60 ↓12.9%	0.1089 ↓14.9%	29.11 ↓22.1%	18.35 ↓24.2%	0.0783 ↓39.5%	41.04 ↓8.4%	23.69 ↓11.4%	0.1072 ↑18.4%
CAST (SIGGRAPH'22)	37.58	26.13	0.1002	22.35	14.37	0.0784	43.07	25.41	0.0692
+ VQGAN	37.22 ↓1.0%	25.37 ↓2.9%	0.1065 ↑6.3%	23.70 ↑6.0%	14.80 ↑3.0%	0.0417 ↓46.8%	40.61 ↓5.7%	23.56 ↓7.3%	0.0684 ↓1.2%
+ CLoSeR (ours)	36.11 ↓3.9%	24.50 ↓6.2%	0.1057 ↑5.5%	23.55 ↑5.4%	14.68 ↑2.2%	0.0417 ↓46.8%	40.28 ↓6.5%	22.86 ↓10.0%	0.0841 ↑21.5%
AesPA-Net (ICCV'23)	43.28	30.42	0.1313	41.97	28.53	0.1258	41.52	25.11	0.0955
+ VQGAN	44.25 ↑2.2%	30.48 ↑0.2%	0.1450 ↑10.4%	26.59 ↓36.6%	16.74 ↓41.3%	0.0464 ↓63.1%	40.50 ↓2.5%	23.78 ↓5.3%	0.0629 ↓34.1%
+ CLoSeR (ours)	45.05 ↑4.1%	31.07 ↑2.1%	0.1314 ↓0.0%	27.35 ↓34.9%	17.23 ↓39.6%	0.0497 ↓60.5%	38.49 ↓7.3%	21.97 ↓12.5%	0.0727 ↓23.9%
DiffuseIT (ICLR'23)	47.13	32.27	0.1598	36.19	23.06	0.0826	58.04	35.86	0.1858
+ VQGAN	48.91 ↑3.8%	32.70 ↑1.3%	0.1110 ↓30.5%	30.57 ↓18.4%	18.27 ↓26.2%	0.0646 ↓27.9%	46.33 ↓20.2%	26.91 ↓25.0%	0.0739 ↓60.2%
+ CLoSeR (ours)	39.38 ↓16.4%	26.46 ↓17.7%	0.0913 ↓42.9%	32.06 ↓11.4%	19.24 ↓16.6%	0.0557 ↓32.6%	41.86 ↓27.9%	23.70 ↓33.9%	0.0807 ↓56.6%
InST (CVPR'23)	57.89	38.57	0.2226	35.04	26.13	0.0818	–	–	–
+ VQGAN	46.46 ↓19.7%	32.11 ↓16.7%	0.0957 ↓57.0%	31.24 ↓10.8%	20.01 ↓23.4%	0.0783 ↓4.3%	–	–	–
+ CLoSeR (ours)	47.23 ↓18.4%	26.46 ↓31.4%	0.0913 ↓58.7%	29.72 ↓15.2%	19.08 ↓27.0%	0.0779 ↓4.8%	–	–	–
StyleID (CVPR'24)	35.60	23.78	0.1198	26.58	17.44	0.0235	44.67	26.97	0.1546
+ VQGAN	35.96 ↑1.0%	23.31 ↓2.0%	0.1080 ↓9.8%	22.05 ↓17.0%	13.71 ↓21.4%	0.0235 ↓0.0%	41.01 ↓8.2%	23.96 ↓11.2%	0.0643 ↓58.4%
+ CLoSeR (ours)	34.19 ↓4.0%	22.36 ↓6.0%	0.0966 ↓19.4%	24.57 ↓7.6%	15.43 ↓11.5%	0.0159 ↓32.3%	36.14 ↓19.1%	20.87 ↓22.6%	0.0725 ↓53.1%
AttenDist (CVPR'25)	33.95	26.13	0.1349	28.16	18.74	0.0506	43.50	30.98	0.1501
+ VQGAN	34.26 ↑0.9%	24.66 ↓5.6%	0.1162 ↓13.9%	22.32 ↓20.7%	13.87 ↓26.0%	0.0271 ↓46.4%	46.47 ↑6.8%	28.43 ↓8.2%	0.0798 ↓46.8%
+ CLoSeR (ours)	33.18 ↓2.3%	23.84 ↓8.8%	0.1046 ↓22.5%	21.95 ↓22.1%	13.65 ↓27.2%	0.0255 ↓49.6%	39.59 ↓9.0%	23.89 ↓22.9%	0.0843 ↓43.8%

Table 6: Validation of Positional Encoding (PE) with PCK (↑).

