Style-CCL: Content-Preserving Style Transfer via Curriculum Continual Learning

Shiwen Zhang [1] **Haoyuan Wang** [1] **Xianghao Zang** [1] **Haibin Huang** [1] **Chi Zhang** [1] **Xuelong Li** [1]

Abstract

Content-Preserving Style transfer, given content and style references, remains challenging for Diffusion Transformers (DiTs) due to entangled content and style features. With a reverse triplet synthesis pipeline to build a million-scale training set and a dual-branch Style-Content DiT (SC-DiT) that decouples style and content via separate ROPE embeddings and causal masking, we observe that such a one-stage training paradigm on mixed style categories causes semantic styles to dominate, hindering texture style learning, and harming content preservation. To address these issues, we propose Style-CCL, a Multi-Stage Curriculum Continual Learning framework that trains SC-DiT from semantic (easy) to texture (hard) styles, and from clean to synthetic data, with Random Memory Rehearsal across stages to avoid catastrophic forgetting. Extensive experiments demonstrate that our Style-CCL achieves state-of-the-art performance in three core metrics: style similarity, content consistency, and aesthetic quality.

1. Introduction

Image customization and editing with multiple references (Wu et al., 2025; Labs, 2025) with diffusion transformers (Labs, 2024; Peebles and Xie, 2023; Esser et al., 2024) has achieved great progress. However, there is still significant scope for improving the effects of content-preserving style transfer (Gatys et al., 2016). Current style transfer models (Wang et al., 2023a; Zhang et al., 2025a) suffer from the leakage/invasion issue (subject/background/facial identities from style reference are over-transferred, polluting the characteristics of content reference) and struggle to keep multiple characteristics in complex content reference. In addition, the generated images of style transfer models often exhibit low aesthetic merit.

[1]Institute of Artificial Intelligence (TeleAI), China Telecom.

This work was finished in September 2025. We did not release the preprint until today.

In order to tackle aforementioned problems, we introduce decoupled style branch and content branch to DiT, termed SC-DiT , by utilizing VAE encoder (Kingma and Welling, 2014) to extract visual features for both branches and apply style RoPE and content RoPE (Su et al., 2024) to each branch to distinguish style and content with causal attention. We collected and synthesized [style reference, content reference, target] triplets to train SC-DiT on FLUX-dev (Labs, 2024).

However, during training SC-DiT with various styles in one stage, we observe some surprising phenomena: First, semantics-related style transformations (e.g. 2D/3D cartoon, simple line drawings, vector design, etc) and texture-related style transformations (e.g. oilpainting, dense line drawings, texture materials, etc) contradict each other. In particular, it turns out that semantics-related style transformations hinders the learning of texture-related style transformations, even very long training time could not alleviate the issue. Second, content characteristics fail to be well-preserved, as synthetic triplets compromise the integrity of clean triplets with respect to precise content preservation. Third, for those particular style categories which contain both semantic-related and texture-related style transformations in the same style reference image, our model always learns semantic-related style transformation in early iterations and texture-related style transformations in late iterations. *Observation 3 clearly demonstrates that SC-DiT has the capability of learning texture-related styles when it is trained on only one style category, while Observation 1 indicates that such a capability is weakened and interfered when semantic and texture style categories are trained together.*

In order to tackle these problems, We propose a Multi-stage Style Curriculum Continual Learning (Style-CCL) paradigm. First, we introduce a theoratical tool, Local Intrinsic Dimensionality (LID) (Tempczyk et al., 2022; Kamkari et al., 2024), to estimate the complexity of style images. With a rough ranking of LID scores by our LID Estimator, we divide the style categories into semantic-related styles and texture-related styles with an approximate boundary. Then we apply our Style Curriculum Continual Learning to gradually learn these subsets from easy to hard, from clean to noisy (Bengio et al., 2009), without catastrophic forget-

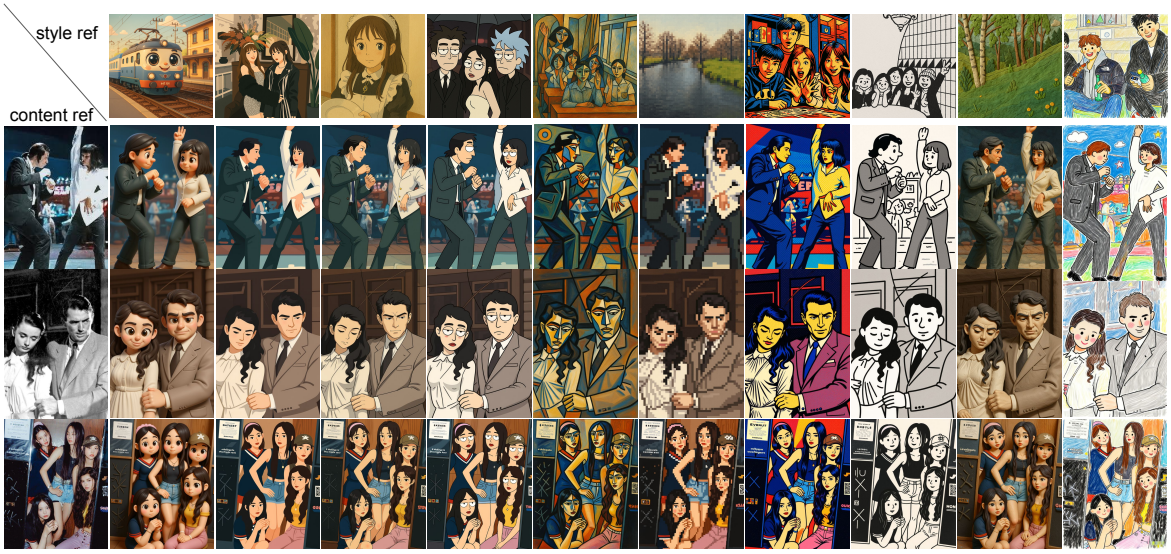


Figure 1. Style-CCL accepts style and content references for content-preserving style transfer, while maintaining high aesthetics merit.

ting. We show the capability of our Style-CCL in Figure 1.

Our main contributions are:

1. We observed that traditional one-stage training paradigm for Conditional DiT with style and content references causes contradiction between semantic-related styles and texture-related styles, where semantic-related styles hinder the learning of texture-related styles. The characteristics of content reference is also not well-preserved in such one stage training.
2. We propose a Multi-Stage Style Curriculum Continual Learning (Style-CCL) to tackle the aforementioned problems and introduce Random Memory Rehearsal to avoid catastrophic forgetting. Our model could smoothly learn thousands of style categories with Style-CCL paradigm and preserve complex characteristics in content reference without subject confusion/mixture.
3. Our Style-CCL achieves new state-of-the-art results in terms of style similarity, content preservation and aesthetics score through quantitative evaluation and user study.

2. Related Work

Zero-Shot Style Transfer with Conditional DiT SD3 (Esser et al., 2024) and FLUX (Labs, 2024) improves text-to-image task significantly by scaling up DiT (Peebles and Xie, 2023; Dosovitskiy et al., 2020; Vaswani et al., 2017) parameters, outperforming previous UNet structures (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2023). However, Such DiT models lacks disentangle properties (Zhang et al., 2020a;b; Zhang, 2022; Zhang et al., 2023; 2025a; Zhang, 2024; Zhang et al., 2026b). With these new powerful text-to-image DiT models, OminiControl (Tan et al., 2024) and EasyControl (Zhang et al., 2025b) enable Conditional Image Generation by concatenating condition image with text condition and noisy latent in self attention modules. Qwen Image Edit (Wu et al., 2025) could handle multiple reference images for subject-driven customization. However, by the time this paper was done (September, 2025), Qwen Image Edit does not support subject+style references, neither does FLUX-Kontext (Labs et al., 2025). OmniConsistency (Song et al., 2025) trained a separate content consistency branch and relies on external Style Loras (Hu et al., 2021) to conduct style transfer with content preservation. Instead, our model unifies content preservation and style transfer capability in one unified model, which is capable to handle universal style categories without the need for Style Loras trained on one specific style.

