ACID-Style: An Adaptive Condition Injection Diffusion Model for Arbitrary Style Transfer

Ting Yang[1], Siyu Yang[1], Xiyao Liu[1*], Songtao Wu[2], Gerald Schaefer[3], Kuanhong Xu[2], Hui Fang[3*]

[1]School of Computer Science and Engineering, Central South University
[2]Sony R&D Center China, Sony (China) Limited
[3]Department of Computer Science, Loughborough University
234712251@csu.edu.cn, 224712192@csu.edu.cn, lxyzoewx@csu.edu.cn, songtao.wu@sony.com,
gerald.schaefer@ieee.org, kuanhong.xu@sony.com, h.fang@lboro.ac.uk

Abstract

Arbitrary style transfer (AST), a popular AI-powered photo editing function, aims to strike an optimal balance between content and style injection from two images in order to generate a novel high-fidelity stylised image. Recently, diffusion models have been applied to AST due to their high generation quality as well as flexibility to embed conditions. However, these models are still not satisfactory and may exhibit inferior performance compared to non-diffusion based methods. This is due to the diffusion process not being purposely designed for AST, leading to suboptimal solutions to trade-off content preservation and style embedding. In this paper, we propose ACID-Style, a novel adaptive condition injection diffusion-based AST framework for improved content/style feature injection to address this research challenge. Using two lightweight adapters, a content and a style injection module, and an adaptive injection mechanism, our approach is able to fully exploit a pre-trained stable diffusion model for AST-specific adaptation and our diffusion model thus learns the most effective timing for content and style injection in the diffusion sampling process. Comprehensive evaluations demonstrate that our method achieves superior style transfer performance, both quantitatively and qualitatively, compared to other state-of-the-art style transfer methods.

Code — https://github.com/the-fall-moon/ACID-Style

1 Introduction

Arbitrary style transfer (AST) blends the semantic structure of a content image with the stylistic features of a style image to produce a high-fidelity stylised result (Jing et al. 2019; Huang and Belongie 2017). AST has flourished, driven by widespread interest in the artistic and practical applications it offers including aesthetics filtering in photography (Yim et al. 2020), architectural rendering (Del Campo et al. 2019), and data augmentation (Zhou et al. 2021). Classical AST methods such as AdaIN (Huang and Belongie 2017), StyTR^2 (Deng et al. 2022), AesPA (Hong et al. 2023), and CAP-VSTNet (Wen, Gao, and Zou 2023), have shown impressive performance founded on advanced deep learning architectures and training strategies.

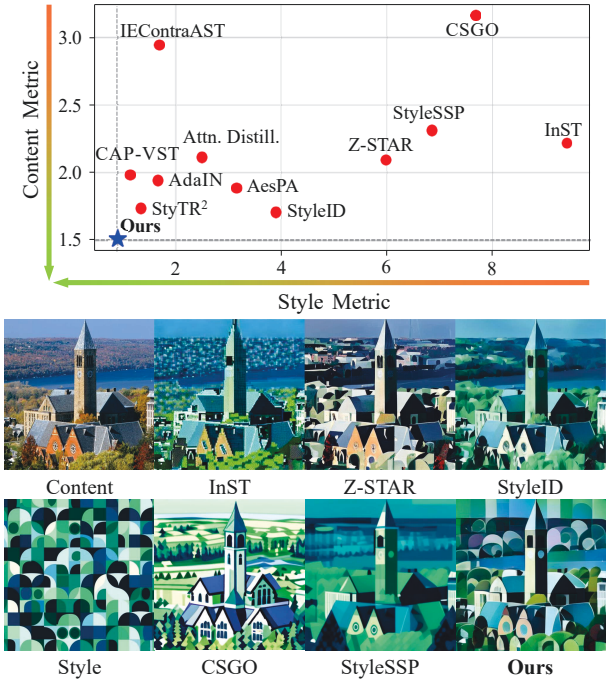

Figure 1: Top: Our proposed diffusion model achieves a superior trade-off compared to other state-of-the-art methods due to the better condition injection mechanism. Bottom: An example of the results obtained by our method in comparison to diffusion-based approaches, exhibiting its excellent content preservation and effective style injection. Note that spatial consistency constraints are not enforced in our method as they are not essential for artistic expressiveness.

Diffusion models have gained increasing traction in AST due to their flexibility to embed conditions in its sampling process and their ability to produce high-fidelity images. InST (Zhang et al. 2023), the first diffusion-based AST uses rare textual token to represent the style image and injects its projected textual feature into a diffusion model via a cross attention mechanism. For improved style representation embedding, in (Deng et al. 2024; Chung, Hyun, and Heo 2024), diffusion features from the style image interact

*Corresponding author.

with their counterparts from the content image using attention re-arrangement for style injection. In contrast, other approaches (Xing et al. 2024; Xu et al. 2025) inject both content and style features as conditions into the diffusion model to balance content preservation and style injection.

Despite the excellent visual quality of diffusion-based methods, they are less effective in style embedding compared to other classical methods as illustrated in Figure 1. This is due to the diffusion denoising process not being able to optimally balance the two conditions without disentangled representations since denoising is mainly designed for high-quality image synthesis with better data distribution alignment in the original image space (Ho, Jain, and Abbeel 2020). Therefore, enhancing the trade-off between content preservation and style injection by learning more representative features is crucial to fully unlocking the potential of diffusion models for AST.

