

Supplementary Material of “STAR: Spatial-Temporal Augmentation with Text-to-Video Models for Real-World Video Super-Resolution”

Anonymous ICCV submission

Paper ID 11854

A. Perception-Distortion Trade-Off

The trade-off between perception and distortion [1] is a widely recognized challenge in the super-resolution domain. Thanks to our *DF Loss*, our method can easily control the model to favor either fidelity or perceptual quality in the generated results. We can adjust the hyper-parameter β in the $b(t)$ to achieve this goal. The total loss in our STAR is:

$$\mathcal{L}_{total} = \mathcal{L}_v + b(t)\mathcal{L}_{DF}, \quad (1)$$

The $b(t)$ can be written as follows:

$$b(t) = \beta \cdot \left(1 - \frac{t}{t_{max}}\right), \quad (2)$$

Where t is the timestep and β is the hyper-parameter that adjusts the weight between \mathcal{L}_v and \mathcal{L}_{DF} , which we set to 1 by default. From equations (1) and (2), we can observe that a larger β increases the weight of the DF loss at each timestep, thereby further enhancing the fidelity of the results. In contrast, a smaller β reduces the influence of the DF loss at each timestep, allowing the v-prediction loss to have a greater impact and produce more perceptual results. The $b(t)$ - t curves under different β are shown in Figure 1.

We conduct experiments under these settings to demonstrate the ability to achieve the perception-distortion trade-off. The quantitative results are shown in Table 1. From Table 1, we can observe that increasing β improves the PSNR and E_{warp}^* , leading to better fidelity. Conversely, decreasing β reduces the LPIPS score, indicating better perceptual quality.

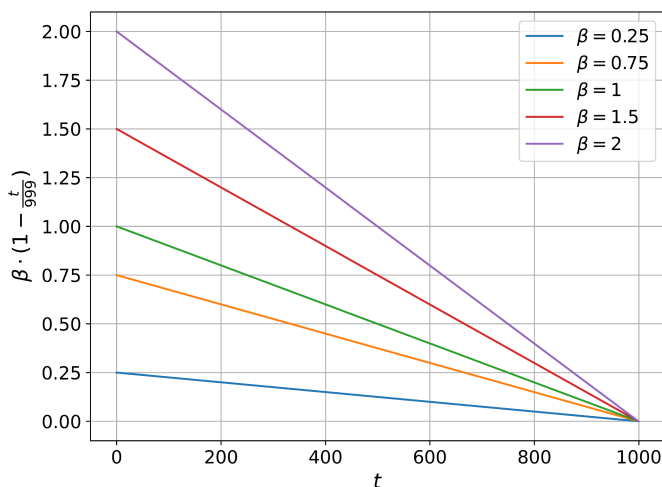


Figure 1. Ablation on $b(t)$. Higher hyper-parameter β produces results with greater fidelity, while lower β emphasizes more perceptual quality.

Table 1. Qualitative comparison under different β of $b(t)$.

β	PSNR \uparrow	LPIPS \downarrow	$E_{warp}^* \downarrow$
0.25	23.55	0.1825	2.88
0.75	23.76	0.1842	2.74
1.0	23.91	0.1885	2.68
1.5	24.08	0.2272	2.53
2.0	24.41	0.3339	2.21

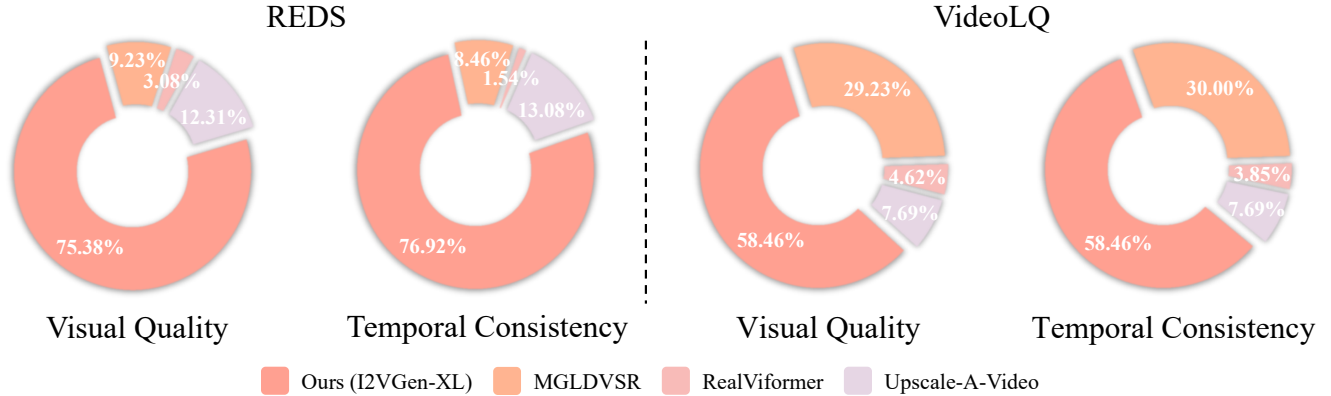


Figure 2. User study results. Our STAR is preferred by human evaluators for both visual quality and temporal consistency.



