

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

## –Supplementary Materials–

Huining Li<sup>1,2</sup> Bangzhen Liu<sup>1,4†</sup> Rui Yang<sup>3</sup> Yang Zhou<sup>1,5</sup> Chenshu Xu<sup>1</sup>  
Xufang Pang<sup>6</sup> Shengfeng He<sup>1†</sup>

<sup>1</sup>Singapore Management University <sup>2</sup>Baidu Inc <sup>3</sup>Huaqiao University

<sup>4</sup>City University of Hong Kong <sup>5</sup>South China University of Technology

<sup>6</sup>Shenzhen Institute of Artificial Intelligence and Robotics for Society

In this supplementary material, we provide more implementation details (Section 1) and conduct more detailed ablation studies on the hyperparameters declared in the main paper (Section 2). We further demonstrate the superiority of our VideoSketcher via subjective assessments from randomly invited audiences (Section 3). We also present additional video sketch stylization results from our VideoSketcher in Section 4 for a more comprehensive understanding.

## 1. Additional Implementation Details

All our experiments were conducted on a single NVIDIA RTX3090 GPU. For the evaluated videos, we first cropped and resized each frames of these videos to the resolution of 512×512. All the videos are trimmed to 50 frames for fairly quantitative evaluation, while videos containing fewer than 50 frames remain at their original maximum frame count.

The selected videos cover a broad range of semantic categories, including humans, animals, and vehicles. Given that sketch stylization typically emphasizes foreground content, we extract the primary foreground subjects from each video using segmentation. To ensure a comprehensive and fair evaluation, each video is stylized using 12 distinct sketch styles. These template sketch-style images are collected from the 4SKST dataset [8] and public authorized online collections, encompassing various categories such as pencil sketches, brush-style drawings, and manga-style outlines, covering a diverse range of stroke patterns and sketching styles.

For comparison with single-style image-based transfer methods, we utilize the first 50 frames of each video. In contrast, for video-based style transfer methods, we adopt 16-frame clips to match the temporal window of the video diffusion model employed in AnyV2V. The style categories include single-line drawings, traditional sketch images, and anime-inspired sketch aesthetics, representing a diverse range of sketch styles for evaluation in style transfer.

## 2. Additional Experiments and Analysis

### 2.1. Comparisons With Zero-Shot Video Stylization Methods

For zero-shot video style transfer, we compare three video-to-video style transfer methods: Vid2vid-zero [6], Text2Video-Zero [5], and Pix2Video [3], all of which rely on textual prompt-based guidance. In our experiments, sketch-style prompts are generated using ChatGPT to ensure consistency and relevance across comparisons.

Figure 1 presents qualitative comparisons with zero-shot video stylization methods, including Vid2Vid-Zero and Pix2Video. These approaches exhibit limited capacity to render sketch-like appearances. For instance, both methods retain the coloration of the swan’s beak, failing to reproduce the monochromatic, line-based characteristics typical of sketch styles. Text2Video-Zero yields more variable results, occasionally introducing visual noise and inconsistencies across frames. In contrast, our method more faithfully adheres to the semantics of the reference sketch, producing stylizations that are both visually coherent and stylistically consistent.

### 2.2. Comparisons With Bessel Function-Based Sketch Video Generation Methods

In addition, we compare with the Bessel function-based sketch video generation SVS [11].

To assess performance against Bessel-based stylization, we generate reference images using manually applied Bessel stroke styles due to current limitations in automatic generation. As illustrated in Figure 2, our method achieves superior structural preservation and stylistic fidelity, even under significant motion, consistently maintaining temporal coherence across frames.

Furthermore, we compare our approach with the Bessel function-based method in Table 1. Our method shows substantial improvements, excelling in temporal consistency, structural preservation (LPIPS), and overall style quality (FID).

<sup>†</sup> Corresponding authors:

bangzliu@cityu.edu.hk, shengfenghe@smu.edu.sg.

