

Fine-grained and Adaptive Style Transfer in Training-free Diffusion Models



Yebin Ahn, Dong-Jin Kim

School of Data Science, Hanyang University

Introduction

Goal of Style Transfer

- Preserve the structure of content image
- Effectively apply the visual characteristics of style image

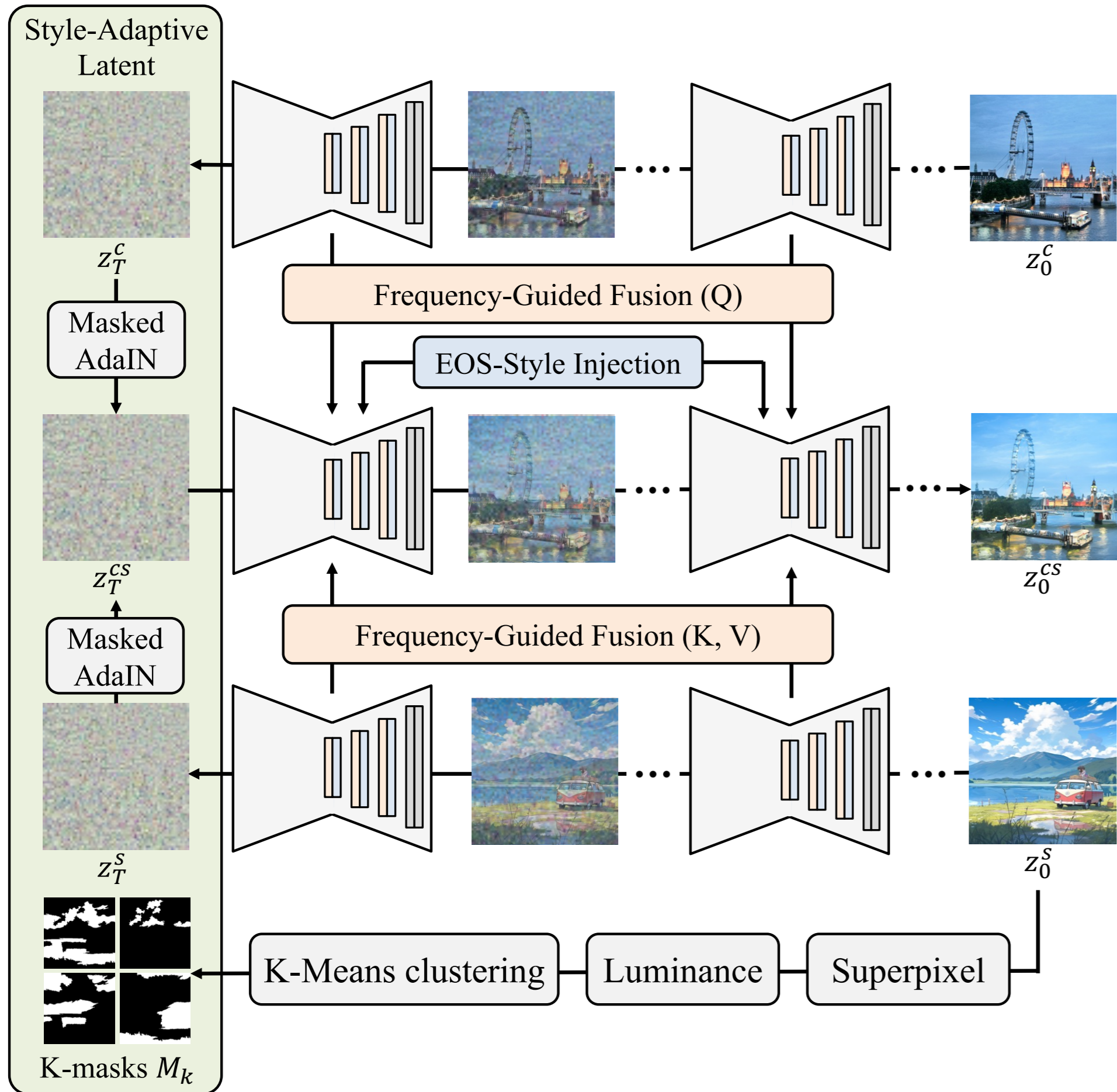
Limitations in StyleID

- (1) AdaIN to generate an initial latent
→ Initial latent is processed as whole, making it **difficult to reflect fine-grained style information**.
- (2) Query linear combination in attention swapping
→ Weighting scale leads to a **trade-off** between style emphasis and structure preservation.



