

VIRES: Video Instance Repainting via Sketch and Text Guided Generation

Supplementary Material

Shuchen Weng^{1,2†‡} Haojie Zheng^{3,4†} Peixuan Zhang⁵ Yuchen Hong^{1,2}
 Han Jiang³ Si Li⁵ Boxin Shi^{1,2*}

¹State Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

²National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

³OpenBayes Information Technology Co., Ltd. ⁴School of Software and Microelectronics, Peking University

⁵School of Artificial Intelligence, Beijing University of Posts and Telecommunications

scweng@baai.ac.cn, suimu@stu.pku.edu.cn, {pxzhang, lisi}@bupt.edu.cn
 yuchenhong.cn@gmail.com, hahn@openbayes.com, shiboxin@pku.edu.cn

A. Appendix

A.1. Variations of Sequential ControlNet

We present four typical architectures of Sequential ControlNet in Tab. S1, where "Conv" denotes a convolutional layer, "Block" means a residual block, and "Down" is a down-sampling layer. Numbers in brackets indicate the input and output channel dimensions, respectively. To determine the optimal architecture under constrained computational resources, we train these variations for 10K steps, excluding the sketch attention and sketch-aware encoder. Quantitative results on VIRESET are shown in Tab. S2, and the best-performing architecture is selected for our model.

Table S1. Architecture of Sequential ControlNet variations.

	Ours (VIRES)	Variation_1	Variation_2	Variation_3
block_1	Conv (3, 72)		Conv (3, 36)	Conv (3, 72)
	Conv (72, 72)		Conv (36, 36)	Conv (72, 72)
	Down (72, 144)	Down (3, 64)	Down (36, 72)	Down (72, 144)
block_2	Conv (144, 144)	Conv (64, 64)		Conv (144, 144)
	Conv (144, 288)	Conv (64, 128)		Conv (144, 288)
	Conv (288, 288)			
	Block (288, 288)	Block (128, 128)	Block (72, 72)	Block (288, 288)
	Down (288, 576)	Down (128, 128)	Down (72, 288)	Down (288, 576)
block_3	Block (576, 576)	Block (128, 256)	Block (288, 288)	Block (576, 576)
	Down (576, 1152)	Down (256, 256)	Down (288, 1152)	Down (576, 1152)
block_4	Block (1152, 1152)			
	Conv (1152, 1152)	Block (256, 256)	Conv (1152, 1152)	Conv (1152, 1152)
	Conv (1152, 1152)		Conv (1152, 1152)	Conv (1152, 1152)
	Conv (1152, 1152)	Conv (256, 1152)	Conv (1152, 1152)	Conv (1152, 1152)

A.2. Validation of additional conditions

VIRES is designed specifically for sketch-based video instance repainting, adopting the Standardized Self-Scaling (SSS) to extract condition features. To explore its generalization to other conditioning signals, we augment our dataset with edge maps [1] and depth maps [6] to pro-

[†] Equal contribution.

[‡] Now at Beijing Academy of Artificial Intelligence.

* Corresponding author.

Table S2. Quantitative experiment results of Sequential ControlNet variations. All scores except PSNR are percentages.

Method	PSNR ↑	SSIM ↑	WE ↓	FC ↑	TC ↑
Variation_1	18.54	69.12	5.36	90.79	15.85
Variation_2	18.49	68.63	5.47	89.14	15.83
Variation_3	18.64	69.89	5.40	90.94	15.85
Ours (VIRES)	19.93	72.85	5.33	91.07	15.87

vide the structure guidance. We then retrain VIRES from scratch on these conditions. Due to limited computational resources, we train for 10K steps with edge maps as a preliminary investigation and 30K steps with depth maps due to its slower convergence. Finally, we compare two VIRES variants: one using SSS and the other using simple addition for feature extraction. As shown in Tab. S3, SSS provides considerable advantages for edge maps, but only marginal improvements for depth maps. We suggest that this difference arises because both sketch and edge maps have high-contrast transitions between black lines and the white background, allowing self-scaling to effectively capture structure details. In contrast, depth maps are smoother and lack such sharp transitions, limiting the benefits of the self-scaling operation. We further present the qualitative results in Fig. S1.

