Content-Adaptive Style Transfer: A Training-Free Approach with VQ Autoencoders

Jongmin Gim*[0009−0005−9082−7427], Jihun Park*[0009−0004−1072−1239], Kyoungmin Lee*[0009−0008−2581−3610], and Sunghoon Im†[0000−0001−9776−8101]

DGIST, Daegu, Republic of Korea
{jongmin4422, pjh2857, kyoungmin, sunghoonim}@dgist.ac.kr

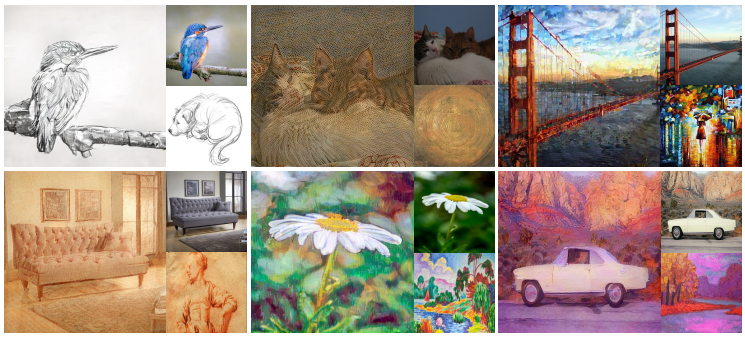

Fig. 1. Results of our Content-Adaptive Style Transfer (CAST). CAST produces the stylized image (left), given content image (top right), and style image (bottom right).

Abstract. We introduce Content-Adaptive Style Transfer (CAST), a novel training-free approach for arbitrary style transfer that enhances visual fidelity using vector quantized-based pretrained autoencoder. Our method systematically applies coherent stylization to corresponding content regions. It starts by capturing the global structure of images through vector quantization, then refines local details using our style-injected decoder. CAST consists of three main components: a content-consistent style injection module, which tailors stylization to unique image regions; an adaptive style refinement module, which fine-tunes stylization intensity; and a content refinement module, which ensures content integrity through interpolation and feature distribution maintenance. Experimental results indicate that CAST outperforms existing generative-based and traditional style transfer models in both quantitative and qualitative measures.

Keywords: Style transfer· VQ-VAE · Training-free.

* Equal Contribution
† Corresponding author

1 Introduction

Style transfer represents a cutting-edge area in image editing research, with widespread applications in cinematography, fashion, interactive gaming, and more. Its primary goal is to adapt the style of an input image while preserving the original content's integrity, a technique popularized by the work [19]. This seminal work spurred further research into learning-based methods [27,45,46,17,30,7] and arbitrary style transfer [25,21,36,34,58,14]. Recent developments have significantly benefited from the integration of attention mechanisms [48,16], which enhance the transfer process by enabling more precise semantic correspondences between style and content images. This is achieved through sophisticated manipulations of the keys and values within attention operations [36,53,34,14,23,13]. Additionally, there has been a concerted effort to incorporate adversarial strategies [20] into style transfer frameworks to further refine visual fidelity of the resulting images [4,50].

Recently, generative models, especially diffusion models [41,22], have excelled in style transfer by aligning closely with target domain distributions [9,52,29,26,57]. However, their *long inference time and large model size* remain significant drawbacks. As an effective alternative, vector quantized-based autoregressive models [47,18] that generate images through discrete token prediction have been gaining attention. These models have been rapidly applied across various fields, including image generation [11,6], editing [28,5], and style transfer [24,40,51,8]. They offer superior performance and faster processing speeds compared to diffusion models when trained on the same datasets. However, autoregressive models still face challenges: they often require *fine-tuning for specific domains* [51,40] or *additional model training* [8,24], which can be both cumbersome and time-consuming.

In response to these challenges, we introduce the Content-Adaptive Style Transfer (CAST), a novel, training-free approach for arbitrary style transfer. This method is specifically designed to apply coherent styles to corresponding regions within the content image, utilizing a feature space within decoding blocks. Inspired by the hierarchical manner in which humans typically perceive or create images [43], our approach initiates by capturing the global structure of the image through vector quantization. Subsequent local details are then refined through our decoder, which incorporates our newly designed style injection module into the multi-scale quantization autoencoder described in [43]. This allows for operation across multiple scales, significantly enhancing the style transfer process. Importantly, this approach utilizes the high visual fidelity of vector quantization-based generative models without relying on traditional training methods.

Our style injection module consists of three main components: *Content-consistent style injection*, *Adaptive style refinement*, and *Content refinement*. The *Content-consistent style injection* is based on the assumption that recognizing and adapting to the unique regions within the content image during stylization will preserve essential content information and yield more natural results. Operating within the VAR decoder's self-attention block, this technique clusters the content image's queries and categorizes them into distinct regions. Style im-

age features are then utilized as keys and values in the self-attention process, allowing for the injection of similar styles into similar content, thereby enhancing the naturalness of the stylized outcomes. The *adaptive style refinement* component adjusts the stylization intensity in various areas of the image, building upon the outcomes of content-consistent style injection. It intensifies stylization where the initial style application is underrepresented and conserves well-stylized areas, thus refining the overall aesthetic effect. This technique not only complements but also enhances the results achieved by the initial style injection, ensuring a balanced and refined visual presentation. The *content refinement* component addresses distortions in content information that can arise from the prior two processes. By employing interpolation within the decoder's residual blocks and maintaining pixel distribution in the feature map, it effectively preserves the integrity of the original content, all without the need for additional training.

