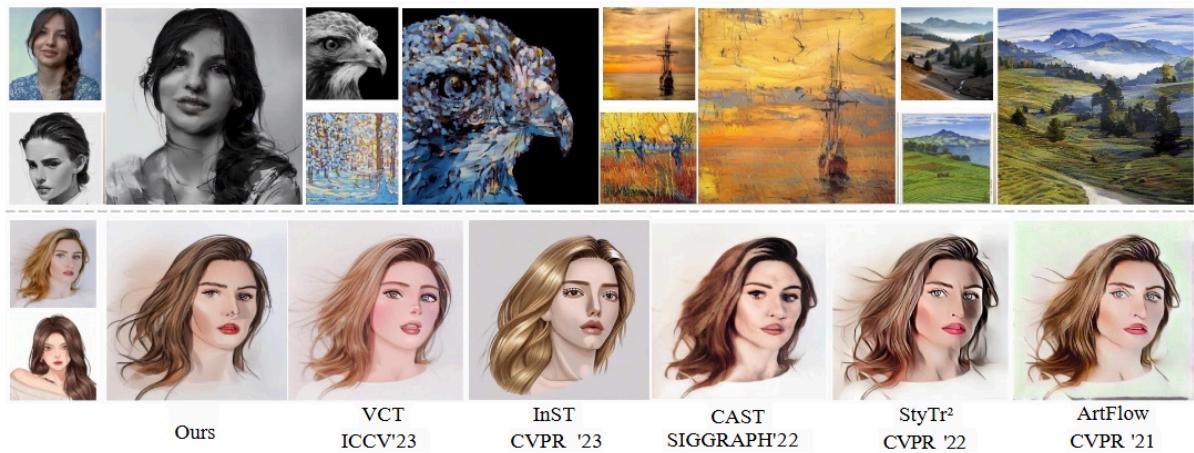


Z*: Zero-shot Style Transfer via Attention Reweighting

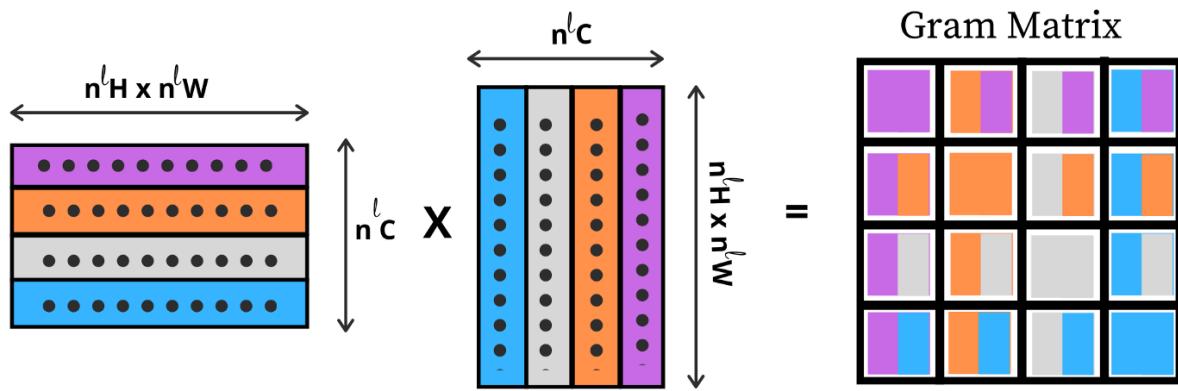
Style Transfer



- **Objective:** content image의 layout은 유지하면서, style image와 유사한 style을 모방 (stylized image)
- Given style image,
 - text로 content image generation
 - reference image (content image)로 content preservation
- Training vs -free