Table S3. Quantitative experiment results of VIRES variants using edge and depth maps. All scores except PSNR are percentages.

Method	PSNR ↑	SSIM ↑	WE ↓	FC ↑	TC ↑
Repainting with edge maps					
W/o SSS	19.43	72.01	6.31	90.04	15.83
W/ SSS	20.48	74.28	6.00	90.49	16.16
Repainting with depth maps					
W/o SSS	18.04	66.16	5.86	91.28	15.98
W/ SSS	18.39	67.28	5.80	91.32	16.14

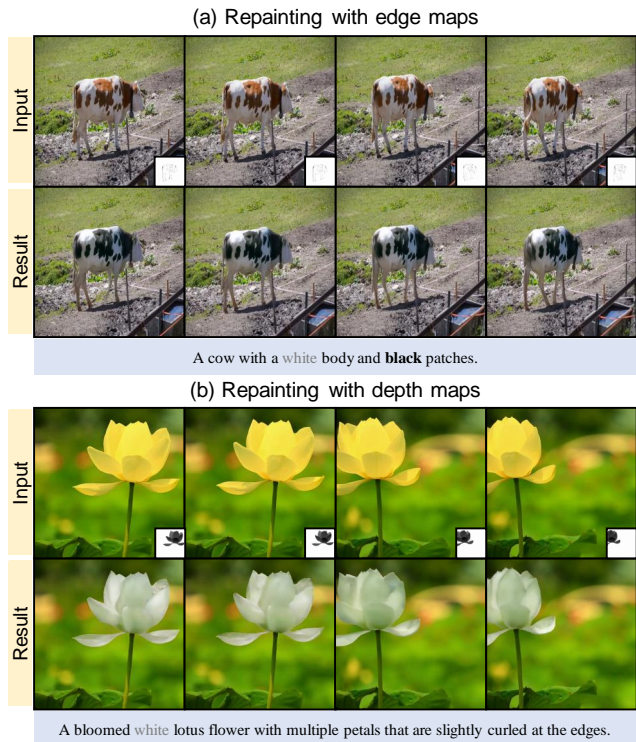


Figure S1. Examples of repainting with additional conditions.

Table S4. Quantitative experiment results of VIREs using sketch guidance of varying sparsity. All scores except PSNR are percentages.

Method	PSNR \uparrow	SSIM \uparrow	WE \downarrow	FC \uparrow	TC \uparrow
$D = 1$	21.87	73.41	5.12	92.21	16.15
$D = 2$	22.64	74.69	5.14	92.21	16.17
$D = 4$	23.24	76.00	5.13	92.18	16.16
$D = 8$	23.36	76.51	5.14	92.19	16.18
$D = 16$	23.67	77.15	5.12	92.19	16.19
$D = 51$	23.87	77.87	5.09	92.23	16.19

A.3. Robustness of sparse conditions

VIREs allows users to provide sparse sketch guidance for video instance repainting, minimizing user effort. To investigate the impact of sketch guidance sparsity on repainting performance, we evaluate VIREs on the VIREsET, using varying interval indices $d \in \{0, \dots, 4\}$, corresponding to $D = 2^d$ sketch frames, and a full set of $D = 51$. As shown in Tab. S4, even with a single sketch frame, VIREs can produce high-quality results (PSNR and SSIM) with robust temporal consistency (WE and FC).

A.4. Compatibility with DiT backbone

In this paper, we build VIREs upon the pre-trained OpenSora v1.2 [9]. Given that many text-to-video models [8, 10] share a similar DiT backbone architecture, our proposed modules offer potential compatibility with these approaches, including the Sequential ControlNet for layout initialization (Sec. 4.2), the standardized self-scaling for details capture (Sec. 4.2), the sketch attention (Sec. 4.3) for semantic injection, and the sketch-aware encoder for structure alignment (Sec. 4.4). We believe our work will inspire further research on guiding pre-trained text-to-video models and open new avenues for conditional video repainting.