Extensive experiments demonstrate that the proposed method outperforms state-of-the-art style transfer techniques, confirming its effectiveness. In summary, our primary contributions include:

- We propose a training-free arbitrary style transfer method that outperforms all competitive traditional and generative model-based style transfer models.
- We present the Content-consistent style injection technique, designed to stylize images according to the distinct regions of the content image.
- We propose the Adaptive style refinement technique, which effectively refines stylization based on the feature differences with the style image.
- We introduce the Content refinement technique, which efficiently preserves the content information of the stylized image.

2 Related work

2.1 Vector-Quantized image generation

VQVAE [47] pioneer the method for quantizing images into discrete tokens. Subsequently, techniques have emerged that train these models alongside adversarial loss from Generative Adversarial Networks (GANs) [20] to generate images stably [18,11,2]. Furthermore, methods based on bi-directional transformers for predicting masked areas in a non-autoregressive manner—as opposed to autoregressive transformers—have significantly accelerate the inference speed [6]. Recent advancements in vector-quantized image generation have led to significant successes, even surpassing diffusion models in image generation [55,43]. Parallel to remarkable progress in language models [35,15,37], models capable of generating images conditioned on text have emerged, utilizing vector-quantized methods [5,38]. Additionally, vector quantization techniques have been applied across various fields, including processing sequential data like video [55,28] and in applications such as image captioning, Visual Question Answering (VQA), and image recognition [54,59]. These applications demonstrate the versatile application of vector quantization in enhancing data representation and manipulation.

2.2 Traditional style transfer

Image Style Transfer, significantly advanced by [19], represents a landmark in image stylization. This method leverages a pre-trained Convolutional Neural Network (CNN), specifically VGGNet, to extract distinct content and style features. However, traditional style transfer methods are known for their substantial computational demands due to the per-image optimization approach. To address this issue, [25] introduces Adaptive Instance Normalization (AdaIN), which realigns the mean and variance of the source image features with those of the style image, enabling real-time style transfer. Subsequently, [31,32] proposed the Whitening and Coloring Transform (WCT), which aligns the entire covariance matrix of the features to enhance stylization outcomes. With the advent of attention mechanisms in neural networks [48,16], various style transfer models have been developed that leverage these mechanisms to achieve impressive results [16,34,23,14,36,53], demonstrating the evolving nature of style transfer technology.

2.3 Generative model-based style transfer

Generative model-based style transfer has heralded new opportunities through the use of Generative Adversarial Networks (GANs) [20], diffusion models [22], and Vector-Quantized Variational Autoencoders (VQVAE) [18]. Research such as [4,50] has moved beyond mere style and content fidelity by incorporating adversarial loss to enhance visual fidelity. Diffusion-based style transfer models [9,52,44,29,10,1,26] employ latent diffusion models, trained on extensive datasets [39], and incorporate style information during the denoising process via cross-attention. This technique has demonstrated impressive results with a focus on visual fidelity. Another significant advancement in this domain involves the use of vector quantization for style transfer. This method, explored in works such as [8,40,51], combines content output from image tokenizers with style information. Notably, [24] has shown that quantized features maintain high visual fidelity, which significantly enhances the quality of style transfer outcomes. However, earlier models based on vector quantization faced limitations, such as the need for a separate style codebook or extensive fine-tuning.

3 Preliminary: Vector quantization

The multi-scale quantization autoencoder in VAR model [43] consists of a VQ-encoder E and a decoder D with a vector-quantization bottleneck $\mathcal{Q}$. The encoder processes an image I, producing a raw feature map $\tilde{f}$. These features are then quantized through the bottleneck to generate the quantized feature map f, which is subsequently input into the decoder D. The decoder reconstructs the image I' as follows:

$$\begin{gathered} \tilde{f} = E(I),\ f = \mathcal{Q}(\tilde{f}),\ I' = D(f), \\ I \in \mathbb{R}^{H' \times W' \times 3},\ \tilde{f} \in \mathbb{R}^{H \times W \times CH},\ f \in \mathbb{R}^{H \times W \times CH}, I' \in \mathbb{R}^{H' \times W' \times 3}, \end{gathered} \tag{1}$$

