Master's Thesis Presentation for the 143th Dissertation Evaluation

Blended Embedding Guided Style Transfer in Inversion-Based Diffusion

인버전 기반 확산 모델의 혼합 임베딩을 활용한 스타일 변환 연구

Sojeong Kim

sojeong6894@gmail.com

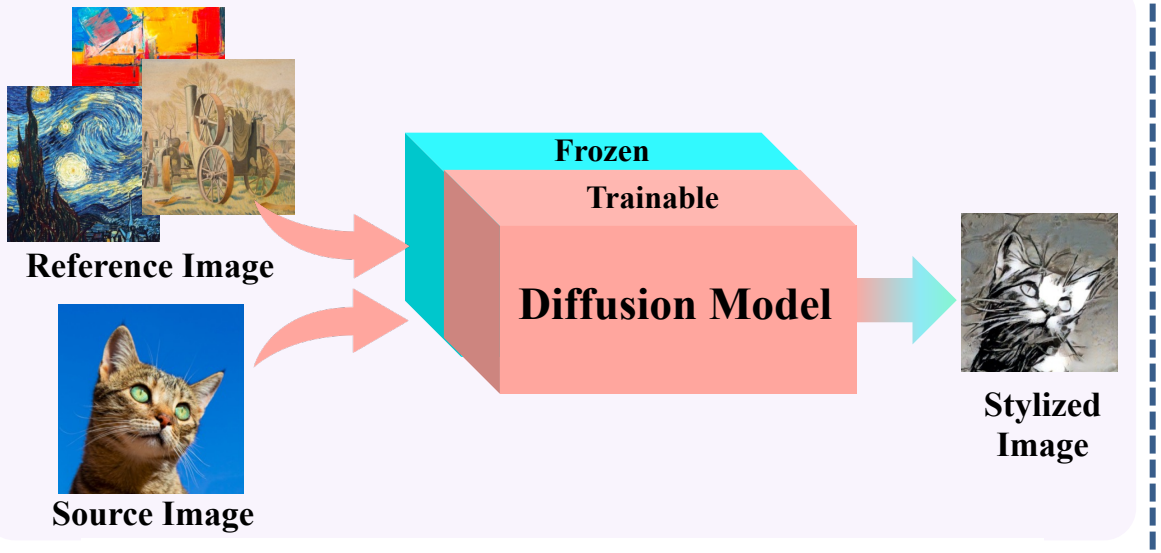
11th April, 2025

Contents

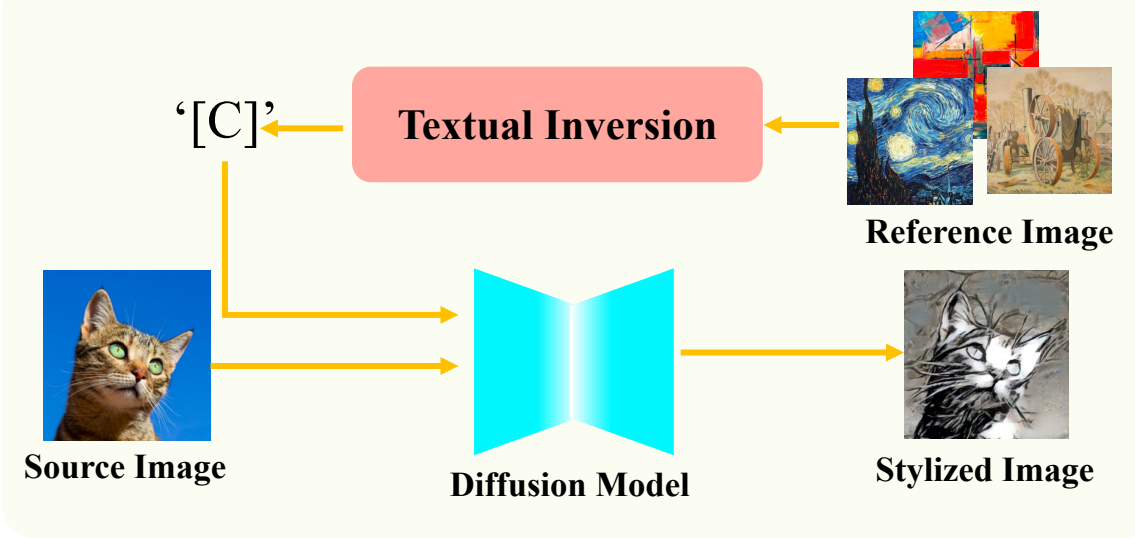
Introduction

Diffusion-based Image Style Transfer

- Diffusion-based image style transfer is increasingly used in applications such as personalized content creation, artistic rendering, and virtual try-on, providing users with intuitive and flexible control over visual styles without requiring expert-level design skills. [1].
- Diffusion-based image style transfer methods are typically categorized into finetune-based and inversion-based approaches, where the former retrains the model with style images and the latter guides stylization using pretrained models with text or image prompts [2].



[Figure 1] Image Style Transfer with Finetune-based methods



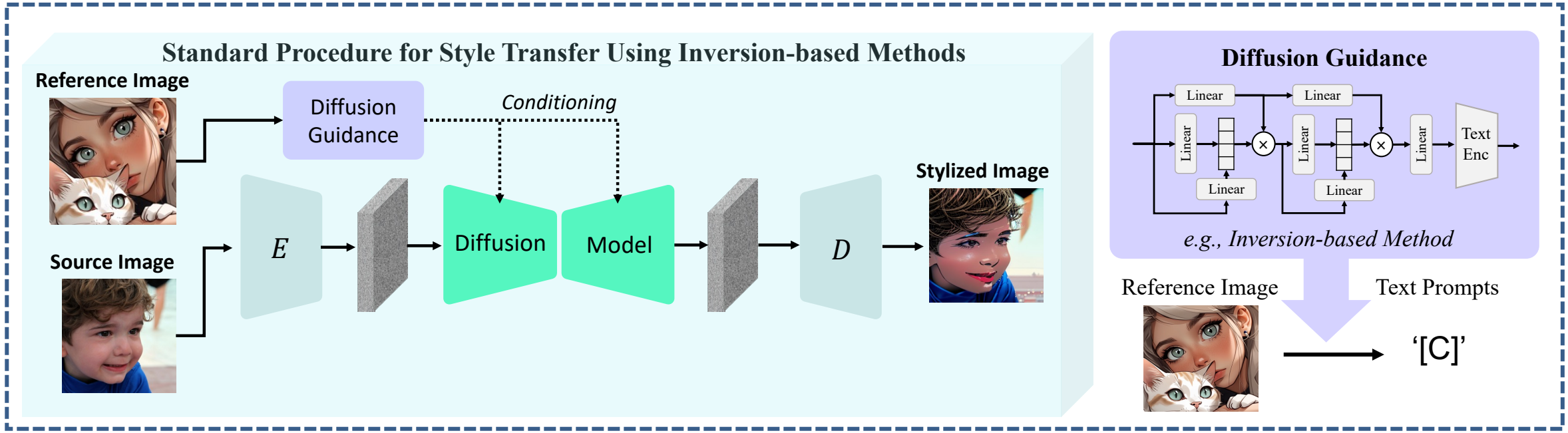
[Figure 2] Image Style Transfer with Inversion-based methods

[1] Liu, WuQin, et al. "FreeStyler: A Free-Form Stylization Method via Multimodal Vector Quantization." International Conference on Computational Visual Media. Singapore: Springer Nature Singapore, 2024.
[2] Qi, Tianhao, et al. "Deadiff: An efficient stylization diffusion model with disentangled representations." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Introduction

Objective and Standard Procedure

- Diffusion-based image style transfer aims to generate high-quality stylized images that preserve input content while reflecting the desired style, with inversion-based methods guiding the process using text embeddings derived from reference images [3].
- Conventional inversion-based methods encode the input into latent space, guide the diffusion process using text embeddings derived from reference images—often through textual inversion—and decode the result into a stylized output via a pre-trained diffusion model [4, 5].



[3] Chung, Jiwoo, Sangeek Hyun, and Jae-Pil Heo. "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.
[4] Zhang, Yuxin, et al. "Inversion-based style transfer with diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
[5] Zhang, Yuxin, et al. "Prospect: Prompt spectrum for attribute-aware personalization of diffusion models." ACM Transactions on Graphics (TOG) 42.6 (2023): 1-14.