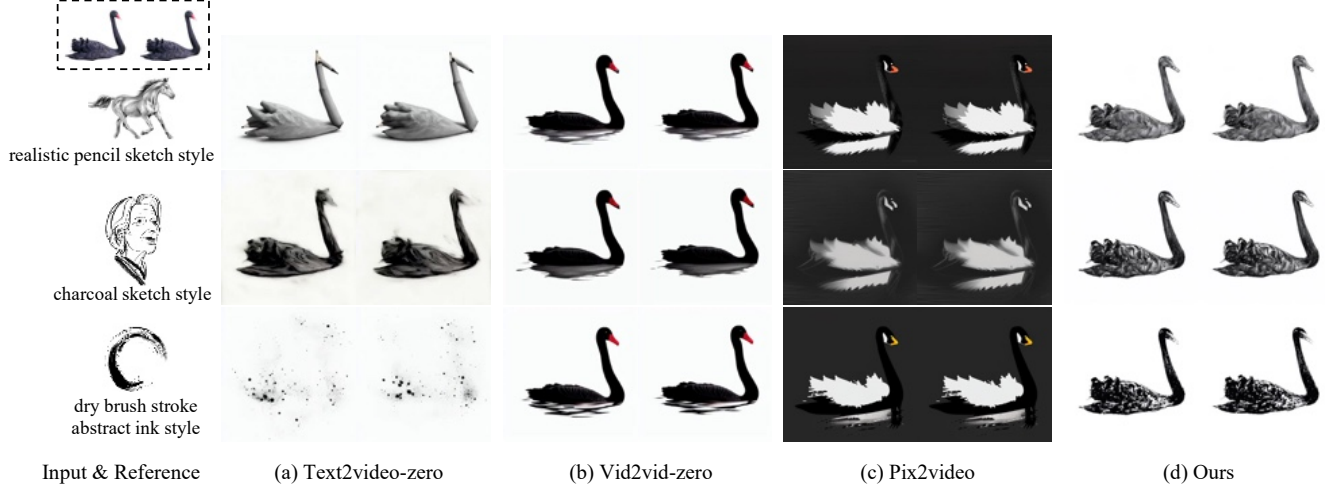


Figure 1. **Comparison with text-driven video style transfer methods.** We compare our model using three different types of sketch styles. Considering the text-guided restriction of these methods, we use the pairing text prompts for sketch transfer. Our model can better capture the sketch semantics of the reference and ensure high-fidelity sketch transfer. Zoom in for a better view.

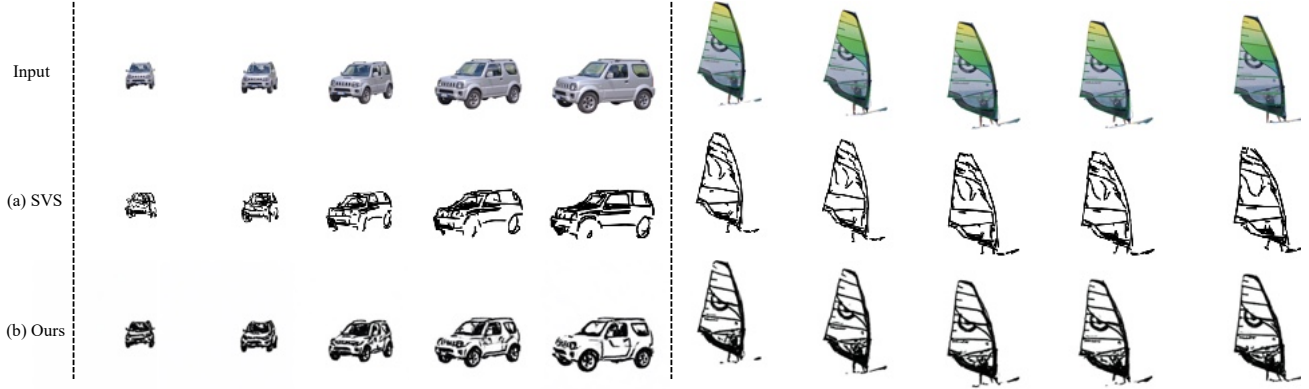


Figure 2. **Qualitative comparison with besel-based method SVS [11].** Our VideoSketcher can better preserve the spatial structure of the original video while achieving more precise and flexible video sketch generation. Zoom in for a better view.