Metrics	AdaAttN		AesPA-Net		AttnDistill		CAST		DiffuseIT	
	+VQ-GAN	+VQ-GAN w/PE	+VQ-GAN	+VQ-GAN w/PE	+VQ-GAN	+VQ-GAN w/PE	+VQ-GAN	+VQ-GAN w/PE	+VQ-GAN	+VQ-GAN w/PE
PCK@5% ↑	0.789	**0.802**	0.841	**0.858**	0.896	**0.901**	0.806	**0.842**	0.741	**0.764**
PCK@7% ↑	0.921	**0.930**	0.942	**0.952**	0.973	**0.973**	0.927	**0.941**	0.900	**0.909**
PCK@10% ↑	0.980	**0.987**	0.984	**0.988**	0.995	**0.995**	0.979	**0.986**	0.974	**0.980**

demonstrate that CLoSeR produces sharper and more stylistically faithful portraits across different iteration counts.

A.2.2 Experiments Results

In this section, we provide a concise yet comprehensive overview of the additional experimental results validating our proposed Continual Learning for Style Refinement (CLoSeR) framework. We compare CLoSeR with state-of-the-art methods, emphasizing its effectiveness in generating high-quality, style-consistent drawings.

Facial Portrait Results. Figure 9 and Figure 10 illustrate the generated artistic styles for facial portraits using various generative methods. CLoSeR demonstrates superior performance in both oil painting and pen drawing styles. For oil paintings, CLoSeR achieves visually appealing results that closely resemble the target style while preserving the identity and structural details of the input faces. Compared to the SOTA methods, CLoSeR avoids overly smoothed or distorted outputs, capturing complex brush strokes and color blending effectively. In pen drawings, CLoSeR produces clear lines and consistent textures, accurately representing the input faces with sharp, well-defined lines.

Natural Scene Results Figure 11 showcases the generated artistic styles for natural scenes based on different artist and generative methods. CLoSeR excels in generating high-quality artistic representations of natural scenes, such as Monet, Van Gogh, and Ukiyo-e styles. For Monet's impressionistic style, CLoSeR captures soft brushwork and natural light effects, producing visually pleasing results. In Van Gogh's post-impressionistic style, CLoSeR effectively reproduces expressive, swirling strokes and vivid color contrasts. For Ukiyo-e, CLoSeR generates flat color regions, strong outlines, and stylized compositions typical of traditional Japanese art. Compared to other SOTA methods, CLoSeR maintains better style consistency and visual fidelity. The continual learn-

Table 7: Catastrophic Forgetting Evaluation on artistic portrait (MetFace, APDrawing, FS2K).

Methods	ArtFID ↓	FID ↓	KID ↓
AdaAttN	41.40	28.15	0.1281
+CLoSeR (Oil)	36.61	24.60	0.1089
+CLoSeR (Oil+Pen)	36.11	24.24	**0.1067**
+CLoSeR (Oil+Pen+Pencil)	**35.44**	**23.69**	0.1106
CAST	37.58	26.13	**0.1002**
+CLoSeR (Oil)	36.22	24.52	0.1055
+CLoSeR (Oil+Pen)	35.79	24.21	0.1045
+CLoSeR (Oil+Pen+Pencil)	**35.53**	**24.03**	0.1093
AesPA-Net	**43.28**	**30.42**	0.1313
+CLoSeR (Oil)	45.34	31.25	**0.1304**
+CLoSeR (Oil+Pen)	45.17	31.07	0.1321
+CLoSeR (Oil+Pen+Pencil)	44.61	30.60	0.1375
QuantArt	32.29	24.43	0.1017
+CLoSeR (Oil)	31.98	23.59	**0.0917**
+CLoSeR (Oil+Pen)	31.89	23.47	0.0929
+CLoSeR (Oil+Pen+Pencil)	**31.83**	**23.41**	0.0958
DiffuseIT	47.13	32.27	0.1598
+CLoSeR (Oil)	39.59	26.60	**0.0913**
+CLoSeR (Oil+Pen)	39.07	26.24	0.0919
+CLoSeR (Oil+Pen+Pencil)	**38.30**	**25.65**	0.0929
StyleID	35.60	23.78	0.1198
+CLoSeR (Oil)	34.19	22.36	**0.0966**
+CLoSeR (Oil+Pen)	**34.14**	**22.31**	0.0971
+CLoSeR (Oil+Pen+Pencil)	**34.14**	**22.31**	0.0971
AttenDistill	33.95	26.13	0.1349
+CLoSeR (Oil)	33.04	23.71	**0.1041**
+CLoSeR (Oil+Pen)	33.00	23.61	0.1043
+CLoSeR (Oil+Pen+Pencil)	**32.87**	**23.50**	0.1089

