Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution

— Supplementary Materials —

Shangchen Zhou* Peiqing Yang* Jianyi Wang Yihang Luo Chen Change Loy
S-Lab, Nanyang Technological University

https://shangchenzhou.com/projects/upscale-a-video

Contents

1. Architecture

1.1. Hyperparameters of Network

Following the hyperparameter table style from Latent Diffusion [8], Table 1 provides an overview of the hyperparameters of the pretrained SD $\times 4$ Upscaler [1] and our inserted temporal layers. Our codes and models will be publicly released.

Table 1. Hyperparameters for the pretrained SD $\times 4$ Upscaler [1] (including U-Net and VAE) and our inserted temporal layers. We train our model on the video patches of size 320×320 with 8 frames.

Hyperparameter	*U-Net*
Training patch shape	$8 \times 320 \times 320 \times 3$
f	4
z-shape	$8 \times 80 \times 80 \times 4$
Channels	256
Depth	2
Channel multiplier	1, 2, 2, 4
Attention resolutions	40, 20, 10
Head number	8
Embedding dimension	1024
CA resolutions	40, 20, 10
CA sequence length	77

Hyperparameter	*VAE*
f	4
Channels	128
Channel multiplier	1, 2, 4

Hyperparameter	*Temporal Layers*
Temporal Attention resolutions	40, 20, 10
Head number	8
Positional encoding	RoPE [9]
3D CNN kernel size	3, 1, 1

2. Dataset

2.1. YouHQ Dataset

In order to enhance the training of our VSR model using higher-quality videos, we collect a large-scale high-definition (1080×1920) dataset from YouTube, consisting of around 37,000 video clips. The YouHQ dataset encompasses a diverse category of scenarios, including street view, landscape, animal, human face, static object, underwater, and nighttime scene. Fig. 1 illustrates the distribution of this dataset.

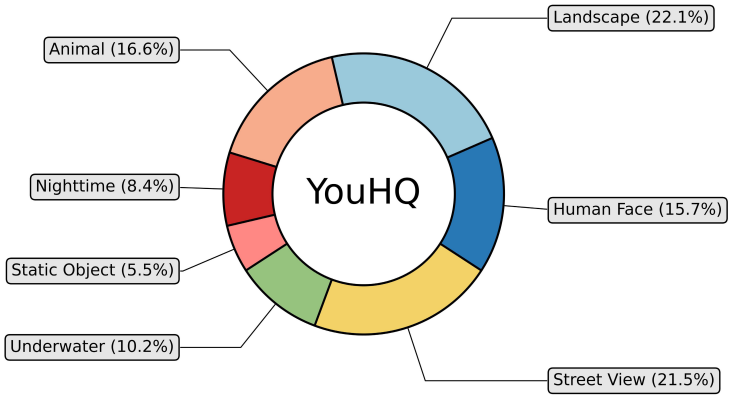


Figure 1. YouHQ Dataset Distribution. It consists of around 37,000 video clips with a diverse category of scenarios, including street view, landscape, animal, human face, static object, underwater, and nighttime scene.

3. More Details on Training and Inference

3.1. Training Strategy for Watermark Removal

We divide the training of the U-Net model into two phases. In the first stage, we train the U-Net using both the WebVid10M [2] and our introduced YouHQ datasets for 70k iterations. In the second stage, to eliminate the impact of watermarks in the WebVid10M data on the results, we conduct an additional 10k iterations of training using only the YouHQ dataset. Fig. 2 showcases the comparisons before and after the second training phase for watermark removal.

Figure 2. Comparison before and after watermark removal (WR) training. In the second stage, watermark removal training is performed only on the YouHQ dataset, effectively removing the watermark introduced by the first-stage training (indicated by the yellow boxes).

3.2. Inference at Arbitrary Resolution and Length

Our model can perform inference on videos of arbitrary scales and lengths. This is achieved by training our model in a patch-wise manner and using the input video as a strong condition. As a result, our model effectively retains its inherent convolutional characteristics. Therefore, it does not impose strict input resolution requirements. Considering memory constraints, we crop the input video into multiple overlapping patches, process them separately, and finally combine the enhanced patches together. Regarding the temporal dimension, at each diffusion step, we cut the video into clips with overlapping frames for inference. The latent features from these overlapping frames are averaged and then passed to the next diffusion step.

3.3. Color Correction

As noted in previous studies [4, 10], diffusion models are prone to experiencing color shift artifacts. To address this issue, we finetune the VAE-Decoder using the input as a condition, which can help maintain consistency in low-frequency information, such as color. Additionally, we have observed that incorporating a training-free *wavelet color correction* module [10] can further enhance color consistency in the results. As shown in Table 2, when applying wavelet color correction, our method yields slightly higher fidelity results, as indicated by improved PSNR, SSIM, and IPIPS scores.