To achieve this, we propose a novel arbitrary style transfer method, ACID-Style, by employing two lightweight adapter modules, one for content and style injection, respectively, and designing an adaptive injection process to empower a pre-trained stable diffusion model (Rombach et al. 2022) so that more representative and disentangled content and style features can be learned for condition injection. Inspired by (Ho, Jain, and Abbeel 2020; Yang et al. 2023), we train these two injection modules in an adaptive manner in order to better align the style transfer process to the diffusion dynamics, which enhances the content structure in the early diffusion stage and gradually enriches the image with stylistic characteristics in its later stage. As shown in Figure 1, our proposed method achieves a better balance of content and style composition when compared to other state-of-the-art (SOTA) diffusion and non-diffusion AST methods.

The main contributions of our ACID-Style model in this paper are:

- We introduce a novel AST framework, that enhances a pre-trained diffusion model with two condition injection modules, enabling the learning of more representative and disentangled embeddings tailored for AST-specific adaptation. To cope with the different characteristics of content and style presentations in an image, we design different mechanisms for their condition injection.
- We propose an adaptive injection strategy in the training stage to optimise the injection timing of content and style conditions. This strategy enables our model to adaptively balance content and style features at different diffusion stages, thus further enhancing the overall stylisation quality.
- We demonstrate that each designed component positively contributes to the style transfer task. When combined, our method achieves a superior balance between content and style compared to other SOTA AST methods, while also being computationally more efficient.

2 Related Work

2.1 Conventional Style Transfer

The pioneering neural style transfer method in (Leon Gatys 2016) uses CNN feature maps, establishing the foundation for arbitrary style transfer research. Subsequent works (An et al. 2021; Chen et al. 2021; Deng et al. 2022; Hong et al. 2023; Zhu et al. 2017) focus on improving efficiency and quality through specialised architectures and loss functions. Huang and Belongie (2017) propose Adaptive Instance Normalization (AdaIN) for real-time transfer by aligning feature statistics between content and style images. Recent advancements introduce GAN-based frameworks with contrastive learning such as IEcontraAST (Chen et al. 2021), vision transformers with stylisation perceptual losses such as StyTR^2 (Deng et al. 2022), and aesthetic-aware patch optimisation methods such as AesPA-Net (Hong et al. 2023), to enhance content-style harmony and visual quality.

2.2 Style Transfer with Pre-trained Models

The emergence of large-scale generative models enables new ways to perform style transfer. Text-inversion approaches, such as InST (Zhang et al. 2023) and DreamStyler (Ahn et al. 2024), map style images into text embeddings, thus enabling prompt-guided generation. StyleDiffusion (Wang, Zhao, and Xing 2023) employs a CLIP (Radford et al. 2021)-based disentanglement loss to effectively separate style attributes from content images to facilitate style transfer. CSGO (Xing et al. 2024) constructs a triplet dataset (content/style/stylised) to train a unified model for image/text-based style transfer, balancing content preservation and stylistic adaptation. StyleSSP (Xu et al. 2025) enhances the initial sampling point of the diffusion model with style information, effectively steering the denoising process toward the target style domain. Attn. Distill. (Zhou et al. 2025) extracts and aligns the attention features between the generated image and the reference images, enabling effective style transfer. InstantStyle (Wang et al. 2024a) achieves zero-shot transfer through innovative CLIP feature subtraction, while StyleAlign (Hertz et al. 2024) ensures consistent stylisation by exchanging self-attention keys and values derived from the DDIM (Song, Meng, and Ermon 2021) inversion process between images.

2.3 Condition Injection in Diffusion Models

Conditioning mechanisms allow a diffusion model to generate images with prior guidance. Early approaches such as (Ho, Jain, and Abbeel 2020) directly concatenate conditional information (e.g., class labels or text embeddings) with its noise input. This paradigm is subsequently extended through classifier guidance (Dhariwal and Nichol 2021) and classifier-free guidance (Ho and Salimans 2021), which leverage either gradient signals from auxiliary classifiers or implicit conditional modelling through joint training.

Recent advances focus on more precise condition injection. ControlNet (Zhang, Rao, and Agrawala 2023) introduces a learnable branch architecture that injects conditions into the denoising network through zero-convolution layers. T2I-Adapter (Mou et al. 2024) achieves enhanced model efficiency by injecting multi-scale conditions into a frozen text-to-image model through a lightweight adapter.

In diffusion-based AST, condition injection plays a crucial role for stylising images. For instance, StyleID (Chung, Hyun, and Heo 2024) and Z-STAR (Deng et al. 2024) both

