STCDiT: Spatio-Temporally Consistent Diffusion Transformer for High-Quality Video Super-Resolution - Supplemental Material -

Junyang Chen Jiangxin Dong Long Sun Yixin Yang Jinshan Pan*
Nanjing University of Science and Technology
https://jychen9811.github.io/STCDiT_page

Overview

In this supplemental material, we first provide more analysis and discussions of our method in Section A. Section B describes the SportsLQ dataset. More experimental results are shown in Section C.

A. In-Depth Analysis of STCDiT

A.1. Effect of Motion-Aware VAE Reconstruction on Video Restoration

Extending the analysis in Section 5.1 of the main paper, we further validate the effectiveness of the motion-aware reconstruction method for video restoration. We compare Wan [8] with a baseline method that replaces the standard VAE with our motion-aware VAE, whose encoded features are subsequently fed into a LoRA-trained DiT [8] for restoration (*Base* for short). Figure 6 shows a visual comparison on a degraded video with severe camera shaking. The comparison results demonstrate that the motion-aware VAE reconstruction method is more effective at producing temporally coherence video latents (e.g., the bricks in Figure 6(d)). Also note that our anchor-frame guided enhancement is able to further improve the restoration performance with sharper details (Figure 6(e)).

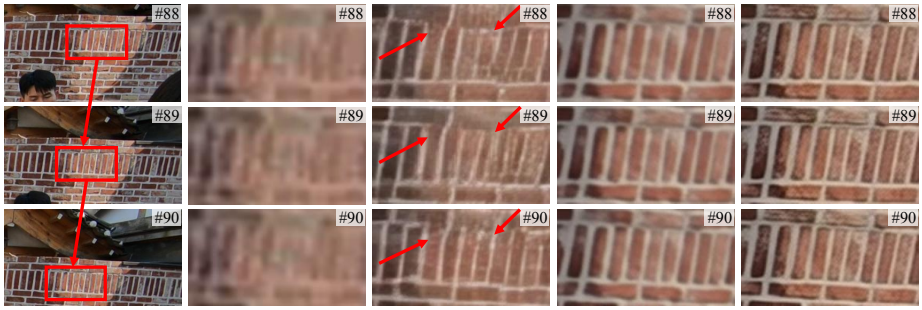

Figure 6. Effectiveness of the motion-aware VAE reconstruction method on a video with camera shaking.

A.2. Effectiveness of Anchor-frame Guided Enhancement

In the main paper, we have compared our method with the LoRA-trained baseline (*i.e.*, *Base*) under the same motion-aware reconstruction setting. The quantitative results in Table 3 show that our approach outperforms *Base* in terms of both DOVER and MUSIQ, demonstrating the effectiveness of the anchor-frame guidance in boosting restoration performance. In this section, we further compare our proposed method against a baseline that replaces the LoRA-trained DiT of *Base* with the fully fine-tuned one ($Base_{w/\,FT}$). Due to GPU resource constraints, the experiments are conducted on the Wan2.1 T2V-1.3B model [8] rather than the Wan 2.1 I2V-14B model [8]. The results in Table 4 show that our method with LoRA yields comparable even better performance compared to the fully fine-tuned baseline, demonstrating the effectiveness of our anchor-frame guided enhancement.

*Corresponding author

Table 4. Effectiveness of the anchor-frame guided enhancement. All methods are trained using the same settings for fair comparison.

Dataset	Method	LIQE [19] ↑	MUSIQ [4] ↑	CLIPIQA+ [9] ↑	MANIQA [15] ↑	FasterVQA [12] ↑	DOVER [13] ↑
REDS30 [5]	$\text{Base}_{\text{w/ FT}}$ (1.3B)	2.283	58.02	0.4477	0.2924	**0.7109**	**40.42**
	Ours (1.3B)	**2.459**	**58.96**	**0.4573**	**0.3271**	0.6956	40.09
RealVSR [16]	$\text{Base}_{\text{w/ FT}}$ (1.3B)	3.8212	69.49	0.5188	0.3658	0.7849	55.13
	Ours (1.3B)	**4.019**	**72.82**	**0.5616**	**0.4142**	**0.7940**	**57.13**

A.3. Visual Comparisons of Temporal Consistency

Similar to [17], we evaluate the temporal consistency property of the restored video. Figure 7 shows the temporal information of the restored results. Compared to other diffusion-based VSR methods [2, 8, 10, 14, 20], our STCDiT generates the video with a better temporal consistency property.