Previous work

1. Gram matrix-based [StyTr2, ArtFlow, Adaln]



- Process: channel flatten → calculate the correlation between channels
 - Measure global style similarity
 - 복잡한 패턴 포착 x (e.g., hair, eyes)
2. Contrastive-based [CAST]
- a. 세밀한 스타일 디테일 x
3. Image-controlled diffusion model [VCT, InST]
- a. 각 style image에 대해 style embedding training
 - b. 콘텐츠를 보존하지 못하면서 입력 스타일에서 벗어나는 결과가 발생

Proposal

1. Training-free
2. Utilize prior knowledge of LDM → attention mechanism → 그럼에도 학습 없이 는 poor content preservation → **attention reweighting**

Method

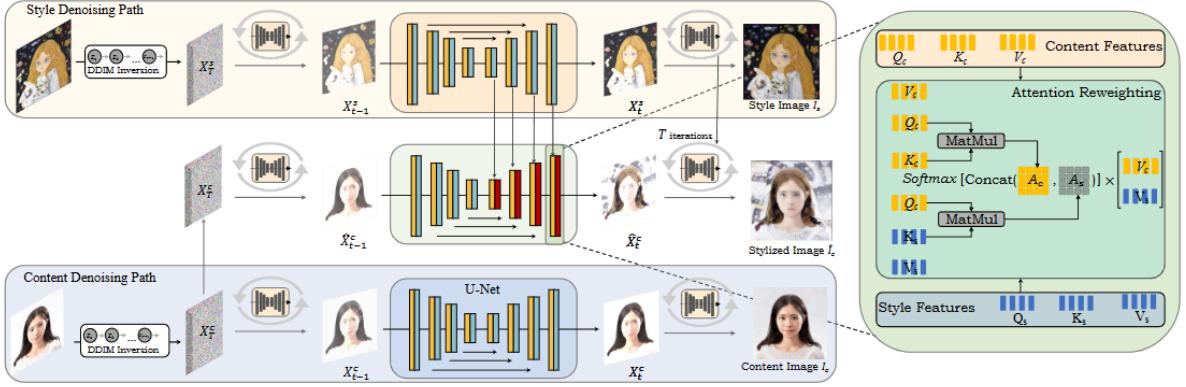


Figure 2. Overall pipeline of our style transfer framework. The stylization process operates in the latent space. We perform DDIM inversion separately for the content and style images. During the denoising process, our Cross-attention Reweighting is employed to integrate style patterns into the content structure. By iteratively performing 50 denoising steps, we are able to achieve the final stylized output.

- Style-cross attention: merge content and style features
 - Content information (e.g., image structure): **Query**
 - Style information (e.g., color, texture, object shape): **Key** and **Value**
1. Naive setting

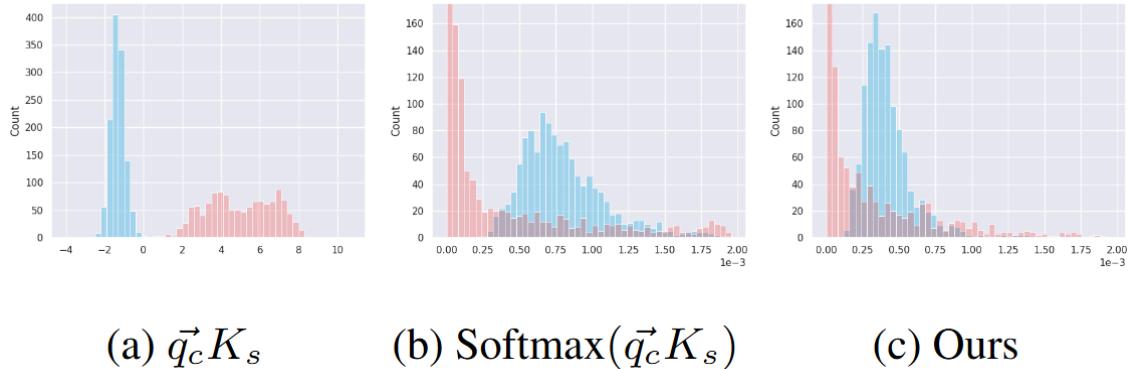
$$\hat{f}_c = \text{Attn}(Q_c, K_s, V_s) = \text{Softmax}\left(\frac{Q_c K_s^T}{\sqrt{d}}\right) V_s. \quad (7)$$

