Aligning Global Semantics and Local Textures in Generative Video Enhancement — Supplementary Material

Zhikai Chen[1], Fuchen Long[2]*, Zhaofan Qiu[2], Ting Yao[2], Wengang Zhou[1]†, Jiebo Luo[3], and Tao Mei[2]
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[2]HiDream.ai Inc. [3]University of Rochester, Rochester, NY USA
czk654@mail.ustc.edu.cn, {longfuchen, qiuzhaofan, tiyao}@hidream.ai
zhwg@ustc.edu.cn, jluo@cs.rochester.edu, tmei@hidream.ai

The supplementary material contains: 1) detailed performance comparisons on three subsets of AIGC-Vid; 2) more comparisons on YouHQ40 [9]; 3) validation on different key-frame selections; 4) validation on different image reference generation approaches; 5) visualization results of GenVE by using different noise levels; 6) analysis of model complexity; 7) more visualization comparisons on the YouHQ40, VideoLQ [2], and AIGC-Vid datasets; 8) hard case analysis with significant scene change.

1. Detailed Performances on AIGC-Vid

We build AIGC-Vid by exploiting three text-to-video (T2V) advances, i.e., ModelScope [8], LaVie [7], VideoCrafter [3] to generate videos with 93 text prompts in VBench [5]. Thus, AIGC-Vid contains 279 videos for evaluation in total. Table 1 details the performances of video enhancement on the three video subsets synthesized by different T2V models. GenVE achieves the best performances across all metrics on the subsets of ModelScope and VdieoCrafter.

2. More Comparisons on YouHQ40

In Table 2, we report the performances on four more evaluation metrics (i.e., E_{warp} [], PSNR, SSIM and LPIPS) across different methods on YouHQ40. GenVE achieves the lowest flow warping error E_{warp} (i.e., 2.188) among all methods, verifying the efficacy of GenVE on keeping better temporal coherence. In terms of PSNR and SSIM, default GenVE may exhibit sub-optimal performances. Nevertheless, we can adjust the degree of video fidelity via using different noising step T_1 in our GenVE. When decreasing T_1 from 900 to 600, GenVE achieves 23.84dB PSNR on YouHQ40, which attains an improvement of 5.30dB.

*Co-corresponding author.

Table 1. Performance comparisons in terms of MUSIQ, DOVER and CLIP-IQA on the three video subsets (ModelScope, LaVie, VideoCrafter) of AIGC-Vid.

Generation	Enhancement	MUSIQ↑	DOVER↑	CLIP-IQA↑
ModelScope	DATSR [1]	47.40	53.63	0.5124
	LaVie-SR [7]	58.43	68.90	<u>0.6080</u>
	Upscale-A-Video [9]	55.08	68.24	0.5803
	VEnhancer [4]	<u>58.89</u>	<u>68.99</u>	0.5460
	GenVE	**59.31**	**69.53**	**0.6107**
LaVie	DATSR [1]	61.20	60.89	0.6123
	LaVie-SR [7]	64.22	**78.76**	0.6603
	Upscale-A-Video [9]	60.58	75.30	0.6207
	VEnhancer [4]	**65.63**	77.86	<u>0.6639</u>
	GenVE	<u>65.44</u>	<u>77.87</u>	**0.6675**
VideoCrafter	DATSR [1]	63.84	57.26	0.6368
	LaVie-SR [7]	<u>68.47</u>	<u>82.88</u>	<u>0.6989</u>
	Upscale-A-Video [9]	64.33	78.99	0.6646
	VEnhancer [4]	68.39	82.54	0.6925
	GenVE	**68.93**	**83.72**	**0.7011**

3. Different Key-frame Selection

To determine the best key frame selection strategy, we experimented by using different key frames in GenVE, and summarize the results in Table 3 . The performance differences between using the first, middle and last frame of the input video are negligible. We speculate that the results are contributed by our masking augmentation which makes GenVE not highly sensitive to the reference visual contents. Thus, we empirically select the first frame for high-quality image reference generation.