Table 1. **Comparison with besel-based method SVS [11].** The best results are marked in **bold**.

Methods	CLIP-Image $\uparrow$	Pixel-MSE $\downarrow$	FID $\downarrow$	LPIPS $\downarrow$	ArtFID $\downarrow$
SVS	<b>0.9855</b>	0.0825	19.0131	0.2053	24.1740
Ours	0.9794	<b>0.0210</b>	<b>15.3837</b>	<b>0.0976</b>	<b>17.9541</b>

### 2.3. The Function of Structural Consistency Query in TLA

In TLA, we employ a structural consistency query  $Q^{sc}$  to preserve the structural consistency of the original video. This query is applied at the intermediate layers of the denoising U-Net, which have the resolutions of  $32 \times 32$  and  $64 \times 64$ , respectively.

The  $\alpha$  in Equation 8 of the main paper controls the degree

of structure preservation of  $Q^{sc}$  by adjusting the proportion of the query  $Q^c$  from the original video. We gradually increase the value of  $\alpha$  from 0 to 1 to investigate its effects under different degrees of query exchange.

As shown in Figure 3, there is no original structure that affects the stylized video generation when  $\alpha = 0$ , leading to a significant distortion of the structure of objects. As  $\alpha$  grows, the structure is more aligned with the original video. However, a complete replacement of the original query ( $\alpha = 1$ ) will also bring redundant patterns from the original video, hindering the expressiveness of the sketch. As a trade-off between structural consistency and strokes' compactness, we choose  $\alpha = 0.6$  in all our experiments to achieve better results.

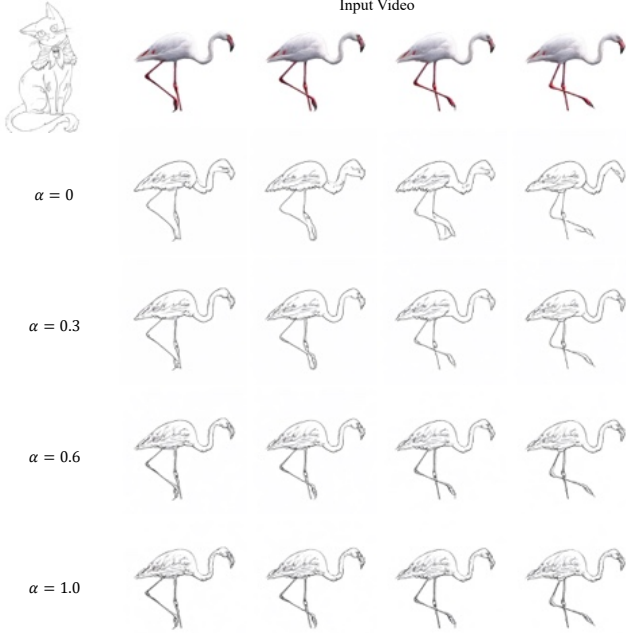


Figure 3. Ablation study of  $\alpha$  in TLA.

## 2.4. The Timing of Applying Style Injection During the Denoising Process

A common consensus of T2I diffusion models is that they focus more on generating the overall semantics of images in the early denoising steps while gradually refining the textures and details in the later periods. Injecting style prematurely may corrupt the overall structure of the image. To visually explain the variations of the attention values across times, we visualize the query of the last self-attention layer in the denoising network of pretrained T2I diffusion model. The queries are mapped into RGB channels by PCA decomposition to observe the structural changes throughout the denoising process. As shown in Figure 4, the object’s structure is not well generated at the early stage ( $t < 0.1T$ ), while it becomes more stable and complete later on. As a result, we deploy a time threshold  $\tau$  in TLA to control the timing of applying TLA for sketch stylization. We also experiment by modifying the value of  $\tau$ , the results are presented in Figure 5. We gradually adjust the timing of applying style injection from 0 to  $0.4T$ . As shown in Figure 5, the earlier injection of the sketch style results in incomplete boundaries on the object’s structure. On the contrary, a late injection exhibits less effect on sketch stylization due to insufficient time for modifying the style. Consequently, the generated video is highly similar to the original one, resulting in inferior stylization effects. In practice,  $\tau$  is empirically set to  $0.1T$ , which can be dynamically extended to balance the trade-off between style transfer and structural preservation.