2. Simple addition: content info를 강화

$$\hat{f}_c = \lambda \cdot \text{Attn}(Q_c, K_s, V_s) + (1 - \lambda) \cdot \text{Attn}(Q_c, K_c, V_c),$$

- Style-cross attention: 스타일을 잘 전달하지만 콘텐츠 구조를 손상시킬 수 있다.
- Content-self attention: 콘텐츠 구조를 잘 보존한다.
- Simple Addition은 이 둘을 적절히 섞어 콘텐츠 보존과 스타일 적용 사이의 균형을 맞추고자 한다.
- λ 를 결정하는 게 매우 어려움.
- 콘텐츠 픽셀이 스타일과 약한 상관관계를 가질 때 ($q_c K_s^T$ 값이 작거나 음수일 때): content structure을 유지하면서 style의 영향은 최소로 해야한다. 즉, 작은 어텐션 가중치를 부여하여 부정적인 영향을 최소화해야 한다.
- 반대로 콘텐츠 픽셀이 스타일과 의미 있는 상관관계를 가질 때 ($q_c K_s^T$ 값이 클 때): 큰 어텐션 가중치를 부여해야 한다. 상관관계가 큰 k_s 의 style로 그릴 것이다.

- 하지만 Softmax 함수는 절대적인 크기를 무시하고 값들의 차이만 증폭시키는 경향으로 인해 약한 상관관계를 가진 style-cross attention에 오히려 큰 어텐션 가중치로 이어지는 비직관적인 결과가 발생할 수 있다.



3. Attention reweighting: λ 값을 어텐션의 Softmax 함수 안쪽에 포함

- a. Softmax로 출력을 정규화하는 동안 content feature 내의 차이와 content-style feature 간의 차이점을 동시에 고려한다.

$$A' = \sigma\left(\left[\lambda \cdot \frac{Q_c K_s^T}{\sqrt{d}}, \quad \frac{Q_c K_c^T}{\sqrt{d}}\right]\right)$$

$$\hat{f}'_c = A' * V'^T = \sigma\left(\left[\lambda \cdot \frac{Q_c K_s^T}{\sqrt{d}}, \quad \frac{Q_c K_c^T}{\sqrt{d}}\right]\right) * \begin{bmatrix} V_s \\ V_c \end{bmatrix}.$$