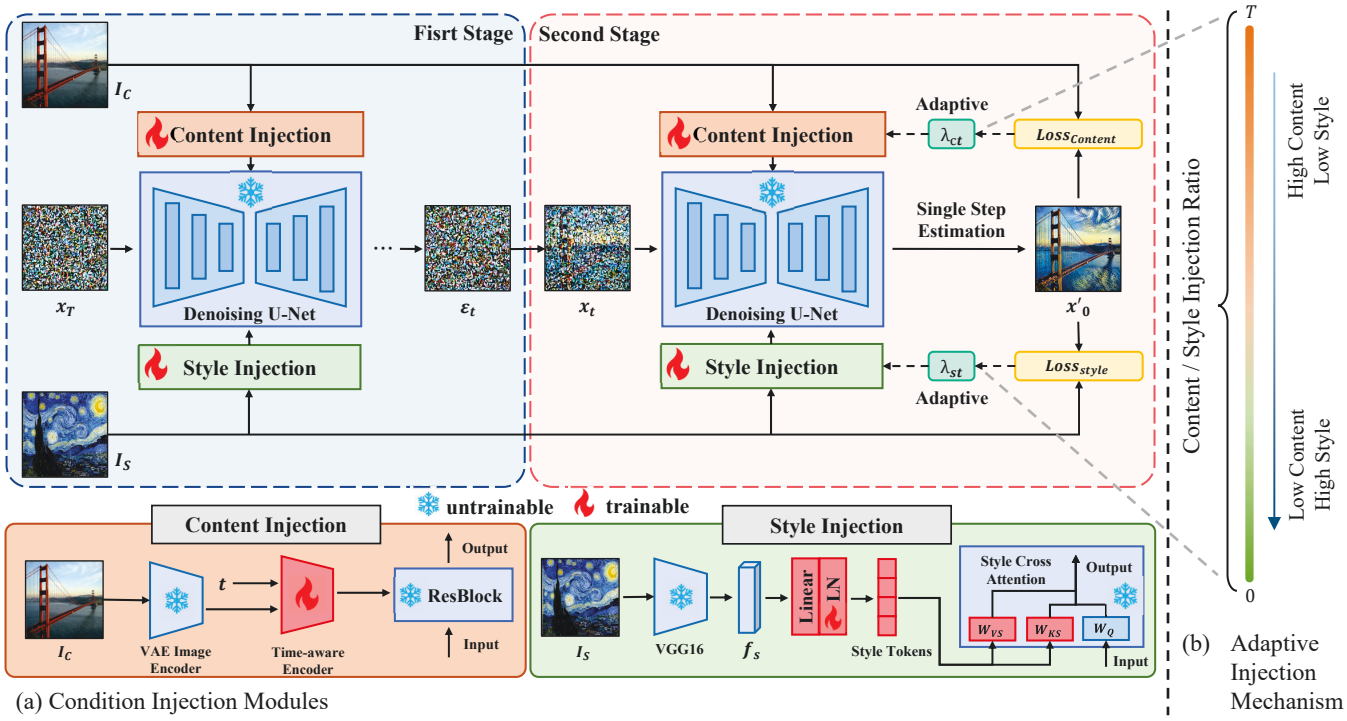

Figure 2: Overview of our proposed method, which comprises two parts: (a) learnable content and style injection modules to extract more representative features, facilitating a pre-trained stable diffusion model to balance content preservation and style embedding; and (b) a dynamic training strategy to align the style transfer process with the diffusion model dynamics.

extract content and style latent representations through inversion operations to enhance self-attention features during sampling. However, how condition injection can be designed and aligned with the diffusion dynamics to achieve a better content/style balance is still under-explored.

3 Proposed Method

An overview of our ACID-Style framework is illustrated in Figure 2. In the following, we first detail the design rationale and implementation of the content/style injection modules, and then explain our proposed condition injection process for an optimal trade-off of content and style composition.

3.1 Injection Modules

To enable an efficient style transfer without the need of diffusion model retraining or its inversion, we propose two separate learnable adapter modules while freezing the weights of a pre-trained stable diffusion model. We use two different mechanisms to inject content and style features, respectively, due to their distinct characteristics.

Content Injection Module For content preservation, i.e. preservation of layout, shape and locations of content in an image, retaining of precise spatial information is required. This is why methods like T2I-Adapter (Mou et al. 2024) and StableSR (Wang et al. 2024b) use a feature addition approach to inject conditions into encoder layers of a stable diffusion model as prior guidance since the addition operation maintains accurate spatial structure information of an image. In our proposed approach, we design a content injection module which shares similarities with the T2I-Adapter and StableSR architectures. Distinct from the T2I-Adapter, we use a time-aware encoder to control content injection by modulating intermediate feature maps $\{F_n^{in}\}_{n=1}^{N}$ from all residual blocks within the U-Net obtained as

$$\hat{F}_n^{in} = (1 + a_n) \odot F_n^{in} + b_n; a_n, b_n = \mathcal{M}_\vartheta^n(F_n^c), \quad (1)$$

where N is the number of the residual blocks in the stable diffusion U-Net, F_n^c denotes content feature maps from time-aware encoder, $\mathcal{M}_\vartheta^n(\cdot)$ denotes small convolutional networks in the time-aware encoder, and a_n and b_n are the affine parameters output by $\mathcal{M}_\vartheta^n(\cdot)$, thus allowing content structure preservation at a finer grained level.

Style Injection Module Inspired by IP-Adapter (Ye et al. 2023), we use a cross-attention mechanism to design our style injection module. Since style is a global property of an image characterised by colour tones, brush strokes, painting techniques, etc. and thus less related to the precise spatial layout of an image. A cross-attention mechanism is thus more suitable for style injection since it modulates features at each layer without considering spatial constraints. For better disentangled style features, we use a pre-trained VGG16 model (Simonyan and Zisserman 2015) to extract features from the style image I_s before projecting them onto their style feature embedding f_s which is tokenised as c_s for style injection. VGG features present diverse low-level local patterns and facilitate better style representation after projection although its style discriminant capability is inferior to

recent network architectures such as CSD (Somepalli et al. 2024). Finally, the tokenised representation is injected into the stable diffusion model via cross attention, formulated as

$$\varphi_{out}^{new} = \text{softmax}\left(\frac{QK_s^T}{\sqrt{d}}\right)V_s, \tag{2}$$

where $Q = \varphi W_Q$, $K_s = c_s W_{KS}$, and $V_s = c_s W_{VS}$, φ is the query feature input, W_Q is the weight matrix of the original projection layers, W_{KS} and W_{VS} are the trainable weight matrices for style injection, and d denotes the dimension of K. To leverage prior knowledge gained during pre-training and speed up the convergence of the style injection module, we initialise the parameters of each style cross-attention layer with those of the corresponding text cross-attention layer. Note that only the parameters of the projection network and W_{KS} and W_{VS} in the style cross-attention are trainable.

