VideoSketcher: A Training-Free Approach for Coherent Video Sketch Transfer

Huining Li[1,2]; Bangzhen Liu[1,4]; Rui Yang[3]; Yang Zhou[1,5]; Chenshu Xu[1]; Xufang Pang[6]; Shengfeng He[1]

[1]Singapore Management University, [2]Baidu Inc, [3]Huaqiao University, [4]City University of Hong Kong
[5]South China University of Technology, [6] Shenzhen Institute of Artificial Intelligence and Robotics for Society

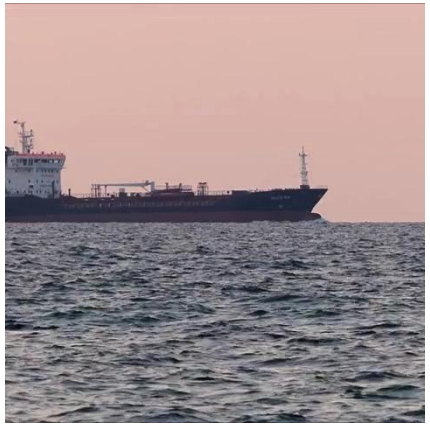
WACV 2026

overview

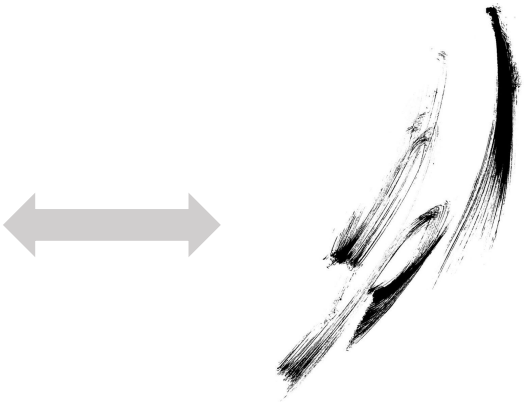
- Motivation
- Limitations of Prior Work
- Method
- Experiments
- Conclusion

Motivation

The Semantic & Structural Gap



Dense & Dynamic Video



Abstract & Sparse Sketch

- **Motivation:**Generating high-quality sketches from video requires a nuanced understanding of **semantic content** and visual **structure**.
- **The Dilemma**:
  - Distribution Mismatch (No color/background)
  - Structural Conflict
- **The Goal**: To achieve **training-free**, **style-controllable** sketch video generation that strictly preserves frame structure while applying specified sketch aesthetics.

Limitations of Prior Work

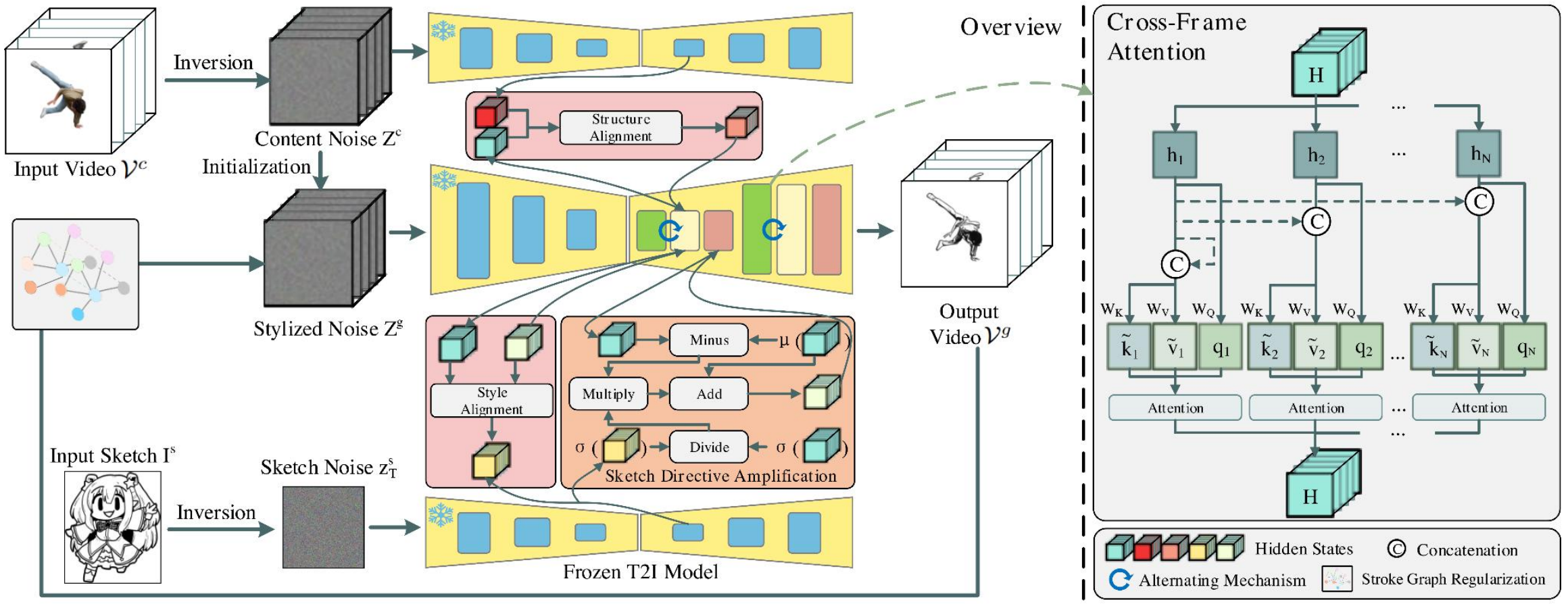
- Limitations of Prior Work
  - **Training-Based Methods** (e.g., ref2sketch): Require expensive training optimization and are restricted to **limited styles**.
  - **Training-Free Methods** (e.g., StyleID): Rely on broad descriptors rather than explicit **sketch-specific semantics**. This semantic gap leads to noticeable **style distortions** and **temporal jitter** across video frames.