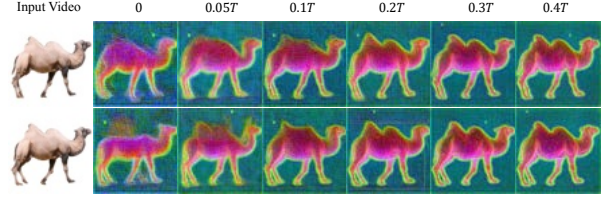


Figure 4. Query visualization in last self-attention layer. The pixels with the same semantics are represented by the same colors (e.g., red for the camel and green for the blank background).

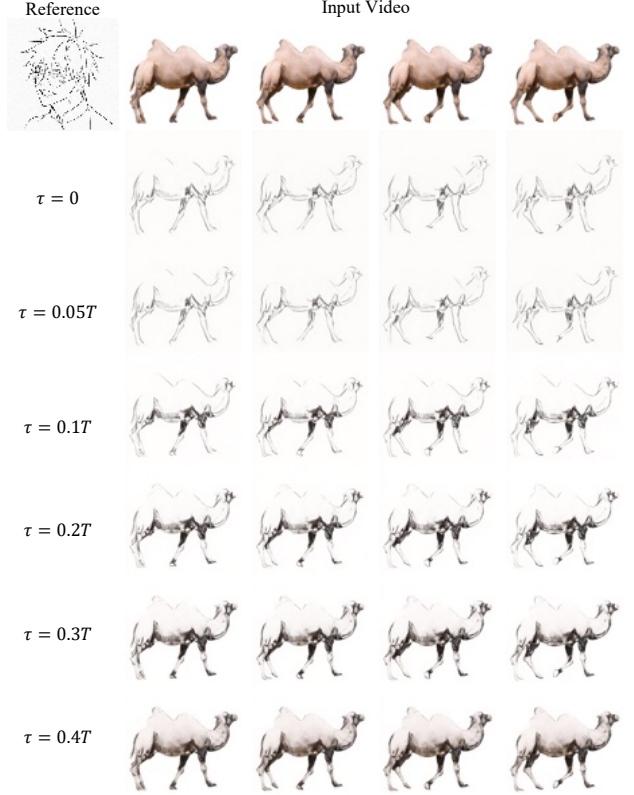


Figure 5. Stylization results of different  $\tau$ . The first row is the input video with the reference sketch.

## 2.5. More Analysis of SDA

We conduct an extra experiment to evaluate the effectiveness of our proposed SDA. For comparison, we employ sketch amplification with several fixed amplification rates  $\gamma$ , such that Equation 13 can be rewritten as:

$$\bar{A}^g = \gamma A^g, \quad \gamma \in [1, 3, 5].$$

The results are shown in Figure 6. As illustrated in the second column of the figure, a small scaling factor leads to excessive blurriness across the entire image. Employing SDA effectively alleviates this issue; however, increasing the scaling factor excessively introduces pronounced artifacts,

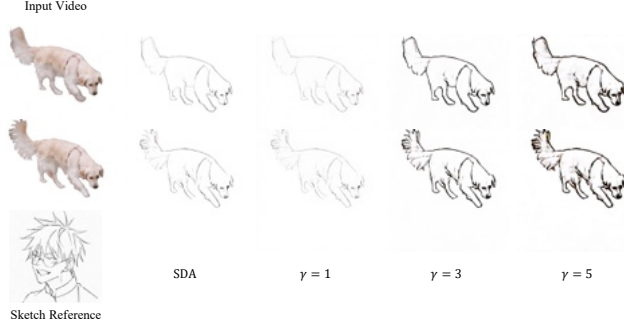


Figure 6. Comparison between SDA and fixed rate amplification.

compromising the overall stylistic integrity.

