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

1 Comparison of our network structure with the baseline network

To facilitate a better understanding how our proposed methods, we visualize the model structure of the baseline VAR with our CAST network as illustrated in Fig. 1. Both the baseline decoder and CAST are initially structured with pairs of residual blocks and self-attention blocks. Following N operations, upsampling is performed, succeeded by another set of N residual block operations.

In CAST, we enhance the baseline's residual and self-attention blocks with additional computations. During the N residual blocks, interpolation is incorporated to preserve content information (RFI). In the self-attention block, we introduce a cross-attention operation, where the content image's features serve as the query, and the style image's features act as the key and value, injecting style information (CSI). Simultaneously, style refinement process is also applied using the adaptive weight, which is calculated based on the difference between the style features and the stylized features (ASR). After upsampling, we integrate the Pixel-Wise Difference Preservation module (CDA) into the N residual blocks to further refine the content information. The subsequent operations after the following upsampling layers are carried out similarly to the baseline.

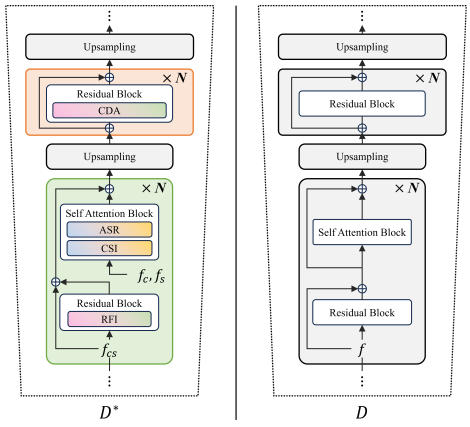

Fig. 1: Comparison of our decoder structure D^* with baseline's decoder structure D.

2 Additional ablation study

Our ablation study highlights the differential impacts of the interpolation weight α than the number of clusters on the style transfer results. As shown in Fig. 2 (Top), a lower interpolation weight ($\alpha = 0.1$) preserves a closer resemblance to the original image, as evidenced by a decrease in LPIPS. Conversely, increasing the weight to ($\alpha = 0.9$) enhances the infusion of style elements, leading to a decrease in ArtFID. In terms of the number of clusters, especially as the number increases, the difference in the overall image quality is trivial as seen in Fig. 2 (Bottom). However, a very low number of clusters, such as 4, can result in poorly defined boundaries within the image. This can adversely affect the clarity and distinction of different regions, which is critical for achieving high-quality style transfers.

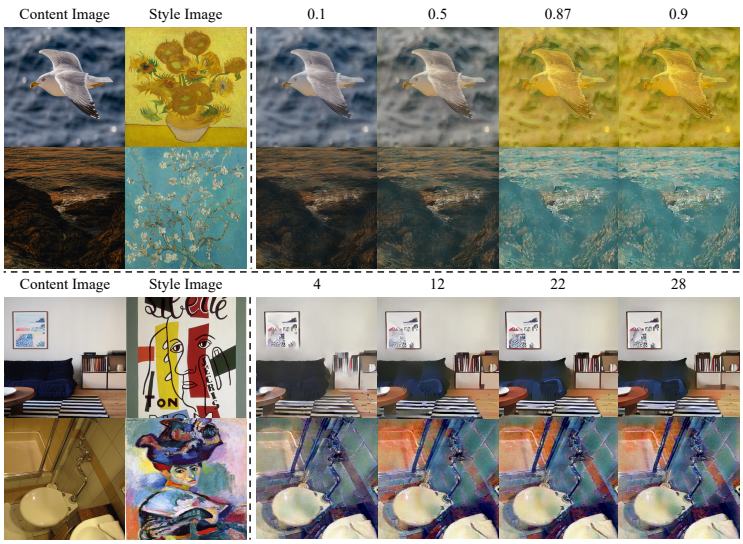

Fig. 2: Comparison of our decoder structure D^* with baseline's decoder structure D.