Input Video | Reference Style | Ref2sketch | StyleID | VideoSketcher(ours)

Our Solution: VideoSketcher

- A novel, training-free framework that strictly maintains temporal coherence and preserves structural integrity.

VideoSketcher Overview



Time-Linked Attention (TLA)

- motivation: preserve **structural integrity** and **transfer local sketch details** from reference images.
- core components:
  - sketch style transfer

$$\hat{h}_i^g = Attn(Q^g, K^s, V^s),$$
$$Q^g = W^Q h_i^g, K^s = W^K h^s, V^s = W^V h^s,$$

  - video structure preservation

$$Q^{sc} = \alpha \times Q^c + (1 - \alpha) \times Q^g$$

  - video temporal consistency

$$\tilde{K}^g = W^K [h_1^g, h_{i-1}^g]$$
$$\tilde{V}^g = W^V [h_1^g, h_{i-1}^g]$$
$$\hat{h}_i^g = Attn(Q^g, \tilde{K}^g, \tilde{V}^g)$$

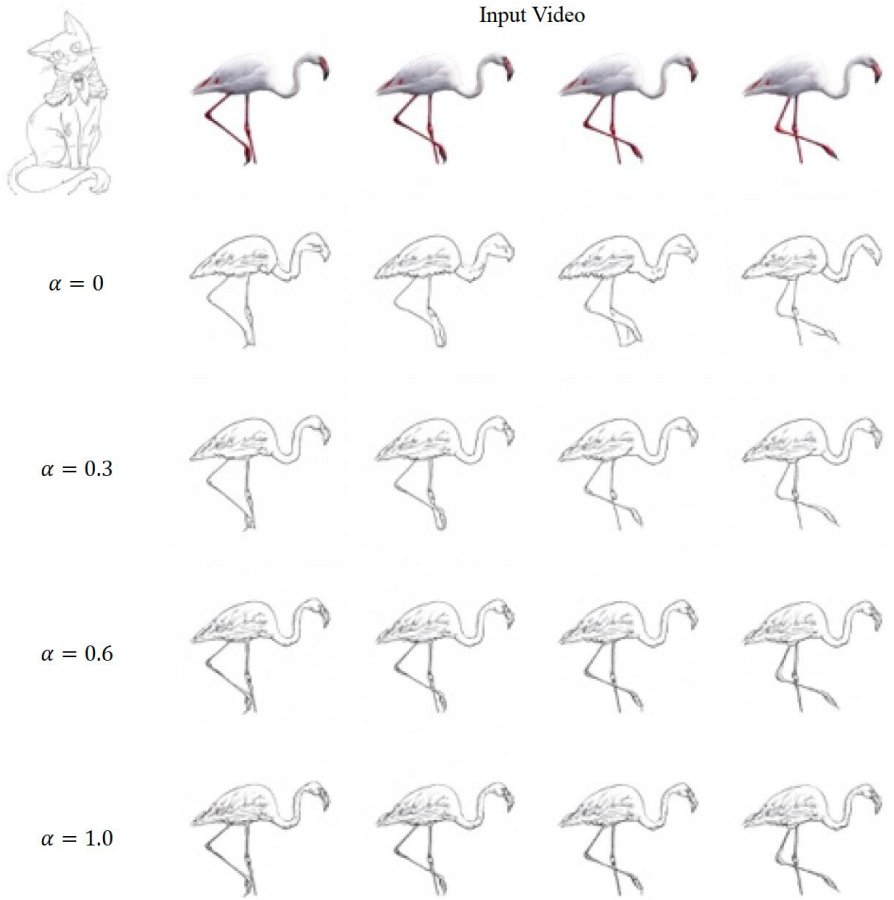
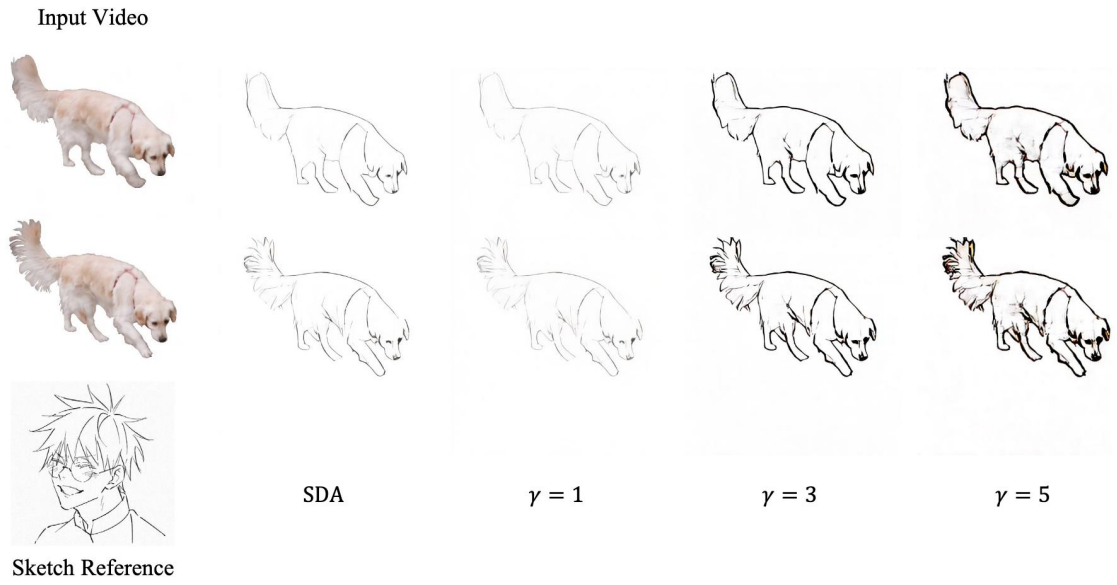


Figure 3. Ablation study of α in TLA.

Sketch Directive Amplification (SDA)

- motivaion: The distortions result from **variance shifts** in the SA's attention map after the key-value exchange.
- idea: a **dynamic adjustment** to focus attention on the semantically rich areas of the sketch image.