3. Style-CCL

3.1. Overview

We begin by introducing our framework for constructing [style reference, content reference, target] training triplets. We then present the architecture of SC-DiT for content-preserving style transfer conditioned on both style and content references. Next, we describe three key empirical observations that expose fundamental limitations of the conventional one-stage training paradigm. Finally, we propose a curriculum continual learning strategy to address these issues.

3.2. Triplet Training Dataset Construction

Unlike subject-driven image pair/triplet data, which naturally exists in videos or photo albums, style triplets are rare in real world. We collected [Style Ref, Content Ref, Target] image triplets from a dataset (Song et al., 2025) sampled from GPT-4O (Hurst et al., 2024) and some Loras from open-source community, and purified them with data cleaning. However, such collection is expensive and we only obtain 30 style categories. The model trained on such limited style categories generalizes poorly to unseen styles. Thus we introduce a reverse triplet synthetic framework inspired by (Wang et al., 2023b) to generate training triplets from style images in-the-wild (Li et al., 2024), where different style images are organized into noisy style clusters. The synthesis framework is shown in Figure 2, where we specifically trained an image editing model on FLUX-dev to convert stylized image into photographic images. Due to page limit, we elaborate implementation details in the appendix. For simplicity, we call the collected clean triplet dataset D_{pure} and the synthetic dataset D_{synth} in the following sections. With a full matching strategy, we have around 330k triplets in D_{pure} and 1 million triplets in D_{synth}, together containing more than 1k style clusters.

3.3. SC-DiT with Style Reference and Content Reference

We extend FLUX-dev (Labs, 2024) with a style-condition branch and a content-condition via separated Loras (Hu et al., 2021) and RoPE (Su et al., 2024), shown in Figure 3. Taking the query of Double-Stream Block in FLUX for example, formally, we denote Z_t, Z_n, Z_s, Z_c as intermediate representations of text, noise, style and content, respectively. W_{Q_n} is the query matrix of the noisy image branch from the DoubleStreamBlock, with W_{Q_t} being the text branch. We introduce A_s, B_s as Lora for style injection, A_c, B_c as Lora for content injection. They are attached to W_{Q_n}.

$$Q_t = W_{Q_t} Z_t, Q_n = W_{Q_n} Z_n \tag{1}$$

$$Q_s = W_{Q_n} Z_s + B_s A_s Z_s, Q_c = W_{Q_n} Z_c + B_c A_c Z_c \tag{2}$$

$$\mathbf{Q} = [Q_t, Q_n, Q_s, Q_c] \tag{3}$$

We have Q_t, Q_n, Q_s, Q_c concatenated to form the overall query $\mathbf{Q}$. Although queries from each branch are concatenated, Style Lora and Content Lora only operate on style feature and content feature themselves, without affecting text branch and noise branch. Similarly, we could obtain the overall key $\mathbf{K}$ and value $\mathbf{V}$. For SingleStreamBlock, the process is easier because there is only one W_Q without distinguishing text and noise.

Then we apply Causal Attention with $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ and a Causal Mask M, shown in Figure 3. With the mask M, we forbid the query from style and content to text and noise, and forbid the interaction between style and content. For M, we set the white blocks in Figure 3 to 0, black blocks in Figure 3 to $-\infty$.

$$O = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + M)\mathbf{V} \tag{4}$$

Furthermore, inspired by (Tan et al., 2024; Zhang et al., 2025b), we rescale the style reference of $H_s \times W_s$ and content reference of $H_c \times W_c$ to a fixed height H and width W with ratio:

$$s_s^h = H_s/H, s_s^w = W_s/W, s_c^h = H_c/H, s_c^w = W_c/W \tag{5}$$

we set the position encoding of content branch PE_c and style branch PE_s by:

$$PE_s[i,j] = [s_s^h \times i + \Delta, s_s^w \times j] \tag{6}$$

$$PE_c[i,j] = [s_c^h \times i, s_c^w \times j] \tag{7}$$

where we set $\Delta = H$ empirically. The designs of PE_s and PE_c are different because the content reference should spatially align with the target yet the style reference should not, thus we add an offset to the style PE.

The optimization objective is based on rectified flow-matching (Liu et al., 2022; Lipman et al., 2022; Esser et al., 2024):

$$L = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,I)} \left\| v_\theta(x_t, t, c_s, c_c, c_p) - (\epsilon - x_0) \right\|_2^2 \tag{8}$$

where x_t represents the image features at time t; c_s, c_c, and c_p are the style, content, prompt conditioning inputs ; v_θ denotes the velocity field; x_0 is the target image feature; and ϵ is the noise.

3.4. Three Observations

We initially trained SC-DiT in a single stage on a mixture of D_{pure} and D_{synth}. However, we observed a surprising phenomenon: texture-related styles were consistently poorly learned, regardless of training time or data scale. For simplicity, we refer to styles without significant textures as semantic-related style transformations. As illustrated in Observation 1 of Figure 4, given the same content reference, semantic styles in the top row are transfered reasonably well, whereas texture styles in the bottom row are clearly deficient in strokes and textures (please zoom in for details). In the first column of the bottom row, prominent strokes are largely ignored. The second and third columns lack visible brushwork, and the fourth column appears overly smooth, missing the clay-like textures present in the style reference. We could also observed that the fine-grained characteristics could not be well-preserved in some cases, for example, the

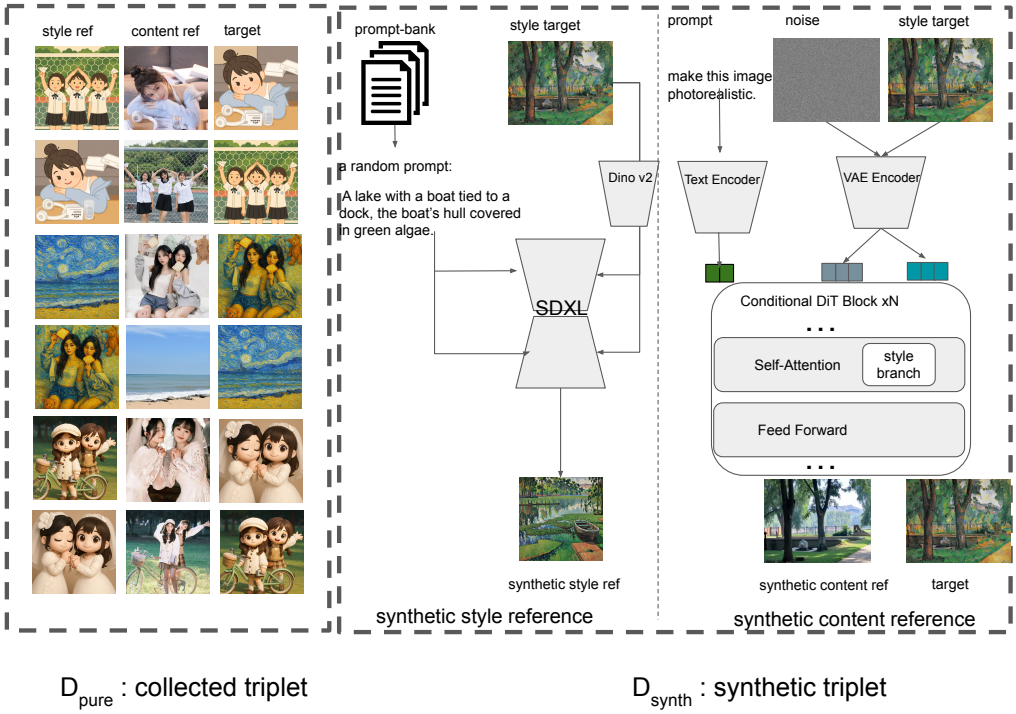


Figure 2. Collected triplets D_{pure} and synthetic triplets D_{synth}

first column in the semantic-style row alters the skin color and clothes (the skew pattern is turned to horizontal) of the left person, the oilpainting in the texture-style could not preserve the facial identities of the content reference. Based on such evidence, we have