Introduction

Problem and Strategy

- Existing inversion-based method [4] often struggle to preserve object identity and apply the reference style consistently, mainly because v isual and textual embeddings from source and reference images are applied or guided separately during the generation process.
- We propose a novel framework that blends visual and textual embeddings from both source and reference images, enabling diffusion mod els to adaptively apply the reference style while preserving the identity of the source image resulting in a more harmonious stylization.

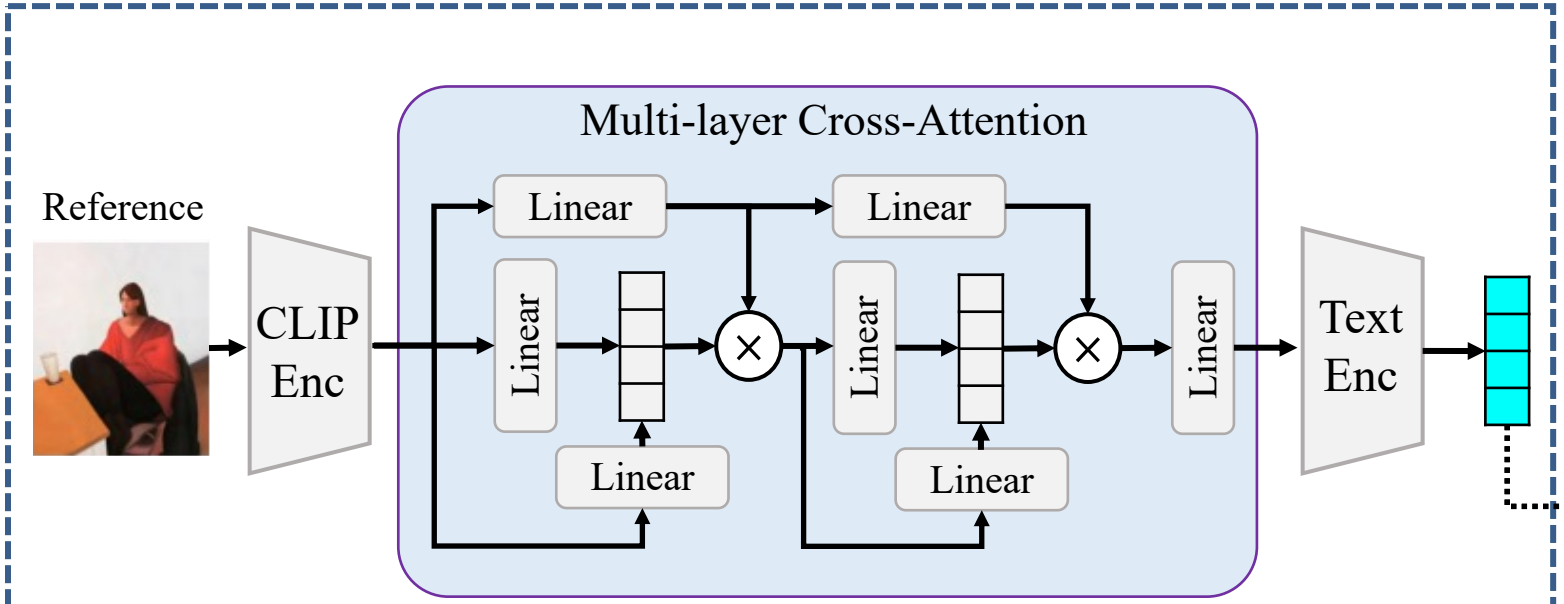
[Figure 4] Stylization Results under Different Style Guidance Levels for Source-Reference Pair Images

[4] Zhang, Yuxin, et al. "Inversion-based style transfer with diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

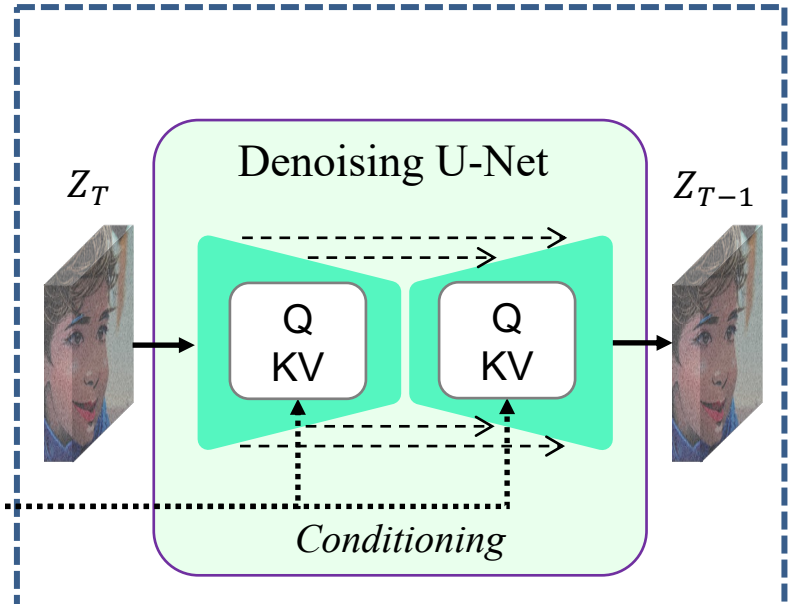
Related Work

InST [4]

- This study proposes an image style transfer model that effectively extracts key style information from artistic images by applying a multi-layer cross-attention mechanism to the image embeddings obtained from a CLIP encoder.
- To preserve the semantics of the source image, a stochastic inversion method is introduced, where random noise is first added to the source image, and the predicted noise from the denoising U-Net is reused as the initial input during generation to maintain content consistency.



[Figure 5] Textual Inversion Architecture for Style Image Embedding



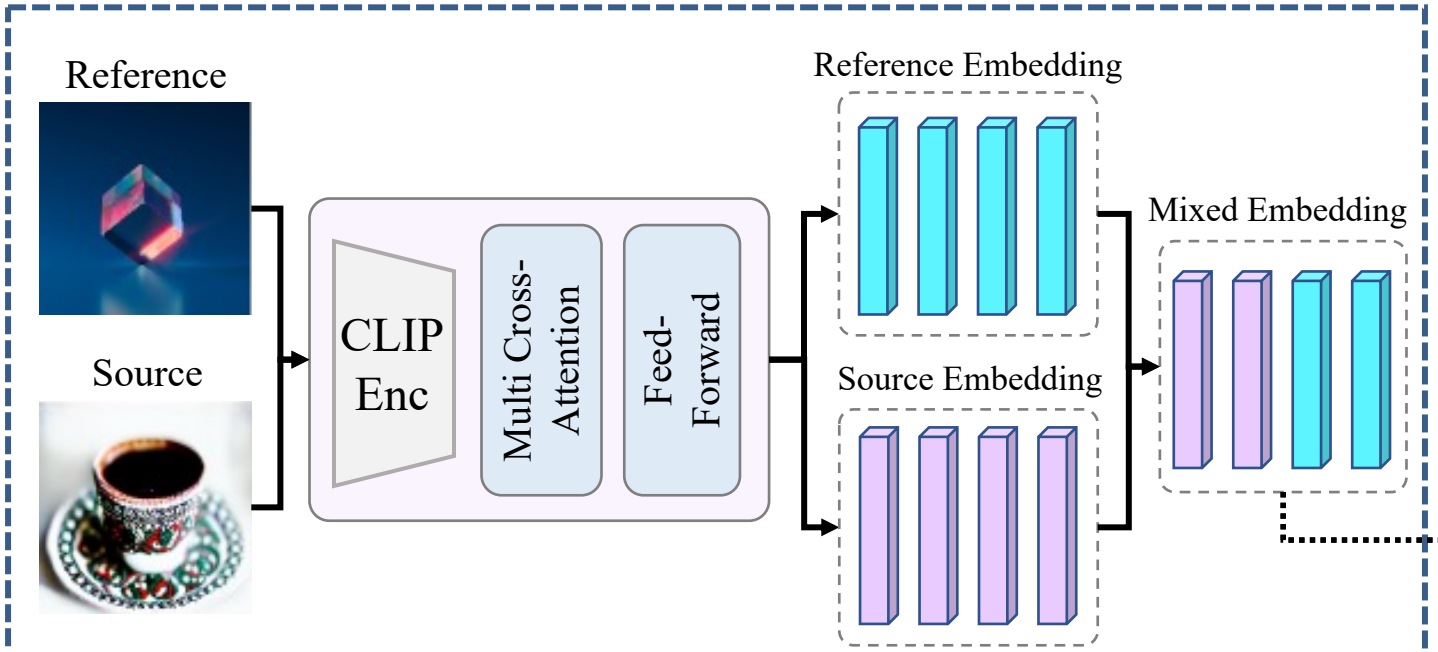
[Figure 6] Denoising Process in Latent Space