A.5. Organization of supplementary video

We provide a supplementary video to dynamically showcase our repainting results. The video is structured as follows: (i) **Typical application scenarios**: We demonstrate four typical repainting scenarios and compare our results with relevant methods [2–5, 7]. Instance repainting/replacement results are shown in Fig. S2, and instance generation/removal results are in Fig. S3. (ii) **Sketch-to-video generation and inpainting**: We demonstrate sketch-to-video generation and conditional video inpainting, comparing our results with VideoComposer [5], as it is the only relevant method that supports this functionality. Results are shown in Fig. S4. (iii) **Sparse sketch guidance**: We showcase sparse sketch guidance, repainting two distinct variations of the same video using only two different first sketch frames. This functionality is not supported by existing methods. Results are shown in Fig. S5. (iv) **Long-duration video repainting**: We demonstrate repainting on a long-duration (13-second) video, with representative frames presented in Fig. S6. (v) **Comparison and ablation study**: Finally, the video includes comparisons with other methods (Sec. 5.1) and additional ablation studies (Sec. 5.2). To improve visual clarity and facilitate detailed comparison, the video playback speed is halved ($2\times$ slower).

References

- [1] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. 1
- [2] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. RAVE: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *CVPR*, 2024. 2
- [3] 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 *ICCV*, 2023.
- [4] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang.

VidToMe: Video token merging for zero-shot video editing. In *CVPR*, 2024.

- [5] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional video synthesis with motion controllability. In *NeurIPS*, 2024. 2
- [6] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1
- [7] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, 2023. 2
- [8] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2
- [9] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all, 2024. 2
- [10] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024. 2