Training the Injection Modules Appropriately training the adapters is crucial to succeed in injecting more representative and disentangled representations for better style transfer effect. Given a content image I_c and a style image I_s, we use our proposed modules to inject their features into a stable diffusion model and generate a stylised image I_{cs}. We then update the two adapters by minimising the loss

$$\mathcal{L}_{total} = \lambda_{ct}\mathcal{L}_{ca} + \lambda_{st}\mathcal{L}_{sa}, \tag{3}$$

where $\mathcal{L}_{ca}$ is the content loss and $\mathcal{L}_{sa}$ is the style loss, and λ_{ct} and λ_{st} are weights to balance the two terms (which are explained in Section 3.2).

For the content loss $\mathcal{L}_{ca}$, we introduce a threshold strategy to reduce the high attention bias towards main objects induced by a pre-trained VGG19, and define the loss as

$$\mathcal{L}_{ca} = \frac{1}{N_l}\sum_{i=0}^{N_l} \|\mathcal{T}(\phi_i(I_{cs})) - \mathcal{T}((\phi_i(I_c))\|_2, \tag{4}$$

where $\phi_i(\cdot)$ are the feature maps from the i-th layer of the VGG19, and N_l denotes the number of layers. $\mathcal{T}(\cdot)$ is our proposed threshold strategy which clips values in each feature map below the 90-th percentile to mitigate the imbalance of stylisation between primary and other objects and thus yield more harmonious results.

The style loss $\mathcal{L}_{sa}$ is defined as

$$\mathcal{L}_{sa} = \frac{1}{N_l}\sum_{i=0}^{N_l}\sum_{f\in(\mu,\sigma)} \|f(\phi_i(I_{cs})) - f(\phi_i(I_s))\|_2, \tag{5}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and variance of the extracted features.

For improved efficiency and efficacy, we adopt a two-stage training strategy (Xu et al. 2024) to update the injection module parameters. In the first stage, we randomly select $t \in [1, T]$, where T is the total number of sampling steps, and perform gradual denoising from $x_T \sim \mathcal{N}(0, \mathrm{I})$ to calculate x_t as

$$\begin{aligned} x_t = &\sqrt{\overline{\alpha}_t}\left(\frac{x_{t+1} - \sqrt{1-\overline{\alpha}_{t+1}}\varepsilon_t}{\sqrt{\overline{\alpha}_{t+1}}}\right) \\ &+ \sqrt{1-\overline{\alpha}_t - \sigma_{t+1}^2}\varepsilon_t + \sigma_{t+1}\varepsilon, \end{aligned} \tag{6}$$

Algorithm 1: Injection module training algorithm.

Input: conditional diffusion model ε_θ, no. of sampling steps T, hyper-parameter λ, content image I_c, style image I_s
Output: optimised model ε_θ^*

$x_T \sim \mathcal{N}(0, \mathrm{I}), t \in [1, T]$;
without gradient descent:
  for $j = T$ to t **do**
    $\varepsilon_j = \varepsilon_\theta(x_{j+1}, j+1, I_c, I_s)$;
    predict x_j using Eq. (6);
  end for
with gradient descent:
  $\varepsilon_{t-1} = \varepsilon_\theta(x_t, t, I_c, I_s)$;
  $x_0' = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon_{t-1}\right)$;
  $I_{cs} = \text{VAE decoder}(x_0')$;
  calculate $\mathcal{L}_{ca}, \mathcal{L}_{sa}$ using Eq. (4) and Eq. (5);
  calculate $\lambda_{st}, \lambda_{ct}$ using Eq. (8);
  $\mathcal{L}_{total} = \lambda_{ct}\mathcal{L}_{ca} + \lambda_{st}\mathcal{L}_{sa}$;
  update θ;

where $\varepsilon_t = \varepsilon_\theta(x_{t+1}, t+1, I_c, I_s)$ denotes the noise predicted by our diffusion model with condition injection, and σ_t is a preset variance constant associated with t.

In the second stage, we input x_t to the U-Net and obtain the predicted ε_{t-1} to estimate x_0' as

$$x_0' = \frac{1}{\sqrt{\overline{\alpha}_t}}(x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon_{t-1}). \tag{7}$$

With the pre-trained VAE Decoder, we decode x_0' into I_{cs} to calculate the above style transfer loss.

Algorithm 1 details the training process in pseudo-code.