3 Additional comparison

In Tab. 1, we conduct further comparisons with AesPA-Net [2] and StyleID [1] which had shown the best performance in the quantitative comparisons of the main paper. For these evaluations, we use a diverse set of images, randomly sampling 20 different content images from the MS-COCO dataset [4] and 40 different style images from the Wikiart dataset [6]. This set is different from the one evaluated in the main paper, providing a broader basis for assessing the

robustness and effectiveness of our approach compared to these state-of-the-art methods.

Table 1: Quantitative comparison of our method with state-of-the-art generative model-based style transfer model [1] and traditional style transfer model [2]

Metric	**Ours**	StyleID	AesPA-Net
ArtFID ↓	**31.976**	33.100	37.700
FID ↓	**19.850**	20.146	22.882
LPIPS ↓	**0.533**	0.5653	0.5786

4 Comparison with state-of-the-art text-driven style transfer models

With the advent of the CLIP model [5], text-driven style transfer has become a burgeoning field within style transfer. This approach enables style transfer solely through textual descriptions, thus removing the need for reference style images. We evaluate our model against prominent text-driven style transfer models such as CLIPstyler [3], StylerDALLE [8], and ZeCon [9], as illustrated in Fig. 3. These models use text descriptions that correspond to famous paintings for style transfer. Despite producing visually striking results, our comparative analysis indicates that these text-driven models often fall short in faithfully capturing and reflecting the intended style as described by text alone. This limitation underscores the inherent challenges of conducting style transfer without tangible visual references, where the specificity and depth of a style image play crucial roles in the accuracy and quality of the style transformation.

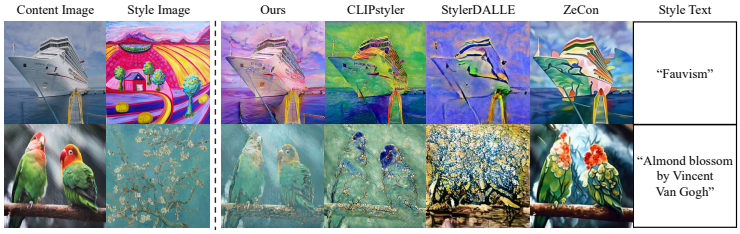

Fig. 3: Quantitative comparison with state-of-the-art text-driven style transfer models

5 Limitations of CAST

We have successfully achieved style transfer by using our proposed decoder D^*, but there are inherent limitations. First, our approach is designed based on the VAR model [7], a multi-scale quantization autoencoder. This dependency inherently ties the quality of our style transfer to the image reconstruction capabilities of the VAR autoencoder. As shown in Fig. 4 (Top), particularly when trained on datasets like ImageNet, exhibits difficulties in reconstructing images from out-of-distribution domains that require fine details, such as images containing small texts. Additionally, the manipulation of embedded image features in our method can lead to the loss of fine details, such as text and intricate figures, thus compromising the fidelity of the style transfer results. These issues can be mitigated with an improved VQ autoencoder or extensive datasets. Second, our method does not allow users to explicitly specify which parts of the content image should adopt certain styles. This limitation leads to unintended stylistic effects, such as text inadvertently turning red, as shown in Fig. 4 (Bottom). This occurs when transforming a content image to resemble a style image without the ability to target specific regions for style application.

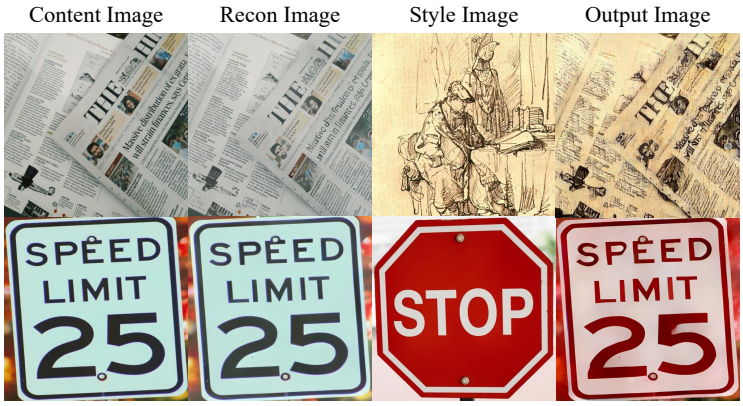

Fig. 4: Failure cases of CAST.