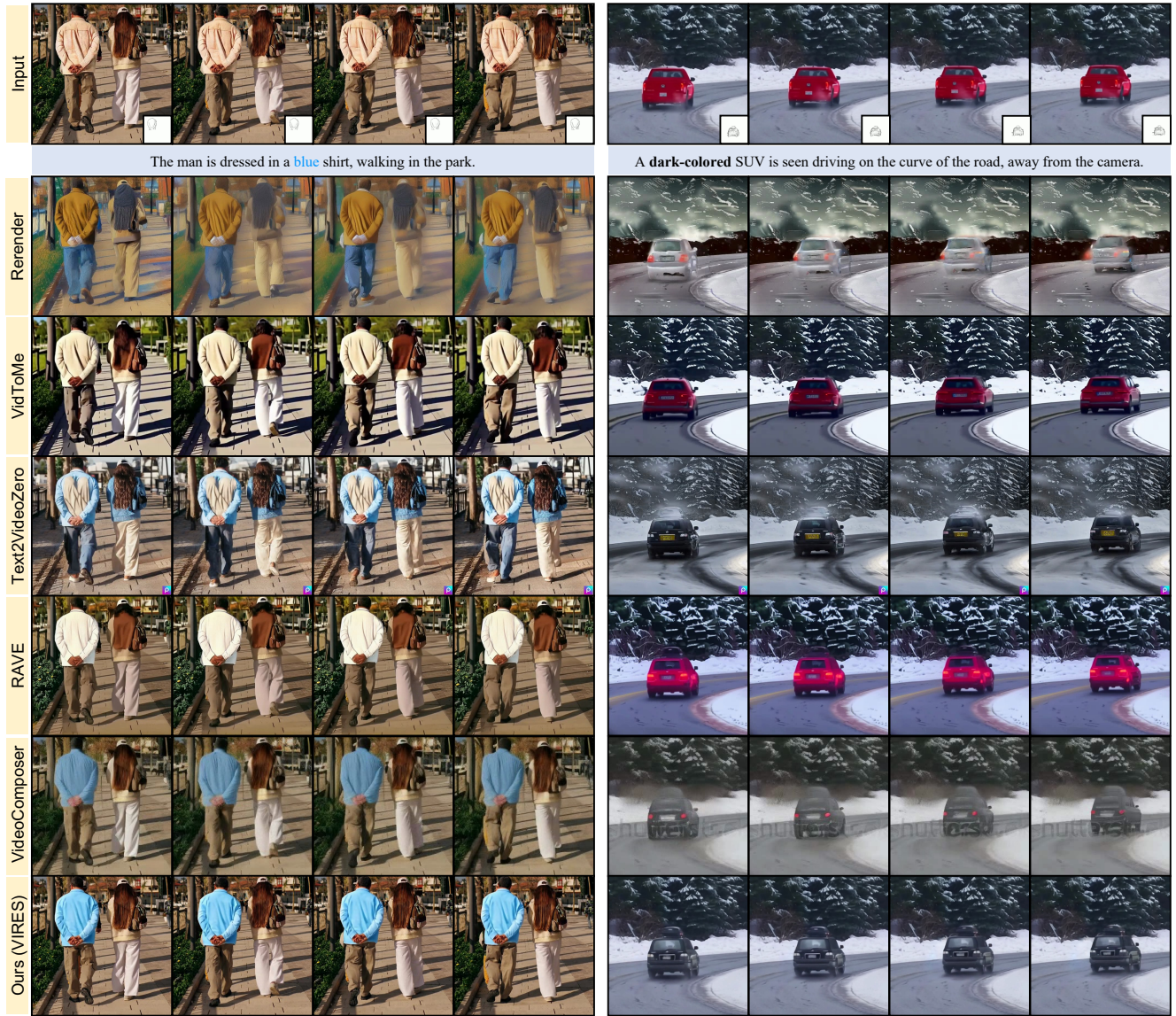


Figure S2. Typical application scenarios. **Left:** Video instance repainting. **Right:** Video instance replacement.