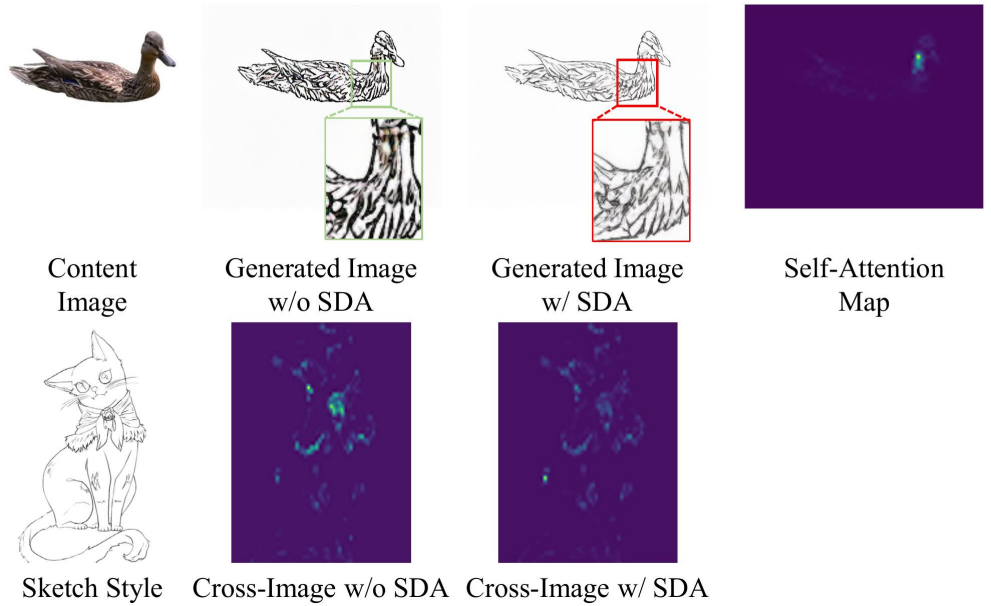


Figure 6. Comparison between SDA and fixed rate amplification.

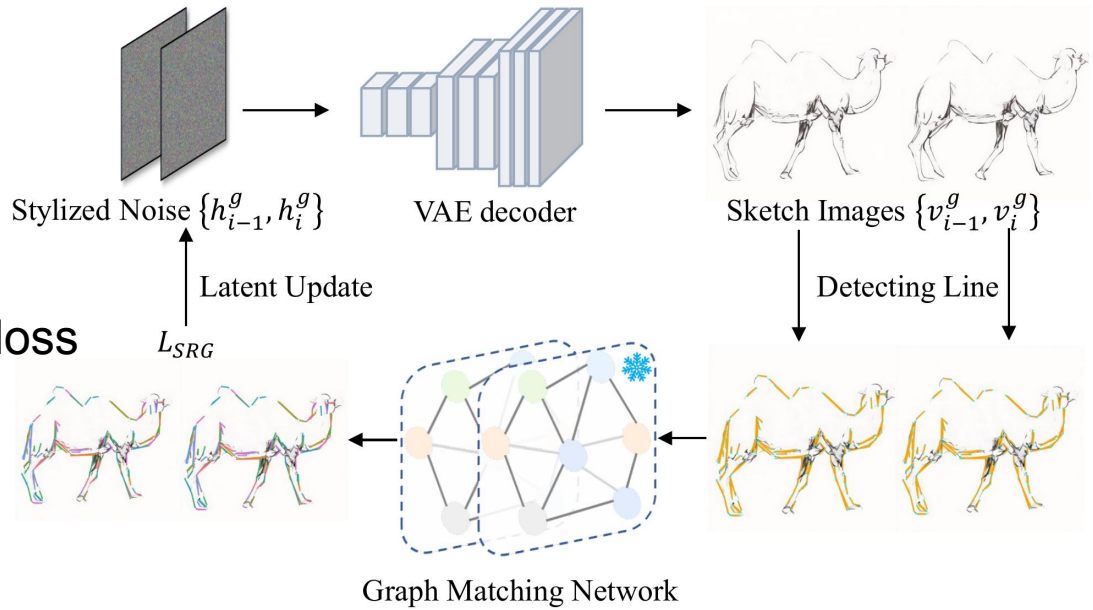
Stroke Graph Regularization (SGR)

- motivation: traditional query-based approaches could compromise **structural consistency** across video frames, particularly under variable lighting conditions.
- Idea: explicit structure regularization at the pixel space.

$$L_{SGR} = \lambda_1 \sum_{m=1}^{N_l} \sum_{p_x \sim l_m^i, p_y \sim l_m^{i-1}} ||p_x - p_y||_2^2 + \lambda_2 \sum_{n=1}^{N_j} ||j_n^i - j_n^{i-1}||_2^2$$