Observation 1: *Semantics-related style transformations hinder the learning of texture-related style transformations when they are mixed and trained in one stage.*

Observation 2: *Characteristics of content reference could not be preserved well when D_{pure} and D_{synth} are trained in one stage.*

We further quantitatively validate **Observation 1** and **Observation 2** in Table 1. To further investigate these phenomena, we selected styles that combine both semantic and texture transformations (right side of Figure 4). In the first example, characters are transformed into a cartoon illustration (semantic) with strong pencil strokes (texture). In the second example, a 3D cartoon style (semantic) is combined with a clay material (texture). We trained SC-DiT on each of these styles *individually* and inspected style transfer results at different iterations. We consistently found that semantic transformations are learned in early iterations, whereas texture transformations emerge in late stages. This experiment demonstrates that without significant interference from semantic styles, SC-DiT is capable of learning texture styles. It also indicates that the complexity of texture style is higher thus harder to learn than semantic styles.

Ranking Style Complexity with FPLID Empirically, we choose to train an Fokker–Planck Local Intrinsic Dimensionality estimator (FPLID) (Kamkari et al., 2024) in FLUX VAE space to approximately measure and rank the complexity of style images. We calculate the Spearman's Correlation Coefficient and Significance on the sorted FPLID of a series of 20 style categories and 35 human users' averaged ranking results. With $\rho = 0.9718$, $p = 0.0007$, we confirm that LID ranking has a strong correlation with style complexity perceived by huamn. Due to space limit, we elaborate the theoretical and experimental details of FPLID in the appendix . Concretely, we train a small DDPM U-Net (Ho et al., 2020) on target images from D_{pure} and D_{synth} in FLUX VAE(Kingma and Welling, 2014) latent space:

$$LID(x, t_0) = D - \sqrt{1-\bar{\alpha}_{t_0}}\, tr(\nabla_x \epsilon(\sqrt{\bar{\alpha}_{t_0}}x, t_0)) + \|\epsilon(\sqrt{\bar{\alpha}_{t_0}}x, t_0)\|_2^2 \tag{9}$$

where x denotes FLUX VAE Latent, from D_{pure} and D_{synth}, t_0 is the timestep for evaluating LID, with $D = 16 \times 64 \times 64 = 65536$. With β_t being the Diffusion process hyper-parameter, $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_t$. The notation tr denotes trace operation, ∇_x denotes the differentiation operator with respect to x and ϵ is the Diffusion UNet.

We observe that semantic style transformations consistently exhibit lower LID, while texture style transformations have higher LID. Please zoom into the second row to see the fine textures on the skin in the last three images, especially

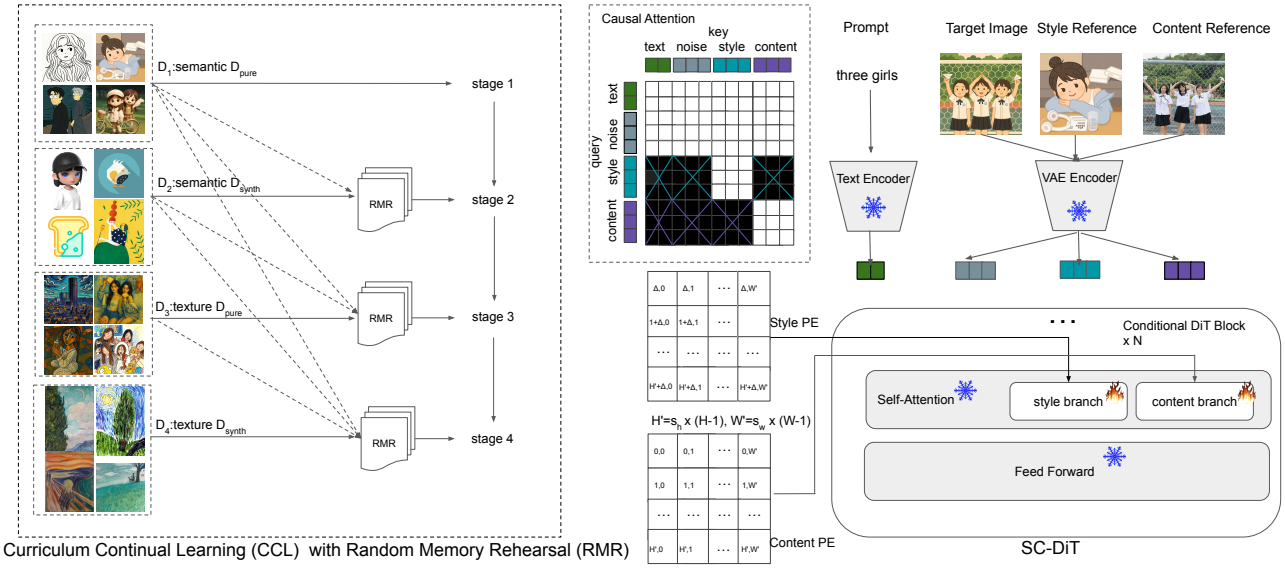


Figure 3. Multi-Stage Style Curriculum Continual Learning and the structure of SC-DiT.

the last one, where the style imitates canvas-like texture. Additional examples are provided in the appendix.Thus we obtain

Observation 3: *Semantic-related style transformations are learned in early stage, have low Local Intrinsic Dimensionality. Texture-related style transformations are learned in late stage, have high Local Intrinsic Dimensionality.*

3.5. Style Curriculum Continual Learning

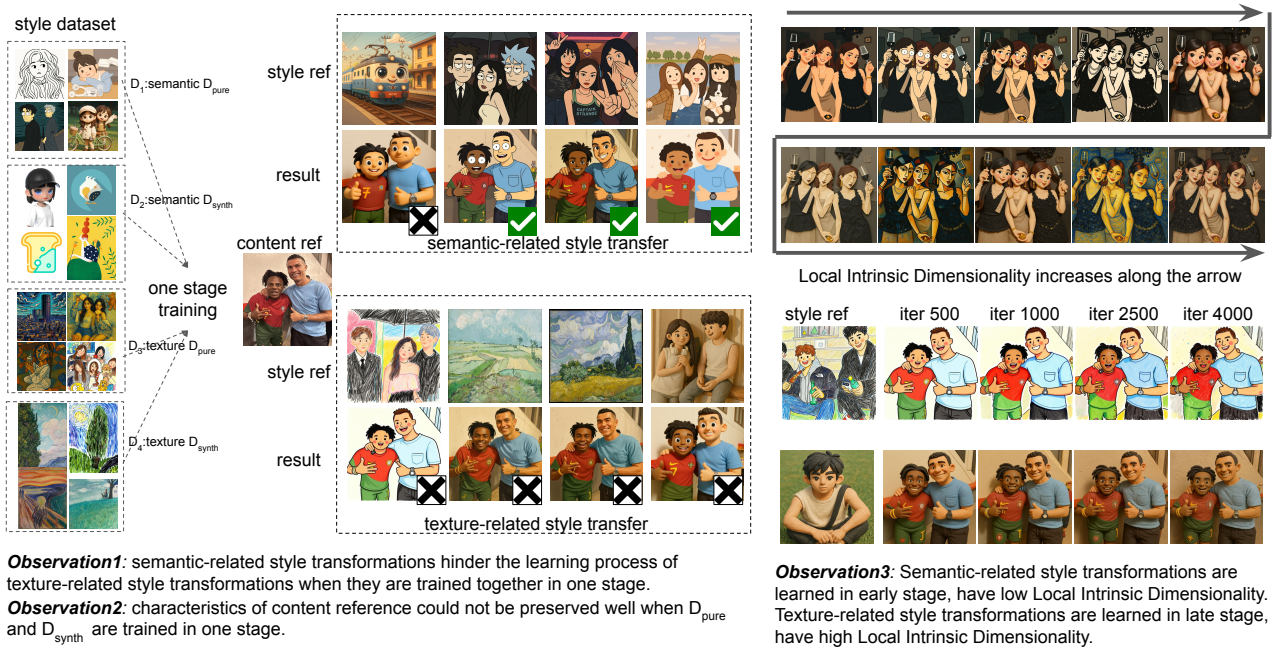
Inspired by our *Observation 3*, with collected dataset D_{pure} and synthetic dataset D_{synth}, we design a novel training paradigm, called **Style Curriculum Continual Learning** (Style-CCL), to tackle the severe problem of *Observation 1*, which causes our SC-DiT performing poorly on texture-related styles. *Observation 3* indicates that with VAE encoder to extract style features, semantic styles are easier to learn and texture styles are harder to learn, which inspires us to apply curriculum learning to separate the training of semantic styles and texture styles.