3.2 Adaptive Injection Mechanism

In the diffusion process, structural information is generated in the early stages while stylistic detail is enriched in the later stages (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021). To investigate this dynamic, we calculate the ratio between content loss $\mathcal{L}_{ca}$ and style loss $\mathcal{L}_{sa}$ by estimating the fully denoised image x_0' at each denoising step t. As illustrated in Figure 3, more content information is recovered in the early stages of the diffusion process, indicated by

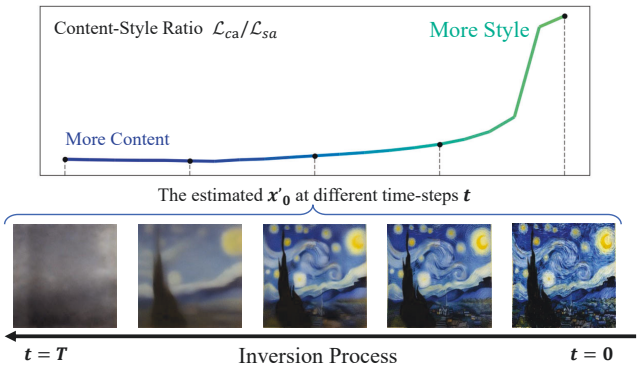

Figure 3: Example how the content-style ratio and the denoised image evolve.

	objective measures					user study		
	$\mathcal{L}_c \downarrow$	$\mathcal{L}_s \downarrow$	content score ↑	style score ↑	ArtFID ↓	content ↑	style ↑	aesthetics ↑
AdaIN	1.9440	1.6575	0.6981	0.3897	22.6152	5.96	6.14	5.96
IEContraAST	2.9588	1.6722	0.7701	0.4635	25.3747	7.68	7.54	7.02
StyTR^2	1.7336	1.3461	0.7675	0.4013	23.5439	7.84	6.62	7.24
AesPA-Net	1.8837	3.1454	0.7260	0.4569	23.6483	6.82	5.54	6.24
CAP-VST	1.9792	1.1453	0.6862	0.3654	23.1461	6.88	7.32	6.78
InST	2.2153	9.4207	0.6997	0.2954	27.2756	6.28	4.36	6.60
Z-STAR	2.0916	5.9811	0.6806	0.3974	22.5900	5.34	3.94	4.08
StyleID	1.7009	3.9107	0.7414	0.4540	26.3915	7.22	6.80	7.06
CSGO	3.1642	7.6781	0.6619	0.4876	27.1926	6.72	4.86	7.04
StyleSSP	2.3522	6.8628	0.6748	0.5062	27.6572	6.44	4.92	6.38
Attn. Distill.	2.1074	2.4997	0.6750	**0.7075**	**19.0274**	7.06	8.12	7.32
Ours	**1.5017**	**0.8853**	**0.7753**	0.4825	21.1256	**8.84**	**8.46**	**8.78**

Table 1: Results, both objective evaluation measures and from the user study, for all methods. Best results are bolded.

lower ratio values, while the style loss dramatically drops towards late stages of the diffusion. We therefore use this finding to set our diffusion loss weights to align with the dynamics of the diffusion model. Given $t \in [1, T]$, we set λ_{ct} and λ_{st} as

$$\begin{aligned} \lambda_{ct} &= 1 - \lambda\sqrt{\overline{\alpha}_{t-1}}, \\ \lambda_{st} &= \lambda\sqrt{\overline{\alpha}_{t-1}}, \end{aligned} \tag{8}$$

where λ is used to trade-off the content and style losses, and $\overline{\alpha}_t$ is a preset constant associated with the diffusion model's dynamics.

4 Experiment Results

4.1 Experimental Settings

Implementation Details We run our experiments on 6 × RTX 3090 GPU and build upon the v2.1 checkpoint of Stable Diffusion (Rombach et al. 2022). Unless stated otherwise, we use DDIM (Song, Meng, and Ermon 2021) sampling strategy with a total of 20 time steps. We adopt the AdamW optimiser (Loshchilov and Hutter 2017) with the learning rate is set to 0.0001. We use a batch size to 4 and set the batches of accumulated gradients to 4. Our injection modules are trained for 45 epochs.

Dataset For training, we use MS-COCO (Lin et al. 2014) as the content dataset and WikiArt (Phillips and Mackintosh 2011) as the style dataset. Following (Deng et al. 2022), we randomly crop all images to a fixed resolution of 256×256 during training, while we select images from a broader range of data sources and crop images to 512×512 during testing.

Baselines We evaluate our proposed method in comparison with eleven state-of-the-art methods, including five conventional style transfer methods, namely AdaIN (Huang and Belongie 2017), IEContraAST (Chen et al. 2021), StyTR^2 (Deng et al. 2022), AesPA (Hong et al. 2023), and CAP-VST (Wen, Gao, and Zou 2023), and six diffusion-based methods, namely InST (Zhang et al. 2023), Z-STAR (Deng et al. 2024), StyleID (Chung, Hyun, and Heo 2024), CSGO (Xing et al. 2024), StyleSSP (Xu et al. 2025), Attn. Distill. (Zhou et al. 2025). For a fair comparison, we use the official implementations of all baseline methods with their recommended configurations.

4.2 Quantitative Analysis

We evaluate the style transfer performance in terms of content and style preservation measures, and the inference time to judge computational efficiency.

Style Transfer Performance To evaluate style transfer performance, we follow the evaluation settings of StyTR^2 (Deng et al. 2022). We generate 800 stylised images, using 20 randomly selected content images and 40 style images covering sketches, watercolours, oil paintings, abstract art, and Japanese ukiyo-e.

We use the style transfer metrics $\mathcal{L}_c$ and $\mathcal{L}_s$ from StyTR^2 (Deng et al. 2022) to measure the alignment of the generated images with content and style as well as two CLIP-based metrics founded on a CLIP ViT-L/14 image encoder, namely the style score from the recently proposed CSD (Somepalli et al. 2024) and the content score defined as $\cos(E(I_{cs}), E(I_c))$, where $E(\cdot)$ encodes an image into an embedding with the CLIP ViT-L/14 model. We also use ArtFID (Wright and Ommer 2022) to assess the overall style transfer quality with consideration of both content preservation and style rendering.