1. Content -style 간의 상관관계가 약할경우: $-\inf \rightarrow$ self-attention 과 동일 \rightarrow 원본 content info 보존

2. Content -style 간의 상관관계가 강할경우: maximum value $q_c k_s$ 와 $q_c k_c$ 가 비슷할때, style과 content를 적절히 섞어서 반영

b. $\lambda = 1.2$

i. 커질수록, style을 더 강하게 반영

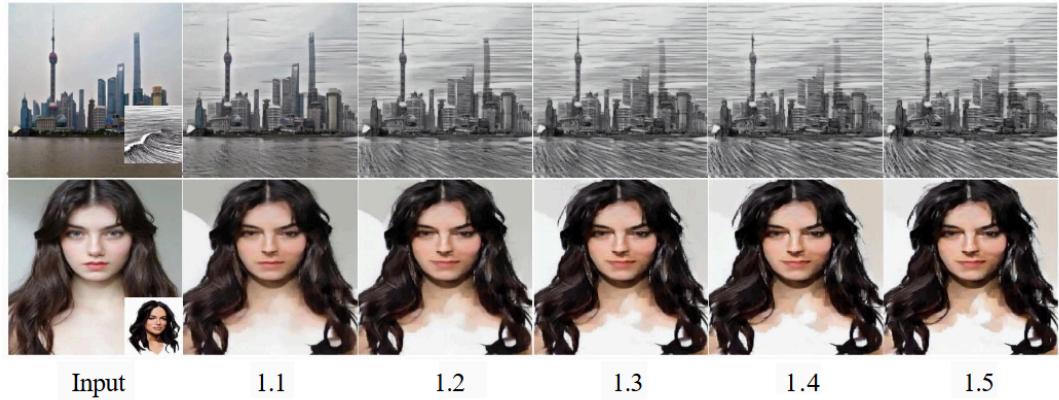
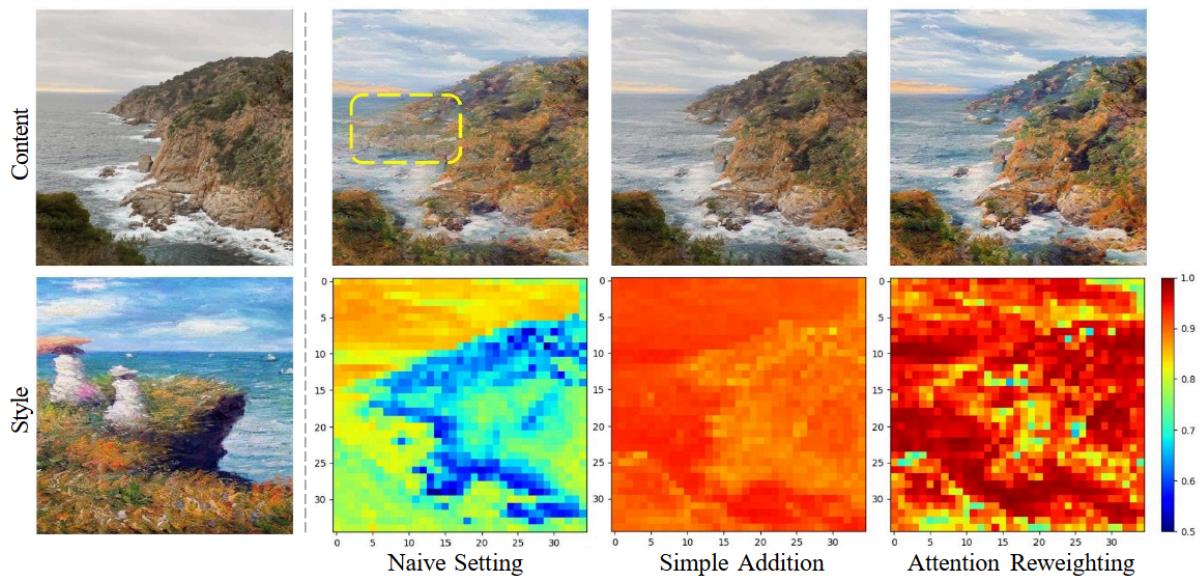


Figure 10. The effect of λ in Eq. (12).