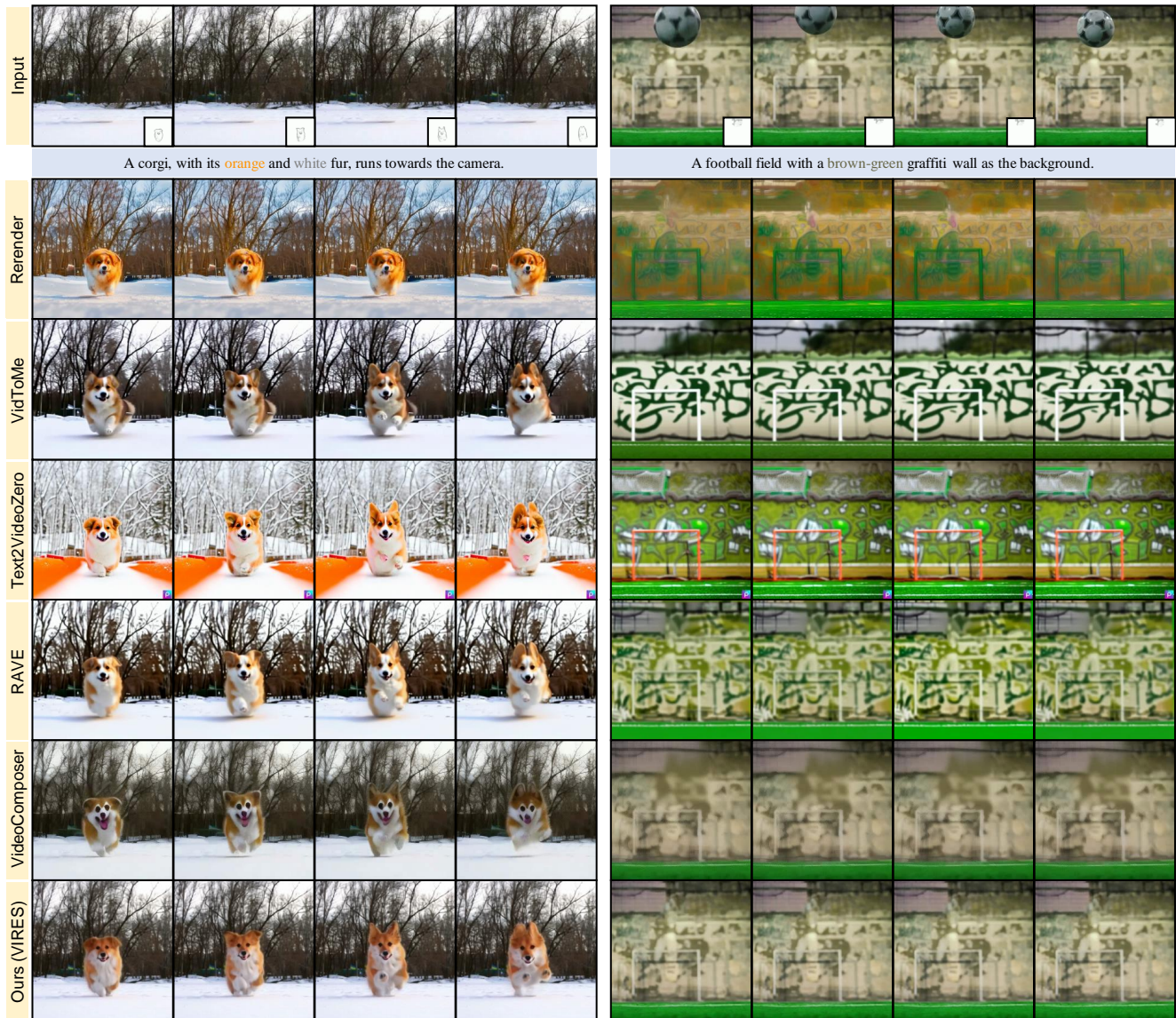


Figure S3. Typical application scenarios. **Left:** Custom instance generation. **Right:** Specified instance removal.