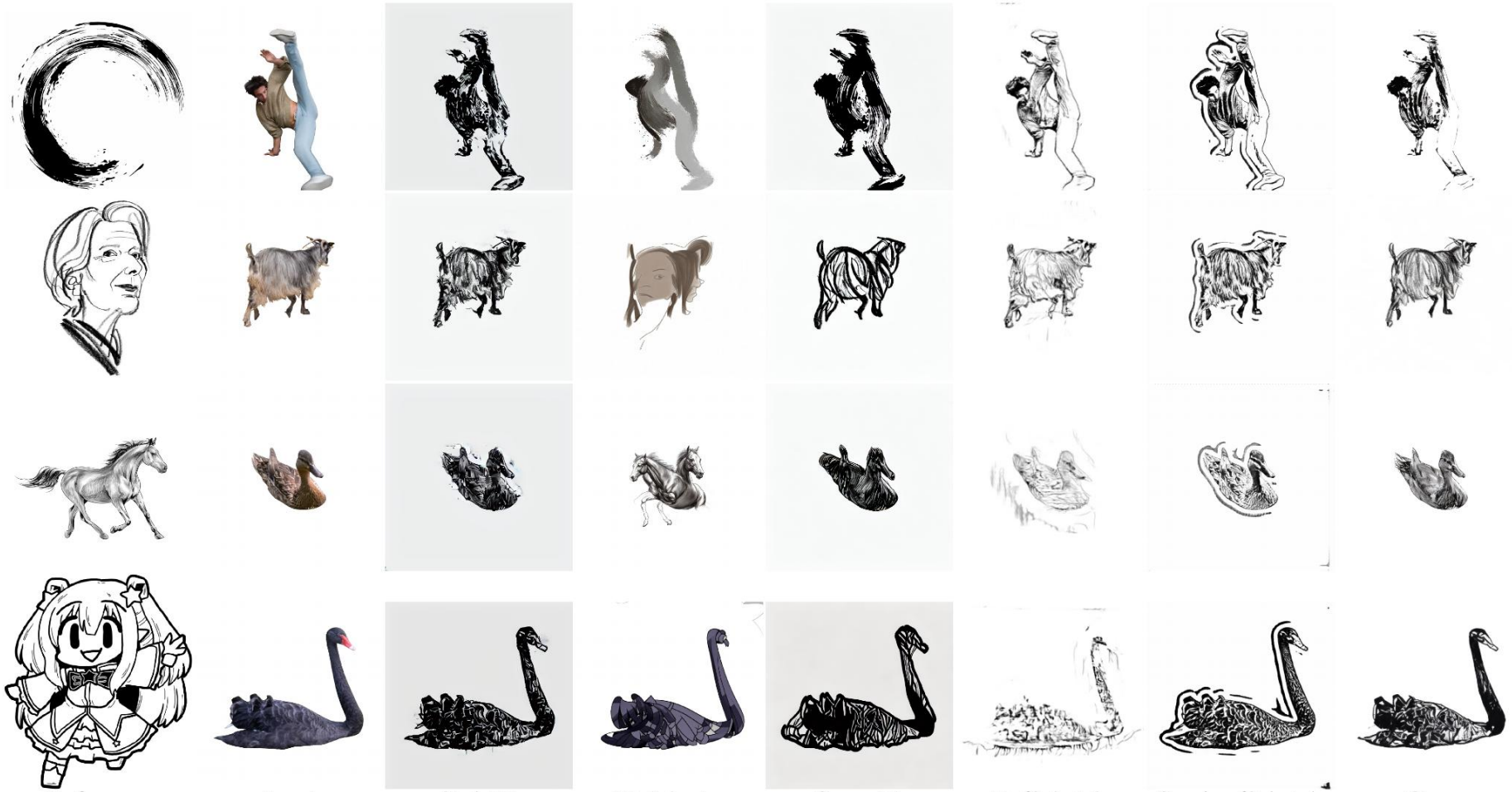
stroke structures loss

keypoints loss



Experiments & Results

- Comparison with Single-image Sketch Style Transfer

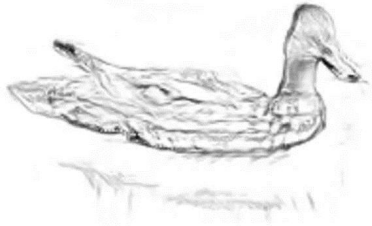


Input Video

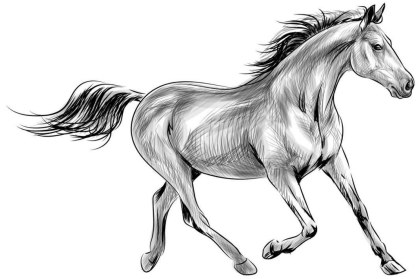
Cross-IA