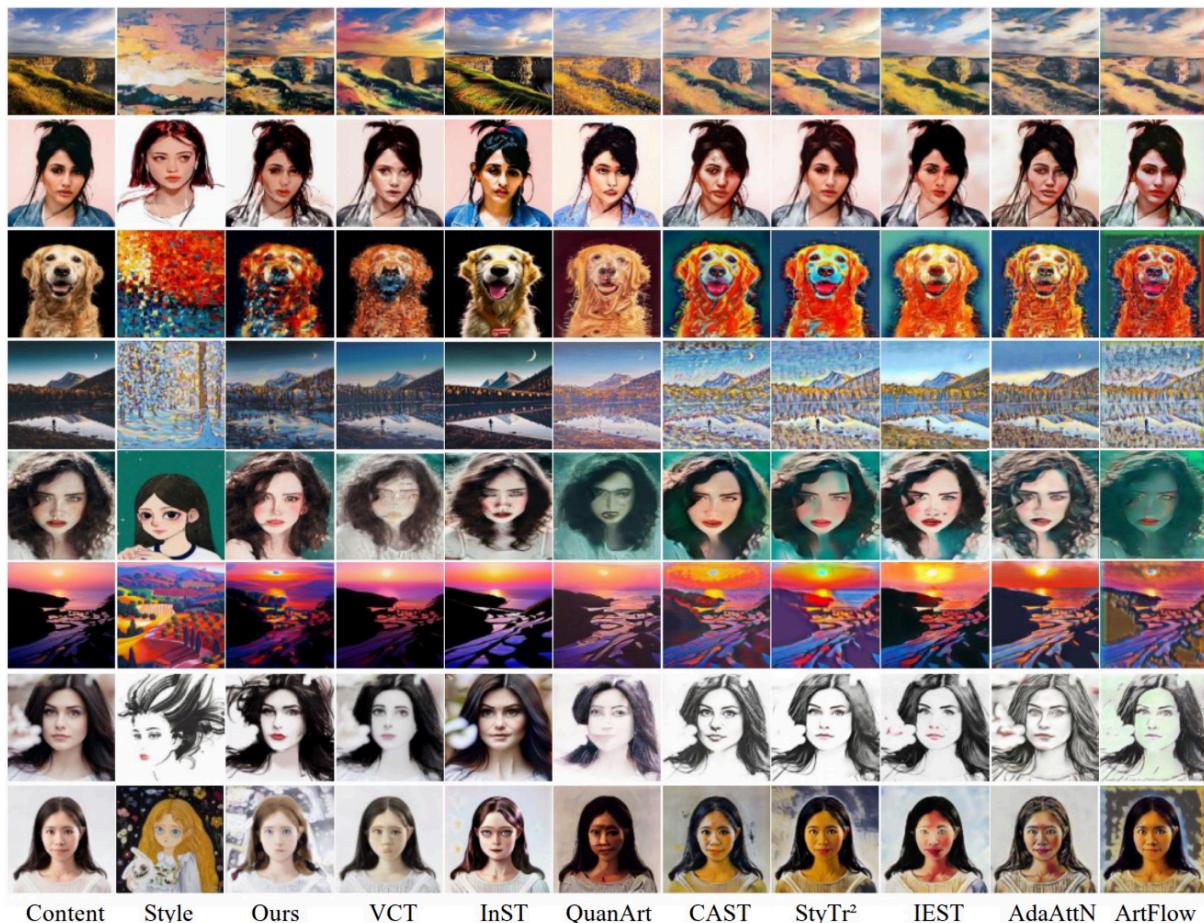
- Content self attention: preserve structure
 - Standard self-attention을 사용하기에 style-cross attention만 알면된다.
- Heat map
 - Cross-attention과 self-attention간의 cosine similarity: low score \rightarrow loss of content info
 - Naive setting: content info를 무시하면서 style pattern을 강조
 - Simple addition: 과도한 content preservation
 - Attention reweighting: 적절



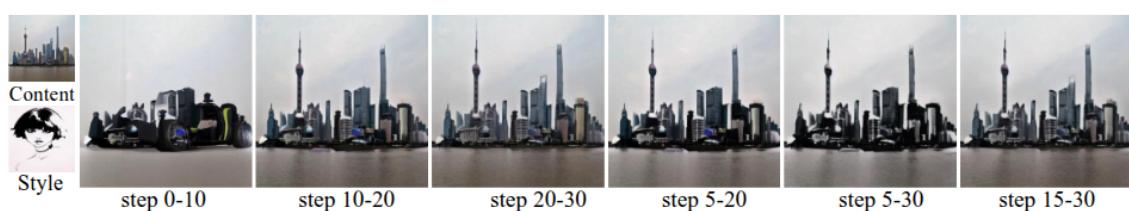
Experiments

- Setting

- Stable Diffusion v1.5 with null text
- Denoising step: 30
- Cross-attention reweighting between layers 10 - 15 during the 5th - 30th steps
- Qualitative evaluation



- Ablation study
 - Different denoising steps



- Different U-Net layers



- Different attention arrangement