4. Different Image Reference Generation

To investigate the influence of image reference generation, we conduct another run of replacing our semantics-aligning image diffuser with a CNN-based super-resolution (SR) model, i.e., ESRGAN [6]. As shown in the lower part of

Table 2. Performance comparisons in terms of E_{warp}, PSNR, SSIM and LPIPS on YouHQ40.

Metrics	LaVie-SR[7]	Upscale-A-Video[9]	VEnhancer[4]	GenVE
$E_{warp}(\times 10^{-3})$ [] $\downarrow$	7.777	2.401	2.230	2.188
PSNR $\uparrow$	22.32	25.83	17.63	18.45
SSIM $\uparrow$	0.527	0.733	0.471	0.484
LPIPS $\downarrow$	0.476	0.268	0.541	0.518

Table 3, GenVE outperforms such "CNN-based SR + Diffusion" paradigm across all three metrics on YouHQ40, verifying the efficacy of semantics-aligning image diffuser for image reference enhancing. Compared to CNN-based SR models, diffusion models excel at synthesizing images with superior aesthetic quality and finer details, which benefits the texture aligning and transferring in video enhancement.

5. Different Noise Levels

In addition to the investigation of noise level n on the image reference as mentioned in Sec. 4.3 of the main paper, we further study the noise level T_1 on the up-sampled low-quality video latent in the forward diffusion process. Figure 1 showcases five video enhancement results of GenVE by using different noise levels T_1 on YouHQ40. As observed in the figure, exploiting higher noise levels makes GenVE generate more fine-grained details, and achieves better visual quality. Nevertheless, when the noise level T_1 equals to $1,000$, GenVE tends to hallucinate some extra visual contents (e.g., messy textures of the house wall in the 1st and 2nd cases), which jeopardizes the spatial fidelity. Therefore, we empirically set T_1 as 900 to seek a good balance between video quality and spatial fidelity.

6. Model Complexity

We report the model complexity comparisons in Table 4. All models are tested on YouHQ40 with a single NVIDIA H100 GPU. Compared to the best competitor VEnhancer, GenVE leads a comparable TFLOPs/runtime of 102.3/3.37s per denoising step. The extra TFLOPs/runtime brought by image diffusion and augmentation in GenVE are 6.0 and 0.04s, accounting for a small part of computation.

7. More Visualization Comparisons

In Figure 2, 3 and 4, we further present six visualization cases on three datasets for comparison. Overall, the videos generated by GenVE attain better visual quality and contain more fancy local details. We also provide a video demo of the comparisons (file "GenDemo.mp4") for more details.

8. Hard Case Analysis

Additionally, we provide two hard cases with significant scene change (files of "HardCase1.mp4" and "HardCase2.mp4"). In each video, the left part is the low-quality video input and the right part is the enhanced result by GenVE. The region in the red box is presented in the zoom-in view. As shown in the two cases, GenVE well enhances videos under extreme scenarios, verifying the merit of our semantic and texture alignment. Note that there are some minor inconsistencies when handling over-exposed video as shown in the case of "HardCase2.mp4". We will further explore the feasibility of introducing additional lighting-aware conditions in our model to address this issue in the future.

Table 3. Ablation studies of using different key frames and image reference generation approaches on YouHQ40.

Ablation	Setting	MUSIQ ↑	DOVER ↑	CLIP-IQA ↑
Key-frame selection	1st frame (GenVE)	69.93	83.50	0.596
	15th frame	70.05	83.48	0.596
	30th frame	69.93	83.56	0.595
Image reference generation	ESRGAN [6]	68.19	80.12	0.557

Table 4. Comparisons of model complexity on YouHQ40. (All runs are tested on a single NVIDIA H100 GPU.)