Shown in Figure 3, according to the sorted FPLID scores, we divide D_{pure} and D_{synth} into four subsets, D_1: semantic styles from D_{pure}, D_2: semantic styles from D_{synth}, D_3: texture styles from D_{pure}, D_4: texture styles from D_{synth}. The division boundary of semantic and texture styles is manually set according to the sorted FPLID scores, The semantic/texture boundary of D_{pure} and D_{synth} are set separately. Due to the intrinsic vagueness of style definition, we do not aim to and cannot precisely classify semantic and texture styles. Instead, the FPLID ranking and semantic/texture boundary are just helpful approximations to ease the training. We train D_1 and D_2 first, and train D_3 and D_4 later. However, the sequential multi-stage training on $\{D_1, D_2, D_3, D_4\}$ leads to catastrophic forgetting problem (Robins, 1995), with styles learned in early stages gradually forgotten in late stages. In addition, the content preservation is gradually weakened since the subset D_1 and D_3 from D_{pure} have higher characteristics consistency than D_2 and D_4 from D_{synth}. Thus we introduce Random Memory Rehearsal across Curriculum Learning stages, by randomly sampling the same amount of training data, with a fixed hyperparameter rehearsal rate R, from each style cluster in previous stages, and mix these previous samples with training data from current stage. The Random Memory Rehearsal is shown in Algorithm 1 and Style-CCL in Algorithm 2.

4. Experiments

Implementation Details. We adopt FLUX dev 1.0 (Labs, 2024) as base model for SC-DiT. The ranks for Style Branch Lora and Content Branch Lora are 128. We apply gradient checkpointing (Griewank and Walther, 2000) to save memory thus our model could easily be trained with short sides of both $512\times$ and $1024\times$. Our model is trained with 4 H100 GPUs, batch size is 1 for each GPU, learning rate is 1e-4.

Evaluation Benchmark. We select 50 style references and 40 content references, mutually pair each of them to generate 2000 style-content pairs for testing. We further select 10 style references and 10 content references as validation set. The style references cover diverse style genres and the content references include different number of persons with diverse gestures, scenes/buildings and subjects in



Observation1: semantic-related style transformations hinder the learning process of texture-related style transformations when they are trained together in one stage.
Observation2: characteristics of content reference could not be preserved well when D_{pure} and D_{synth} are trained in one stage.

Observation3: Semantic-related style transformations are learned in early stage, have low Local Intrinsic Dimensionality. Texture-related style transformations are learned in late stage, have high Local Intrinsic Dimensionality.

Figure 4. Our key observations of SC-DiT.

Algorithm 1 Random Memory Rehearsal (RMR)

Input: The previous triplet dataset $D_{pre} = \{S_1, S_2, ..., S_{N_{pre}}\}$ containing N_{pre} style clusters, and the current triplet dataset $D_{cur} = \{S'_1, S'_2, ..., S'_{N_{cur}}\}$ containing N_{cur} style clusters, and a fixed rehearsal sampling rate R.
Output: $D_{rmr} = \{T_1, T_2, ..., T_{N_{pre}}, S'_1, S'_2, ..., S'_{N_{cur}}\}$ containing $N_{rmr} = N_{pre} + N_{cur}$ style clusters.
Procedure:
$D_{rmr} \leftarrow \{\}$
for $S_i \in D_{pre}$ **do**
  Given a fixed rehearsal sampling rate R, we have sampling number k for each style cluster, $k \leftarrow \sum_{i}^{N_{cur}} |S'_i| \times R/N_{pre}$
  Randomly sample triplets $T_i = \{s_i^1, s_i^2, ..., s_i^k\}$ from S_i
  Insert T_i to D_{rmr}
end for
return D_{rmr}

Algorithm 2 Style Curriculum Continual Learning (Style-CCL)

Input:
D_1: semantic styles from D_{pure},
D_2: semantic styles from D_{synth},
D_3: texture styles from D_{pure},
D_4: texture styles from D_{synth}.
Output: SC-DiT$_{final}$

Procedure:
Train SC-DiT$_1 \leftarrow \{D_1$, FLUX-dev$\}$
Train SC-DiT$_2 \leftarrow \{$RMR(D_1, D_2), SC-DiT$_1\}$
Train SC-DiT$_3 \leftarrow \{$RMR(D_1, D_2, D_3), SC-DiT$_2\}$
Train SC-DiT$_4 \leftarrow \{$RMR(D_1, D_2, D_3, D_4), SC-DiT$_3\}$
SC-DiT$_{final} \leftarrow$ SC-DIT$_4$
return SC-DiT$_{final}$

complex scenarios. The images are of different aspect ratios. we show the details of these benchmarks in the appendix.

Evaluation Metrics. We evaluate our method with the following metrics. For **Style Consistency**, we use CSD Score (Somepalli et al., 2024) to measure the style similarity between the style reference and the generated image. For **Aesthetics**,we use the LAION Aesthetics Predictor (Schuhmann and Beaumont, 2022) to estimate the aesthetic quality of the generated image. For **Content Preservation**, we propose a new *Content Preservation Cut-Off Score* (CPC Score) with a style consistency threshold. Intuitively, a model that simply replicates the content reference without transferring style would receive an artificially high content score. To avoid this, we first use Qwen-VL (Bai et al., 2025) to generate a detailed caption T_{vlm} for the content reference image I_{content}, and compute the CLIP score (Radford et al., 2021) between T_{vlm} and the generated image I_{res}. We then compute the CSD Score between I_{style} and I_{res}; if this score falls below a threshold, the CLIP score is set to zero as a penalty.

$$CPC@thresh = \begin{cases} CLIP(I_{res}, T_{vlm}), & \text{if } CSD(I_{res}, I_{style}) >= \text{thresh} \\ 0, & \text{if } CSD(I_{res}, I_{style}) < \text{thresh} \end{cases} \quad (10)$$

