

# ZERO-SHOT VIDEO RESTORATION AND ENHANCEMENT WITH ASSISTANCE OF VIDEO DIFFUSION MODELS

## – SUPPLEMENTARY MATERIAL –

**Anonymous authors**

Paper under double-blind review

This supplementary file provides details that were not presented in the main paper due to page limitations. In the following, we first provide the details of the experiment settings. Then we present more comparison results. Hereafter, we provide an ablation study on the fusion ratio strategy. Finally, a demo for comparing video results is given.

### 1 EXPERIMENT SETTINGS

For the video restoration tasks, we add zero-shot video deblurring for comparison. Following Cao et al. (2025), we collected 10 ground truth (GT) videos from the dataset REDS Nah et al. (2019). For the degradation of video deblurring, we follow Kwon & Ye (2024a;b) and use the temporal uniform blur kernel. For all diffusion models, the prompts are null texts. The  $M$  in COT-Based Fusion Ratio Strategy is set to 4. To accelerate the inference time, we apply the strategy every 10 timesteps and then apply the obtained  $\lambda^{F1}$ ,  $\lambda^{F2}$  and  $\lambda^F$  to the following 9 timesteps. Our method can be combined with the aggregation sampling in Wang et al. (2024) to test on higher-resolution videos. For long video, we separate it into different video clips, neighbouring video clips share one overlapping frame. When our Temporal-Strengthening Post-Processing finishes the process of the previous video clip, the last/shared frame is used as the image condition in the next video clip to maintain the long-range temporal consistency. All experiments were conducted on a 96G H20 GPU.

### 2 COMPARISON WITH STATE-OF-THE-ART METHODS

For zero-shot video deblurring, we compare our method with supervised (sup.) method VRT Liang et al. (2024) and zero-shot video restoration method SVI Kwon & Ye (2024a) and VISION-XL Kwon & Ye (2024b). Since SVI only releases the code for zero-shot video deblurring, we do not compare it on other restoration tasks. Table 1 lists the quantitative results on the evaluation data for zero-shot video deblurring. It can be observed that our method outperforms the compared SOTA methods on all six metrics.

We also conducted a user study to further evaluate the visual quality and temporal consistency. To facilitate the comparison, we only compared the zero-shot video super-resolution results of PSLD, VISION-XL, and PSLD+ZVRV on eight groups of videos. Each user is asked to evaluate the visual quality and temporal consistency of the video with a score ranged from 1 to 3, where 3 indicates good quality and 1 indicates bad quality. The average score for PSLD, VISION-XL, and PSLD+ZVRV is 2.07, 1.05 and 2.91, respectively, which demonstrates that our method has much better visual quality and temporal consistency. We also provide a demo video in the supplementary materials.

The inference time of our method is contingent upon that of the employed T2V and I2V models. In the future, we will explore acceleration techniques for T2V and I2V models to expedite our method.

### 3 ABLATION STUDY

In this section, we conduct an ablation study on the COT-based fusion ratio strategy (with different hyperparameters  $M$  and  $r$ ) and the fixed fusion ratio strategy. For the fixed fusion ratio strategy, we set  $\lambda^{F1}$ ,  $\lambda^{F2}$ , and  $\lambda^F$  to 0.1, 0.01, and 0.5, respectively, for all timesteps. For hyperparameter  $r$ , the

Table 1: Quantitative comparison with state-of-the-art methods for zero-shot video deblurring. The best results are highlighted in bold and the second best results are underlined. The WE and t-LPIPS values have been multiplied by 100.

Methods	Backbone	PSNR $\uparrow$	SSIM $\uparrow$	CLIP-IQA $\uparrow$	LPIPS $\downarrow$	WE $\downarrow$	FVD $\downarrow$	DOVER $\uparrow$	t-LPIPS $\downarrow$	VMAF $\uparrow$
VRT(sup.)	-	18.98	0.4505	0.2093	0.7155	0.1816	1694.8	7.581	0.86	81.94
PSLD	SDXL	19.69	0.4996	0.1905	0.4546	1.5298	662.0	5.573	3.12	76.02
SVI	-	18.25	0.4839	0.2010	0.5871	1.3463	1218.3	7.202	2.35	79.47
VISION-XL	SDXL	<u>19.82</u>	<u>0.5068</u>	<u>0.2147</u>	<u>0.4232</u>	1.2623	<u>457.7</u>	7.298	1.84	80.66
PSLD+ZVRV	SDXL	<b>20.47</b>	<b>0.5342</b>	<b>0.4813</b>	<b>0.3675</b>	<b>0.1632</b>	<b>326.4</b>	<b>8.456</b>	<b>0.67</b>	<b>84.53</b>

Table 2: Ablation study on the COT-based fusion ratio strategy (with different hyperparameters  $M$  and  $r$ ) and the fixed fusion ratio strategy. The best results are highlighted in bold.

Fusion Ratio Strategy		PSNR $\uparrow$	SSIM $\uparrow$	CLIP-IQA $\uparrow$	LPIPS $\downarrow$	WE $\downarrow$	FVD $\downarrow$	DOVER $\uparrow$	t-LPIPS $\downarrow$	VMAF $\uparrow$
COT-Based Fusion Ratio Strategy	$M=2, r=0.45$	26.74	0.6953	0.8472	0.1649	0.4186	258.5	7.251	0.92	83.42
	$M=3, r=0.45$	27.15	0.7196	0.8599	0.1573	0.3952	240.8	8.854	0.63	85.01
	$M=4, r=0.45$	<b>27.42</b>	<b>0.7388</b>	<b>0.8691</b>	<b>0.1395</b>	<b>0.3755</b>	<b>231.7</b>	<b>9.076</b>	<b>0.41</b>	<b>86.37</b>
	$M=4, r=0.50$	27.20	0.7316	0.8610	0.1432	0.3764	239.5	8.983	0.51	86.04
	$M=4, r=0.40$	27.09	0.7272	0.8623	0.1408	0.3783	234.1	8.779	0.50	85.79
Fixed Fusion Ratio Strategy		26.59	0.6915	0.8347	0.1724	0.4298	265.2	6.451	0.98	82.26

value is halved after each search. Taking  $4\times$  blind video super-resolution as an example, Table 2 lists the quantitative comparison results. It can be observed that the COT-based fusion ratio strategy is more sensitive to the hyperparameter  $M$ . As  $M$  increases, the performance improves as well. Considering the trade-off between speed and performance, we set  $M = 4$ , and we set  $r = 0.45$  since it achieves the best performance when  $M = 4$ . Besides, we apply the COT-based fusion ratio every 10 timesteps and apply the obtained  $\lambda^{F1}$ ,  $\lambda^{F2}$ , and  $\lambda^F$  to the following 9 timesteps. Although applying the COT-based fusion ratio more frequently can result in better performance, we choose to apply it every 10 timesteps considering the trade-off between speed and performance. When  $M$  ranges from 2 to 4, and  $r$  ranges from 0.4 to 0.5, the COT-based fusion ratio strategy consistently outperforms the fixed fusion ratio strategy.

## REFERENCES

- Cong Cao, Huanjing Yue, Xin Liu, and Jingyu Yang. Zero-shot video restoration and enhancement using pre-trained image diffusion model. *AAAI*, 2025.
- Taesung Kwon and Jong Chul Ye. Solving video inverse problems using image diffusion models. *ICLR*, 2024a.
- Taesung Kwon and Jong Chul Ye. Vision-xl: High definition video inverse problem solver using latent image diffusion models. *arXiv preprint arXiv:2412.00156*, 2024b.
- Jingyun Liang, Jiezhong Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.
- Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.