IP-Adapter

Ref2sketch

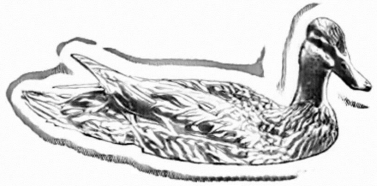


Reference Style



StyleID

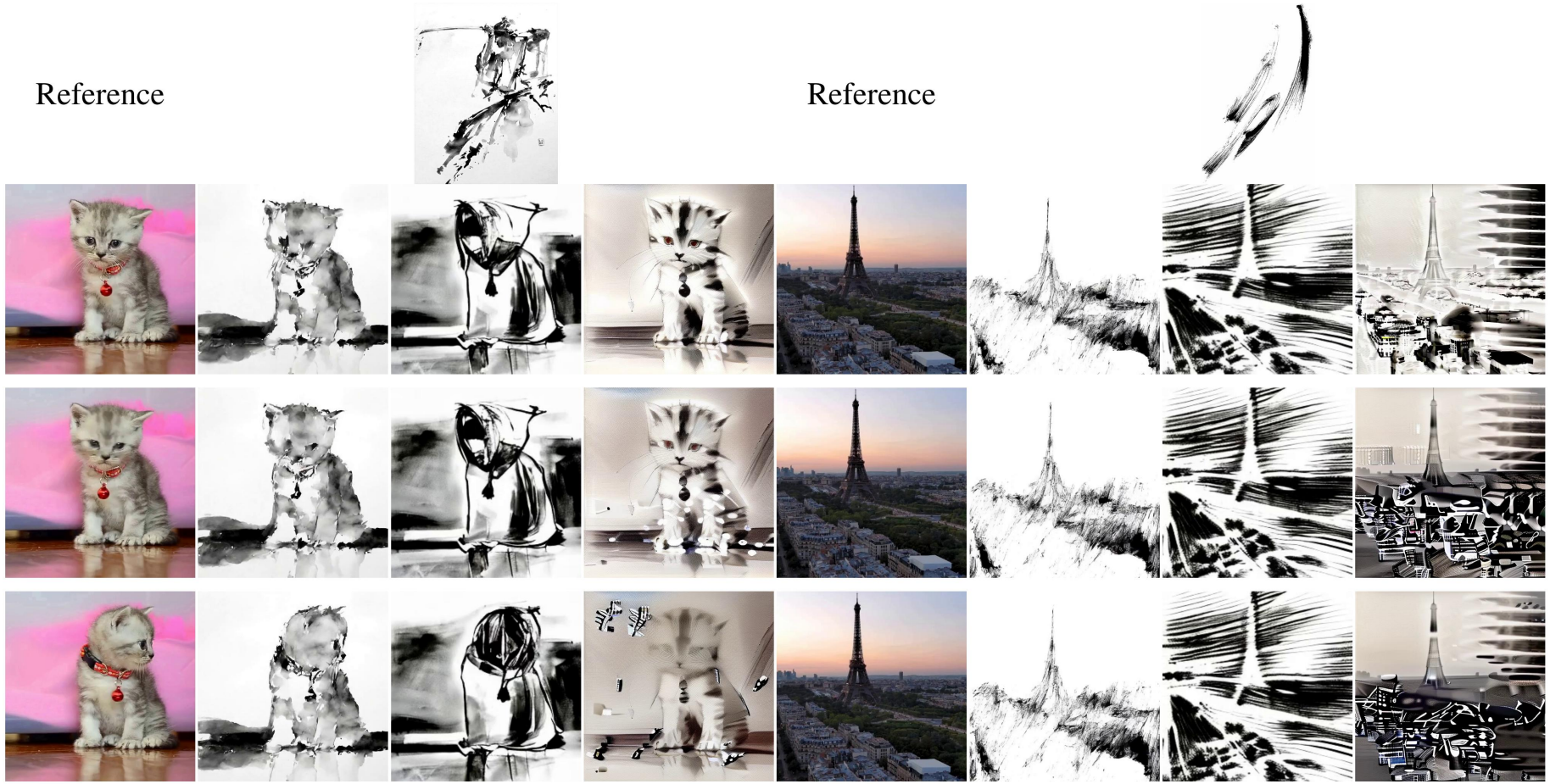
Semi-Ref2sketch



VideoSketcher(ours)

Experiments & Results

- Qualitative comparison of video style transfer results.