[4] Zhang, Yuxin, et al. "Inversion-based style transfer with diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

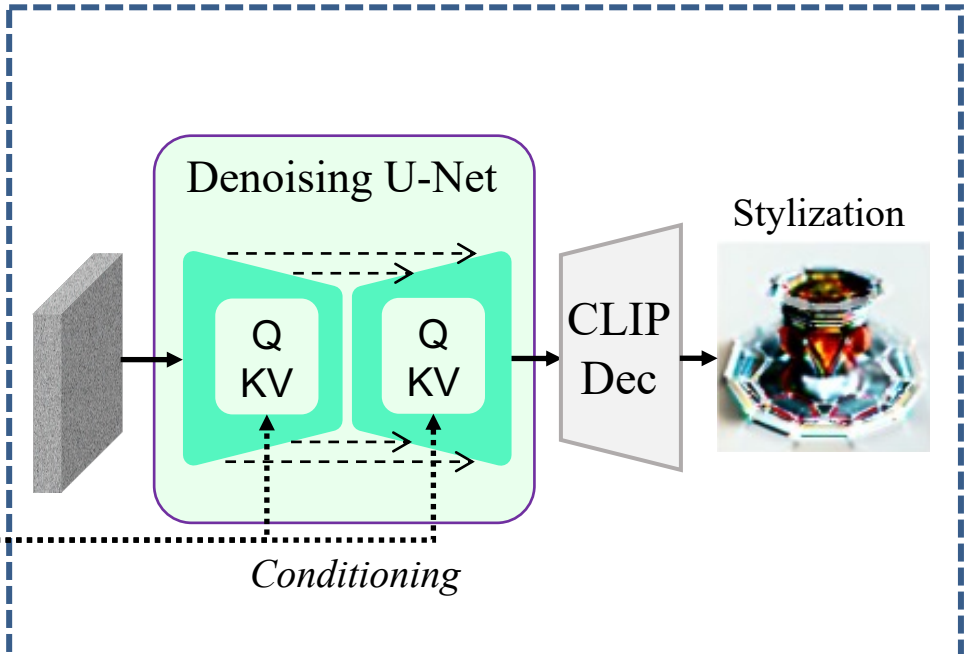
Related Work

ProSpect [5]

- This study proposes Prompt Spectrum (ProSpect), a novel image representation and manipulation framework that disentangles visual attributes from a single image by leveraging a newly introduced Prompt Spectrum Space (P*).
- Prompt Spectrum Space (P*) is proposed to enable stage-wise control of the generation process in diffusion models by dividing the 1000 denoising steps into multiple stages, each conditioned by a distinct textual embedding in the CLIP text-image space.



[Figure 7] Pipeline of ProSpect with Stage-wise Token Embedding Fusion



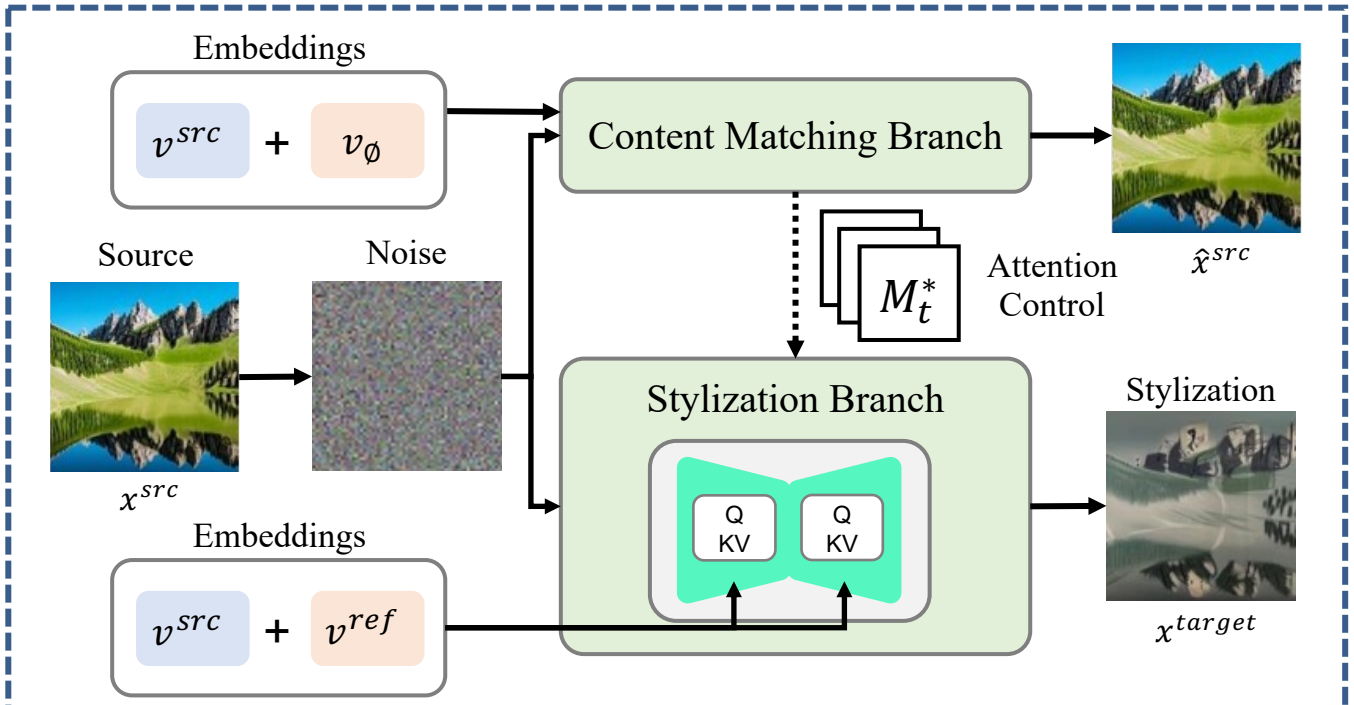
[Figure 8] Image Stylization Pipeline

[5] Zhang, Yuxin, et al. "Prospect: Prompt spectrum for attribute-aware personalization of diffusion models." ACM Transactions on Graphics (TOG) 42.6 (2023): 1-14.