6 User study

We conducted a user study involving 38 participants ranging from their 20s to 50s. The study consisted of 10 questions comparing the stylization results of our method with those of StyleID and AesPA-Net, both of which are known for their competitive performance in qualitative evaluations. As shown in Tab. 2, the results confirms that our model was the preferred choice, receiving the highest preference rate, compared StyleID and AesPA-Net. These findings suggest that our approach is generally favored over the other models, with StyleID showing the closest competition in quantitative results.

Table 2: User preference results.

	Ours	StyleID	AesPA-Net
Preferences (%)	**39.5**	36.6	23.9

7 Various qualitative results

We present additional stylization results of our CAST method in Fig. 6. These results demonstrate that our method effectively transfers style to image areas that correspond with the content region. Additionally, it preserves content information effectively, thereby achieving balanced style transfer results.

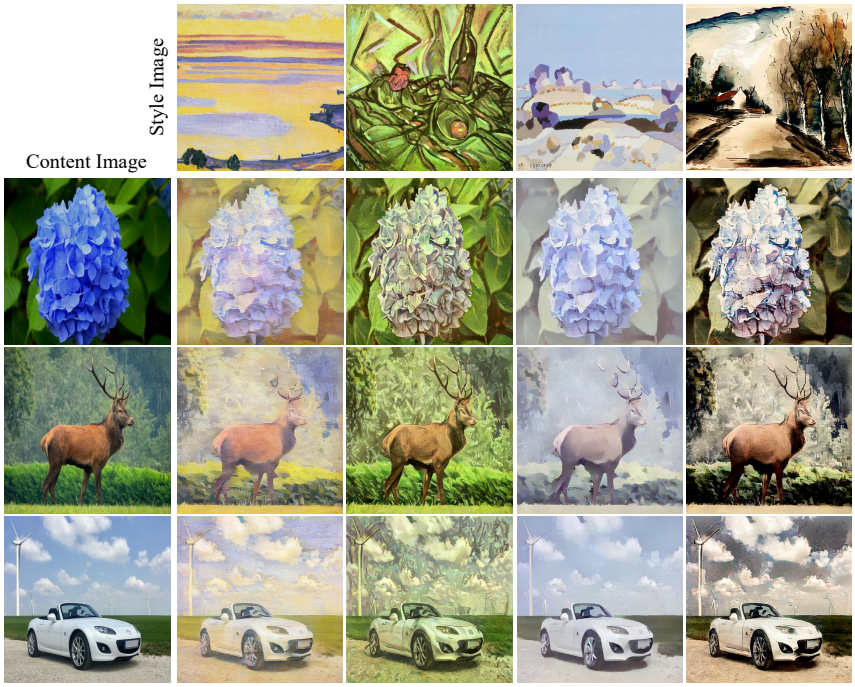

Fig. 5: Various style transfer results of style and content image pairs.

