PS-SR: Pseudo-Single-Step Video Super-Resolution via Speculative Diffusion — CVPR 2026 Supplementary Material*

Aiqiu Wu[1], Zhaofan Qiu[2], Ting Yao[2], and Tao Mei[2]
[1]University of Science and Technology of China [2]HiDream.ai Inc.
wuaiqiu@mail.ustc.edu.cn, {qiuzhaofan, tiyao, tmei}@hidream.ai

The supplementary material contains: 1) details and visualization examples for long video super-resolution using PS-SR; 2) comparison with non-diffusion methods; 3) discussion on performance discrepancies; 4) additional analysis on inference efficiency; 5) choice of hyperparameters for balancing reconstruction and creativity; 6) additional analysis on frequency-domain update rule; 7) more video examples generated by PS-SR.

1. Long Video Super-Resolution

We propose an overlapping clip splitting and merging strategy that breaks the memory limitation of long sequence video, achieving arbitrary-length video super-resolution. Specifically, the input video is first segmented into fixed-length clips with short temporal overlaps, and each clip is independently processed using PS-SR. The processed clips are subsequently concatenated along the temporal dimension to reconstruct long high-quality video.

For overlapping frames, we employ a temporal position-aware averaging mechanism that assigns adaptive weights according to each frame's relative temporal location within the overlap. Let $k \in \{1, ..., L\}$ denote the index of a frame within the overlap, and let $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ denote the preceding clip and the following clip, respectively. The merged frame $\hat{\mathbf{x}}_k$ is computed as:

$$\hat{\mathbf{x}}_k = (1 - w_k)\mathbf{x}_k^{(1)} + w_k\mathbf{x}_k^{(2)}, \tag{1}$$

where $w_k = \frac{k}{L+1}$ gradually increases along the temporal direction, enabling a smooth transition from $\mathbf{x}^{(1)}$ to $\mathbf{x}^{(2)}$ across the overlap. In our implementation, each clip contains 50 frames and the overlap length is set to 3 frames. Visualization examples in Figure 1 show that our strategy produces seamless transitions across clip boundaries.

2. Comparison with Non-Diffusion Methods

We further compare PS-SR with non-diffusion VSR approaches, including RealBasicVSR [1] and RealViformer [6]. As summarized in Table 1, PS-SR delivers the best performance across most metrics on both the synthetic SPMCS dataset [3] and the real-world VideoLQ dataset [1]. These results demonstrate that, compared with non-diffusion methods, PS-SR excels at producing videos with superior perceptual fidelity and finer details, owing to the strong generative prior of video diffusion models.

Table 1. Performance comparisons with non-diffusion methods on synthetic dataset SPMCS and real-world dataset VideoLQ. The best results are **bolded**.

Datasets	Metric	RealBasicVSR [1]	RealViformer [6]	PS-SR (Ours)
SPMCS	PSNR ↑	21.670	22.045	**22.092**
	SSIM ↑	0.5795	0.6013	**0.6287**
	LPIPS ↓	0.5503	0.4456	**0.2940**
	DISTS ↓	0.2940	0.2369	**0.1454**
	CLIP-IQA ↑	**0.4455**	0.3289	0.3686
	MUSIQ ↑	57.287	60.897	**61.004**
	NIQE ↓	6.1658	4.5743	**3.9542**
VideoLQ	CLIP-IQA ↑	0.3070	0.2897	**0.3155**
	MUSIQ ↑	59.348	59.274	**62.091**
	NIQE ↓	4.7659	4.8672	**4.6975**

3. Discussion on Performance Discrepancies

As shown in Table 2, notable discrepancies exist among the results presented in different publications. These variations primarily stem from two factors. First, the degradation pipeline involves stochastic operations (e.g., blur, noise injection, compression, and resizing), which inevitably produce different degraded inputs even under identical settings. Consequently, the low-quality videos used for evaluation may differ across studies. Second, differences in encoding and storage formats may impact video quality that further affect the reported performance.

In our experiments, we adopt the degradation pipeline following previous works [2, 5] to generate low-quality videos in synthetic datasets for evaluation. In addition, we standardize the output format of all generated videos, ensuring fair comparisons across methods.

*This work was performed at HiDream.ai.

Table 2. Performance comparisons of existing VSR methods on SPMCS and UDM10 datasets reported by different publications.