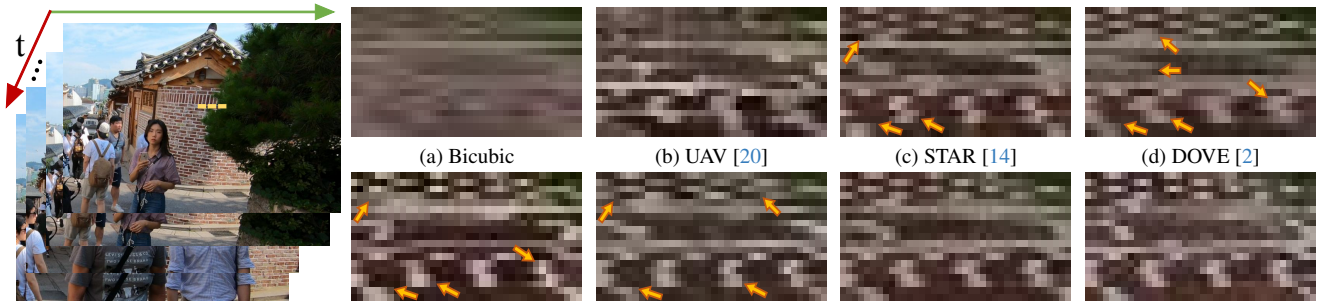

Figure 7. Visual comparisons of the temporal consistency for the restored videos. We visualize the pixels of the selected rows (the dotted yellow line) according to [17].

A.4. Analysis of Anchor-Frame Sampling Setting

For videos with complex motion, segmentation yields numerous base frames. Using all as anchors entails high computational cost, while overly sparse sampling adversely impacts the structural guidance of anchors (see Table 5 on REDS [5]). To balance the efficiency and performance, we uniformly sample 1/4 of these base frames as anchors.

Table 5. Quantitative comparison of different anchor-frame sampling rates.

Sampling rate	1	1/4	1/8	1/16
LPIPS [18] ↓	0.2817	0.2866	0.2928	0.2997
DOVER [13] ↑	42.70	42.94	42.23	41.81
E^*_{warp} [3] ↓	1.21	1.27	1.33	1.36

In addition, to minimize motion misalignment between anchors and other frames in long clips, we increase segmentation density to limit intra-clip motion variation. Table 6 shows that anchor frame guidance improves as the maximum number of frames per clip decreases on UDM10 [7]. We set this maximum to 9 for efficiency.

Table 6. Quantitative comparison of different maximum numbers of frames in each clip.

Frame number	5	9	17	33
LPIPS ↓	0.1645	0.1682	0.1759	0.1863
DOVER ↑	63.97	64.10	63.16	62.57
E^*_{warp} ↓	0.73	0.74	0.77	0.82

A.5. Inference speed and GPU memory of different variants of STCDiT in Table 3

Table 7 reports the computational cost of STCDiT on the first video of RealVSR [16] (1024 × 512, 50 frames). The increased computational cost mainly arises from the enlarged self-attention caused by anchor-token concatenation.

Table 7. GPU memory consumption and inference time. #GPU Mem. denotes the maximum of GPU memory consumption.

Method	Wan [8]	Base	$\text{Base}_{\text{w/ FF}}$	$\text{Base}_{\text{w/ FF\&ACFM}}$	Ours
#Avg.Time per Step [s]	5.92	7.72	10.13	11.43	11.62
#GPU Mem. [M]	36599	37393	38248	38276	38331

A.6. Finetuning VAE with additional module

Table 8 shows that, on REDS dataset, incorporating deformable convolutions or temporal attention at each VAE scale remains insufficient for reconstructing temporally coherent videos, since the temporal scaling operators in VAE struggle to model complex spatial transformations across frames.