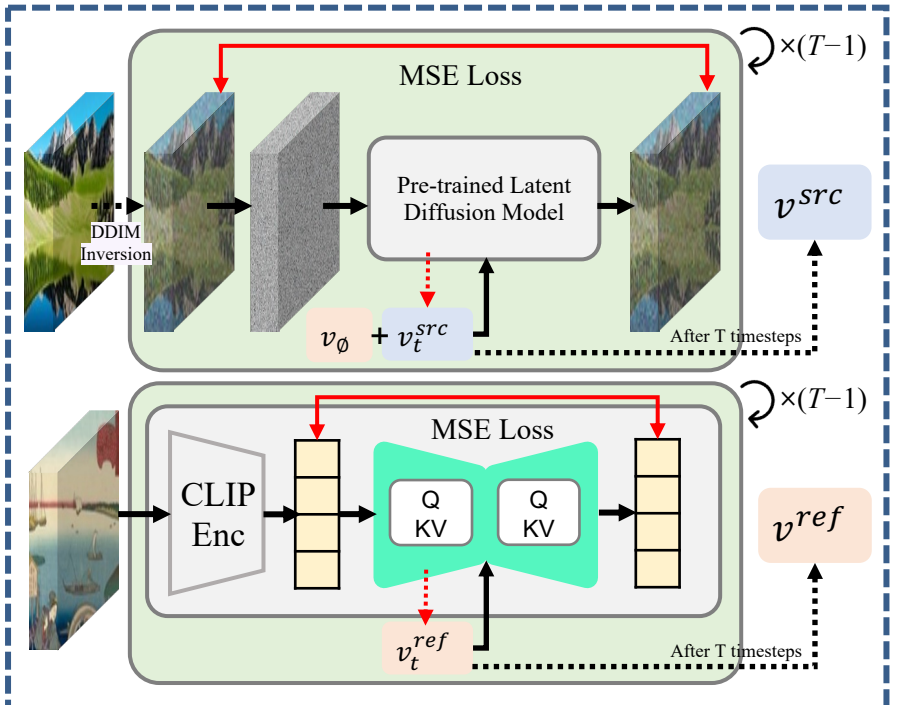
Related Work

VCT [6]

- This paper proposes a method that generates optimized embeddings(v^{src}, v^{ref}) from both a source image and a reference image, and fuses them to effectively guide the stylization process.
- The source embedding(v^{src}) is optimized at each timestep through unconditional DDIM inversion to reconstruct the source image, while the reference embedding(v^{ref}) is learned using a multi-concept strategy to capture the semantic style of the reference image.



[Figure 9] The Overall Visual Concept Translator(VCT) Framework



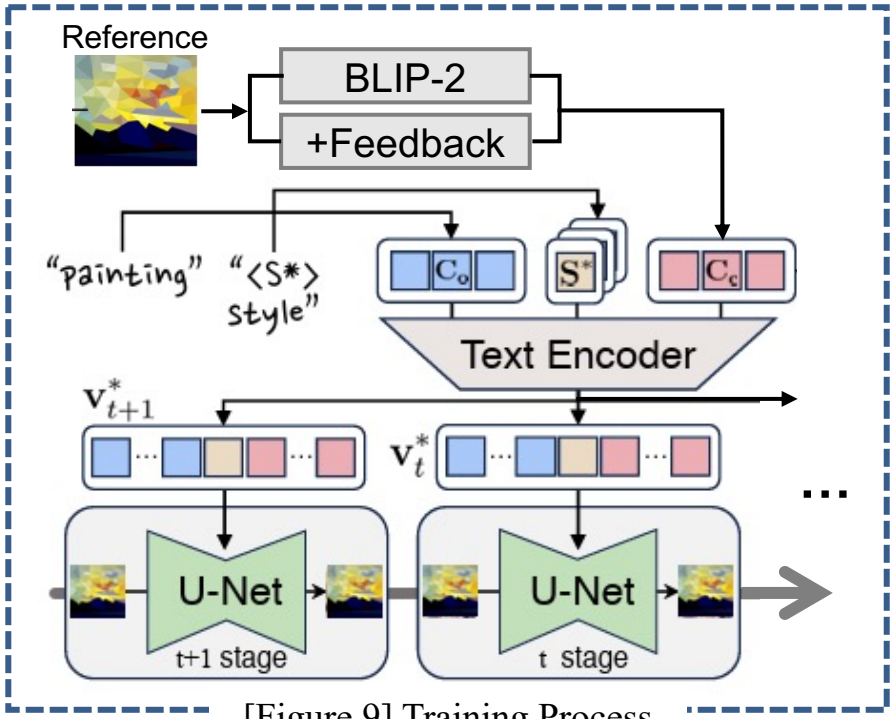
[Figure 10] A Framework of Embeddings

[6] Cheng, Bin, et al. "General image-to-image translation with one-shot image guidance." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

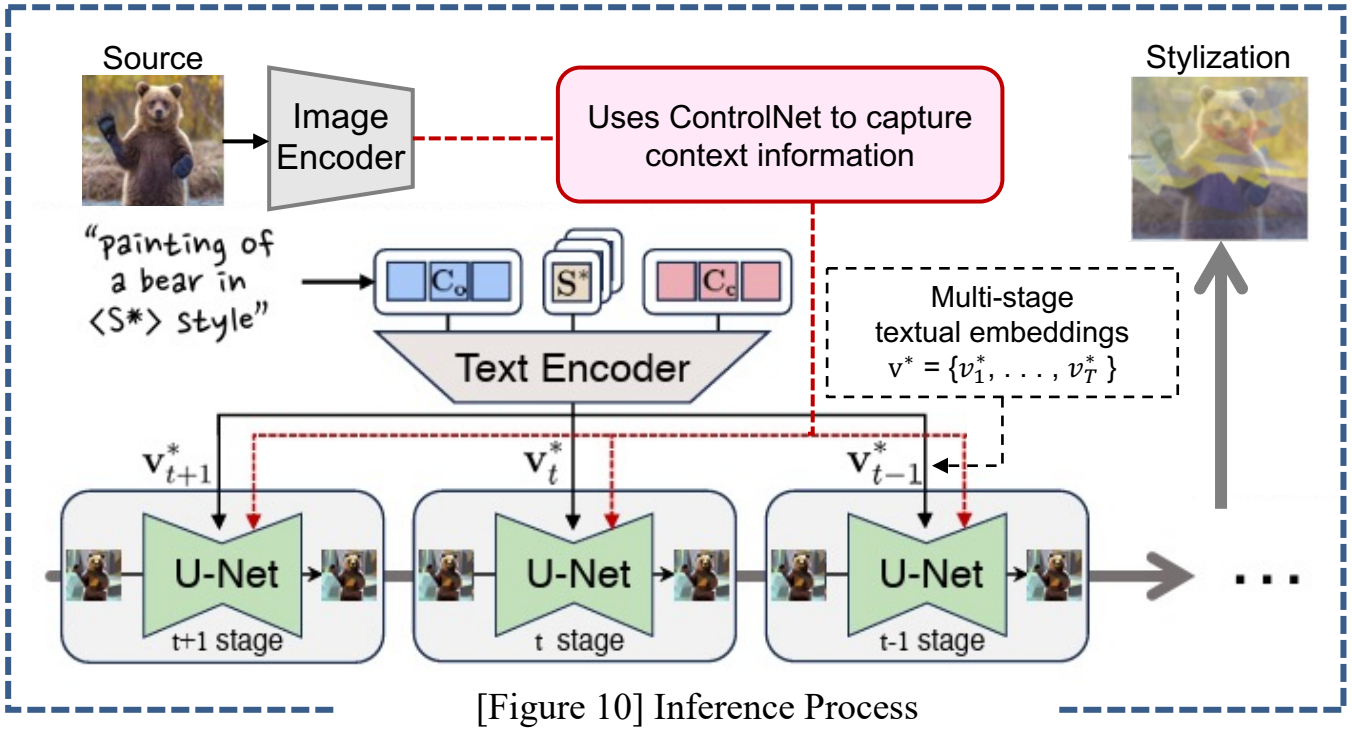
Related Work

Dreamstyler [7]

- This research constructs a training prompt consisting of an opening text, multi-stage style tokens, and a context description generated by BLIP-2 and human feedback, and projects it into stage-wise textual embeddings aligned with denoising timesteps.
- During inference, it encodes the given prompt into multi-stage textual embeddings aligned with the denoising process, and uses ControlNet to capture contextual information from the source image, enabling more precise and visually consistent style transfer.