Model	# of Params	TFLOPs	Memory	Runtime (per step)	MUSIQ ↑
LaVie-SR[7]	0.36B	18.5	19G	0.35s	64.17
Upscale-A-Video[9]	0.36B	18.5	18G	1.00s	63.69
VEnhancer[4]	2.0B	101.1	42G	3.29s	68.99
GenVE	2.0B	102.3	58G	3.37s	69.93

References

[1] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based Image Super-Resolution with Deformable Attention Transformer. In *ECCV*, 2022. 1

[2] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating Tradeoffs in Real-World Video Super-Resolution. In *CVPR*, 2022. 1

[3] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. In *CVPR*, 2024. 1

[4] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. VEnhancer: Generative Space-Time Enhancement for Video Generation. *arXiv:2407.07667*, 2024. 1, 2, 3

[5] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *CVPR*, 2024. 1

[6] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *ICCV*, 2021. 1, 3

[7] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models. *IJCV*, 2024. 1, 2, 3

[8] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. *arXiv:2311.04145*, 2023. 1

[9] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution. In *CVPR*, 2024. 1, 2, 3

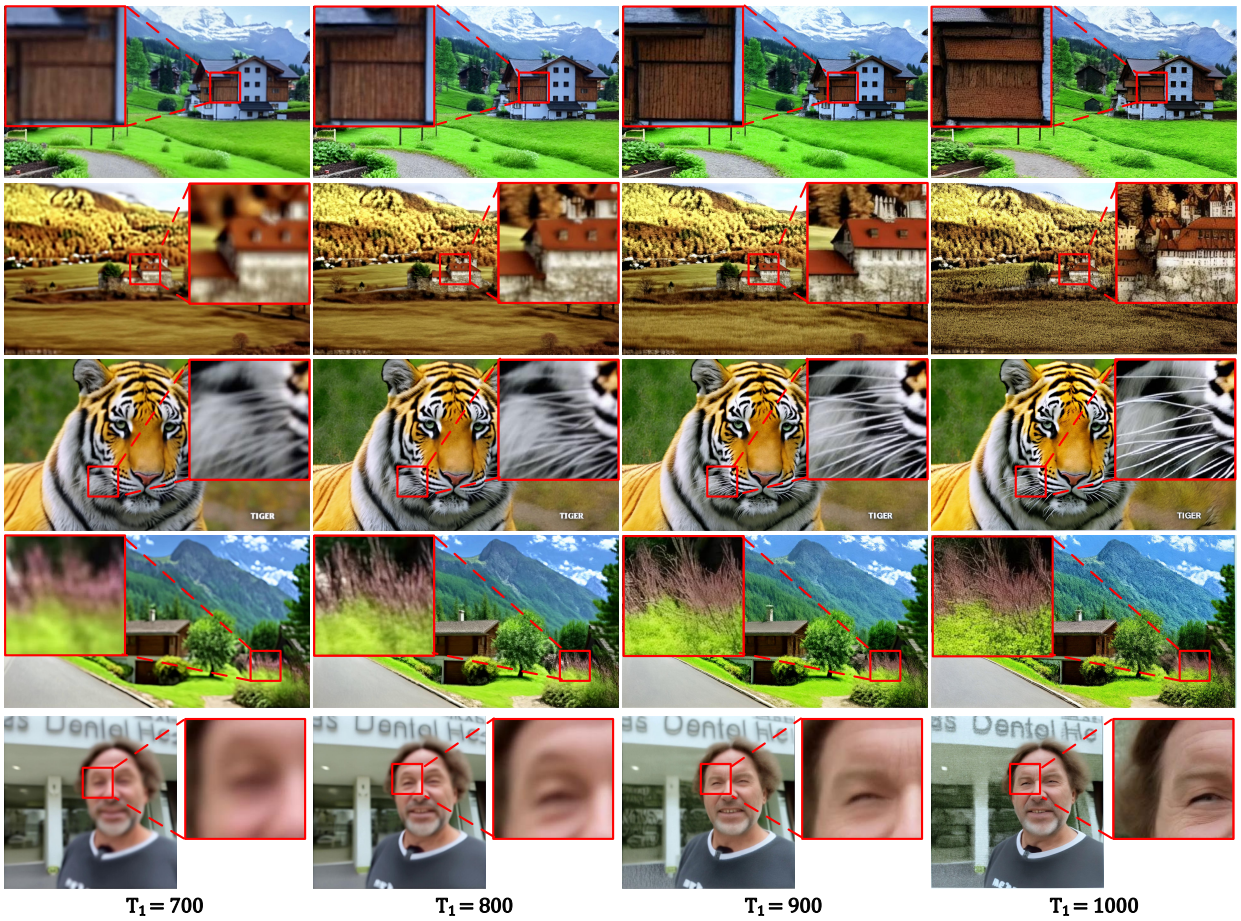

Figure 1. Five video enhancement results of GenVE by using different noise levels T_1 in forward diffusion process on YouHQ40.