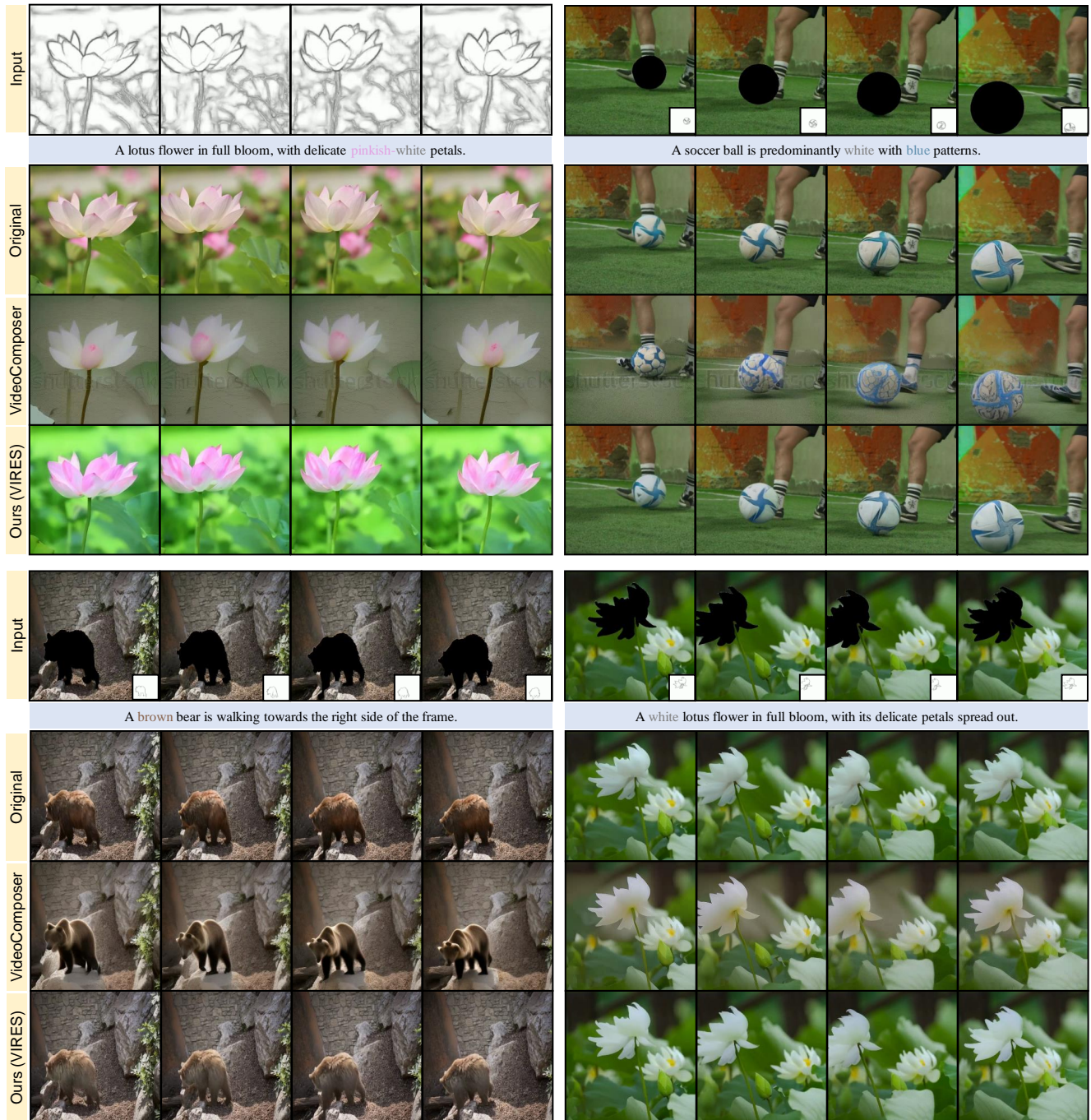


Figure S4. Sketch-to-video generation and inpainting. **Topleft:** Sketch-to-video generation. **Others:** Sketch-to-video inpainting.



Figure S5. Sparse sketch guidance. Repainting the video using different first sketch frames. **Left:** First variation. **Right:** Second variation.

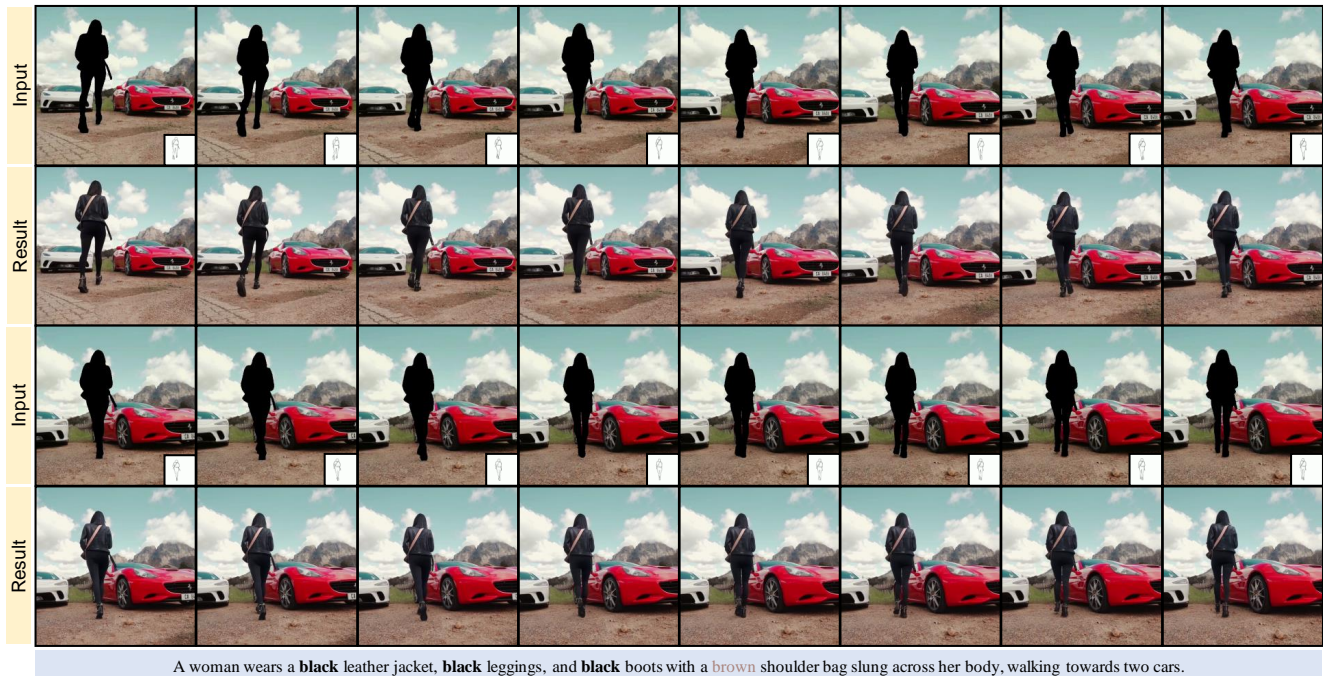


Figure S6. Long-duration video repainting. Restoring the damaged video to clearly depict a realistic female character.