## 2.6. Results with diverse sketch style

To validate our VideoSketcher’s ability to accurately capture the target sketch style, we further perform visual comparison by sketching videos using diverse sketch reference images of the same content. The reference sketches are adopted from the 4SKST dataset [8]. We collect three example reference images for each style type. As shown in Fig. 7, our VideoSketcher successfully disentangles the style and content of the reference sketch, effectively extracting the intrinsic style feature without being affected by the content.

## 3. User Study

We conducted a user study to evaluate the quality of the generated results by selecting 10 representative videos and engaging 50 participants to rank the videos based on four key criteria: (1) consistency with the reference style (SC), (2) preservation of structural integrity from the original video (SI), (3) temporal coherence across frames (TC), and (4) overall visual performance (OVP). The results are reported in Table 2. Our method outperforms existing approaches with the highest average scores across all evaluated metrics, indicating that the sketch stylization of our VideoSketcher is more user-preferable.

Table 2. User study on DAVIS [7]. The best and the second best results are **bolded** and underlined, respectively.

Method	SC (%) ↑	SI (%) ↑	TC (%) ↑	OVP (%) ↑
StyleID [4]	5.2	<b>8.0</b>	5.5	6.0
Cross-IA [1]	<u>10.8</u>	3.4	4.2	4.2
Ref2sketch [2]	3.5	6.5	5.8	5.0
Semi-ref2sketch [8]	4.5	4.1	<u>10.4</u>	<u>8.5</u>
IP-Adapter [10]	6.0	4.2	3.6	4.5
Ours	<b>70.0</b>	<b>73.7</b>	<b>70.5</b>	<b>71.7</b>

## 4. More Results

We exhibit more sketch stylization results with references selected from the 4SKST dataset [8] in Figure 8, Figure 9, and Figure 10. The example videos are collected from the DAVIS dataset, the LOVEU-TGVE dataset [9], and open available websites. We recommend zooming in for better visualization.

## References

- [1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH Conf.*, pages 1–12, 2024. 4
- [2] Amirsaman Ashtari, Chang Wook Seo, Cholmin Kang, Sihun Cha, and Junyong Noh. Reference Based Sketch Extraction via Attention Mechanism. *ACM Trans. Graph.*, 41(6):1–16, 2022. 4
- [3] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Int. Conf. Comput. Vis.*, pages 23206–23217, 2023. 1
- [4] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8795–8805, 2024. 4
- [5] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Int. Conf. Comput. Vis.*, pages 15954–15964, 2023. 1
- [6] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-P2P: Video Editing with Cross-attention Control. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8599–8608, 2024. 1
- [7] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 724–732, 2016. 4
- [8] Chang Wook Seo, Amirsaman Ashtari, and Junyong Noh. Semi-supervised reference-based sketch extraction using a contrastive learning framework, 2024. 1, 4
- [9] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, Rui He, Feng Hu, Junhua Hu, Hai Huang, Hanyu Zhu, Xu Cheng, Jie Tang, Mike Zheng Shou, Kurt Keutzer, and Forrest Iandola. Cvr 2023 text guided video editing competition, 2023. 4
- [10] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 4
- [11] Yudian Zheng, Xiaodong Cun, Menghan Xia, and Chi-Man Pun. Sketch Video Synthesis, 2023. 1, 2