Datasets	Metric	STAR					SeedVR			
		(by us)	(by [7])	(by [8])	(by [2])	(by [4])	(by us)	(by [7])	(by [8])	(by [4])
SPMCS	PSNR↑	21.437	21.240	23.350	20.730	22.58	20.738	21.220	24.310	20.78
	SSIM↑	0.5653	0.5441	0.6747	0.4890	0.609	0.5901	0.5672	0.7257	0.575
	LPIPS ↓	0.4220	0.5257	0.2670	0.6060	0.420	0.3313	0.3488	0.2165	0.395
	DISTS ↓	0.2179	0.2872	0.1209	0.3420	0.229	0.1461	0.1611	0.0970	0.166
UDM10	PSNR↑	23.635	23.470	25.730	24.451	24.66	22.860	23.390	25.680	24.29
	SSIM↑	0.7334	0.6804	0.7907	0.7140	0.747	0.7211	0.6843	0.7753	0.731
	LPIPS ↓	0.3433	0.4242	0.2061	0.4170	0.359	0.2796	0.3583	0.1915	0.264
	DISTS ↓	0.1730	0.2156	0.0978	0.2300	0.195	0.1301	0.1339	0.0845	0.124

4. Additional Analysis on Inference Efficiency

Figure 2 compares the inference time of traditional diffusion and proposed speculative diffusion under different numbers of sampling steps. While the inference time of both methods increases approximately linearly with the number of steps, our speculative diffusion significantly reduces time consumption, yielding about a $3\times$ speedup. This efficiency gain stems from delegating refinement steps to the lightweight draft model instead of repeatedly invoking the full-architecture base model. Consequently, this speculative diffusion mechanism enables PS-SR to perform multi-step sampling while maintaining a computation cost comparable to single-step inference.

5. Balance Reconstruction and Creativity

The reconstruction fidelity and visual creativity of PS-SR are jointly governed by two key hyperparameters: the number of speculative sampling steps T and the refinement strength α. The sampling steps T control the degree of iterative refinement. Increasing T enhances sharpness and fine-grained details, whereas small T preserves stronger input-output consistency. As shown in Figure 3, the output appears smooth and slightly blurred when $T = 1$. As T increases, the results become sharper but introduce unrealistic artifacts. We set $T = 4$ as an optimal point that aligns well with human preference.

The refinement strength α controls the interpolation between the previous estimate and the current refinement prediction. As illustrated in Figure 4, a small α keeps the output close to the previous step's result and thus favors reconstruction fidelity, while a larger α shifts the output toward the refined prediction and injects richer high-frequency details. Based on human evaluation, we empirically choose $\alpha = 0.6$ to balance fidelity and perceptual quality.

6. Frequency-Domain Update Analysis

To better assess the effectiveness of the proposed frequency-domain update rule in mitigating semantic shift, we visualize the results obtained with and without this component. As shown in Figure 5, disabling the frequency-domain update rule (w/o FDU) produces sharper but semantically inconsistent outputs, leading to noticeable deviations from the ground truth—for example, shifts in the car color and leaf margins. In contrast, enabling the frequency-domain update rule (w/ FDU) effectively suppresses semantic drift and preserves input-output consistency.

7. More Video Examples

We provide more video examples in project page (https://waq2001.github.io/PS-SR-page/) to illustrate high-quality videos generated by PS-SR.

References

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

[2] 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 *NeurIPS*, 2025. 1, 2

[3] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 1

[4] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv:2506.05301*, 2025. 2

[5] Zhongdao Wang, Guodongfang Zhao, Jingjing Ren, Bailan Feng, Shifeng Zhang, and Wenbo Li. Turbovsr: Fantastic video upscalers and where to find them. In *ICCV*, 2025. 1

[6] Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *ECCV*, 2024. 1

[7] Ziqing Zhang, Kai Liu, Zheng Chen, Xi Li, Yucong Chen, Bingnan Duan, Linghe Kong, and Yulun Zhang. Infvsr: Breaking length limits of generic video super-resolution. *arXiv:2510.00948*, 2025. 2

[8] Weisong Zhao, Jingkai Zhou, Xiangyu Zhu, Weihua Chen, Xiao-Yu Zhang, Zhen Lei, and Fan Wang. Realisvsr: Detail-enhanced diffusion for real-world 4k video super-resolution. *arXiv:2507.19138*, 2025. 2