Table 8. Quantitative comparison of VAE variants with different operators.

Method	Baseline	w/ deformable convolution	w/ temporal attention layer	Motion-aware VAE
PSNR ↑	27.22	27.34	27.29	31.42
SSIM ↑	0.7802	0.8031	0.7789	0.8924
E^*_{warp} ↓	1.76	1.69	1.77	1.34

A.7. Comparison with DLoRAL

Table 9 shows that our approach outperforms DLoRAL [6] in terms of LPIPS, DOVER and E^*_{warp}.

Table 9. Quantitative comparison with DLoRAL [6].

Datasets	REDS				UDM10				RealVSR			
Metrics	LPIPS ↓	CLIPIQA+ ↑	DOVER ↑	E^*_{warp} ↓	LPIPS ↓	CLIPIQA+ ↑	DOVER ↑	E^*_{warp} ↓	LPIPS ↓	CLIPIQA+ ↑	DOVER ↑	E^*_{warp} ↓
DLoRAL	0.2901	0.4595	32.33	1.71	0.1956	0.5032	54.66	0.81	0.1757	0.5418	53.31	1.57
Ours	0.2866	0.4728	42.94	1.27	0.1682	0.5234	64.10	0.74	0.1553	0.5393	61.57	1.53

A.8. Differences in Network Details between STCDiT and STCDiT-tiny

The difference between STCDiT and STCDiT-tiny lies in how to obtain the video features $\mathbf{F}^V$ (Figure 8). For STCDiT, similar to Wan 2.1 I2V-14B [8], it derives the video feature $\mathbf{F}^V$ by concatenating the LQ video latent $\mathbf{Y}$, the noise latent $\mathbf{N}$, and an all-one mask $\mathbf{M} \in \mathbb{R}^{C\times F'\times H\times W}$ along the channel dimension, followed by a patchify operation (Figure 8 (a)). For STCDiT-tiny, to obtain the video feature $\mathbf{F}^V$, the LQ video latent $\mathbf{Y}$ and the noise latent $\mathbf{N}$ are fed into two separate patchify operations. The patchify operation used for processing $\mathbf{Y}$, denoted as LQ patchify, is initialized with the weights of the original video patchify used in Wan 2.1 T2V-1.3B [8]. Then, the output of the LQ patchify is passed through a zero-initialized linear layer, which helps stabilize the training process. Finally, we obtain the video feature $\mathbf{F}$ by fusing the LQ features and noise features output from the linear layer and the video patchify operation via element-wise addition (Figure 8 (b)).

B. Description about the SportsLQ Dataset

In Section 4 of the main paper, to complement the lack of human activity scenes in existing real-world benchmarks [1, 16], we collect a dataset of 20 LQ videos with various sports, consisting of basketball, soccer, table tennis, and other sports. In addition, each video has a resolution of 720×1280 pixels. We present more examples in Figure 9.

C. Quantitative Comparisons

In this section, we present more visual comparisons with state-of-the-art methods [2, 8, 10, 11, 14, 17] on various benchmarks (*i.e.*, VideoLQ [1], REDS [5], RealVSR [16] and SportsLQ). Figures 10-12 show that our STCDiT and STCDiT-tiny can generate temporally consistent video content with faithful structures (*i.e.*, letters, numbers, windows, buildings, arms, and basketball net).