Proposed Framework

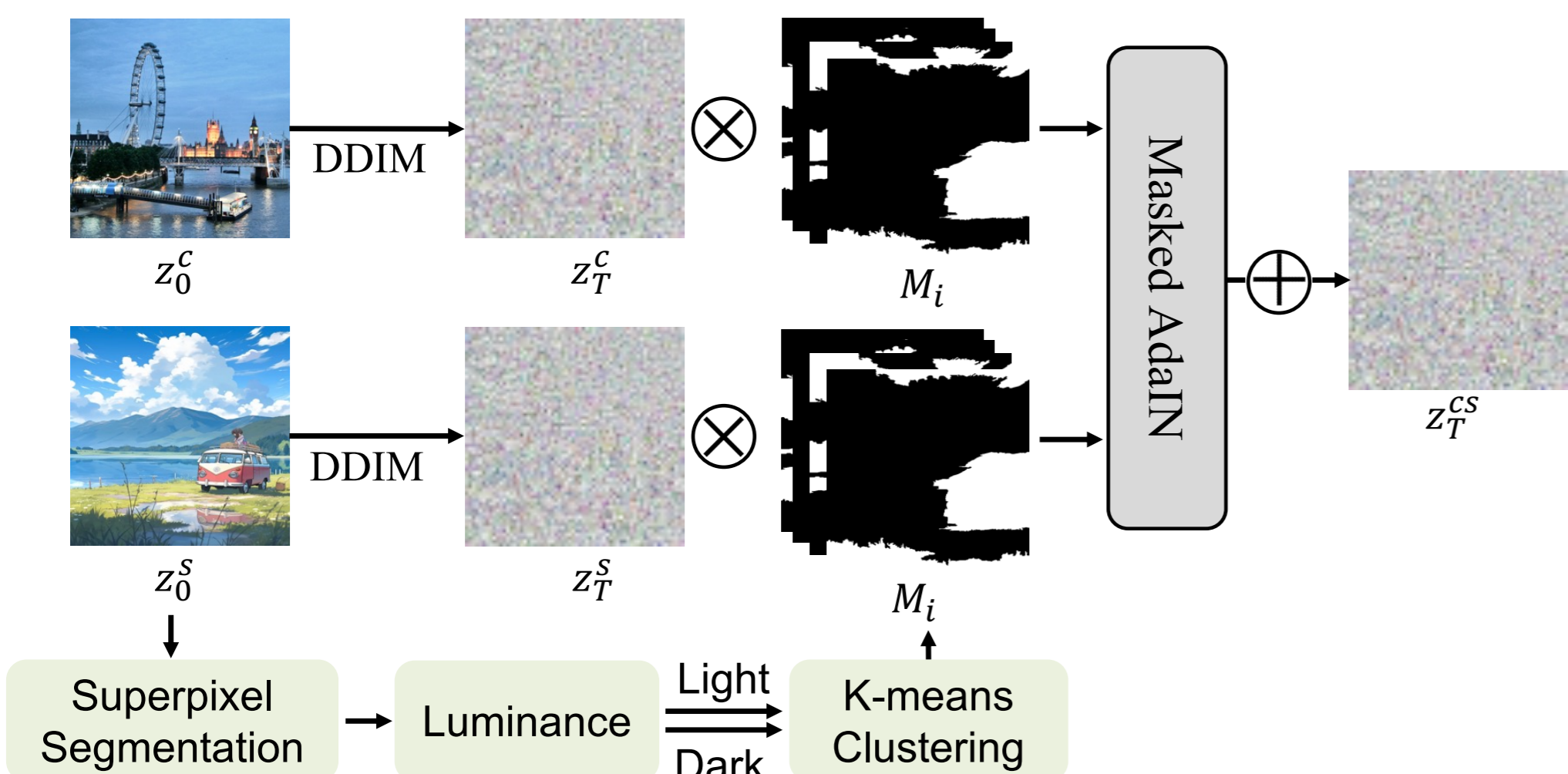
Overall Framework



Style Adaptive Latent

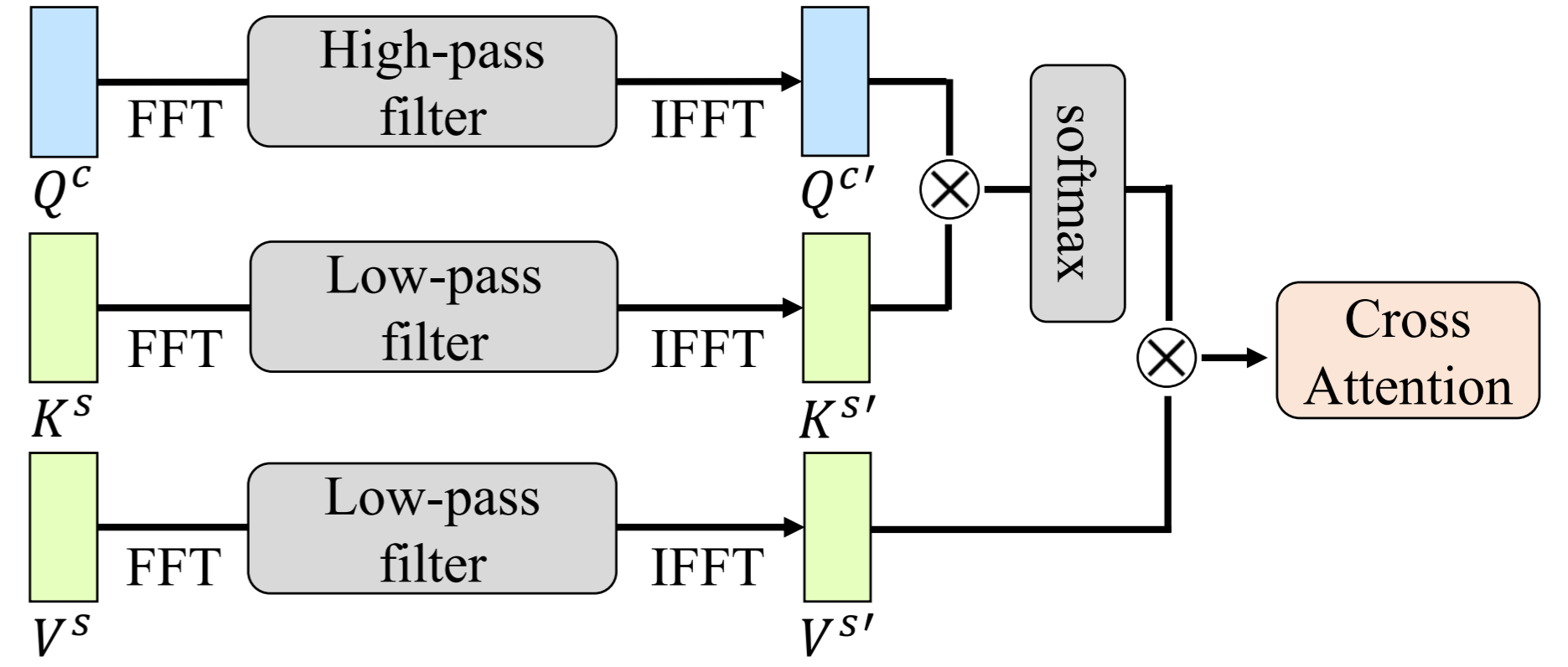
- Extracting Masks: **superpixel segmentation, luminance, k-means clustering**
- Obtain an initial latent through masked AdaIN

$$z_T^{CS} = \sum_{i=1}^n \left(\frac{z_{T,M_i}^C - \mu(z_{T,M_i}^C)}{\sigma(z_{T,M_i}^C)} \sigma(z_{T,M_i}^S) + \mu(z_{T,M_i}^S) \right)$$