Quantitative Results & Ablation Study

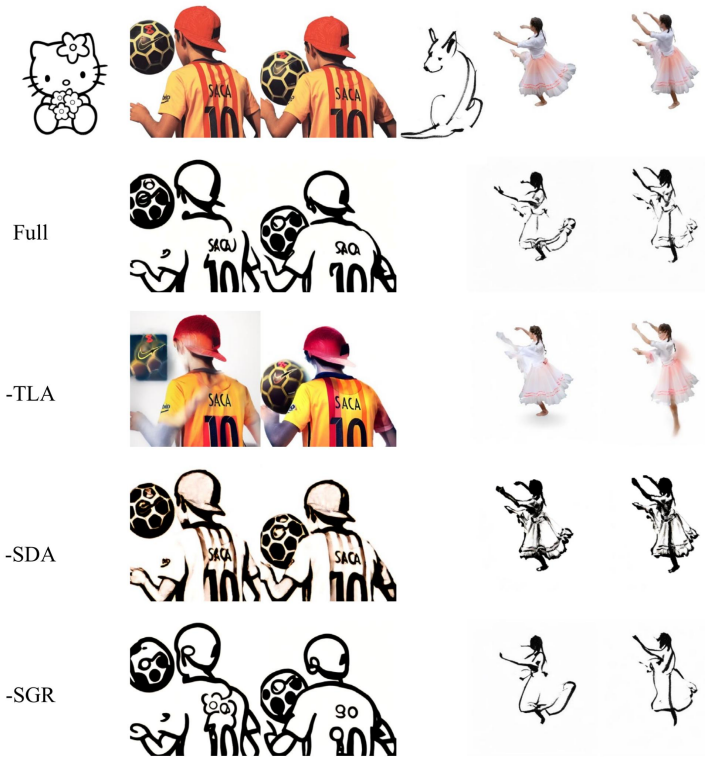
- Quantitative comparison

Methods	CLIP-Image ↑	Pixel-MSE ↓	FID ↓	LPIPS ↓	ArtFID ↓
StyleID	0.9575	0.0324	35.4799	0.1860	42.1789
IP-Adapter	0.9390	0.0210	**24.7112**	0.1330	**29.4180**
Cross-IA	0.9568	0.0755	25.6296	0.1941	31.6175
Ref2sketch	0.9571	**0.0158**	31.8608	0.1927	39.1200
Semi-ref2sketch	0.9736	0.0320	27.8413	0.1872	34.3170
Ours	**0.9740**	0.0193	27.8128	**0.0862**	31.3526

Quantitative comparison with style transfer methods on reference-guided sketch style transfer.

Methods	CLIP-Image ↑	Pixel-MSE ↓	FID ↓	LPIPS ↓	ArtFID ↓
AnyV2V	0.9489	0.1078	26.5821	0.7316	47.7184
Videocomposer	**0.9746**	0.0321	**22.4721**	0.7181	40.4158
Ours	0.9720	**0.0219**	24.735	**0.5703**	**40.3031**

Comparison with style transfer methods on video style transfer



Ablation Study

Conclusion & Future Work

- **Key Contributions**: Proposed VideoSketcher, a novel training-free framework for coherent video sketch transfer.
  - Designed TLA (temporal consistency), SDA (edge reinforcement), and SGR (stroke continuity) to mitigate spatial artifacts and preserve fine details.

- Limitations & Future Work:
  - Current performance may be limited by **highly dynamic content** and extremely abstract styles.
  - Future work will focus on adapting to greater **scene variability** and diverse animation effects.

Code available

- the code and resources available at: https://github.com/lihuining/VideoSketcher