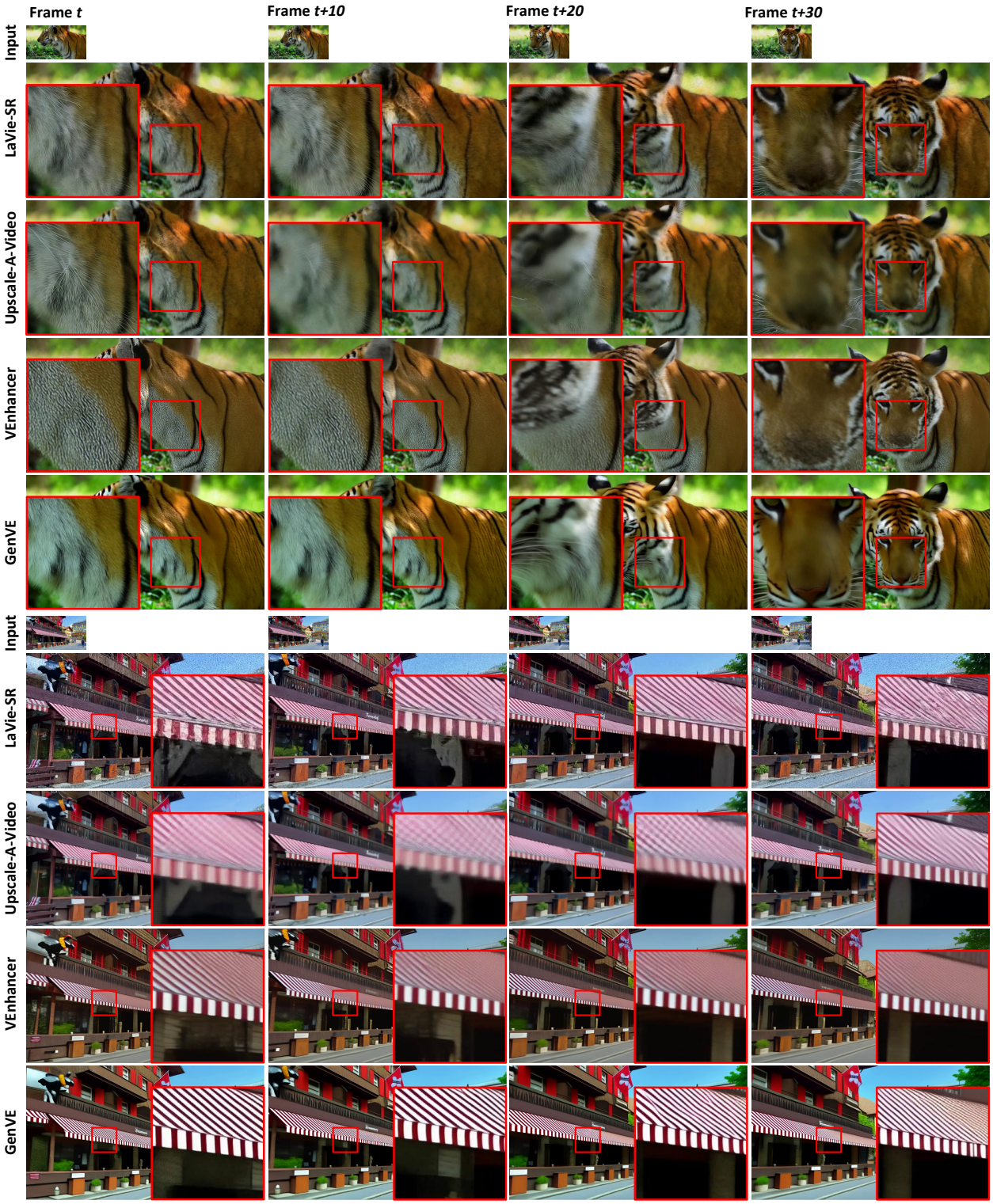

Figure 2. Two visual examples of video enhancement results by different approaches on videos from YouHQ40. The region in the red box is presented in the zoom-in view for comparison.