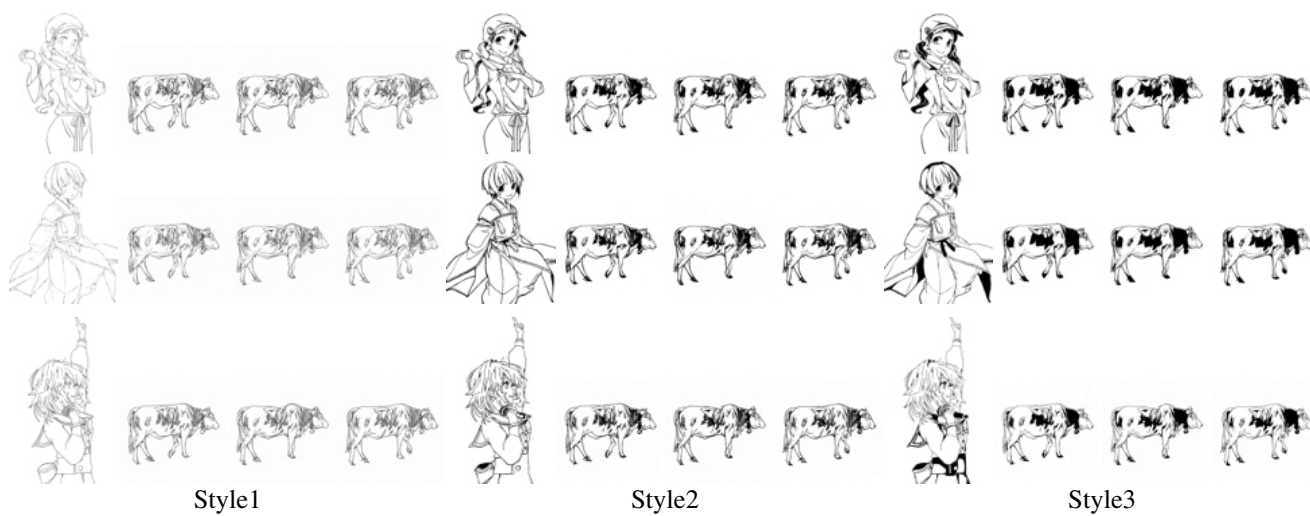


Figure 7. Sketch stylization results with different style references of the same content. Please zoom in for a better view.