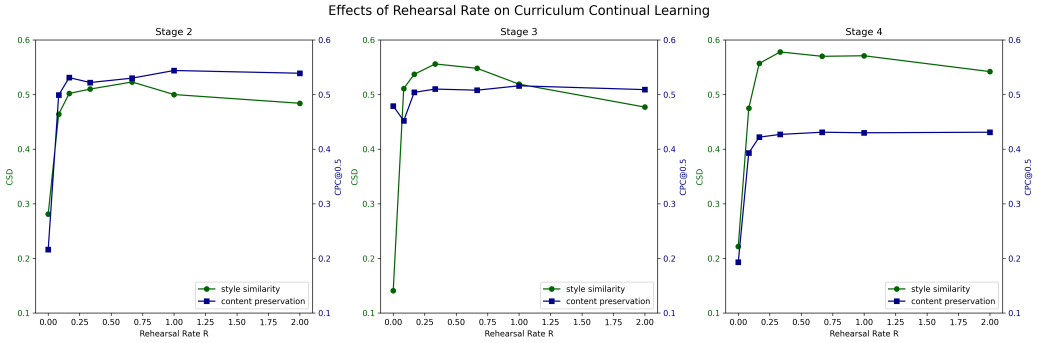


Figure 5. Quantitative ablation studies on the effects of Continual Learning and Rehearsal Rate. We train 7 models with different Rehearsal Rates R and validate the CSD Score and CPC Score on the validation set of each stage. When $R = 0$, no Curriculum Learning is applied.

4.1. Ablation Study

Importance of Multi-Stage CCL We conduct quantitative evaluations for Observation 1 and Observation 2 in Table 1, demonstrating that texture styles interfere the semantic styles. Shown in Figure 6, we qualitatively compare the effects of number of CCL stages.

Training Strategy	Semantic Style ↑	Texture Style ↑	Overall Style ↑	Content Preservation ↑
One Stage	0.571	0.117	0.344	0.298
Two Stages	0.574	0.526	0.557	0.392
Four Stages	**0.595**	**0.561**	**0.578**	**0.427**

Table 1. Quantitative ablation studies on the multi-stage CCL training strategy. We found that one stage training with mixed semantic and texture styles causes low style similarity for texture styles, which quantitatively validates our Observation 1.

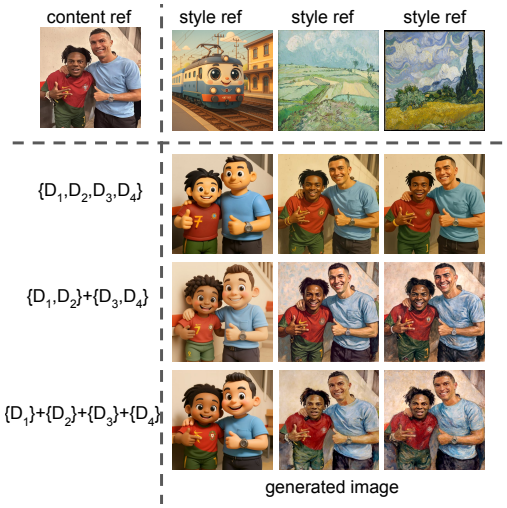


Figure 6. Importance of Multi-Stage CCL.

When SC-DiT is trained in one stage with a naive mixture of $\{D_1, D_2, D_3, D_4\}$, we could find in Figure 6, the T-shirt of the left person is inconsistent with the content reference (the green part of the T-shirt should be skew) and his skin color is not correctly preserved. The oil painting styles could not be correctly illustrated and characteristics of these two persons obviously change.

When SC-DiT is trained in two stage CCL $\{D_1, D_2\} + \{D_3, D_4\}$ (we apply Random Memory Rehearsal in the second stage, which is not explicitly written in Figure 6 for simplicity), we could find that SC-DiT could simultaneously learn semantic styles and texture styles. However, the characteristics still could not be preserved well. The green part of the T-shirt, worn by the left 3D-cartoon person, is still not skew. The facial identities of the persons in oil painting results are still not similar enough with the content reference. This is due to fact that clean collected triplets and noisy synthetic triplets are trained together, which has a negative impact on the characteristics consistency.

When SC-DiT is trained with four-stage CCL $\{D_1\} + \{D_2\} + \{D_3\} + \{D_4\}$ (algorithm 2), we could find SC-DiT performs well on both semantic styles and texture styles. Furthermore, the characteristics of content references could be well-preserved.

Continual Learning and Random Memory Rehearsal We thoroughly explore the effects of Rehearsal Rate R of Style Curriculum Continual Learning. We quantitatively measure style similarity with CSD Score and content preservation with CPC Score on the validation set of each stage in Table 4. We found that with an increasing R, the style similarity first increases and then decreases, while the content preservation keeps increasing then gradually saturates. We choose to set the Rehearsal Rate R to $1/3$.

We plot the effects of Random Rehearsal Rate R of each stage in Style-CCL on the validation set in Figure 5. We do not show the first stage because it is a normal training without continual learning. We observe some interesting phenomena in Figure 5:

- With an increasing R, the style similarity first increases then decreases. When $R = 0$ there is no continual learning thus some styles in previous stage will be

Figure 7. Qualitative Comparison with State-of-the-art Style Transfer Models.

Model	Style Similarity CSD Score↑	Content Preservation CPC Score@0.5 ↑	Content Preservation CPC Score@0.3:0.9 ↑	Aesthetic Score↑
OmniStyle	0.447	0.194	0.163	5.881
OmniGen-v2	0.462	0.243	0.166	5.843
DreamO	0.402	0.193	0.102	6.149
StyleID	0.453	0.190	0.180	5.749
StyleShot	0.450	0.227	0.116	5.740
InstantStyle	0.397	0.189	0.134	5.464
StyleSSP	0.494	0.291	0.207	5.130
CSGO	<u>0.535</u>	<u>0.379</u>	<u>0.224</u>	5.969
Style-CCL (ours)	**0.561**	**0.401**	**0.236**	**6.297**

Table 2. Quantitative comparison of our Style-CCL with previous state-of-the-art style transfer methods. The best score is stressed by bold font and the second best score is marked by underline.

forgotten, which leads to very low style similarity in every stage. However, when R keeps increasing, style data from previous stages dominate, which hinders the learning of data from current stage, thus style similarity gradually decreases.

- With an increasing R, the trending of content preservation is different for different stages. In stage 2 and stage 4, where noisy synthetic data dominates the current stage, we could observe content preservation keeps increasing with R and gradually saturates. In stage 3, where clean data dominates the current stage, increasing R leads to no significant fluctuation.

Semantic Texture Boundary We rank the FPLID score of all style clusters in training set by averaging each cluster. We train a model with semantic/texture data with style-ccl for each boundary and measure the models' performance with CSD score and CPC@0.5 score on validation set. Shown in Figure 8, we epirically set the semantic-texture boundary to FPLID=4000. This is never meant to be a precise division. Instead, it is just to ease the training process for Style-CCL.

4.2. Comparison with State-of-the-art Methods

Quantitative Comparison We quantitatively compare our Style-CCL with multiple current state-of-the-art style transfer models in Table 2, including UNet-based, DiT-based from the aspects of style similarity, content preservation and aesthetics score.

Qualitative Comparison We present qualitative visual comparison with state-of-the-art style transfer models on diverse style references from our test benchmark in Figure 7, where our Style-CCL tackles both semantic-related and texture-related style transfer and generate images with high aesthetics.

User Study. We employ 20 human evaluators to pick one best performance model from the candidates regarding style similarity, content consistency, aesthetics and their overall choices. The user study result is presented in Table 3 in percentage format.