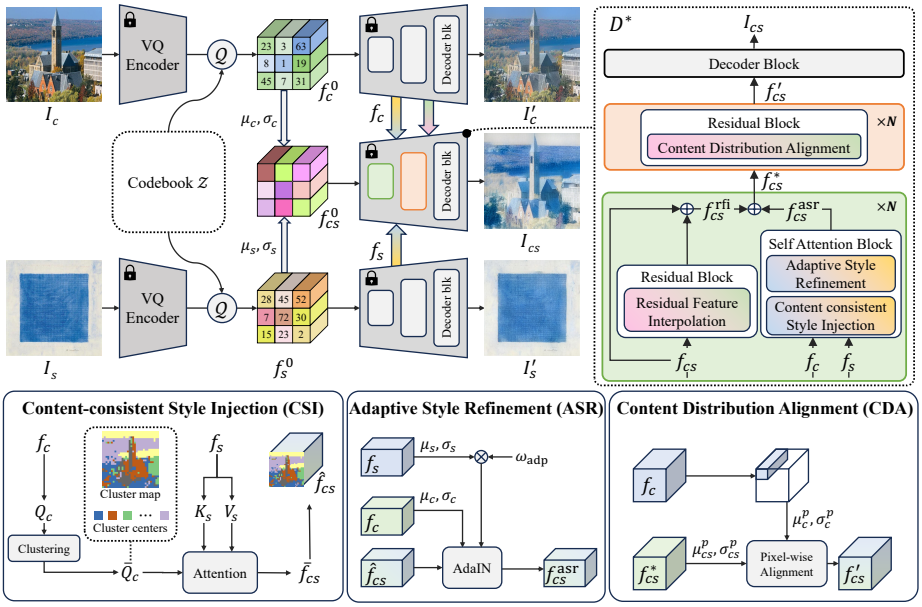

Fig. 2. Overall pipeline of CAST. The initial vector-quantized features f_c^0 and f_s^0 are extracted from the content image I_c and the style image I_s using a VQ Encoder and then combined using AdaIN to create the initial stylized feature f_{cs}^0. This feature is input to the decoder D^*, initially passing through N green blocks that incorporate CSI and ASR modules, along with residual feature interpolation. The enhanced feature, f_{cs}^*, then progresses through the N orange blocks equipped with a CDA module to produce f_{cs}'. In the subsequent process, the feature passes through the same structure as decoder D, culminating in the stylized image I_{cs}.

where H, W, and CH denote the height, width and channel of feature maps, respectively. H' and W' represent the height and width of images. Given the raw feature $\tilde{f}$, the vector-quantization bottleneck $\mathcal{Q}$ quantizes this feature to produce f, aligning each vector with the nearest vector in the learned discrete codebook. The codebook $z \in \mathbb{R}^{K \times CH}$ consists of K vectors as follows:

$$f^{h,w} = z^{\dot{k}}, \text{ where } \dot{k} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmin}} \|\tilde{f}^{h,w} - z^k\|_2, \\ \forall h = \{1,\dots,H\}, \ \forall w = \{1,\dots,W\}. \tag{2}$$

4 Method

4.1 Overall pipeline

Our model employs a VQ-encoder E and a vector-quantization bottleneck $\mathcal{Q}$ as dervied from VAR's multi-scale quantization autoencoder [43]. We design a new

structure for the decoder D^* as illustrated in Fig. 2. The decoder receives the content feature f_c^0, style feature f_s^0 and the initial stylized feature f_{cs}^0 as inputs. These features are defined as follows:

$$I_{cs} = D^*(f_{cs}^0, f_c^0, f_s^0), \text{ where } f_c^0 = \mathcal{Q}(E(I_c)), \ f_s^0 = \mathcal{Q}(E(I_s)), \tag{3}$$

where f_{cs}^0 is constructed using AdaIN [25] as follows:

$$f_{cs}^0 = \left(\frac{f_c^0 - \mu_c}{\sigma_c}\right) \cdot \sigma_s + \mu_s, \tag{4}$$

where σ_c and σ_s denote channel-wise standard deviation of content feature f_c and style feature f_s, respectively; μ_c, μ_s represent the channel-wise mean of content feature f_c and style feature f_s, respectively. The initial stylized features f_{cs}^0 are then transformed into high fidelity stylized features $\{f_{cs}^n\}_{n=1}^{2N}$ by being sequentially processed through $2N$-number of decoder blocks (green, orange blocks in Fig. 2). This transformation enhances the detail of the stylization, enabling more refined and visually appealing outputs. Starting from the subsequent section, we will simplify the notation by removing the block-specific term n from the feature f^n and denote it as f to streamline the equations and discussions.

4.2 Content-consistent Style Injection (CSI)

To achieve content-consistent style transfer, we utilize content features as queries and style features as keys and values inspired by the successful stylization techniques detailed in [9]. For all content-consistent style injection module in N green blocks, the queries of content features Q_c and the keys and values of style features K_s and V_s are extracted as follows:

$$Q_c = f_c \cdot W^q, \ K_s = f_s \cdot W^k, \ V_s = f_s \cdot W^v, \tag{5}$$

where W^q, W^k, and W^v are the weight matrices that transforms a feature f to the query, key and value, respectively. To inject the appropriate style into each region of a content image, we construct the set of cluster centroids $\{\bar{Q}_c^k\}_{k=1}^K$ by averaging the clustered content image queries Q_c as follows:

$$\bar{Q}_c^k = \frac{1}{|C_k|} \sum_{Q_c^{h,w} \in C_k} Q_c^{h,w}, \tag{6}$$

where C_k denotes the set of content image queries that belong to cluster k and K is the number of clusters. We then extract stylized clustered features $\bar{f}_{cs}^k$ through a cross-attention operation between the set of cluster centroids $\{\bar{Q}_c^k\}_{k=1}^K$ and the style image keys K_s and values V_s as follows:

$$\bar{f}_{cs}^k = \text{Softmax}\left(\frac{\bar{Q}_c^k K_s^T}{\sqrt{n_k}}\right) V_s, \tag{7}$$

where n_k denotes the dimension of the projected feature. Finally, we map the stylized clustered features $\bar{f}_{cs}^{k}$ into the image coordinates to produce the content-consistent style injected feature $\hat{f}_{cs} \in \mathbb{R}^{H \times W \times CH}$ for each block as follows:

$$\hat{f}_{cs}^{h,w} = \bar{f}_{cs}^{\dot{k}}, \text{ where } \dot{k} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmin}} \|Q_c^{h,w} - \bar{Q}_c^{k}\|_2, \\ \forall h = \{1,\dots,H\}, \ \forall w = \{1,\dots,W\}. \tag{8}$$