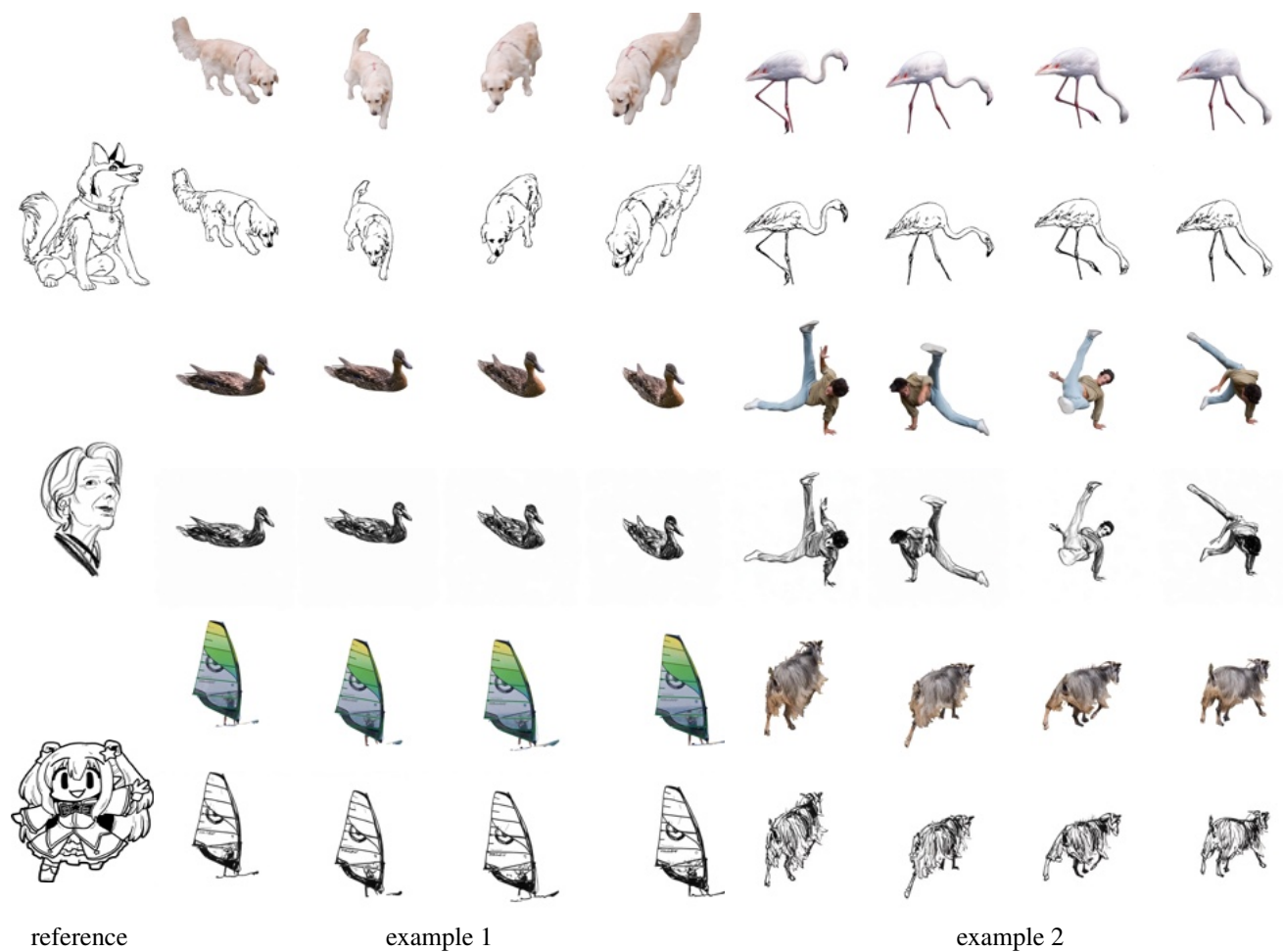


Figure 8. More results with diverse sketch style.