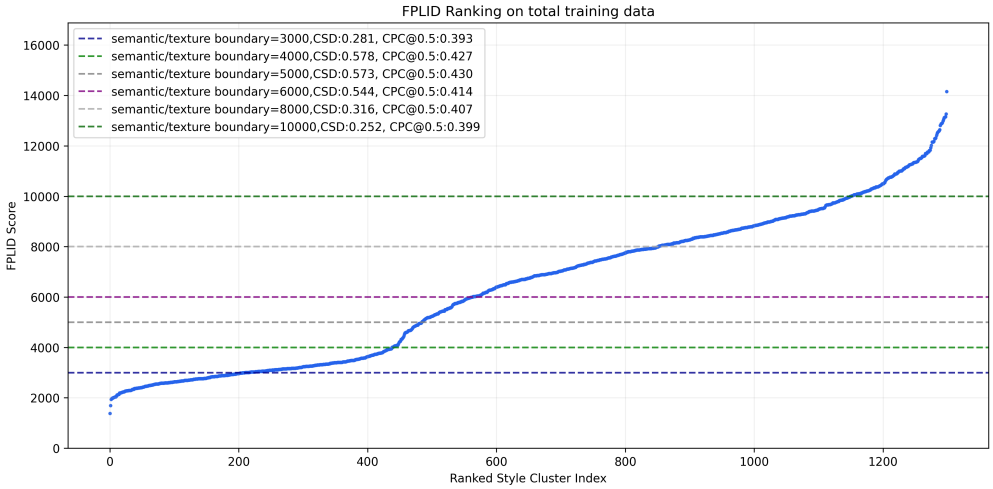


Figure 8. We rank all style clusters by the average of FPLID in each cluster. Various semantic/texture boundaries are experimented, and we epirically set it to be 4000.

Model	Style	Content	Aesthetics	Overall
StyleShot	0.75%	0.25%	1.75%	0.25%
InstantStyle	3.25%	1.25%	0.25%	0.50%
StyleSSP	3.75%	4.50%	8.25%	2.25%
StyleID	2.00%	22.25%	2.50%	3.75%
CSGO	5.50%	6.25%	1.00%	5.25%
OmniStyle	6.50%	7.50%	4.25%	5.50%
USO	8.50%	27.75%	10.00%	9.75%
Style-CCL (ours)	**69.75%**	**30.25%**	**72.00%**	**72.75%**

Table 3. User Study.

4.3. Limitations of Style-CCL-FLUX 1.0

Please note these limitations are for Style-CCL FLUX only. We observe two main limitations, as shown in Figure 11. First, in crowded scenes with many people, it may fail to preserve all individuals or their characteristics and it also struggles with rare, fine-grained out-of-distribution styles (e.g., Chinese ceramic art). In fact, our QwenStyle (Zhang et al., 2026a) and TeleStyle series (Zhang et al., 2026d;c) have tackled these issues.

5. Transfer Style-CCL to stronger foundation models

We have successfully transfer Style-CCL to Qwen-Image-Edit series (2509, 2511) (Wu et al., 2025) in December 2025. Our first version of Style-CCL model is TeleStyle V1 (QwenStyle) (Zhang et al., 2026a;d), released and open-sourced in Jan 2026, demonstrating strong generalization capability, high style similarity, content consistency and aesthetic merits. TeleStyle V1 (QwenStyle) established new state-of-the-art content-preserving style transfer performance in open-source models. We show a few examples in Figure 9. QwenStyle could generalize to unseen styles in Figure 10. In June 2026, we release TeleStyle V2 (Zhang et al., 2026c), achieving style transfer performance on par with top close-source model, gemini-3-pro-image-preview (nano banana pro) (Team, 2025). Beyond content-preserving style transfer task, TeleStyle series are also general text-guided image editing models on par with Qwen-Image-Edit (Wu et al., 2025) via Distribution-Matching-Distillation (Yin et al., 2024; Fan et al., 2026).

6. Conclusion

We observed that the one-stage training paradigm of SC-DiT suffers from semantic-texture interference and characteristics shifting. Thus we present Style-CCL, a Multi-Stage Style Curriculum Continual Learning framework content-preserving style transfer to tackle these problems. Our Style-CCL achieves new state-of-the-art performance on style similarity, content preservation and aesthetics score.

7. Appendix

7.1. Limitations of Style-CCL-FLUX 1.0

We observe two main limitations, as shown in Figure 11. First, in crowded scenes with many people, our model could be unstable thus may fail to preserve all individuals or their characteristics from the content reference. For example, in the first row, the 3D cartoon style could not preserve the correct number of characters. However, in the second row, the pixel effect could maintain the correct number of people. Second, our model may also struggle with some fine-grained out-of-distribution styles. For example, the third and fourth rows demonstrate the style transfer effects with Chinese ceramic art style reference. The style could not be precisely reproduced and looks more like a comic-book style,though the color is corretly transferred.

Figure 9. We transfer Style-CCL algorithm to Qwen-Image-Edit-2509 to train QwenStyle. The vanilla QIE-2509 is incapable of content-preserving style transfer. Our QwenStyle established new state-of-the-art on this task.

Figure 10. Out-of-Distribution cases for QwenStyle. The "failed" cases are not even wrong, since these style references could be interpreted as "photo-realism" styles. If one needs to transfer the material, perhaps it is better to use prompt directly.

7.2. Quantitative experiments on Multi-Stage CCL and Rehearsal Ratio

We show the quantitative ablation of rehearsal rate in Table 4

7.3. Estimating Style Complexity with FPLID

We further analyze this behavior using FPLID, the Fokker–Planck Local Intrinsic Dimensionality estimator (Kamkari et al., 2024). For a given disjoint union of manifolds and a point x on this union, the Local Intrinsic Dimensionality (LID) of x is defined as the dimension of the submanifold that contains x, which intuitively reflects the minimal number of variables needed to distinguish x from nearby samples. Higher LID indicates higher image complexity. The FPLID formulation leverages the Fokker–Planck equation to dramatically reduce the computational cost of normal bundle–based estimators (Tempczyk et al., 2022; Stanczuk et al., 2022), requiring only a single sample to estimate LID.

Concretely, we train a small DDPM U-Net (Ho et al., 2020) on target images from D_{pure} and D_{synth} in FLUX VAE(Kingma and Welling, 2014) latent space, enabling diffusion-model–based LID estimation for FLUX latents. With Variance-preserving DMs, score-matching formation(Song et al., 2020) of FPLID (Kamkari et al., 2024) is

$$\begin{aligned} FPLID(x,t_0) &= D + \left(1 - e^{-B(t_0)}\right)\left(\operatorname{tr}\left(\nabla s\left(e^{-\frac{1}{2}B(t_0)}x, t_0\right)\right) + \left\|s\left(e^{-\frac{1}{2}B(t_0)}x, t_0\right)\right\|_2^2\right) \\ &= D + \sigma^2(t_0)\left(\operatorname{tr}\left(\nabla s\left(\psi(t_0)x, t_0\right)\right) + \left\|s\left(\psi(t_0)x, t_0\right)\right\|_2^2\right) \end{aligned} \tag{11}$$

re such that

$$f(x,t) = -\frac{1}{2}\beta(t)x, \quad \text{and} \quad g(t) = \sqrt{\beta(t)}, \tag{12}$$

where β is a positive scalar function (Song et al., 2020),

$$\psi(t) = e^{-\frac{1}{2}B(t)}, \quad \text{and} \quad \sigma^2(t) = 1 - e^{-B(t)}, \tag{13}$$

$$B(t) \coloneqq \int_0^t \beta(u)du. \tag{14}$$

Since DDPM (Ho et al., 2020) and Score Matching (Song et al., 2020) could be mutually converted by, $t/T = t \in$

	Rehearsal Ratio R	0	1/12	1/6	1/3	2/3	1	2
Stage 2	Style Similarity on $D_1 + D_2 \uparrow$	0.281	0.464	0.502	0.510	**0.523**	0.500	0.484
	Content Preservation on $D_1 + D_2 \uparrow$	0.216	0.499	0.531	0.522	0.530	**0.544**	0.539
Stage 3	Style Similarity on $D_1 + D_2 + D_3 \uparrow$	0.141	0.511	0.537	**0.556**	0.548	0.519	0.477
	Content Preservation on $D_1 + D_2 + D_3 \uparrow$	0.479	0.452	0.504	0.510	0.508	**0.516**	0.509
Stage 4	Style Similarity on $D_1 + D_2 + D_3 + D_4 \uparrow$	0.222	0.475	0.557	**0.578**	0.570	0.571	0.542
	Content Preservation on $D_1 + D_2 + D_3 + D_4 \uparrow$	0.193	0.393	0.422	0.427	**0.431**	0.430	**0.431**

Table 4. Quantitative ablation studies on the effects of Continual Learning and Rehearsal Rate. We train 7 models with different Rehearsal Rates R and validate the CSD Score and CPC Score on the validation set of each stage. When $R = 0$, no Curriculum Learning is applied.