4.3 Adaptive Style Refinement (ASR)

While content-consistent style transfer effectively aligns style features with corresponding content regions, this approach can sometimes result in the loss of fine detailed information from style sources. This occurs because the region-wise application tends to homogenize features within each region, potentially oversimplifying or averaging out intricate textures and subtle variations. To tackle this issue, we introduce an Adaptive Style Refinement module that refines the style information in channels where the stylized features $\hat{f}_{cs}$ are not adequately representative of the style features f_s. The refinement process begins by calculating an adaptive weight w_{adp} that quantifies the degree of difference between $\hat{f}_{cs}$ and f_s as follows:

$$w_{\text{adp}} = S\left(\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(f_s - \hat{f}_{cs}\right)\right), \ w_{\text{adp}} \in \mathbb{R}^{CH}, \tag{9}$$

where S denotes min-max scaling function. A large adaptive weight indicates that the corresponding channel requires additional style information, while a small adaptive weight suggests that no further style information is needed for that channel. Lastly, we apply AdaIN [25] combined with the adaptive weight to refine the style as follows:

$$f_{cs}^{\text{asr}} = w_{\text{adp}}\left(\left(\frac{\hat{f}_{cs} - \mu_c}{\sigma_c}\right) \cdot \sigma_s + \mu_s\right). \tag{10}$$

4.4 Content refinement

While Sec. 4.2 and Sec. 4.3 effectively injects style into each distinct region, we observe that this often results in the dilution of essential content information. To address this challenge and enhance content preservation during the stylization process, we introduce a content refinement module.

Residual Feature Interpolation (RFI) We leverage the architecture of the decoder, particularly noting how the self-attention block and the residual block are connected via skip-connections. These connections allow for the preservation and gradual integration of detailed content information, which is crucial for the stable reconstruction of the image. Inspired by the methodologies discussed in [44], we employ the residual block connected to the self-attention as a

mechanism for content refinement. This is achieved by interpolating the stylized features f_{cs} with the original content features f_c, enhancing the content fidelity in the final output. The interpolation operation performed within the residual block is defined as follows:

$$f_{cs}^{\text{rfi}} = \text{ResBlock}\big(\alpha \cdot f_{cs} + (1-\alpha) \cdot f_c\big), \tag{11}$$

where α represents the interpolation weight, which determines the balance between the stylized and original content features. The function ResBlock$(\cdot)$ denotes a convolutional layer equipped with a residual block that facilitates the integration of these features. The interpolated features f_{cs}^{rfi} are added to the adaptively refined stylized feature f_{cs}^{asr} to obtain features f_{cs}^{*} that well preserve both content and style information as follows:

$$f_{cs}^{*} = f_{cs}^{\text{asr}} + f_{cs}^{\text{rfi}}. \tag{12}$$

This feature is then used as the input f_{cs} in Eq. (11) for the subsequent block $(n+1)$, which shares the same structure as the previous n, and this process is repeated for N blocks. This configuration fine-tunes the extent to which content features influence the overall composition, ensuring that critical attributes of the content are retained while embracing the desired stylistic transformations.

Content Distribution Alignment (CDA) To ensure that the essential content information does not undergo distortion during the stylization process in the N-number of green blocks, we introduce a Content distribution alignment technique that operates in the middle decoder blocks. This approach is particularly vital as standard stylization techniques often alter the spatial relationships and gradients within the content, which can lead to a loss of defining features and perceived sharpness. The proposed technique aims to maintain the original differences between adjacent pixels in the content image, thereby preserving the structural and textural integrity of the content. This is achieved by aligning the pixel-wise distribution of the stylized features f_{cs}^{*} with those of the original content features f_c. The alignment process involves adjusting the mean and standard deviation of the stylized features to match those of the content features as follows:

$$\begin{aligned} f'_{cs} &= \left(\frac{f_{cs}^{*} - \mu_{cs}^{p}}{\sigma_{cs}^{p}}\right) \cdot \sigma_{c}^{p} + \mu_{c}^{p}, \\ \text{where } \mu_{c}^{p} = \frac{1}{CH}\sum_{ch=1}^{CH} f_{c}^{ch}, \sigma_{c}^{p} &= \sqrt{\frac{1}{CH}\sum_{ch=1}^{CH}(f_{c}^{ch} - \mu_{c}^{p})^2}, \end{aligned} \tag{13}$$

where μ_{cs}^{p}, μ_{c}^{p} are the pixel-wise means of the stylized and content features, respectively, and σ_{cs}^{p} and σ_{c}^{p} are the corresponding standard deviations. This feature f'_{cs} is then utilized as the input f_{cs}^{*} in Eq. (13) for the subsequent block $(n+1)$, which maintains the same structure as the previous n. This iterative process is executed through a total of N blocks.