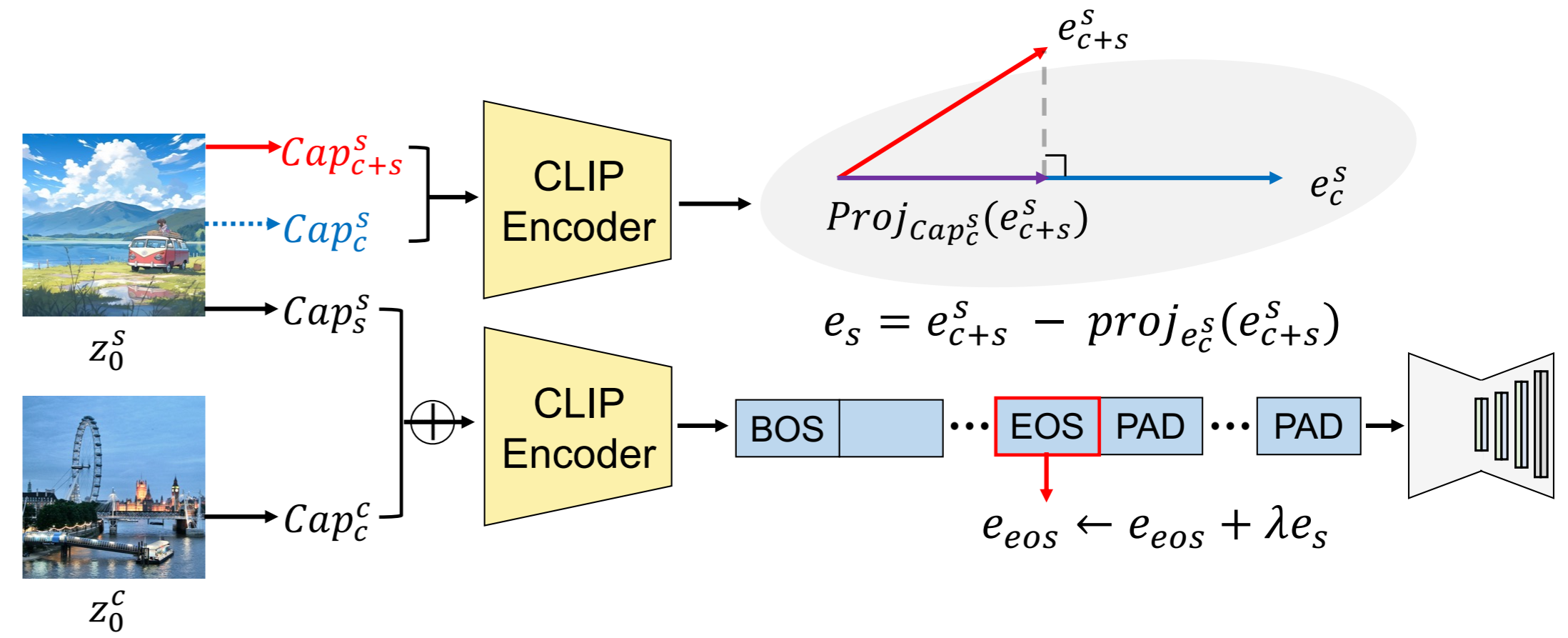
Frequency-Guided Fusion

- **Query**: applying **high-pass filter** for content preservation
- **Key, Value**: applying **low-pass filter** for style reflection