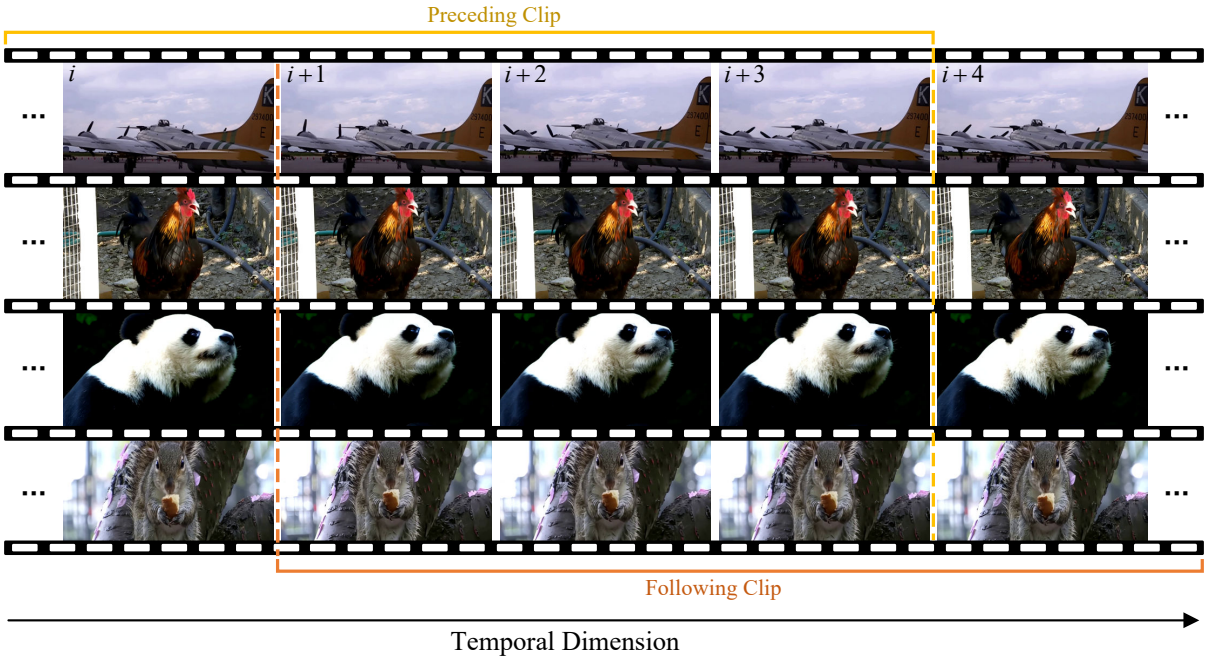

Figure 1. Visualization examples of long video super-resolution results produced by PS-SR.

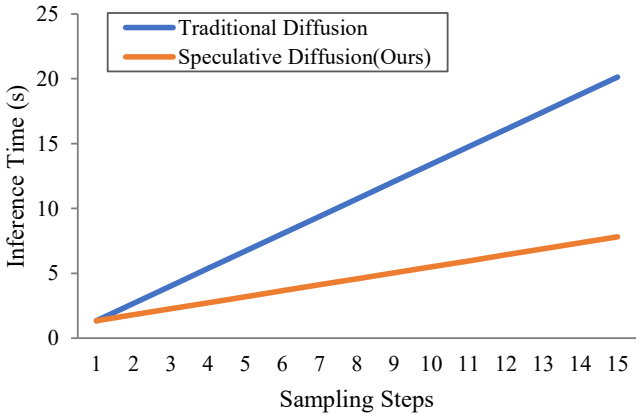

Figure 2. Inference time comparison between traditional diffusion and our speculative diffusion across different sampling steps, evaluated on 29-frame 720 $\times$ 1280 videos using an NVIDIA A800 GPU.

Figure 3. A visualization example of PS-SR variants by using different speculative steps T.