Subsequently, the stylized feature f'_{cs} from the 2N-th block is processed in the same manner as the original decoder D, resulting in the synthesis of the stylized

image I_{cs}. This ensures seamless integration and refinement of style throughout the network. The overall pipeline of the proposed stylization process is illustrated on the right side of Fig. 2.

5 Experiment

5.1 Implementation Details

We conduct all experiments resizing images size by 512×512 and using the frozen multi-scale quantization autoencoder proposed by VAR [43], trained on the ImageNet dataset [12]. The autoencoder features a codebook size of 4,096 and a quantized feature size of (32, 32, 32). The each number of blocks in the decoder N is set to 3. We defaulted the number of clusters to 22 and the interpolation weight α to 0.87. All experiments were performed on a single NVIDIA A6000 GPU.

5.2 Experimental setup

Traditional style transfer models often employ style and content losses for training and evaluation, based on the method [19]. However, these metrics can lead to overfitting on style images and might not fully account for visual fidelity, potentially resulting in unfair comparisons. Thus, we utilize a recently proposed metric, ArtFID [49], which was validated through a large-scale user study. ArtFID complements traditional style and content loss evaluation schemes, which is computed as follows:

$$\text{ArtFID} = (1 + \text{ LPIPS}(I_c, I_{cs}))(1 + \text{FID}(I_s, I_{cs}))), \tag{14}$$

where I_c and I_s denote the content and style images, respectively, and I_{cs} denotes stylized image. $\text{LPIPS}(I_c, I_{cs})$ measures content fidelity between the content and stylized images, while $\text{FID}(I_s, I_{cs})$ assesses style fidelity. Evaluations are performed under the same conditions as those in StyleID [9], using content images from the MS-COCO dataset [33] and style images from the WikiArt dataset [42]. All images were center-cropped to a size of 512×512 pixels. For quantitative evaluation, we generate 800 stylized images by applying style transfers to 20 content images and 40 style images from each dataset, following the method used by StyleID [9].

5.3 Quantitative comparison with state-of-the-art style transfer models

To evaluate our model, we conduct quantitative assessments against seven state-of-the-art style transfer models—AesPA-Net [23], StyTR^2 [13], EFDM [56], MAST [14], AdaAttn [34], AdaConv [3], and AdaIN [25]—as well as four generative model-based and one vector quantization based style transfer models, including four diffusion-based models (DiffuseIT [29], InST [57], StyleID [9], DiffStyle [26])

Table 1. Quantitative comparison of ours with state-of-the-art traditional style transfer models. The symbol ↓ indicates lower values are better.

Metric	**Ours**	AesPa-Net	StyTR^2	EFDM	MAST	AdaAttn	AdaConv	AdaIN
ArtFID ↓	**28.370**	31.420	30.720	34.605	31.282	30.350	31.856	30.933
FID ↓	**17.788**	19.760	18.890	20.062	18.199	18.658	19.022	18.242
LPIPS ↓	**0.5100**	0.5135	0.5445	0.6430	0.6293	0.5439	0.5562	0.6076
time (s) ↓	0.730	0.286	0.306	**0.045**	1.149	0.115	0.055	0.046
#Param (M) ↓	108.95	24.20	48.34	**7.01**	17.12	26.57	62.83	**7.01**

Table 2. Quantitative comparison of ours with state-of-the-art generative model-based style transfer models. The symbol * denotes for 256×256 resolution reproduction model.

Metric	**Ours**	DiffuseIT*	InST	StyleID	DiffStyle*	QuantArt
ArtFID ↓	**28.370**	40.721	40.633	28.801	41.464	35.747
FID ↓	**17.788**	23.065	21.571	18.131	20.903	23.558
LPIPS ↓	0.5100	0.6921	0.8002	0.5055	0.8931	**0.4556**
time (s) ↓	0.730	518.811	3.930	10.905	105.785	**0.138**
#Param (M) ↓	**108.95**	552.81	1497.59	1066.24	553.84	112.35

and one vector quantization-based model (QuantArt [24]). Note that the data in Tab. 1 and Tab. 2 are sourced from the work [9], with the exception of the data for the QuantArt, which we obtain independently.

Comparison with traditional style transfer In Tab. 1, we present a quantitative comparison with leading traditional style transfer models. Our results show that our method more accurately reflect the style from the style image (measured by FID) and preserve the content from the content image (measured by LPIPS) more effectively than traditional methods. This confirms that our approach achieves a balanced style transfer as indicated by ArtFID.

Comparison with generative model-based style transfer To further validate the efficiency of our proposed CAST method, we present a quantitative comparison with contemporary generative model-based style transfer models in Tab. 2. The results reveal that, with the exception of [9] and [24], most generative model-based style transfer approaches struggle to preserve content information adequately, as indicated by LPIPS scores. They also fall short in accurately reflecting the intended style, as measured by FID scores. Specifically, [24] maintains content fidelity effectively but lacks style incorporation, leading to unbalanced visual results. While [9] yields results that are closest to our CAST in terms of style-content balance, our method surpasses it in terms of inference time, the efficiency of parameter usage, and overall performance as measured by ArtFID, thus demonstrating enhanced effectiveness and usability.