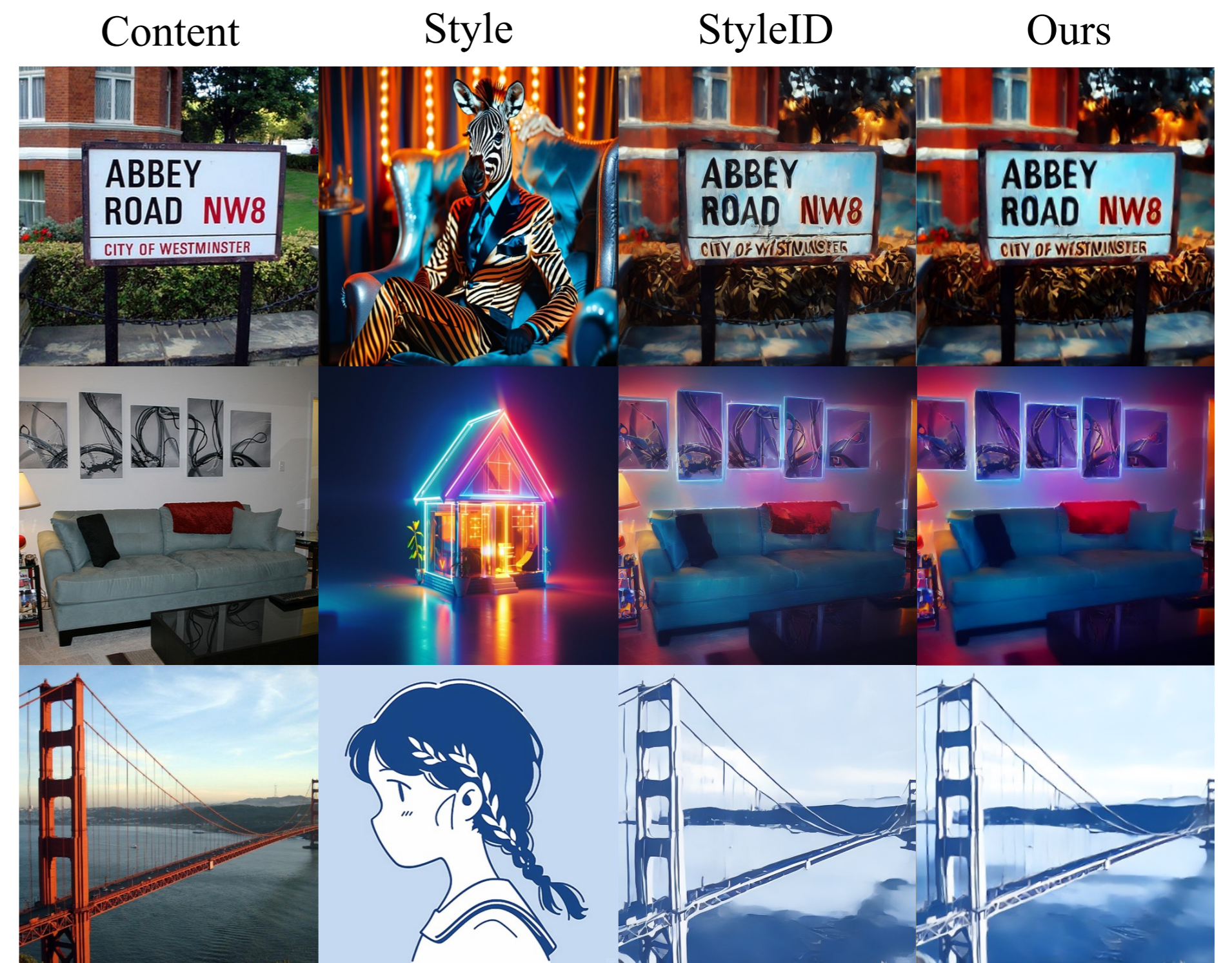
Style-Adaptive Injection

- Extracting **pure style** representation via **Gram-Schmidt**
- Adding style vector to the **EOS token**

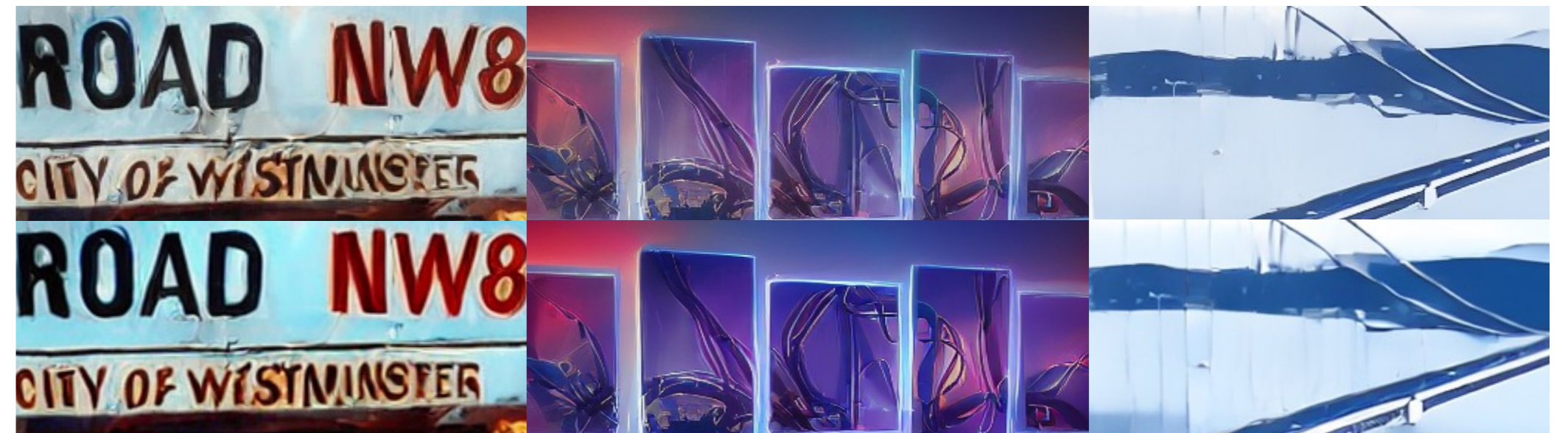


Experiments

Qualitative Results



- Detail View – Top: StyleID / Bottom: Ours



Quantitative Results

	ArtFID ↓	FID ↓	LPIPS ↓	CLIP-I ↑
InST	32.6979	20.7841	0.5010	0.8287
StyleID	25.1609	16.6291	0.4272	0.7678
Z*	23.8693	15.7551	0.4246	0.7566
Ours	22.8755	15.1699	0.4147	0.8401

Ablation Study

SAL	FGF	ESI	ArtFID ↓	FID ↓	LPIPS ↓	CLIP-I ↑
	✓	✓	23.2217	15.3028	0.4244	0.8305
✓		✓	24.5808	16.5302	0.4022	0.8491
✓	✓		23.3486	15.4810	0.4167	0.8349
✓	✓	✓	22.8755	15.1699	0.4147	0.8401