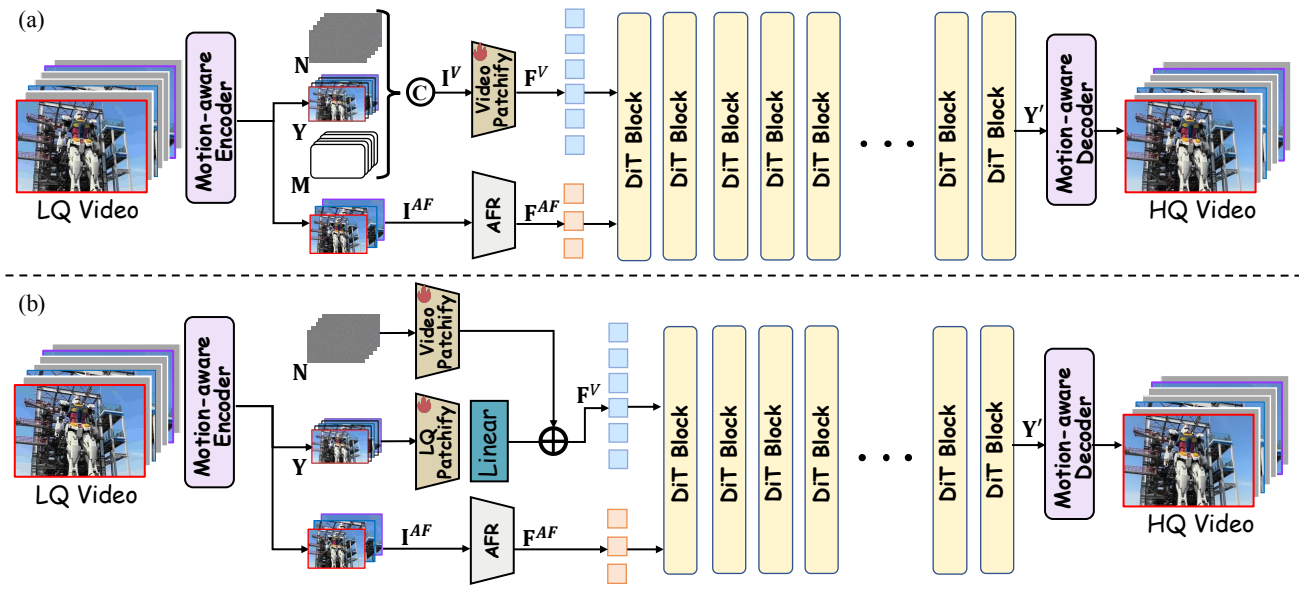

Figure 8. **Differences in network details between STCDiT and STCDiT-tiny in obtaining the video feature $\mathbf{F}^V$.** (a) STCDiT obtains $\mathbf{F}^V$ by feeding the concatenation of the LQ video latent $\mathbf{Y}$, the noise latent $\mathbf{N}$, and a binary mask $\mathbf{M}$ into a patchify operation. (b) STCDiT-tiny obtains $\mathbf{F}^V$ by processing $\mathbf{Y}$ and $\mathbf{N}$ separately and then fusing the resulting features via element-wise addition.

(a) Basketball (b) Table Tennis (c) American Football

(d) Soccer (e) Figure Skating (f) Cycling

(g) Swimming (h) Running (i) Ski Jumping

Figure 9. Examples from the SportsLQ dataset.