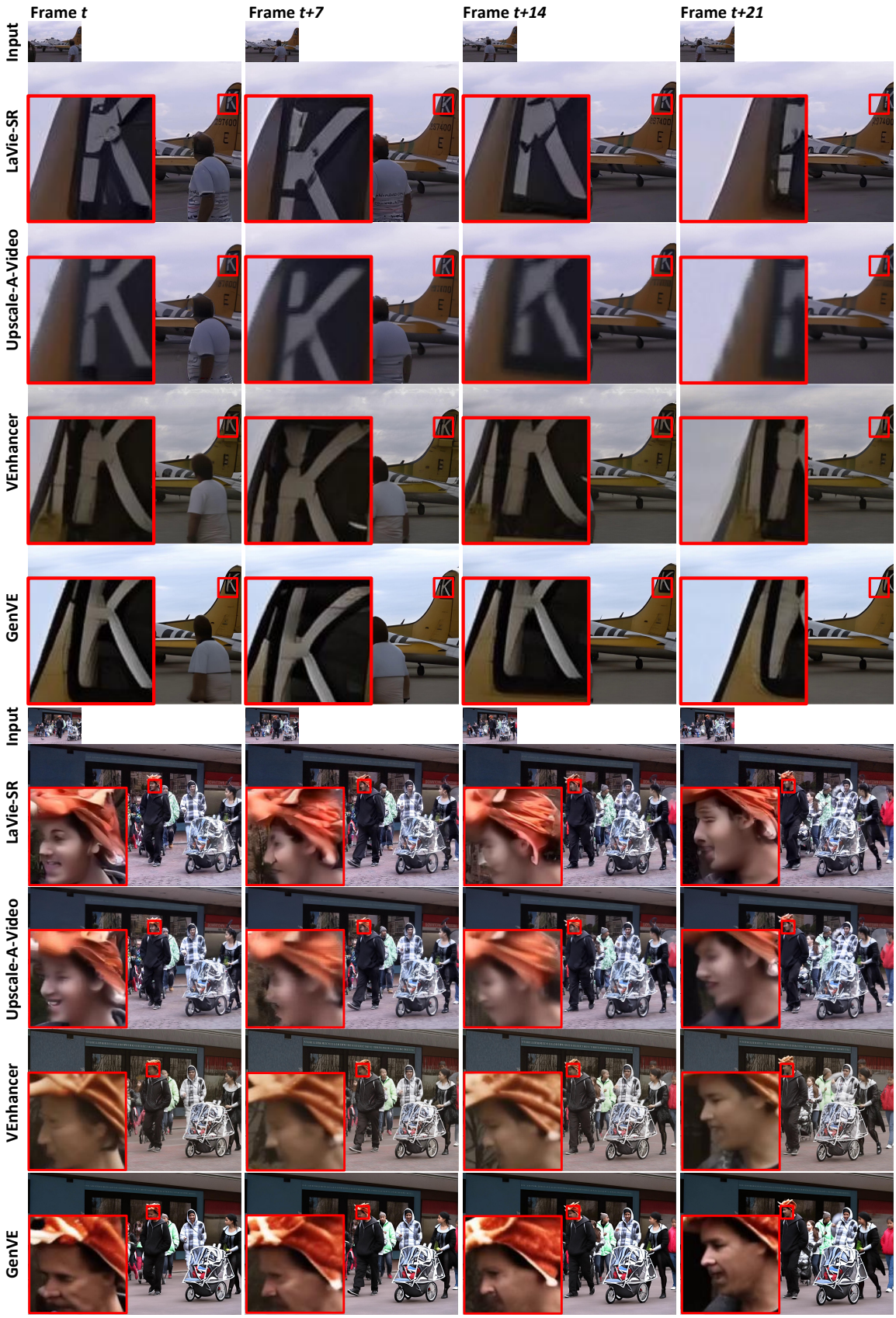

Figure 3. Two visual examples of video enhancement results by different approaches on videos from VideoLQ. The region in the red box is presented in the zoom-in view for comparison.