[Figure 9] Training Process



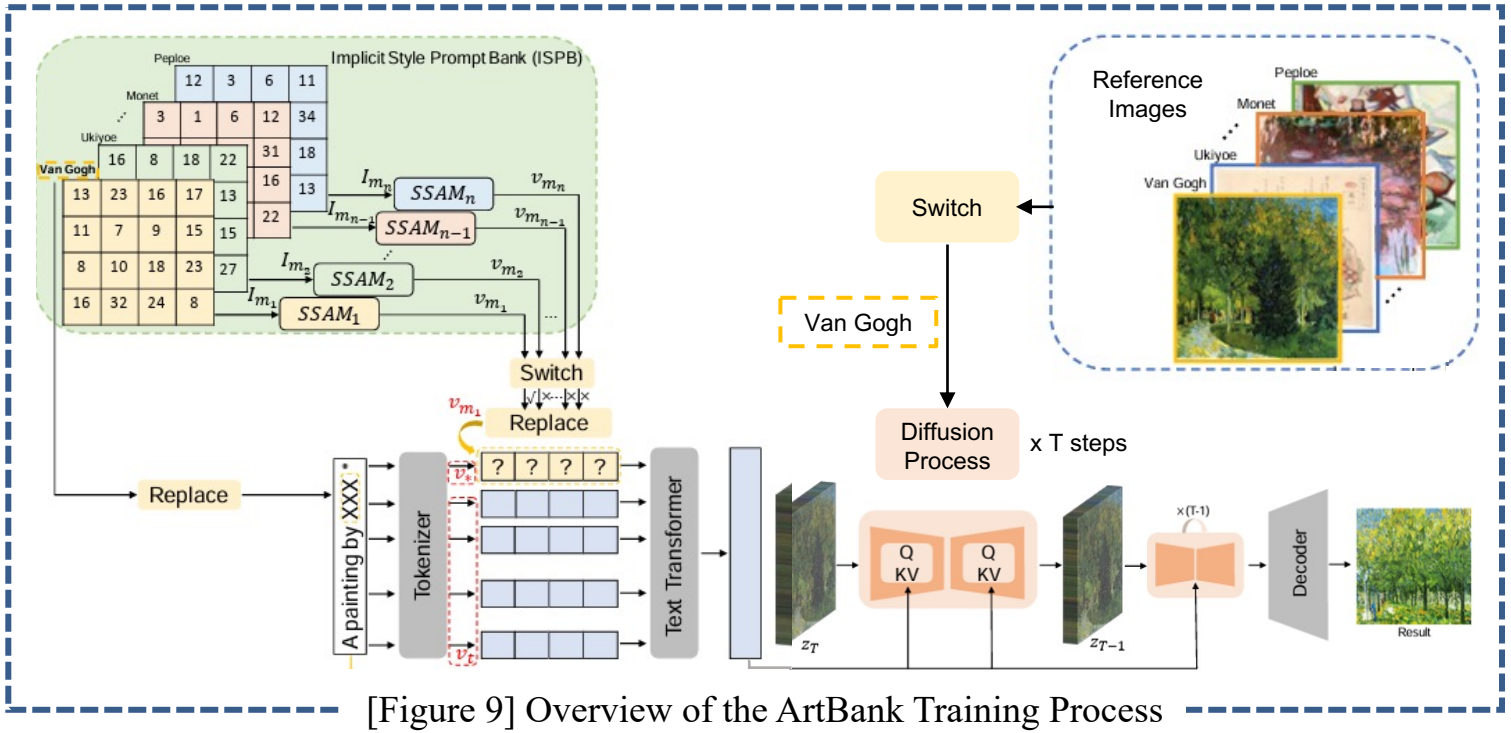
[Figure 10] Inference Process

[7] Ahn, Namhyuk, et al. "Dreamstyler: Paint by style inversion with text-to-image diffusion models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 2. 2024.

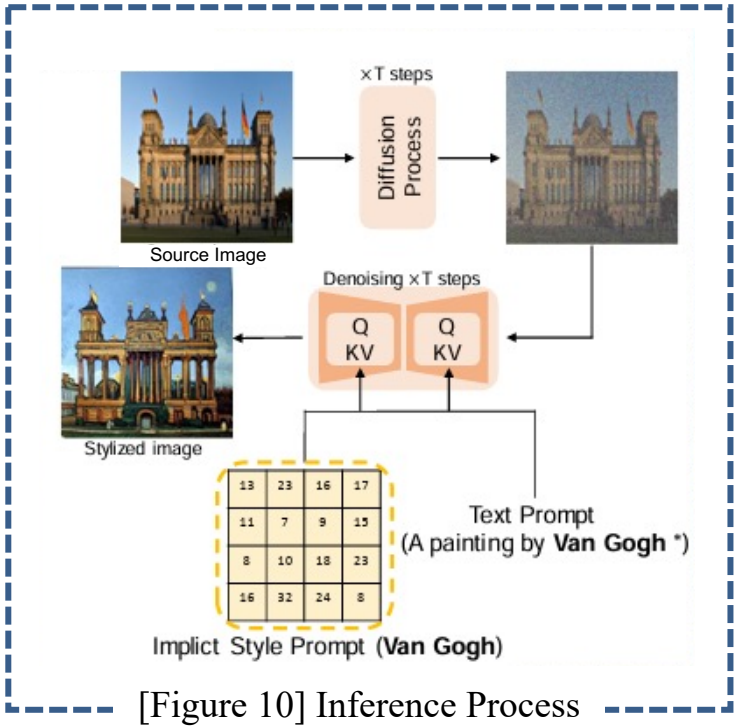
Related Work

ArtBank [8]

- This study proposes a novel framework that leverages the prior knowledge of a pretrained Stable Diffusion model and a trainable Implicit Style Prompt Bank (ISPB) to generate highly realistic stylized images while preserving content structure.
- To enhance training efficiency and expressiveness, a Spatial-Statistical Self-Attention Module (SSAM) is introduced to project learnable style parameters into the embedding space, enabling effective conditioning from diverse artistic styles.



[Figure 9] Overview of the ArtBank Training Process



[Figure 10] Inference Process

[8] Zhang, Zhanjie, et al. "Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 7. 2024.

Related Work

Common Drawback and Solution

- Existing methods struggle to effectively capture visual style information from reference images by relying solely on textual embeddings, often applying styles even to essential regions of the source image that should be preserved, resulting in visually disharmonious outputs.
- By blending visual embeddings from the reference image and textual embeddings from the source image into each type of embedding, the method enables diffusion models to apply styles adaptively while preserving source semantics, resulting in more harmonious outputs.

[Table 1] Counterpart Summary

Model	Author	Publication	Summary	Drawback
InST	Zhang, Yuxin, et al.	IEEE/CVF conference on computer vision and pattern recognition	• It extracts style information via multi-layer cross-attention on CLIP embeddings, and preserves content by reusing predicted noise.	• limitations in capturing visual characteristics through textual embeddings and lacks adaptation of style information to each source image.
ProSpect	Zhang, Yuxin, et al.	International Conference on Machine Learning	• It disentangles visual attributes from a single image by introducing a stage-wise Prompt Spectrum Space to guide diffusion generation with distinct textual embeddings.	• The visual style attributes for each embedding stage are not clearly defined, may vary by dataset, and often lead to severe distortion of the source image.
VCT	Cheng, Bin, et al.	IEEE/CVF international conference on computer vision	• It fuses timestep-wise optimized source embeddings and multi-concept reference embeddings to effectively guide the generation process.	• Limitations in capturing visual style information through textual embeddings in reference images, which lead to poor stylization.
Dreamstyler	Ahn, Namhyuk, et al.	AAAI Conference on Artificial Intelligence	• It generates stage-wise embeddings from a structured prompt and uses ControlNet to enable precise, consistent style transfer from the source image.	• Inadequate capture of visual style information from the reference image and distortion of the source image often result in poor stylization.
ArtBank	Zhang, Zhanjie, et al.	AAAI Conference on Artificial Intelligence	• It combines a pretrained diffusion model with a trainable style prompt bank and a self-attention module to efficiently generate realistic stylized images while preserving content structure.	• Failure to accurately extract style from arbitrary or rare reference images, combined with source image distortion, often leads to poor stylization.

Proposed Method

Notation and Definition

- Each data sample includes source and reference visual embeddings (x^{src}, x^{ref}) and their corresponding textual embeddings (v^{src}, v^{ref}) along with a noise map z_t and added Gaussian noise ϵ_t at timestep t.
- The CLIP encoder $E(I)$ extracts image embeddings, and blending is modulated by scaling vectors γ, and $\lambda \in [0, 1]$, allowing diffusion models to adaptively fuse style and content information for coherent stylization.