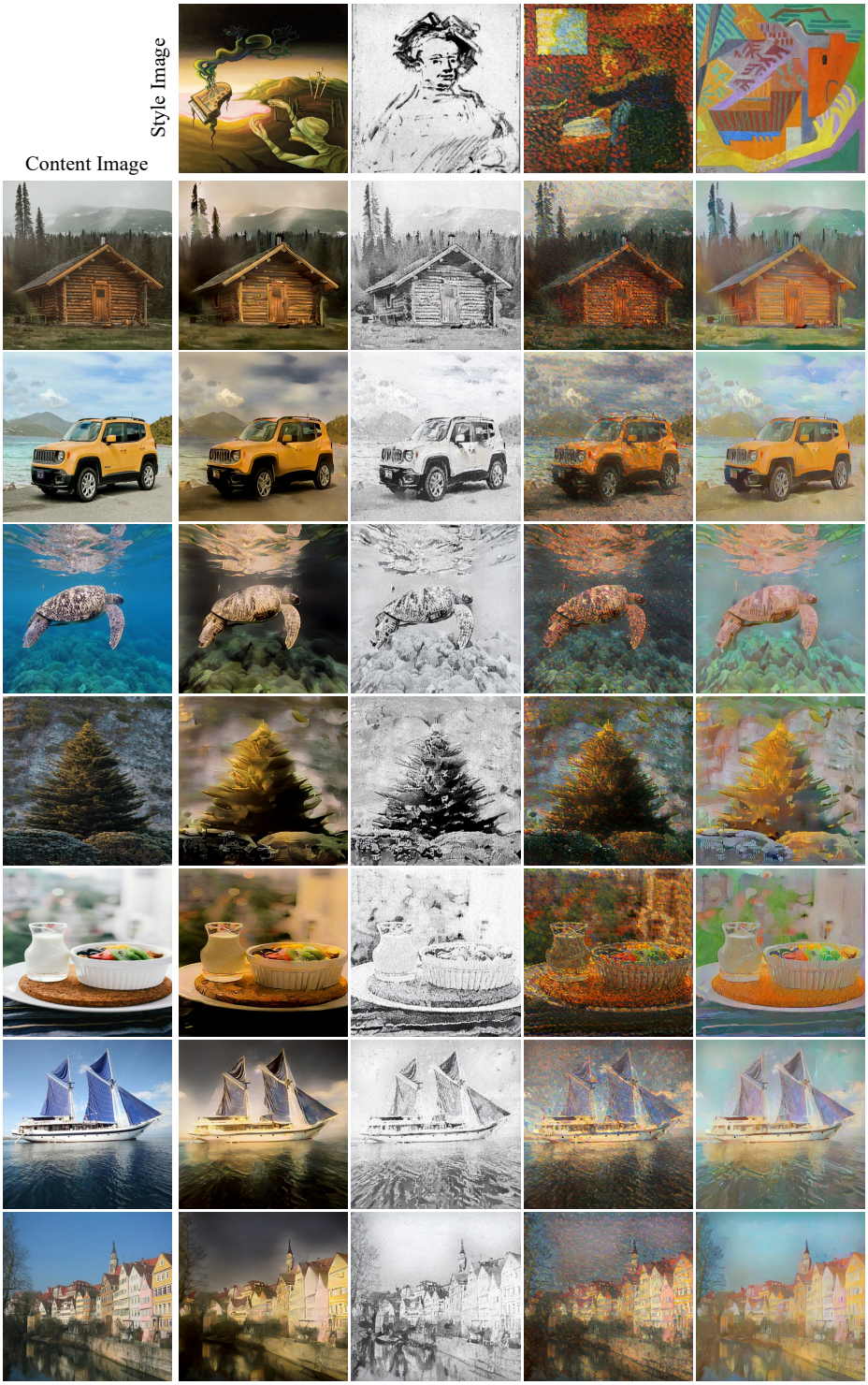

Fig. 6: Various style transfer results of style and content image pairs.

References

1. 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, 3
2. 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, 3
3. Kwon, G., Ye, J.C.: Clipstyler: Image style transfer with a single text condition (2022) 3
4. 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) 2
5. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 3
6. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork (2018) 2
7. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction (2024) 4
8. Xu, Z., Sangineto, E., Sebe, N.: Stylerdalle: Language-guided style transfer using a vector-quantized tokenizer of a large-scale generative model (2023) 3
9. 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) 3