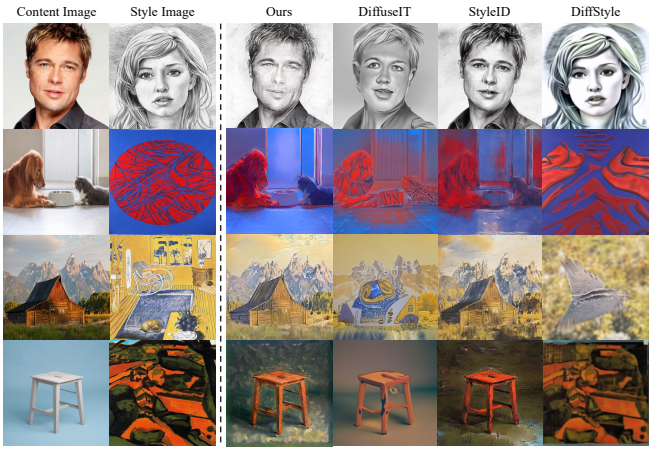

Fig. 3. Qualitative comparison with state-of-the-art generative model-based style transfer models.

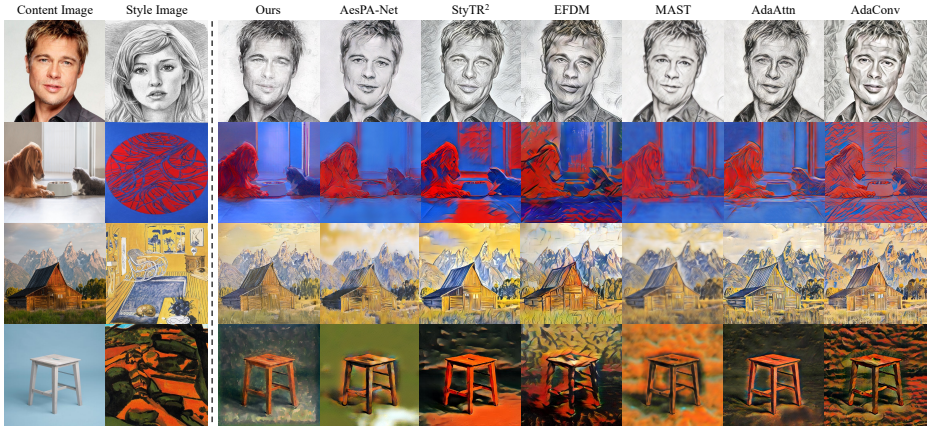

Fig. 4. Qualitative comparison with state-of-the-art traditional style transfer models.

5.4 Qualitative comparison with state-of-the-art style transfer models

Comparison with traditional style transfer To showcase the effectiveness of our proposed method, we present a qualitative comparison with state-of-the-art traditional style transfer models in Fig. 4. The results from this comparison clearly demonstrate that our CAST model excels at preserving content integrity while effectively reflecting the desired style across different regions. This leads to a more cohesive and naturally stylized outcome across the entire image, surpassing the capabilities of traditional models which often struggle to balance style application with content preservation.

Comparison with generative model-based style transfer Similarly, we also conduct a qualitative comparison with leading generative model-based style transfer approaches, as illustrated in Fig. 3. These results reveal that many generative models either compromise too much content detail or do not adequately capture the style essence, often resulting in a mismatch between style application and content integrity. In comparison, while StyleID [9] shows results that most closely resemble those of our CAST, it occasionally loses content details or applies styles inconsistently, leading to less natural-looking outcomes. Our model, by contrast, maintains a superior balance between style fidelity and content preservation, consistently delivering more natural and aesthetically pleasing results.

5.5 Ablation Study

To validate the efficacy of our proposed techniques and ensure they perform as designed, we conduct an ablation study. The results of this study are both quantitatively and qualitatively analyzed, with quantitative outcomes presented in Tab. 3 and qualitative observations illustrated in Fig. 5. Notably, the quantitative results shown in Tab. 3-(e) highlight the comprehensive impact when combining all proposed methods: Content-consistent Style Injection (CSI), Adaptive Style Refinement (ASR), Residual Feature Interpolation (RFI), and Content Distribution Alignment (CDA). In Tab. 3-(a), the initialization using AdaIN demonstrates effective style reflection while preserving content. The results from Tab. 3-(b) reveal that our CSI module significantly enhances both FID and LPIPS scores by ensuring content-consistent stylization, although it indicates that optimal stylization has not been fully realized. Tab. 3-(c) shows that the ASR module addresses the stylization shortfall noted previously, albeit at a slight cost to content fidelity. Further analysis in Tab. 3-(d) and (e) affirm that RFI and CDA effectively maintain the overall and detailed content information, respectively.

Qualitative assessments in Fig. 5 complement these findings. Fig. 5-(a) shows an under-stylized image that lacks content consideration. Fig. 5-(b) displays natural stylization with differentiated styles applied to various content regions, though still slightly lacking in depth. Fig. 5-(c) confirms the effective compensation for this deficiency. Fig. 5-(d) verifies the preservation of overall content across the image, and Fig. 5-(e) demonstrates the precise adjustment of additional content details. These visual results confirm that our proposed methods function according to design specifications.