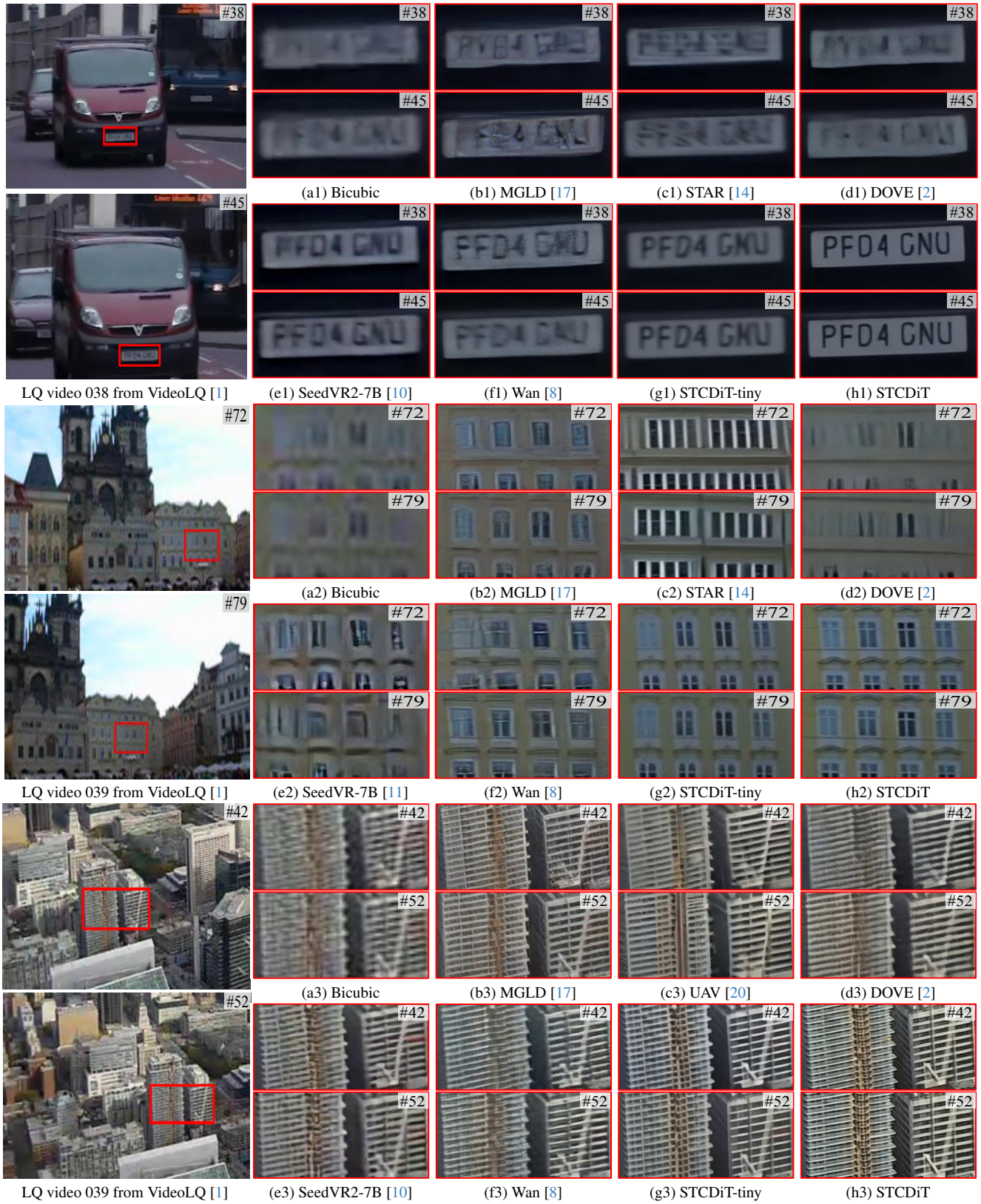

Figure 10. VSR results ($\times 4$) on the real-world benchmark [1]. We provide two types of video motion scenarios. The first row provides a vehicle approaching scenario, and the second and third rows provide surround-view scenarios. Compared to competing methods, our method restores videos that are both temporally consistent and structurally faithful.

Figure 11. The first and second rows show VSR results ($\times 4$) on the real-world benchmark [1] and the synthetic benchmark [5], respectively, while the third row shows VSR results ($\times 1$) on the real-world benchmark [16]. Compared to competing methods, our method restores videos with better temporal consistency and structural fidelity.