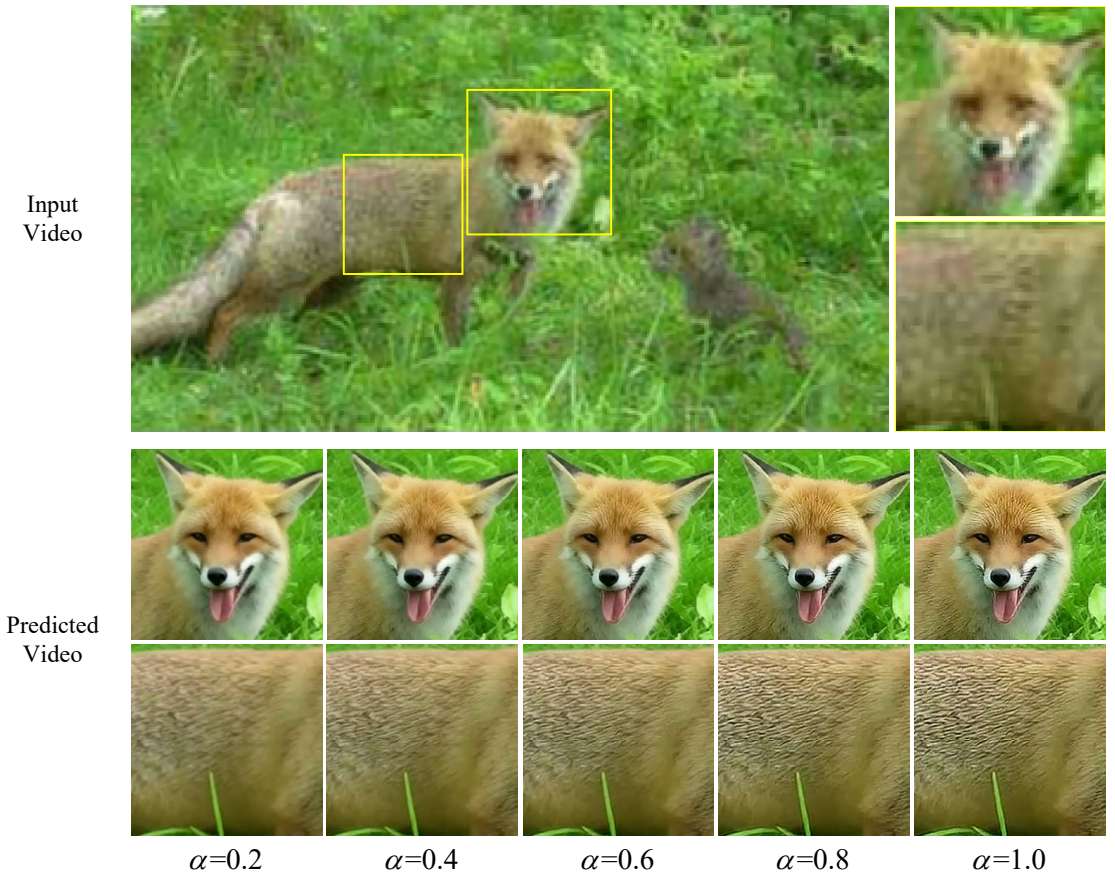

Figure 4. A visualization example of PS-SR variants by using different refinement strength α.

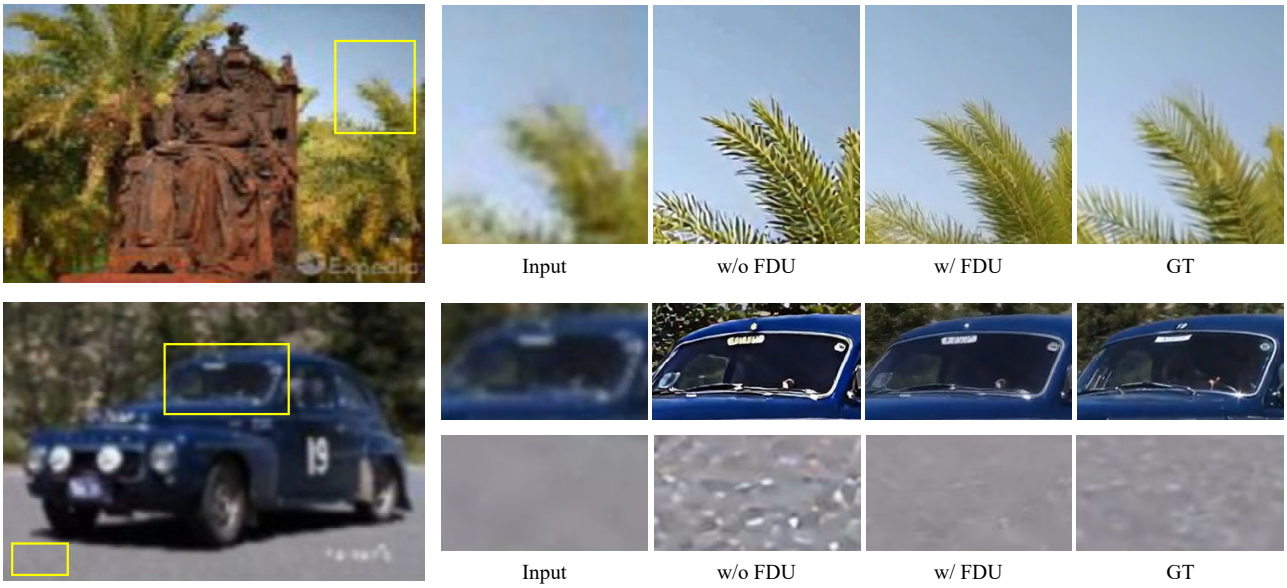

Figure 5. Visualization examples showing the effectiveness of frequency-domain update rule in PS-SR.