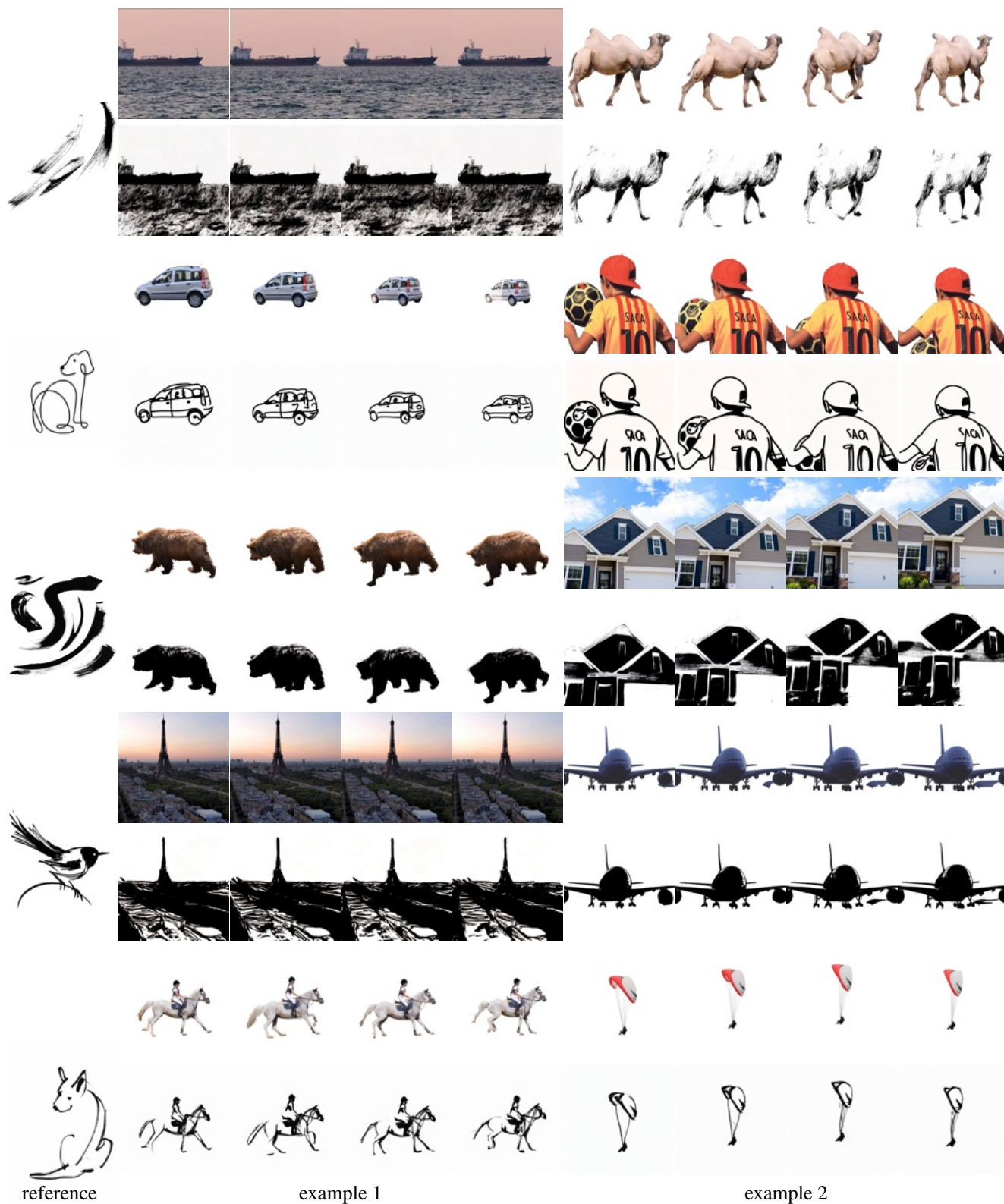


Figure 9. More results with diverse sketch style.

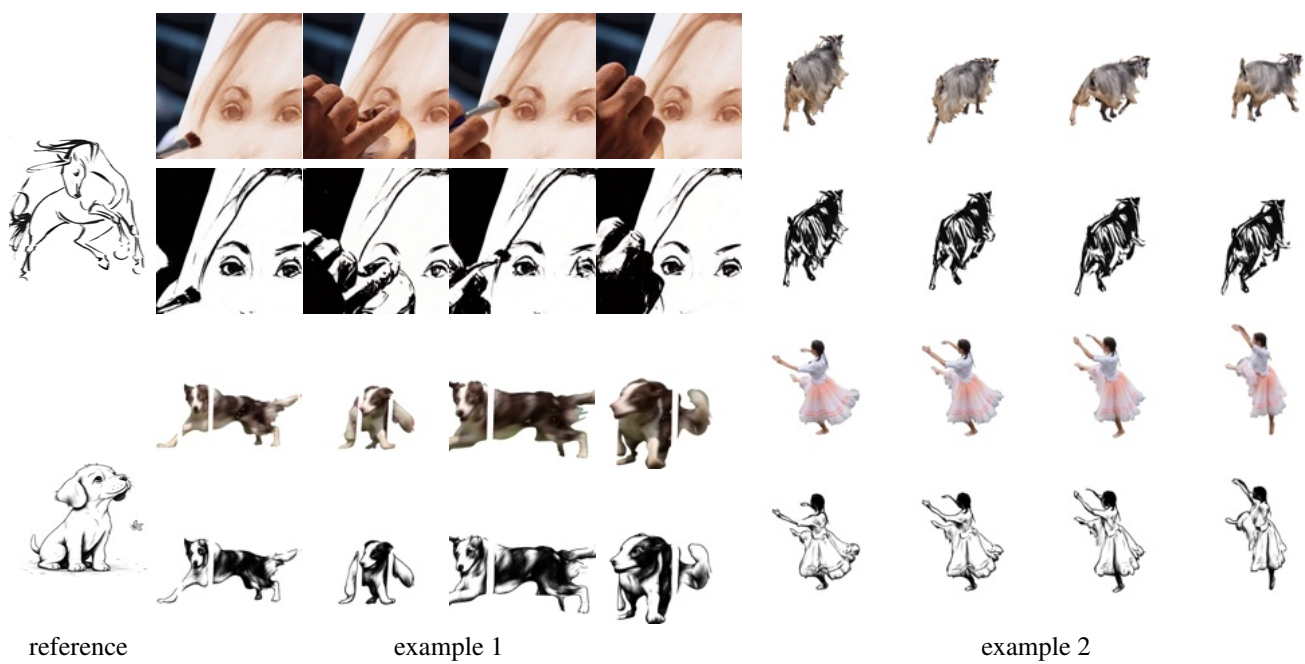


Figure 10. More results with diverse sketch style.