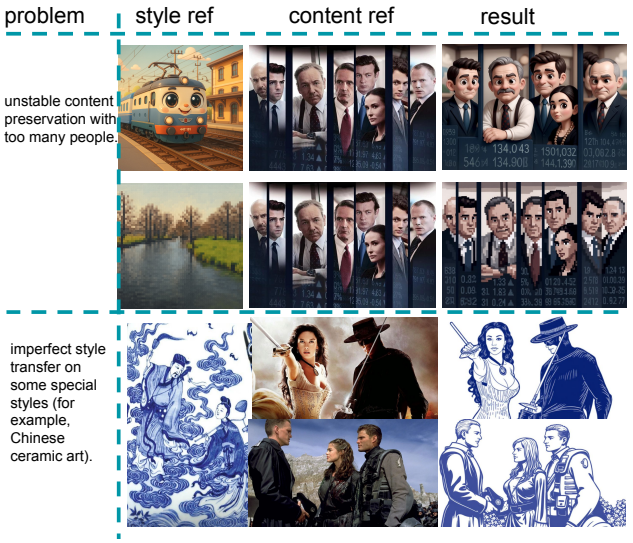


Figure 11. Limations: Our model is unstable when there are too many people in the content reference and style fidelity decreases on some specific style genres.

$[0, 1] \to t \in \{0, 1, \ldots, T\}$, $x_{t/T} \to x_t$, $\beta(t/T) = \beta(t) \to \beta_t$, $\psi(t/T) = \psi(t) \to \sqrt{\bar{\alpha}_t}$, $\sigma(t/T) = \sigma(t) \to \sqrt{1 - \bar{\alpha}_t}$, $\hat{s}(x, t/T) = \hat{s}(x, t) \to -\epsilon(x, t)/\sqrt{1 - \bar{\alpha}_t}$,
where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_t$. Thus we could have

$$FPLID(x, t_0) = D - \sqrt{1 - \bar{\alpha}_{t_0}} tr\left(\nabla \epsilon(\sqrt{\bar{\alpha}_{t_0}} x, t_0)\right) + \|\epsilon(\sqrt{\bar{\alpha}_{t_0}} x, t_0)\|_2^2, \tag{15}$$

where x denotes FLUX VAE Latent of a generated image from our SC-DiT, from D_{pure} and D_{synth}, t_0 is the timestep for evaluating LID, with $D = 16 \times 64 \times 64 = 65536$. With β_t being the Diffusion process hyper-parameter, $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_t$. The notation tr denotes trace operation, ∇_x denotes the differentiation operator with respect to x and ϵ is the Diffusion UNet.

7.4. Relative LID Ranking of Style Clusters

We demonstrate the qualitative and quantitative ranking of style complexity with our DDPM LID Estimator trained in FLUX VAE Latent Space in the main paper.

To validate the correlation of LID score and human perception of style complexity, we asked 35 users to rank the complexity order of 20 style clusters. The Spearman's rank-order correlation analysis reveals a very strong positive monotonic relationship between the FPLID ranking scores and human judgements with $\rho = 0.9718$, $p = 0.0007$. This result confirms that the LID ranking order is strongly and significantly aligned with human preferences on style complexity, demonstrating its effectiveness for serving as an indicator for Curriculumn Learning.

We show a qualitative example of ranking a subset of the training set. Different with the ranking in the main paper where we control the variance by using the same content and different styles, such strict scheme is impossible for ranking training data. However, shown in Figure 12, we could still observe the consistent trend from simple to complex in the ranking result, though there are very few outliers. Our manual cut of semantic styles and texture styles locates near the end of the fourth row.

7.5. Training Triplet Dataset Construction

7.5.1. Purified Triplet Matching

We extract style categories from OmniConsistency dataset (Song et al., 2025), which contains 22 style categories from GPT-4O (Hurst et al., 2024). We further collect 8 styles with Loras and internet data. We construct the [style ref, content ref, target] triplet with these data and filter the style similarity with CSD score (Somepalli et al., 2024), content similarity with clip score (Radford et al., 2021), facial similarity with arcface (Deng et al., 2019). Previous state-of-the-art style transfer models often suffer from the human facial identity leakage from style reference to content reference. To address such a problem, we purposely set a high proportion of style references to those images containing faces. For simplicity, we call this purified triplet dataset D_{pure} in the following sections. In later experiments, we found that the facial identity leakage problem is alleviated when we train our model with such triplets. Finally we get around 330k training data for D_{pure}.