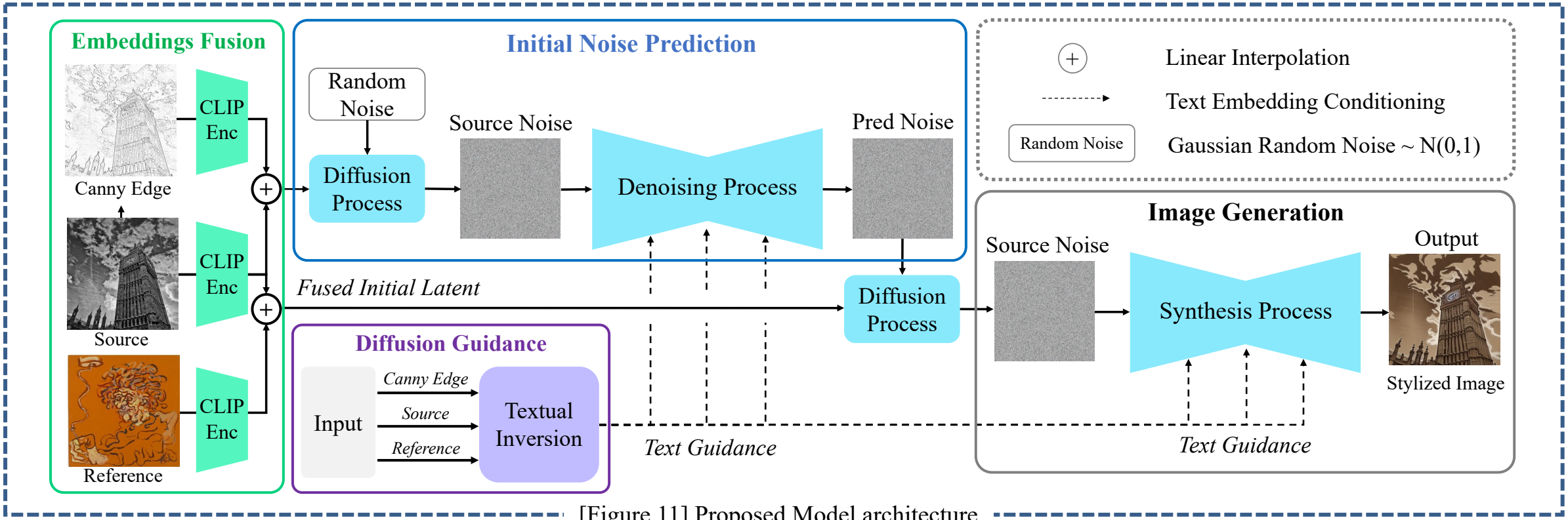
[Table 2] Notation Symbol Definition

Notation	Name	Definition
x^{src}	Visual embedding of source image	A feature representation extracted from the source image using a visual encoder.
x^{ref}	Visual embedding of reference image	A feature representation extracted from the reference image using a visual encoder.
v^{src}	Textual embedding of source image	A textual representation describing the source image, extracted via a text encoder.
v^{ref}	Textual embedding of reference image	A textual representation describing the reference image, extracted via a text encoder.
z_t	A noise map at timestep t	The noise map used by the diffusion model at timestep t.
ϵ_t	The added Gaussian noise at timestep t	Gaussian noise added to the image at diffusion timestep t during the forward process.
γ, λ	Scaling vectors	The scaling vectors in the range [0, 1], used to control the blending intensity between the two embeddings
$E(I)$	CLIP-based image embedding	Image embedding generated by the CLIP encoder

Proposed Method

Procedural Overview

- The proposed method consists of four main components: embeddings fusion, diffusion guidance, initial noise prediction, and image generation branches.
- Blended visual and textual embeddings from the source and reference images, obtained through the embedding fusion and diffusion guidance branches, guide the initial noise prediction and image generation stages, enabling style transfer with more harmonious results.

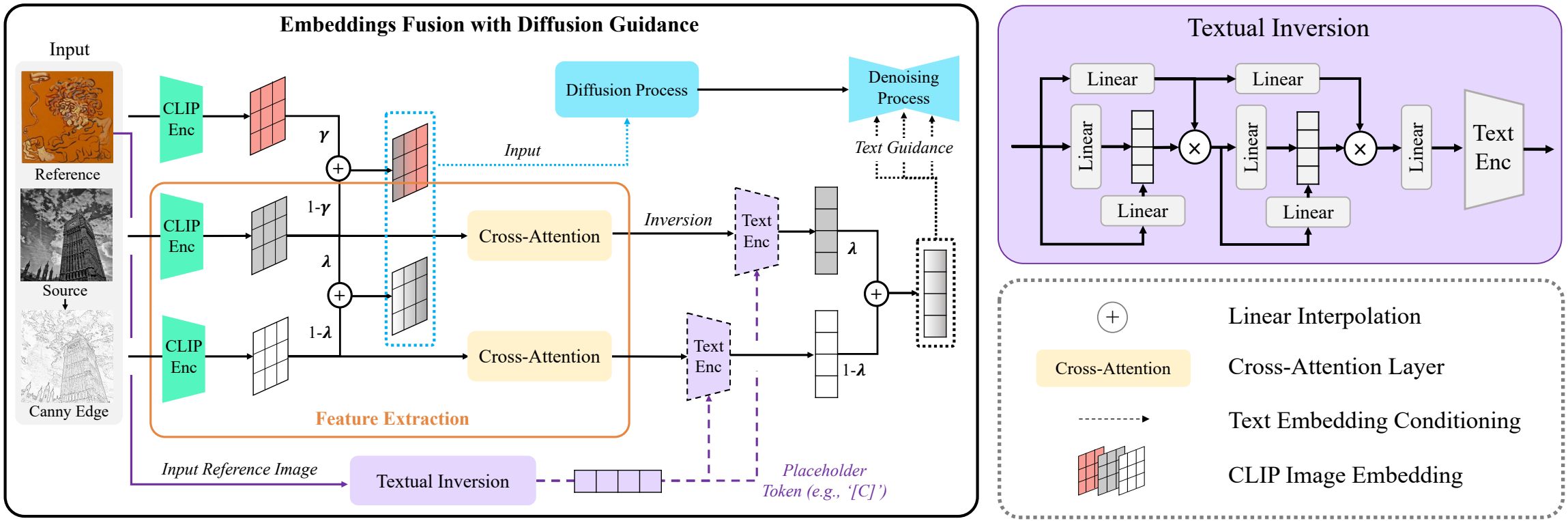


[Figure 11] Proposed Model architecture

Proposed Method

Details

- The embedding fusion process involves a series of linear interpolations between CLIP image and text embeddings derived from the reference image, source image, and Canny edge map, controlled by hyperparameters γ and λ.
- To derive textual embeddings, cross-attention and linear interpolation among the embeddings of the source image, Canny edge map, and reference image are used to generate a new conditioning value for the denoising process.



[Figure 12] Detailed Proposed Model Architecture

Experimental Settings

Dataset

- In the experimental setup, style textual embeddings are trained using 16 reference images randomly selected from the WikiArt dataset [9], with each training run utilizing only a single reference image.
- During the inference stage, stylized images are generated for 100 randomly selected images from the MS-COCO 2014 dataset [10] using each of the pre-trained style textual embeddings, resulting in a total of 1,600 stylized images.