The results, in terms of averages over the stylised images, for all methods are reported in Table 1. As we can see from there, our proposed method yields the best performance on both StyTR^2 style transfer metrics, $\mathcal{L}_c$ and $\mathcal{L}_s$. For the CLIP-based measure, our approach outperforms all other methods in terms of content score, while for style score, it is only outclassed by StyleSSP and Attn. Distill. Additionally, for ArtFID, our method ranks second only to Attn. Distill. However, both StyleSSP and Attn. Distill. introduce excessive semantic information from the style images, resulting in semantically inconsistent stylizations and lower content scores. These results demonstrate that our condition injection modules successfully inject style information without compromising content semantic integrity, outperforming previous SOTA methods.

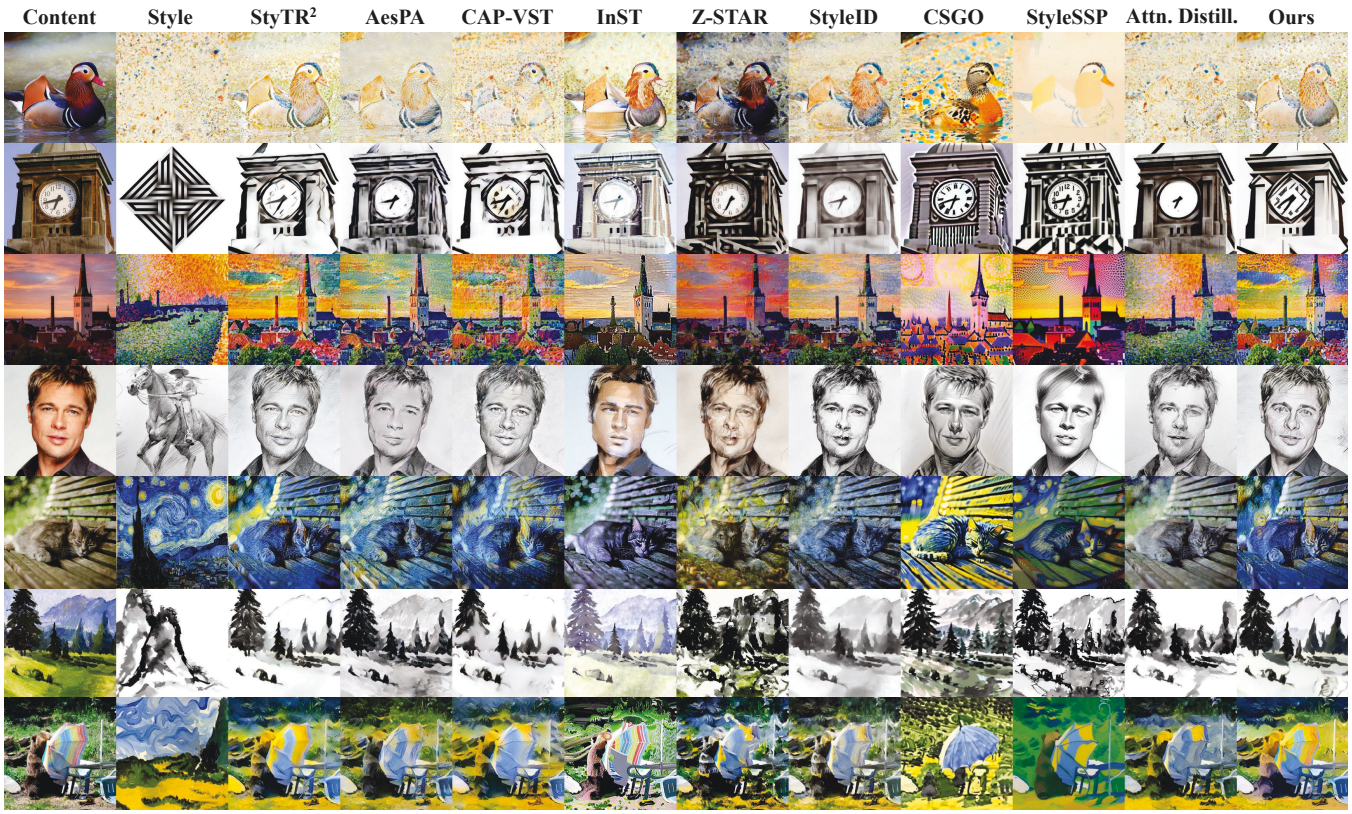

Figure 4: Examples of stylised images obtained by different AST methods.

Inference Time Runtime is also an important aspect of AST methods, especially as pre-trained-based methods may require significant time to generate a styled image. We therefore measure the inference time of our model in comparison to other diffusion-based approaches and present the results in Table 2.

Our model generates a stylised image in just 1.79 seconds and is thus significantly faster than other methods. This efficiency is achieved through our injection modules, which allow the model to efficiently capture both content and style information without requiring either model retraining for each style image as in InST, or diffusion inversion during the generation process for content and style images as in Z-STAR, StyleID, StyleSSP and Attn. Distill.

	time [s]
InST	1976.3
Z-STAR	69.39
StyleID	10.85
CSGO	6.07
StyleSSP	32.24
Attn. Distill.	29.59
Ours	**1.79**

Table 2: Inference time of our proposed method in comparison with other diffusion-based AST approaches.