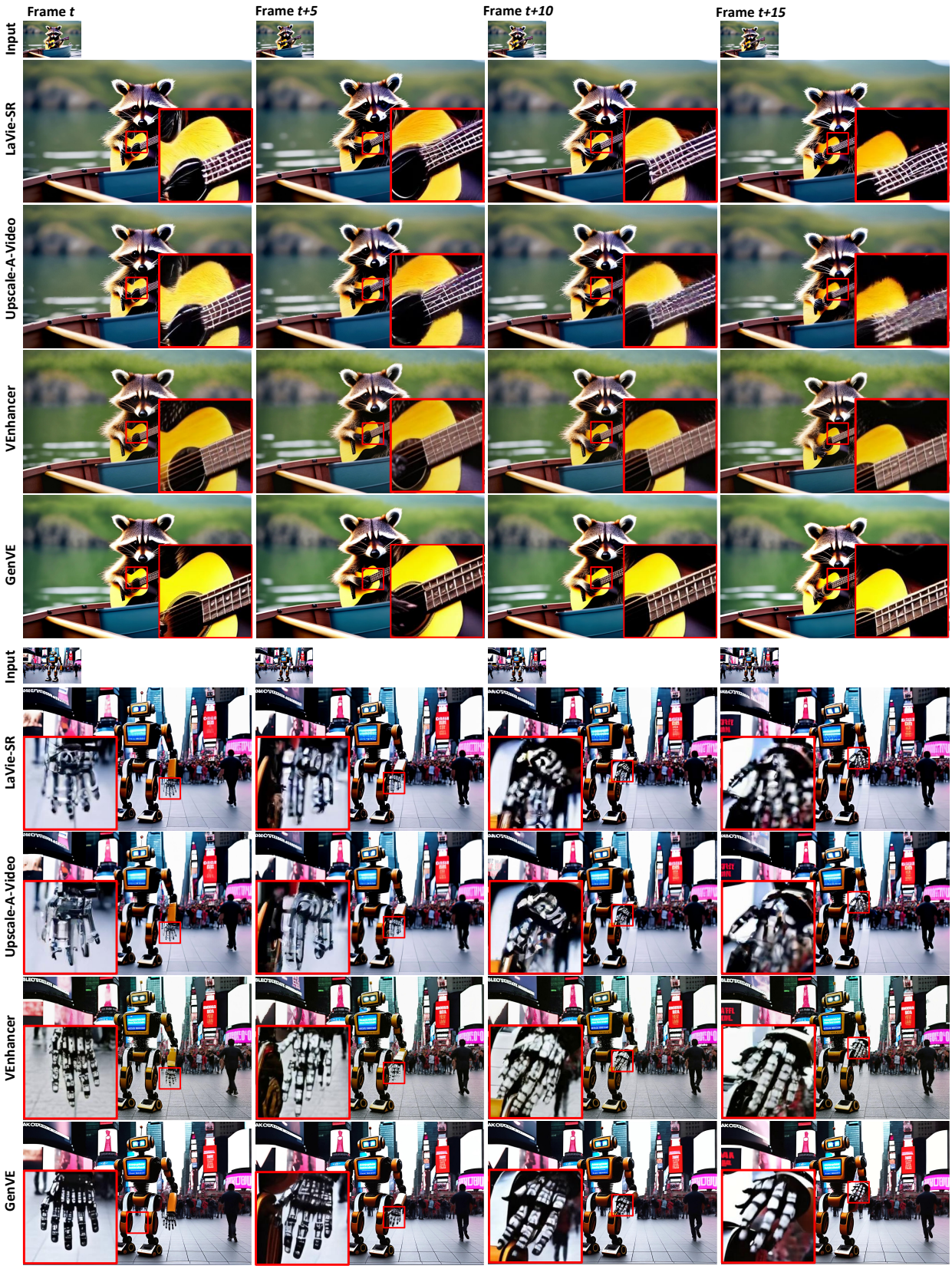

Figure 4. Two visual examples of video enhancement results by different approaches on videos from AIGC-Vid. The region in the red box is presented in the zoom-in view for comparison.