ing approach ensures that CLoSeR refines its understanding of each artistic style, leading to more accurate and consistent results.

A.3 Limitations

While CLoSeR significantly improves geometric consistency and style fidelity in refined outputs, it still inherits certain limitations from the underlying VQ-GAN framework. Specifically, when the input synthesis is severely distorted or lacks semantic structure, the refinement effect becomes limited. Additionally, our current approach focuses primarily on spatial consistency, and may underexplore finer temporal or semantic dynamics in video or cross-modal tasks, which could be addressed in future work.

A.4 Usage of LLM

We employed a large language model (LLM) as an auxiliary tool during the manuscript preparation process. Specifically, the LLM was used to polish the writing, check spelling and grammar errors, and improve the overall clarity and readability of the text. Importantly, the LLM was not involved in designing the methodology, conducting experiments, or analyzing results; all technical contributions, experimental designs, and conclusions were developed solely by the authors. The use of the LLM was limited to language refinement, helping to ensure that the presentation of our work is logically coherent and accessible to a broader research audience.

Table 8: Catastrophic Forgetting Evaluation on natural scenes (Monet, Van Gogh, Ukiyo-e).

Methods	ArtFID ↓	FID ↓	KID ↓
AdaAttN	32.34	19.63	0.0602
+CLoSeR (+Monet)	19.35	10.99	**0.0467**
+CLoSeR (+Monet+VanGogh)	**19.27**	**10.95**	0.0513
+CLoSeR (+Monet+VanGogh+Ukiyo-e)	**19.27**	**10.95**	0.0493
CAST	19.53	11.43	0.0159
+CLoSeR (+Monet)	18.87	10.64	0.0155
+CLoSeR (+Monet+VanGogh)	**18.03**	**10.11**	**0.0117**
+CLoSeR (+Monet+VanGogh+Ukiyo-e)	18.05	10.13	0.0118
AesPA-Net	23.58	13.82	0.0808
+CLoSeR (+Monet)	**21.31**	**12.02**	**0.0639**
+CLoSeR (+Monet+VanGogh)	22.05	12.48	0.0677
+CLoSeR (+Monet+VanGogh+Ukiyo-e)	22.05	12.48	0.0677
StyleID	30.63	18.78	0.0370
+CLoSeR (+Monet)	**19.85**	**11.82**	0.0184
+CLoSeR (+Monet+VanGogh)	20.07	11.98	0.0165
+CLoSeR (+Monet+VanGogh+Ukiyo-e)	19.96	11.91	**0.0159**
AttenDistill	21.22	14.29	0.0489
+CLoSeR (+Monet)	**16.35**	**10.10**	**0.0216**
+CLoSeR (+Monet+VanGogh)	17.03	10.54	0.0267
+CLoSeR (+Monet+VanGogh+Ukiyo-e)	16.97	10.50	0.0257