4.3 Qualitative Analysis

We conduct a visual comparison with other SOTA style transfer methods and perform a user study to gather preferences regarding the different methods.

Visual Evaluation Figure 4 shows a range of examples of the stylised images obtained from different style transfer methods. As is evident from there, our proposed method effectively integrates consistent stylistic characteristics from the style images while preserving the semantic structure of the content images, demonstrating consistency between stylised and reference images. In contrast, StyTR^2, InST, StyleID, and StyleSSP exhibit insufficient style representation. Additionally, CAP-VST, InST, Z-STAR and Attn. Distill. cause partial structure degradation. AesPA, InST, StyleSSP and CSGO fail to accurately capture the artistic textures of the style images. Furthermore, Z-STAR incorrectly retains colour information from the content image, while Attn. Distill. introduces excessive partial semantic information from style images, resulting in stylised outputs that are semantically inconsistent with the content.

User Study We conduct a user study to further compare our method with the other approaches. For this, we randomly select 30 pairs of content and style images, and generate stylised images using all AST methods. Each of the 50 participants (28 males, 22 females, age 17-54) is shown 10 content and style images pairs together with the stylised

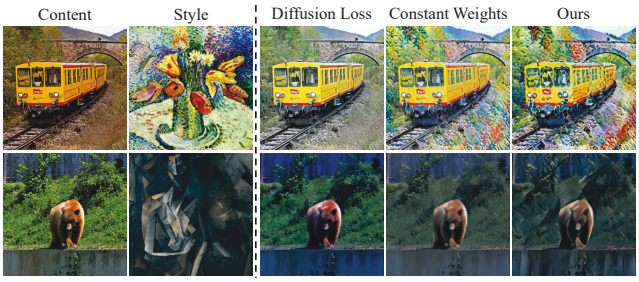

Figure 5: Examples of stylised images to evaluate the effectiveness of the proposed condition injection mechanism.

	$\mathcal{L}_c \downarrow$	$\mathcal{L}_s \downarrow$	CS $\uparrow$	SS $\uparrow$
diffusion loss	0.6251	7.3581	0.9536	0.0974
constant weights	1.5537	1.1059	0.7989	0.3159
ACID-Style	1.5017	0.8853	0.7753	0.4825

Table 3: Ablation results to evaluate the effectiveness of the proposed condition injection mechanism. CS=content score, SS=style score.

image and asked to rate the latter based on three metrics: content preservation, style consistency, and aesthetics. The results, in terms of average scores for all methods, are reported in Table 1. From there, we can see that our method is the most preferred one, yielding the highest scores in all three criteria, thus further highlighting the effectiveness of our proposed approach.

4.4 Ablation Study

We conduct ablation experiments to validate the contribution of our proposed contributions, and evaluate the impact of the hyper-parameter λ on the trade-off between style and content in stylised images.

Condition Injection Mechanism To validate the effectiveness of our proposed framework and its injection mechanism, we conduct experiments across different training methods. Based on the same network model, we evaluate the performance with diffusion loss, with a constant weights, and our proposed approach. For diffusion loss, we follow the training method of (Chen 2023), where the content image I_c and the style image I_s are set to the same image during training, with random dropping of either the content or style condition to train the injection modules. For constant weights, we use our defined losses $\mathcal{L}_{ca}$ and $\mathcal{L}_{sa}$ but use equal, constant weights to combine them.

As shown in Figure 5 and Table 3, the model trained with diffusion loss fails to inject style patterns, while the model trained with constant weights fails to achieve appropriate balance between content and style. In contrast, our model incorporates richer style information while maintaining the content semantic structure. This indicates that our condition injection modules effectively aligns the stylised images with the content and style images. Additionally, our adaptive injection mechanism further enhances the quality of style transfer. These results affirm the contribution of our proposed framework and adaptive injection mechanism.

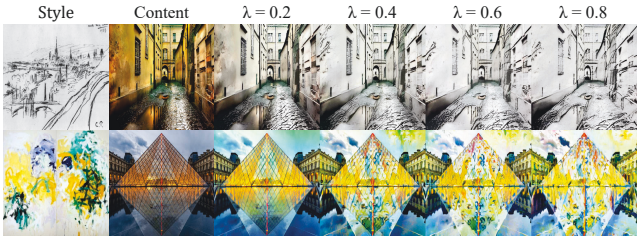

Figure 6: Stylisation examples for different values of hyper-parameter λ.

	$\mathcal{L}_c \downarrow$	$\mathcal{L}_s \downarrow$	CS $\uparrow$	SS $\uparrow$
$\lambda = 0.2$	0.8150	2.7842	0.8948	0.2160
$\lambda = 0.4$	1.2252	1.2672	0.8248	0.3659
$\lambda = 0.6$	1.5017	0.8853	0.7753	0.4825
$\lambda = 0.8$	1.8638	0.6927	0.7257	0.5774

Table 4: Stylisation results for different values of hyper-parameter λ. CS=content score, SS=style score.

Hyper-parameter λ Our framework allows for flexible control over the trade-off between content and style fidelity by adjusting the hyper-parameter λ (as described in Section 3.2). To evaluate the influence of this, we conduct experiments varying $\lambda \in \{0.2, 0.4, 0.6, 0.8\}$ and present the results in Table 4, while some stylised images are shown in Figure 6. As expected, high λ values introduce stronger style features at the expense of content fidelity, while low λ values favour content over style. For a suitable balance to achieve visually appealing style transfer results, we set $\lambda = 0.6$ as our default, which is also what the results in Table 1 are based on.