Figure 3. Qualitative comparisons on synthetic datasets. Our STAR generates more detailed and realistic results. (Zoom-in for best view)

B. More Results

B.1. User Study

To find the human-preferred results between our STAR and other state-of-the-art methods, we conduct a user study that evaluate the results on both real-world and synthetic datasets. Specifically, we use the real-world dataset VideoLQ [2] and the synthetic dataset REDS30 [3]. We select two image-diffusion-model-based methods, Upscale-A-Video [6] and MGLD-VSR [4]; and one GAN-based method, RealViformer [5] for comparison. We invite 12 evaluators to participate in the user study. For each evaluator, we randomly select 10 videos from each dataset and present four results: one from our STAR and three from the compared methods. The evaluators were asked to choose which result had the best visual quality and temporal consistency. The results of the user study are depicted in Figure 2, indicating that our STAR is preferred by most human evaluators for both visual quality and temporal consistency.



Figure 4. Qualitative comparisons on real-world datasets. Our STAR produces the clearest facial details and the most accurate text structure. (Zoom-in for best view)

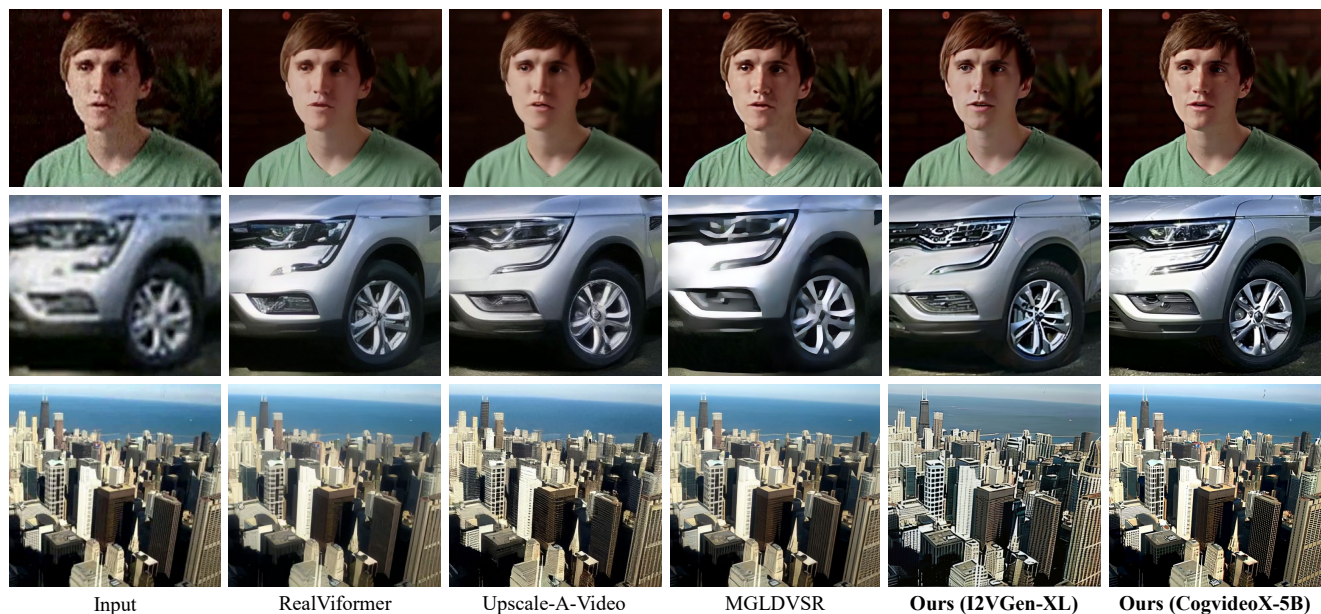


Figure 5. Qualitative comparisons on synthetic and real-world datasets with larger T2V models. Scaling up the T2V model enhances detail and realism in video super-resolution results. (Zoom-in for best view)

B.2. Qualitative Comparisons

We provide more visual comparisons on synthetic and real-world datasets in Figure 3 and Figure 4 to further highlight our advantages in spatial quality. These results clearly demonstrate that our method preserves richer details and achieves greater

029

030

031

realism. To demonstrate the impact of scaling up with larger text-to-video (T2V) models, we present additional results in Figure 5. It is evident that scaling up the T2V model further improves the restoration effect, indicating that a large and robust T2V model can serve as a strong base model for video super-resolution.

B.3. Video Demo

We provide a demo video [STAR-demo.mp4] in the supplementary material, showcasing the temporal and spatial advantages of our proposed STAR more intuitively. This video includes additional results and comparisons on synthetic, real-world, and AIGC videos.

References

- [1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 1
- [2] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, pages 5962–5971, 2022. 2
- [3] 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 *CVPRW*, pages 0–0, 2019. 2
- [4] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. 2024. 2
- [5] Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. *ECCV*, 2024. 2
- [6] 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*, pages 2535–2545, 2024. 2