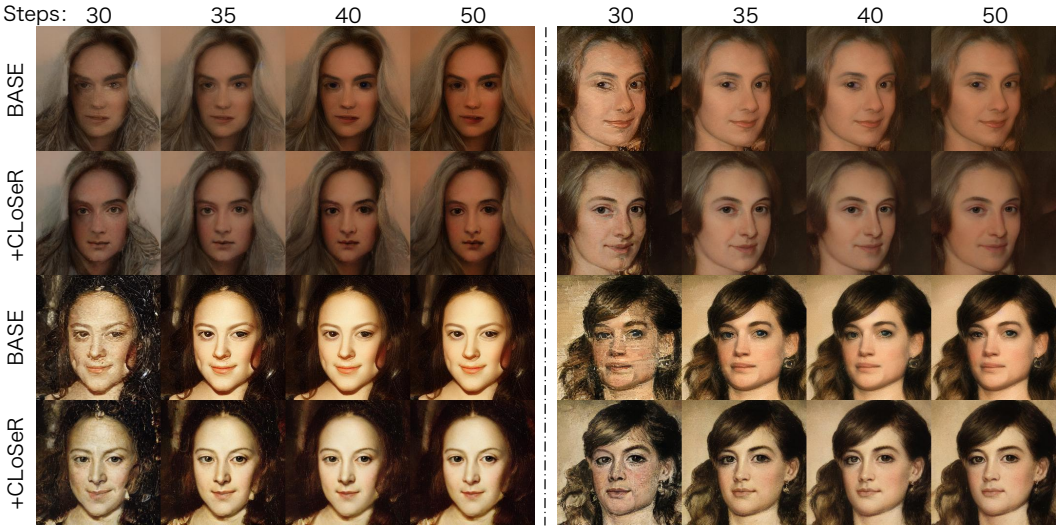

Figure 7: Qualitative evaluation of StyleID refinement with CLoSeR on the MetFace dataset under different sampling steps (30, 35, 40, 50). Compared to the baseline (BASE), CLoSeR produces sharper, more consistent, and stylistically faithful results across all iterations.Please zoom in for details.

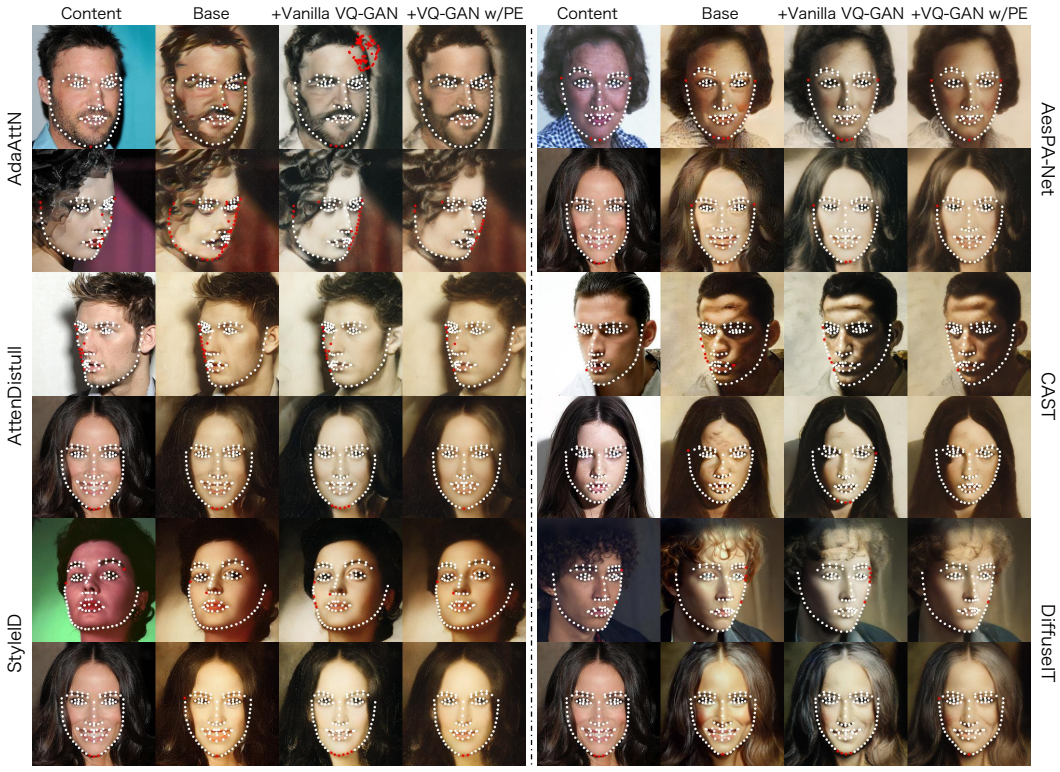

Figure 8: Comparison of facial landmark detection across different generative methods on oil painting portraits. Each column shows the detected landmarks on stylized outputs, highlighting the impact of VQ-GAN and positional encoding (PE) on geometric consistency. Please zoom in for details.