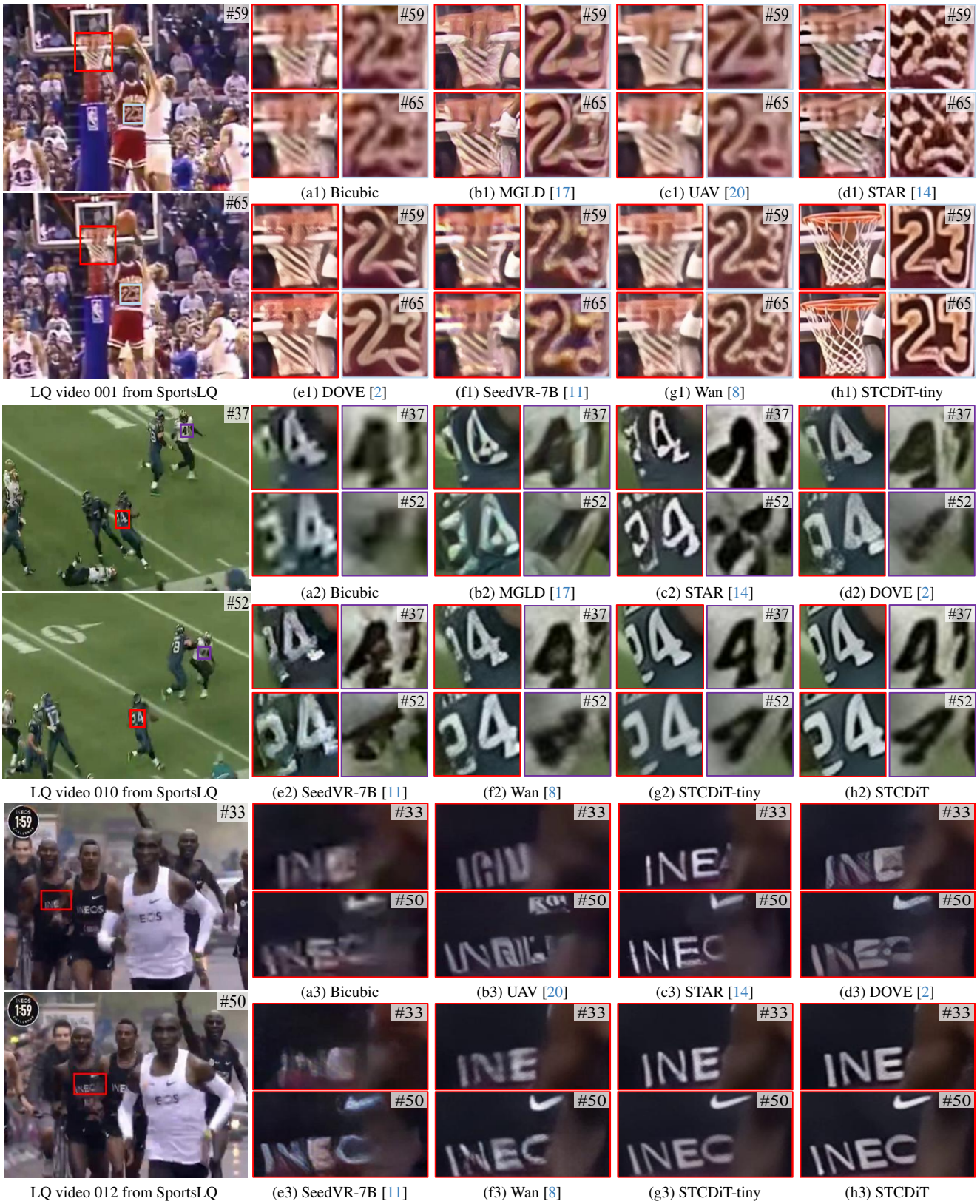

Figure 12. VSR results (×4) on the real-world benchmark (*i.e.*, SportsLQ). Compared to competing methods, our method recovers more faithful structural details.

References

[1] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022. 3, 5, 6

[2] Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. Dove: Efficient one-step diffusion model for real-world video super-resolution. In *NeurIPS*, 2025. 2, 3, 5, 6, 7

[3] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *CVPR*, 2025. 2

[4] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *CVPR*, 2021. 2

[5] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, 2019. 2, 3, 6

[6] Yujing Sun, Lingchen Sun, Shuaizheng Liu, Rongyuan Wu, Zhengqiang Zhang, and Lei Zhang. One-step diffusion for detail-rich and temporally consistent video super-resolution. In *NIPS*, 2025. 3

[7] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 2

[8] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 3, 5, 7

[9] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 2

[10] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv preprint arXiv:2506.05301*, 2025. 2, 3, 5, 6

[11] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *CVPR*, 2025. 3, 5, 6, 7

[12] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE TPAMI*, 2023. 2

[13] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 2

[14] Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *ICCV*, 2025. 2, 3, 5, 6, 7

[15] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 2022. 2

[16] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *ICCV*, 2021. 2, 3, 6

[17] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *ECCV*, 2024. 2, 3, 5, 6, 7

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[19] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, 2023. 2

[20] 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. 2, 5, 6, 7