5 Conclusions

In this paper, we have proposed ACID-Style as a novel style transfer method that allows effective conditional feature learning to disentangle content and style representations that support a pre-trained diffusion model for better style transfer. We introduce two dedicated adapter modules to inject content and style conditions, respectively, along with an adaptive injection mechanism that aligns the embedding process with the dynamics of the diffusion process. Extensive experiments confirm our approach to achieve a balanced visual effect between content preservation and style embedding, outperforming other state-of-the-art approaches.

Although our method achieves visually appealing style transfer effects, some challenges remain to enable its extension to generate consistent stylised videos and to enable localised style transfer based on different semantics within an image. In future work, we therefore aim to extend our method to video style transfer and to investigate the incorporation of semantic information to enable semantic-aware style transfer.

Acknowledgements

This research is supported by the Natural Science Foundation of Hunan Province, China (2022GK5002, 2024JK2015, 2024JJ5440), the National Natural Science Foundation of China (61602527), the Special Foundation for Distinguished Young Scientists of Changsha (kq2209003), the Changsha Municipal Natural Science Foundation (kq2202109), the National Foreign Expert Project (G2023041039L), the 111 Project (D23006), the Foundation of State Key Laboratory of High Performance Computing, National University of Defense Technology (202401-13), and the High Performance Computing Center of Central South University.

References

Ahn, N.; Lee, J.; Lee, C.; Kim, K.; Kim, D.; Nam, S.-H.; and Hong, K. 2024. DreamStyler: Paint by style inversion with text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, volume 38, 674–681.

An, J.; Huang, S.; Song, Y.; Dou, D.; Liu, W.; and Luo, J. 2021. ArtFlow: Unbiased image style transfer via reversible neural flows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 862–871.

Chen, D.-Y. 2023. ArtFusion: Controllable Arbitrary Style Transfer using Dual Conditional Latent Diffusion Models. arXiv:2306.09330.

Chen, H.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D.; et al. 2021. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34: 26561–26573.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8795–8805.

Del Campo, M.; Manninger, S.; Sanche, M.; and Wang, L. 2019. The Church of AI — An examination of architecture in a posthuman design ecology. In *24th Conference for Computer-Aided Architectural Design Research in Asia*, 767–772.

Deng, Y.; He, X.; Tang, F.; and Dong, W. 2024. Z*: Zero-shot Style Transfer via Attention Reweighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6934–6944.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. StyTr^2: Image style transfer with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11326–11336.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Hertz, A.; Voynov, A.; Fruchter, S.; and Cohen-Or, D. 2024. Style aligned image generation via shared attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4775–4785.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hong, K.; Jeon, S.; Lee, J.; Ahn, N.; Kim, K.; Lee, P.; Kim, D.; Uh, Y.; and Byun, H. 2023. AesPA-Net: Aesthetic pattern-aware style transfer networks. In *IEEE/CVF International Conference on Computer Vision*, 22758–22767.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF International Conference on Computer Vision*, 1501–1510.

Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; and Song, M. 2019. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, 26(11): 3365–3385.

Leon Gatys, M. B., Alexander Ecker. 2016. A Neural Algorithm of Artistic Style. *Journal of Vision*, 16(326).

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Phillips, F.; and Mackintosh, B. 2011. Wiki Art Gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3): 593–608.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Somepalli, G.; Gupta, A.; Gupta, K.; Palta, S.; Goldblum, M.; Geiping, J.; Shrivastava, A.; and Goldstein, T. 2024. Measuring Style Similarity in Diffusion Models. arXiv:2404.01292.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Wang, H.; Spinelli, M.; Wang, Q.; Bai, X.; Qin, Z.; and Chen, A. 2024a. InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation. arXiv:2404.02733.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2024b. Exploiting diffusion prior for real-world image

super-resolution. *International Journal of Computer Vision*, 1–21.

Wang, Z.; Zhao, L.; and Xing, W. 2023. StyleDiffusion: Controllable disentangled style transfer via diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 7677–7689.

Wen, L.; Gao, C.; and Zou, C. 2023. CAP-VSTNet: content affinity preserved versatile style transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18300–18309.

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

Xing, P.; Wang, H.; Sun, Y.; Wang, Q.; Bai, X.; Ai, H.; Huang, R.; and Li, Z. 2024. CSGO: Content-Style Composition in Text-to-Image Generation. arXiv:2408.16766.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36.

Xu, R.; Xi, W.; Wang, X.; Mao, Y.; and Cheng, Z. 2025. StyleSSP: Sampling StartPoint Enhancement for Training-free Diffusion-based Method for Style Transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18260–18269.

Yang, X.; Zhou, D.; Feng, J.; and Wang, X. 2023. Diffusion probabilistic model made slim. In *IEEE/CVF Conference on computer vision and pattern recognition*, 22552–22562.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721.

Yim, J.; Yoo, J.; Do, W.-j.; Kim, B.; and Choe, J. 2020. Filter style transfer between photos. In *European Conference on Computer Vision*, 103–119.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023. Inversion-based style transfer with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10146–10156.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *International Conference on Learning Representations*.

Zhou, Y.; Gao, X.; Chen, Z.; and Huang, H. 2025. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18270–18280.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE/CVF International Conference on Computer Vision*, 2223–2232.