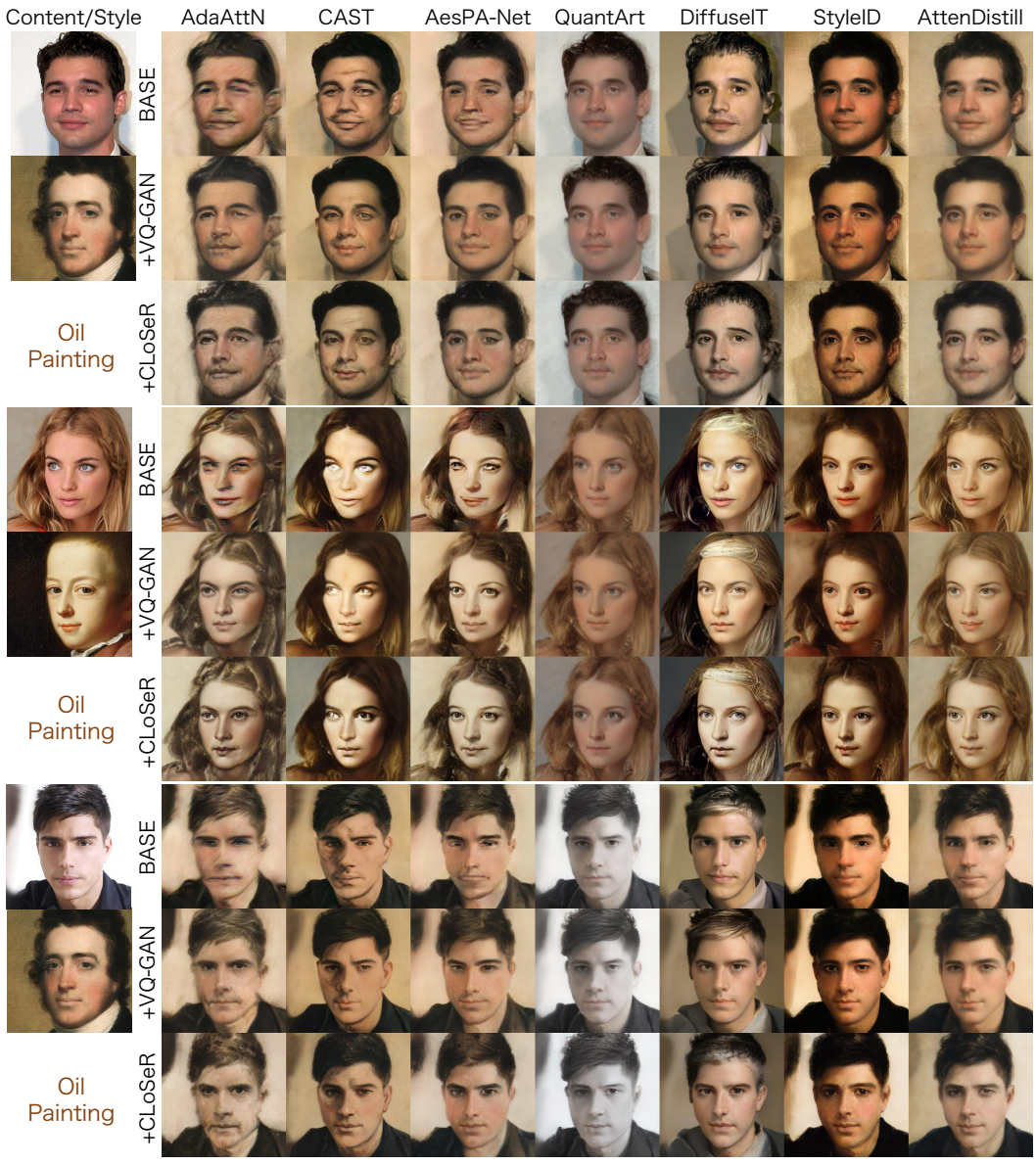

Figure 9: Generated results of artistic styles for oil facial portraits based on different generative methods.Please zoom in for details.