[Table 3] Datasets Information

Dataset	Total Classes	Total Images
WikiArt [9]	27 styles	80,020
MS COCO 2014 [10]	80 classes	82,783 / 40,504 / 40,745 (Train / Valid / Test)

Experimental Settings

Hyperparameter Settings

[Table 4] Hyperparameter Settings

Model	Batch size	Max iteration	Style Guidance	Optimizer	Learning rate	Resolution
InST [4]	1	6100	0.5	Adam	1e-4	512 × 512
ProSpect [5]		1000				
VCT [6]		500	-		5e-4	
Dreamstyler [7]		500			2e-3	
ArtBank [8]		1000	0.5		1e-4	

Experimental Settings

Evaluation Measures

- *VGG Content Loss* [11]

$$\mathcal{L}_{content}(I_{gen}, I_{src}) = \frac{1}{C_l H_l W_l} \left\| \phi_l(I_{gen}) - \phi_l(I_{src}) \right\|_2^2$$

  - I_{gen}: Generated image (stylized image)
  - I_{src}: Source image
  - $\phi_l(\cdot)$: Feature map extracted from the l-th layer of a pre-trained VGG network
  - C_l, H_l, W_l: Number of channels, height, and width of the feature map at layer l
  - $\|\cdot\|_2^2$: L2 norm squared (squared Euclidean distance)

- *VGG Style Loss* [11]

$$\mathcal{L}_{content}(I_{gen}, I_{src}) = \sum_{l \in \mathcal{L}} w_l \cdot \left\| G_l^{gen} - G_l^{ref} \right\|_F^2$$

  - I_{gen}: Generated image (stylized image)
  - I_{ref}: Reference image
  - $l \in \mathcal{L}$: A set of layers in the VGG network used for style comparison
  - w_l: Weight assigned to each layer l
  - G_l^{gen}, G_l^{ref}: Gram matrices of the generated image and the reference image at layer l
  - $\|\cdot\|_F^2$: Frobenius norm squared (measures the squared difference between two matrices)

Experimental Results

Performance Comparison

- The experiment comprehensively evaluates model performance using the WikiArt [9] dataset as reference images and the MSCOCO 2014 [10] dataset as source images, providing insights into the generalization capability across diverse and rare source-reference pairs.
- The proposed method achieves the lowest scores in both VGG content loss and VGG style loss, indicating the most balanced style transfer performance by effectively preserving both semantic structure and stylistic features.

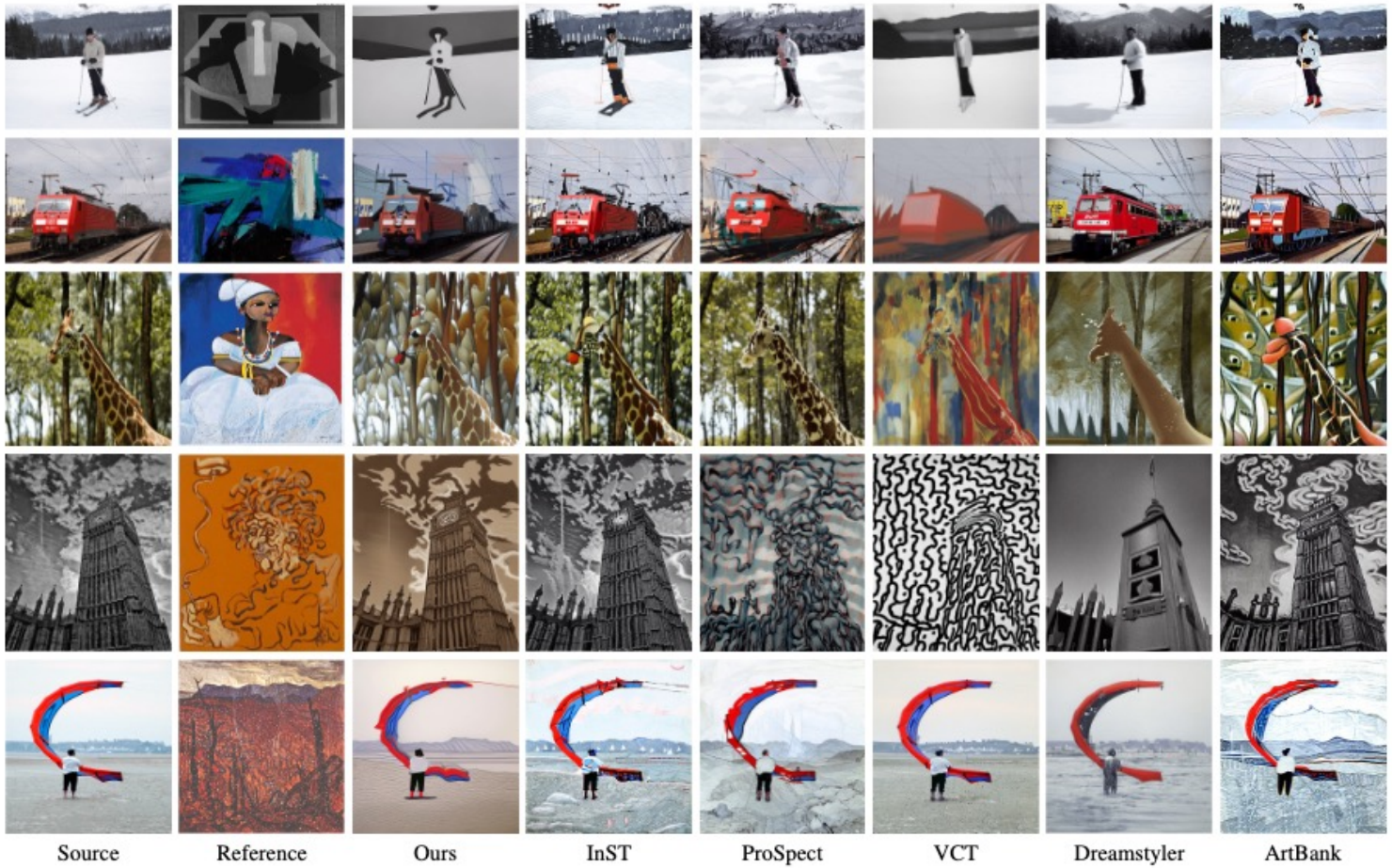
[Table 5] Quantitative Results Evaluated by VGG Content Loss (Semantic Similarity) and VGG Style Loss (Style Consistency)

Model	VGG Content Loss (↓)	VGG Style Loss (↓)
Ours	**169.8570 ± 20.56**	**0.0594 ± 0.04**
InST [4]	238.9704 ± 31.62	0.3438 ± 0.05
ProSpect [5]	299.3260 ± 76.06	0.1397 ± 0.11
VCT [6]	194.7037 ± 4.64	0.0838 ± 0.03
Dreamstyler [7]	178.7759 ± 7.09	0.0932 ± 0.03
ArtBank [8]	205.1655 ± 29.75	0.1722 ± 0.03

Experimental Results

Performance Comparison

[Figure 13] Qualitative comparisons of diffusion-based methods, including our proposed method, InST, ProSpect, VCT, Dreamstyler, and ArtBank.



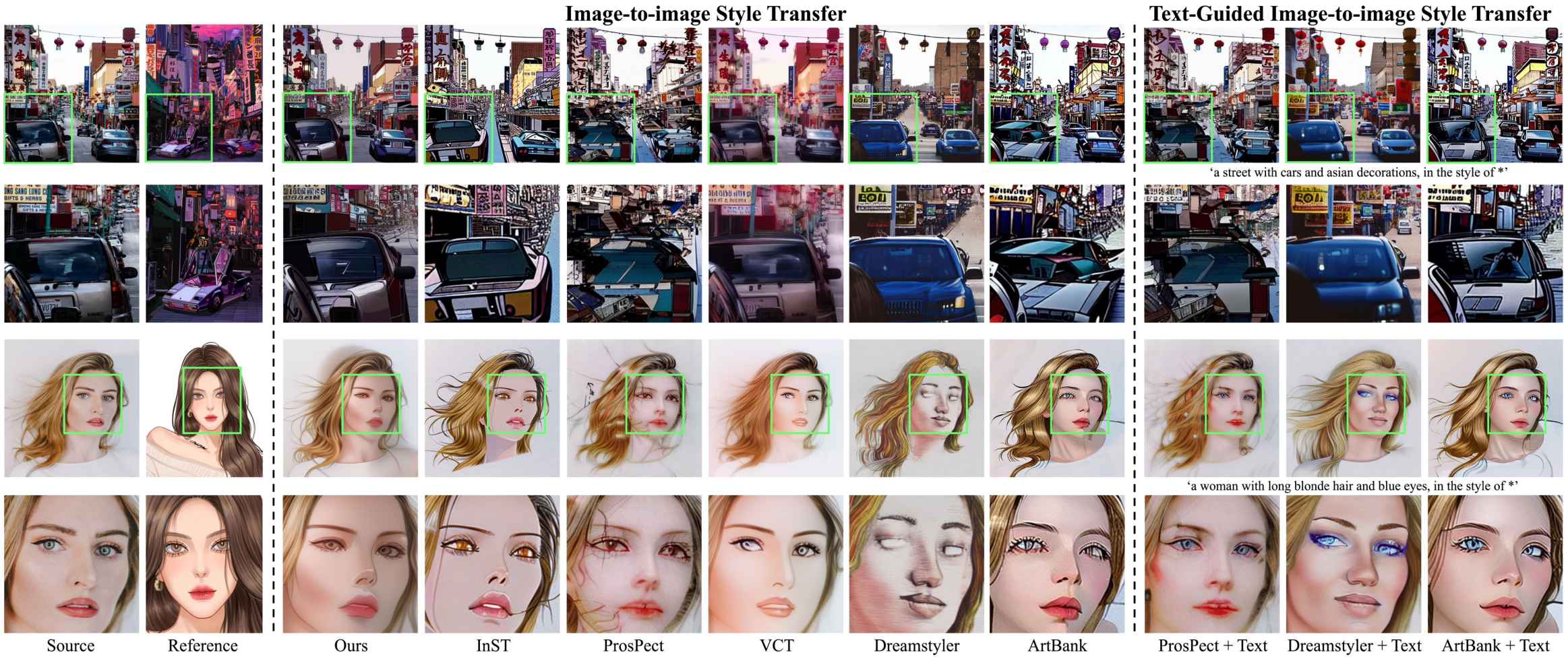
[Table 6] User study results on preferences for the proposed method and other style transfer models.