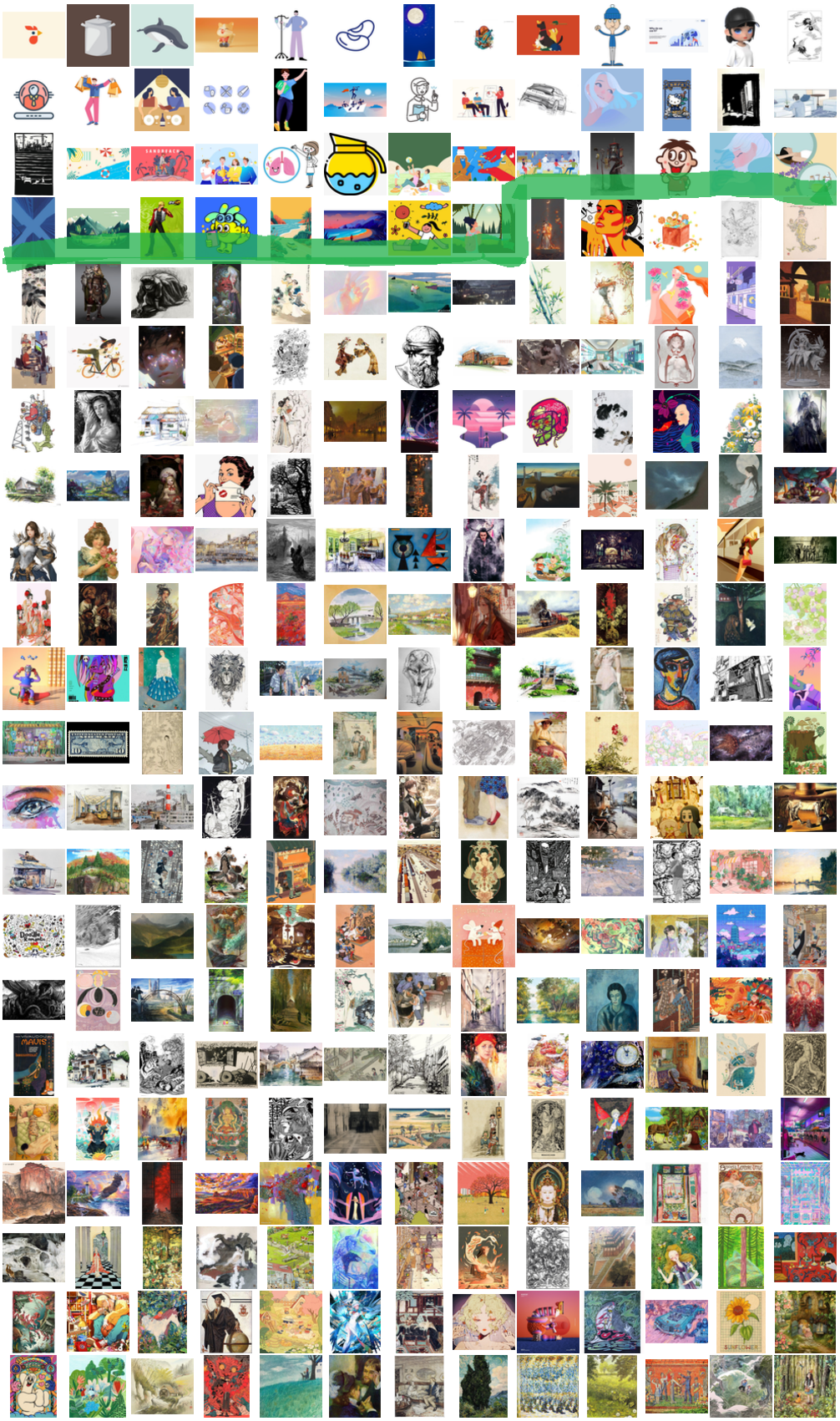


Figure 12. We randomly select one image each from hundreds of styles clusters and rank them with Our DDPM LID Estimator in FLUX VAE Latent Space. The LID scores increase from left to right, up to down.

7.5.2. Reverse Triplet Synthetic Framework

Although the purified triplets are clean and of high quality, the scale of such data is small and the data collecting is expensive. In order to utilize in-the-wild style images on internet, we introduce a reverse triplet synthetic framework inspired by (Wang et al., 2023b), where we reversely generate a style reference and a content reference from target image. Shown in Figure **??**, we use LLM (Bai et al., 2025) to generate a prompt bank, where the prompts focus on subjects instead of person because we utilize an internal SDXL (Podell et al., 2023) style adapter to generate style reference, which still suffers from facial identity leakage problem aforementioned. We randomly select a prompt from the prompt bank and feed it to the SDXL text encoder.

SDXL-based Style Transfer model to create style reference This SDXL-based style transfer model CDST (Zhang et al., 2025a) is trained with 14 million text-image pairs, the same data format like common text-to-image models. This model is good at prompt + style reference customization, but does not perform well on content-preserving style transfer with content reference and style reference. We feed the style target image to Dinov2 (Oquab et al., 2023) to extract image embeddings and input the image embeddings to a style transformer (Wang et al., 2023a) to compress the image embeddings to a fixed length tokens. These tokens are then fed to MLP to align the channel dimension and merged into UNet with learnable cross attentions. During inference, we only feed the compressed tokens to the decoder of UNet, which could effectively isolate style features according to Forgedit (Zhang et al., 2023; Zhang, 2024). Finally, we could get the synthetic style reference.

FLUX-based Photorealistic Converter to create content reference To get the content reference, we train a specific in-context DiT model shown in Figure 2, which accepts a prompt "make this image photographic" and a stylized image. It converts this stylized image into a photographic image while preserving the layout and content of the style image. We train this model with data from D_{pure}, where we reform each triplet [style ref, content ref, target] in D_{pure} to [target, content ref] with target being reference image, content reference being target. We introduce an auxiliary Lora in the FLUX structure to learn such editing capability with content RoPE (Su et al., 2024). The structure and causal attention mechanism is almost the same as SC-DiT in Figure 3, except that there is just one reference image (the stylized target) instead of two.

With such synthetic framework, we utilize images in Style30k(Li et al., 2024) as style target and reversely synthesize 2 million triplet by generating 20 style reference images and 1 content reference image for each style target image. We further utilize CSD score and CLIP score to filter style similarity and content consistency. Finally we have around 1 million synthetic triplets, denoted D_{synth}.

7.6. Computation Cost

Our Style-CCL is trained based on FLUX-dev (Labs, 2024), and trained with 4 H100 GPU with 80GB. We trained the model for 200 hours. The inference speed is around 3.6 seconds for generating a 512×512 image on one H100. QwenStyle and TeleStyle are also trained with 4 H100 for lora models, 8 H100 for complete parameter SFT. The inference of QwenStyle and TeleStyle is very efficient, generating images in 1K resolution in 4 seconds.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41667–41676, 2026.

Leon A Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.

Andreas Griewank and Andrea Walther. Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Transactions on Mathematical Software (TOMS)*, 26(1):19–45, 2000.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Hamid Kamkari, Brendan Ross, Rasa Hosseinzadeh, Jesse Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. *Advances in Neural Information Processing Systems*, 37:38307–38354, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

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.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pages 110–126. Springer, 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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Christoph Schuhmann and Romain Beaumont. Laion-aesthetics. *LAION. AI*, 2022.

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

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.

Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025.

Jan Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Your diffusion model secretly knows the dimension of the data manifold. *arXiv preprint arXiv:2212.12611*, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024.

Gemini Team. Gemini-3-pro-image-preview. 2025.

Piotr Tempczyk, Rafał Michaluk, Lukasz Garncarek, Przemysław Spurek, Jacek Tabor, and Adam Golinski. Lidl: Local intrinsic dimension estimation using approximate likelihood. In *International Conference on Machine Learning*, pages 21205–21231. PMLR, 2022.

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.

Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A unified stylized image generation model. *arXiv preprint arXiv:2309.01770*, 2023a.

Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7677–7689, 2023b.

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.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Shiwen Zhang. Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning. *arXiv preprint arXiv:2203.05928*, 2022.

Shiwen Zhang. Fast Imagic: Solving Overfitting in Text-guided Image Editing via Disentangled UNet with Forgetting Mechanism and Unified Vision-Language Optimization. In *PMLR*, 2024.

Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R Scott, and Limin Wang. V4d: 4d convolutional neural networks for video-level representation learning. In *International Conference on Learning Representations*, 2020a.

Shiwen Zhang, Sheng Guo, Limin Wang, Weilin Huang, and Matthew Scott. Knowledge integration networks for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020b.

Shiwen Zhang, Shuai Xiao, and Weilin Huang. Forgedit: Text guided image editing via learning and forgetting. *arXiv preprint arXiv:2309.10556*, 2023.

Shiwen Zhang, Zhuowei Chen, Lang Chen, and Yanze Wu. Cdst: Color disentangled style transfer for universal style reference customization. *arXiv preprint arXiv:2506.13770*, 2025a.

Shiwen Zhang, Haibin Huang, Chi Zhang, and Xuelong Li. Qwenstyle: Content-preserving style transfer with qwen-image-edit. *arXiv preprint arXiv:2601.06202*, 2026a.

Shiwen Zhang, Yifan Xu, Haibin Huang, Chi Zhang, and Xuelong Li. Telecomposer: Multi-reference-driven image and video customization. *arXiv preprint arXiv:2606*, 2026b.

Shiwen Zhang, Yifan Xu, Haibin Huang, Chi Zhang, and Xuelong Li. Telestyle v2: Beyond content-preserving style transfer with self-distillation and distribution-matching-distillation. *arXiv preprint arXiv:2606*, 2026c.

Shiwen Zhang, Xiaoyan Yang, Bojia Zi, Haibin Huang, Chi Zhang, and Xuelong Li. Telestyle: Content-preserving style transfer in images and videos. *arXiv preprint arXiv:2601.20175*, 2026d.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19513–19524, 2025b.