Figure 10: Generated results of artistic styles for pen-drawing facial portraits based on different generative methods.Please zoom in for details.

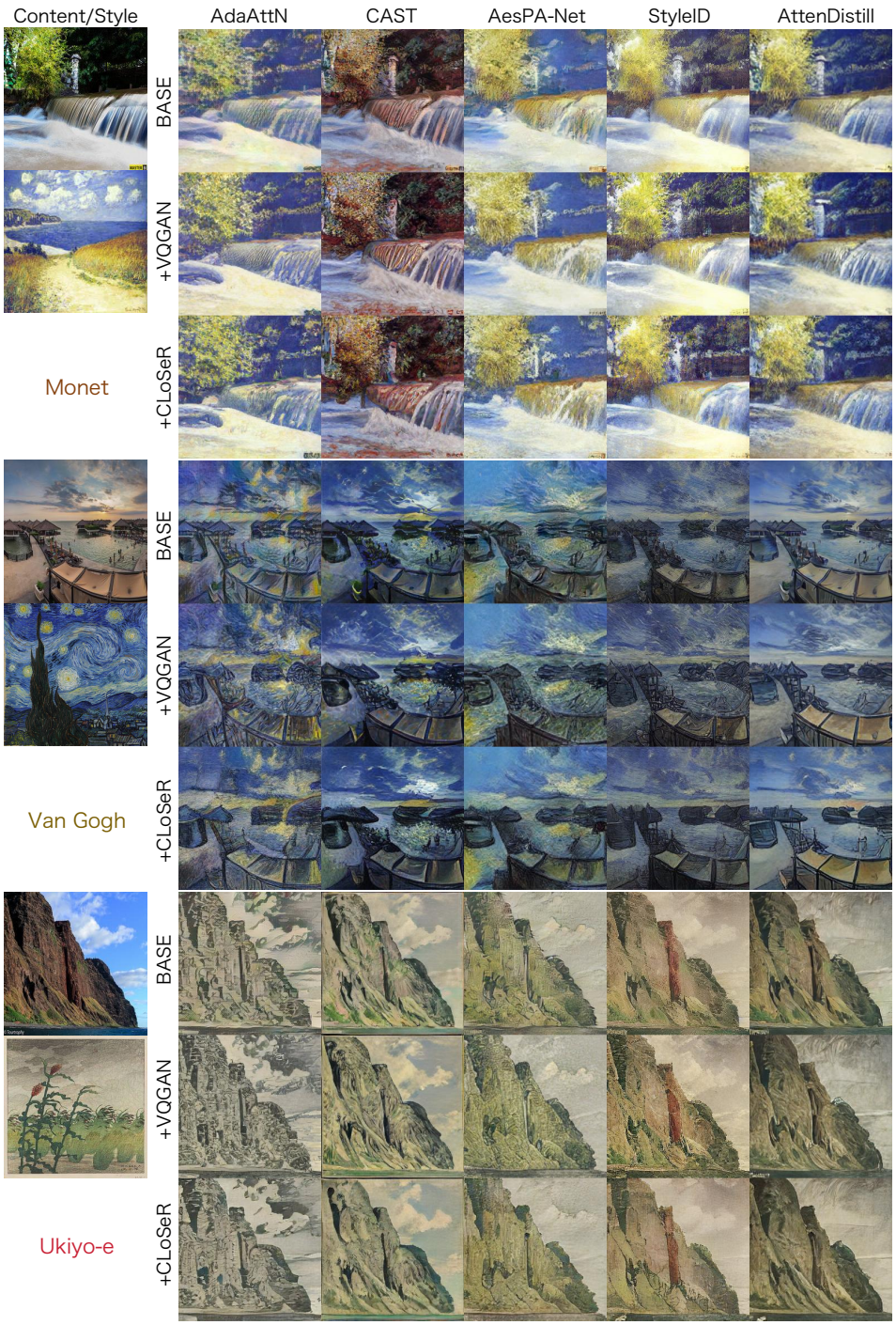

Figure 11: Generated results of artistic styles for natural scenes based on different artist and generative methods.Please zoom in for details.