5.6 Number of cluster

The Content-consistent Style Injection module plays a vital role in enabling region-aware style transfer. To assess how the number of clusters impacts this module, we conduct an additional ablation study. The results, detailed in Tab. 4-(Left), demonstrate that the number of clusters significantly influences both the degree of stylization, as measured by FID, and content preservation, as assessed

Table 3. Ablation study on the initial AdaIN, Content-consistent Style Injection (CSI), Adaptive Style Refinement (ASR), Residual Feature Interpolation (RFI) and Content Distribution Alignment (CDA).

#	Component					Quantitative Metrics		
	init AdaIN	CSI	ASR	RFI	CDA	ArtFID ↓	FID ↓	LPIPS ↓
(a)	✓					31.797	19.909	0.5207
(b)	✓	✓				29.698	18.553	<u>0.5188</u>
(c)	✓	✓	✓			29.546	**16.875**	0.6530
(d)	✓	✓	✓	✓		<u>28.372</u>	<u>17.578</u>	0.5271
(e)	✓	✓	✓	✓	✓	**28.370**	17.788	**0.5100**

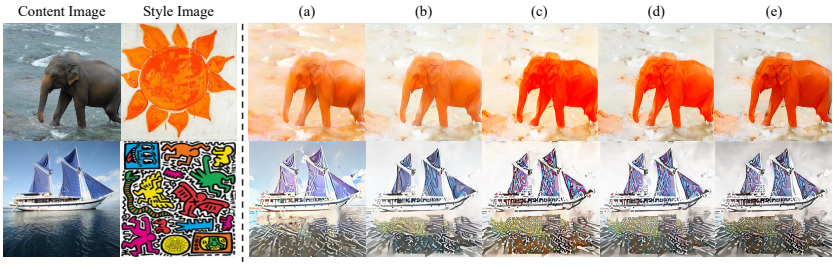

Fig. 5. Qualitative results showcasing the impact of applying/omitting the proposed modules. The results from (a)-(e) correspond to the configurations in Tab. 3

by LPIPS. These findings led us to conduct all subsequent experiments using 22 clusters, a configuration that yielded the most balanced results in style transfer.

5.7 Interpolation weight

To preserve content information that might be lost during the Content-consistent Style Injection and Adaptive Style Refinement phases, we implement interpolation within the residual blocks. To evaluate the effects of different interpolation weights on our model's performance, we conduct further experiments as detailed in Tab. 4-(Right). These results highlight the significant impact of interpolation weight on both the degree of stylization and content preservation. A lower interpolation weight ($\alpha = 0.1$) leads to excellent content preservation as indicated by LPIPS, but it also results in insufficient stylization, as reflected by higher FID scores. Conversely, increasing the interpolation weight enhances stylization at the cost of content preservation. This illustrates a clear trade-off between the two metrics. Based on our findings, we selected an interpolation weight of 0.87 for all experiments, as it offers the most balanced results between stylization and content preservation.

Table 4. Ablation study on the number of clusters (Left) and the interpolation weight (Right). The symbol ↓ indicates lower values are better.

# of Clusters	ArtFID ↓	FID ↓	LPIPS ↓
4	28.508	**17.698**	0.5247
8	28.444	17.775	0.5150
12	28.480	17.833	0.5122
16	28.547	17.883	0.5118
20	28.596	17.927	0.5109
22	**28.370**	17.788	0.5100
24	28.432	17.834	0.5097
28	28.424	17.831	**0.5094**

α	ArtFID ↓	FID ↓	LPIPS ↓
0.1	33.925	26.839	**0.2186**
0.3	33.241	25.856	0.2377
0.5	31.990	23.939	0.2827
0.7	29.822	20.640	0.3781
0.87	**28.370**	17.788	0.5100
0.9	28.398	**17.473**	0.5373

6 Conclusion

In this paper, we present the Content-Adaptive Style Transfer (CAST), a pioneering training-free approach for arbitrary style transfer that capitalizes on vector quantization-based models. This approach includes several innovative techniques. We introduce the Content-consistent Style Injection technique to ensure precise stylization that is tailored to the unique regions of the content image, facilitating a natural integration of style elements. Additionally, we propose the Adaptive Style Refinement technique, which fine-tunes stylization based on the nuanced differences between the style and stylized image features. We also design the Content Refinement technique, which effectively preserves the content information of the stylized image, maintaining the integrity of the original content. Our experimental results demonstrate that CAST achieves state-of-the-art performance in both qualitative and quantitative evaluations compared to existing generative and conventional style transfer models. Looking ahead, there is potential to extend our proposed method to accommodate various input style prompts, such as text, broadening the applicability of our approach to a wider array of artistic and design purposes. This future direction promises to further enhance the capabilities of style transfer technology, making it more versatile and accessible for users seeking to creatively combine textual descriptions with visual content.

Acknowledgments
This work was supported by the 2024 innovation base artificial intelligence data convergence project project with the funding of the 2024 government (Ministry of Science and ICT) (S2201-24-1002) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00210908).