Model	Votes	Percentage
Ours	**375**	**32.6%**
InST [4]	187	16.3%
ProSpect [5]	224	19.5%
VCT [6]	93	8.1%
Dreamstyler [7]	49	4.3%
ArtBank [8]	222	19.3%
Total	1,150	100%

Experimental Results

Performance Comparison

[Figure 14] Qualitative results with zoomed-in views of specific square regions from each first row. Columns 3 to 8 present image-to-image stylized results without context-aware text conditioning, while columns 9 to 11 present text-description-guided stylized results.



Experimental Results

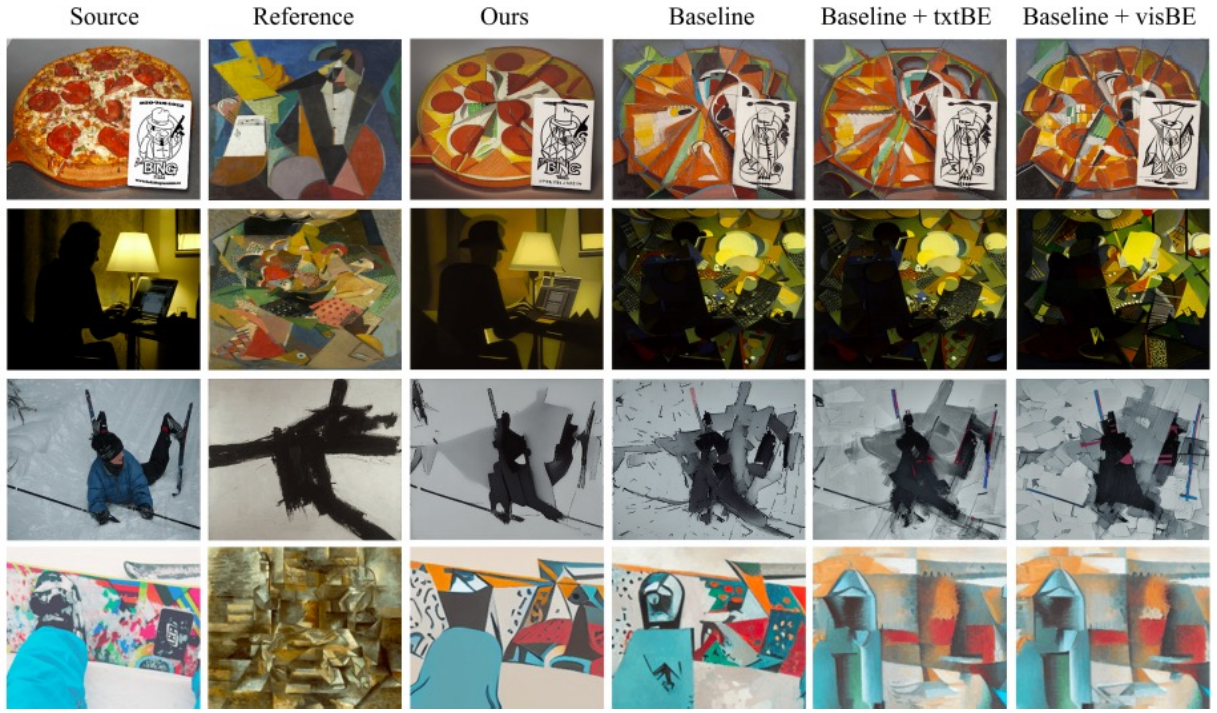
In-depth Analysis

- Quantitative and qualitative comparisons of output image examples produced by four ablation models—No BE (Baseline), BE on textual embeddings only (Baseline + txtBE), BE on visual embeddings only (Baseline + visBE), and BE on both embeddings (Ours)—demonstrate the effectiveness of incorporating both embeddings in improving overall style transfer performance.

[Table 7] Comparison result of four boundary enhancement (BE) settings: No BE (Baseline), BE on textual embeddings only (Baseline + txtBE), BE on visual embeddings only (Baseline + visBE), and BE on both embeddings (ours)

Model	VGG Content Loss (↓)	VGG Style Loss (↓)
Baseline	312.0630 ± 63.38	0.1343 ± 0.06
Baseline + txtBE	309.1944 ± 57.17	0.1418 ± 0.05
Baseline + visBE	301.0851 ± 60.57	0.1385 ± 0.05
Ours	**169.8570 ± 20.56**	**0.0594 ± 0.04**

[Figure 15] Comparison of output image examples produced by four ablation models: No BE (Baseline), BE on textual embedding only (Baseline + txtBE), BE on visual embeddings only (Baseline + visBE), and BE on both embeddings (ours)



Conclusion and Achievement

- The study aimed to address the limitations of conventional style transfer methods, which often enforce full style adaptation and rely solely on textual embeddings to transfer style information, leading to structural distortions and visually disharmonious results by applying styles even to essential regions of the source image.
- To overcome this, a blended embedding strategy combined with a Canny edge-based constraint was proposed, allowing the diffusion model to selectively preserve important source features.
- Experimental results on 1,600 source-reference pairs demonstrate superior performance in both semantic and style preservation, producing more visually coherent outputs.

Achievement

- Conferences
  - A Study of Text Guided Image Generation Based on Diffusion Model, ***IEIE***, 2023
  - An Efficient Fine-Tuning of Generative Language Model for Aspect-Based Sentiment Analysis, ***ICCE***, 2024
  - A Study of Style Transfer based on Text to Image Diffusion Models, ***ICCE***, 2025
- Others
  - ***2023 Singapore K-Tech Festival - AI/ML Innovation Research Center Booth***, Singapore, Singapore, 2023
  - ***2023 SW Education Festival***, Ilsan, Korea, 2023
  - ***The AI Show 2024***, Ilsan, Korea, October 23-25, 2024

Thank you

Reference

[1] Liu, WuQin, et al. "FreeStyler: A Free-Form Stylization Method via Multimodal Vector Quantization." International Conference on Computational Visual Media. Singapore: Springer Nature Singapore, 2024.

[2] Qi, Tianhao, et al. "Deadiff: An efficient stylization diffusion model with disentangled representations." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[3] Chung, Jiwoo, Sangeek Hyun, and Jae-Pil Heo. "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

[4] Zhang, Yuxin, et al. "Inversion-based style transfer with diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[5] Zhang, Yuxin, et al. "Prospect: Prompt spectrum for attribute-aware personalization of diffusion models." ACM Transactions on Graphics (TOG) 42.6 (2023): 1-14.

[6] Cheng, Bin, et al. "General image-to-image translation with one-shot image guidance." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[7] Ahn, Namhyuk, et al. "Dreamstyler: Paint by style inversion with text-to-image diffusion models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 2. 2024.

[8] Zhang, Zhanjie, et al. "Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 7. 2024.

[9] Saleh, Babak, and Ahmed Elgammal. "Large-scale classification of fine-art paintings: Learning the right metric on the right feature." arXiv preprint arXiv:1505.00855 (2015).

[10] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. Springer International Publishing, 2014.

[11] Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.