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 4
2. Cao, S., Yin, Y., Huang, L., Liu, Y., Zhao, X., Zhao, D., Huang, K.: Efficient-vqgan: Towards high-resolution image generation with efficient vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7368–7377 (2023) 3
3. Chandran, P., Zoss, G., Gotardo, P., Gross, M., Bradley, D.: Adaptive convolutions for structure-aware style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7972–7981 (2021) 9
4. Chang, H.Y., Wang, Z., Chuang, Y.Y.: Domain-specific mappings for generative adversarial style transfer (2020) 2, 4
5. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023) 2, 3
6. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022) 2, 3
7. Chen, H., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D., et al.: Artistic style transfer with internal-external learning and contrastive learning. Advances in Neural Information Processing Systems **34**, 26561–26573 (2021) 2
8. Chen, Y.J., Cheng, S.I., Chiu, W.C., Tseng, H.Y., Lee, H.Y.: Vector quantized image-to-image translation. In: European Conference on Computer Vision. pp. 440–456. Springer (2022) 2, 4
9. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer (2024) 2, 4, 6, 9, 10, 12
10. Couairon, G., Grechka, A., Verbeek, J., Schwenk, H., Cord, M.: Flexit: Towards flexible semantic image translation (2022) 4
11. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: Vqgan-clip: Open domain image generation and editing with natural language guidance (2022) 2, 3
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 9
13. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr^2: Image style transfer with transformers (2022) 2, 9
14. Deng, Y., Tang, F., Dong, W., Sun, W., Huang, F., Xu, C.: Arbitrary style transfer via multi-adaptation network (2020) 2, 4, 9
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019) 3
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021) 2, 4
17. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016) 2

18. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 2, 3, 4
19. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 2, 4, 9
20. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014) 2, 3, 4
21. Gu, S., Chen, C., Liao, J., Yuan, L.: Arbitrary style transfer with deep feature reshuffle. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8222–8231 (2018) 2
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020) 2, 4
23. Hong, K., Jeon, S., Lee, J., Ahn, N., Kim, K., Lee, P., Kim, D., Uh, Y., Byun, H.: Aespa-net: Aesthetic pattern-aware style transfer networks (2023) 2, 4, 9
24. Huang, S., An, J., Wei, D., Luo, J., Pfister, H.: Quantart: Quantizing image style transfer towards high visual fidelity (2023) 2, 4, 10
25. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 2, 4, 6, 7, 9
26. Jeong, J., Kwon, M., Uh, Y.: Training-free content injection using h-space in diffusion models (2024) 2, 4, 9
27. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution (2016) 2
28. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., Cheng, Y., Chiu, M.C., Dillon, J., Essa, I., Gupta, A., Hahn, M., Hauth, A., Hendon, D., Martinez, A., Minnen, D., Ross, D., Schindler, G., Sirotenko, M., Sohn, K., Somandepalli, K., Wang, H., Yan, J., Yang, M.H., Yang, X., Seybold, B., Jiang, L.: Videopoet: A large language model for zero-shot video generation (2023) 2, 3
29. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation (2023) 2, 4, 9
30. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Diversified texture synthesis with feed-forward networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3920–3928 (2017) 2
31. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms (2017) 4
32. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A closed-form solution to photorealistic image stylization (2018) 4
33. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015) 9
34. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer (2021) 2, 4, 9
35. OpenAI: Gpt-4 technical report (2023) 3
36. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks (2019) 2, 4
37. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023) 3
38. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation (2021) 3

39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022) 4
40. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., Hao, Y., Essa, I., Rubinstein, M., Krishnan, D.: Styledrop: Text-to-image generation in any style (2023) 2, 4
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022) 2
42. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork (2018) 9
43. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction (2024) 2, 3, 4, 5, 9
44. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation (2022) 4, 7
45. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.: Texture networks: Feed-forward synthesis of textures and stylized images. arXiv preprint arXiv:1603.03417 (2016) 2
46. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 2
47. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 2, 3
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023) 2, 4
49. Wright, M., Ommer, B.: Artfid: Quantitative evaluation of neural style transfer (2022) 9
50. Xu, W., Long, C., Wang, R., Wang, G.: Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer (2021) 2, 4
51. Xu, Z., Sangineto, E., Sebe, N.: Stylerdalle: Language-guided style transfer using a vector-quantized tokenizer of a large-scale generative model (2023) 2, 4
52. Yang, S., Hwang, H., Ye, J.C.: Zero-shot contrastive loss for text-guided diffusion image style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22873–22882 (2023) 2, 4
53. Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y.J., Wang, J.: Attention-aware multi-stroke style transfer (2019) 2, 4
54. Yu, L., Cheng, Y., Wang, Z., Kumar, V., Macherey, W., Huang, Y., Ross, D.A., Essa, I., Bisk, Y., Yang, M.H., Murphy, K., Hauptmann, A.G., Jiang, L.: Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms (2023) 3
55. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., Gong, B., Yang, M.H., Essa, I., Ross, D.A., Jiang, L.: Language model beats diffusion – tokenizer is key to visual generation (2023) 3
56. Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization (2022) 9
57. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models (2023) 2, 9
58. Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–8 (2022) 2
59. Zhu, L., Wei, F., Lu, Y.: Beyond text: Frozen large language models in visual signal comprehension (2024) 3