4. More Results

4.1. User Study

For further comprehensive comparisons, we carried out a user study that evaluated the results of both real-world and AIGC videos. We included four different methods in this study, consisting of two diffusion-based image super-resolution methods, *i.e.*, StableSR [10] and SD $\times$4 Upscaler [1], along with a CNN-based video super-resolution method, *i.e.*, RealBasicVSR [3]. We invite a total of 20 participants for this user study. Each volunteer was presented with a set of 10 randomly selected video triplets, which included an input video, the result obtained from one of the compared methods, and our result. Their task was to choose the visually superior enhanced video from the given options. The user study findings, depicted in Fig. 3, reveal a clear preference among the volunteers for our results over those produced by other methods.

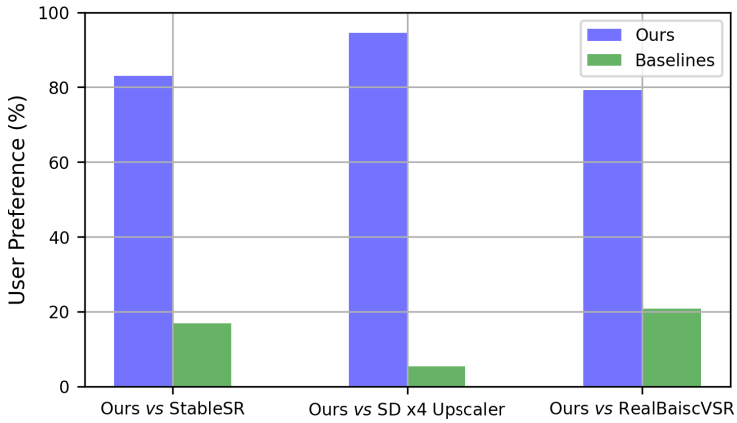


Figure 3. User study results. Our Upscale-A-Video is preferred by human voters over other methods.

4.2. Ablation on Different Pretrained Priors

Recent studies [6, 10] have shown that the large text-to-image Stable Diffusion (SD) [8] is highly effective as a generative prior for blind image restoration tasks. In contrast to these works, we choose to employ a pretrained text-guided image upscaling model, *i.e.*, SD $\times$4 Upscaler [1], as our prior for the video super-resolution (VSR) task. We have also employed Stable Diffusion (SD) as the prior for retraining the network and compared the results of these two different priors for the VSR task. As indicated in Table 2, our model based on the SD $\times$4 Upscaler demonstrates clear advantages in terms of restoration fidelity (PSNR, SSIM, and LPIPS) and temporal consistency (E^*_{warp}). It is important to note that the variant network based on SD exhibits a more noticeable color shift issue after training, which necessitates the use of the 'wavelet color correction' module [10] for correction. However, even with this correction, our model outperforms the variant using SD as the prior. It is worth mentioning that when applying wavelet color correction, our model also achieves higher fidelity results in terms of PSNR, SSIM, and LPIPS. Additionally, Fig. 4 provides visual comparison results for better illustration.

Real-world Video | SD Prior | SD Prior (+ color fix) | Ours

Figure 4. Visual comparison on variant networks with different pretrained priors (*i.e.*, Stable Diffusion (SD) [8] and SD $\times$4 Upscaler [1]). The variant network based on SD often suffers from color shift, requiring additional color correction, and may also lead to unexpected artifacts, such as the child's face in the second example. Furthermore, Our model shows superior generative capabilities compared to the SD-based baseline, *e.g.*, in the first example, our model successfully restores the wall, whereas the SD-based model fails to do so.

Table 2. Ablation study of different pretrained priors, *i.e.*, Stable Diffusion [8] and SD $\times$4 Upscaler [1], on YouHQ40 test set. Our Upscale-A-Video based on the SD $\times$4 Upscaler showcases clear advantages in terms of restoration fidelity (PSNR, SSIM, and LPIPS) as well as temporal consistency (E^*_{warp}).

Metrics	Stable Diffusion	Stable Diffusion (+ color fix)	**SD $\times$4 Upscaler**	**SD $\times$4 Upscaler (+ color fix)**
PSNR ↑	19.03	23.81	25.83	26.07
SSIM ↑	0.590	0.632	0.733	0.737
LPIPS ↓	0.383	0.343	0.268	0.267
E^*_{warp} ↓	1.821	1.707	0.737	0.738

4.3. Ablation on Positions of Recurrent Latent Propagation Module

As discussed in Sec. 3.3 of the main manuscript, it is not necessary to employ the recurrent latent propagation module during every diffusion step in the inference process. Instead, we have the flexibility to choose specific steps for latent propagation and aggregation. Here, we showcase the performance variations when placing this module at different positions, evaluating on the YouHQ40 test set. The results presented in Table 3 indicate that when propagation happens later in the diffusion denoising steps during inference, the warping loss tends to decrease, suggesting better temporal consistency. However, the restoration fidelity also decreases. To balance these factors, we by default choose the middle position for this propagation module. Additionally, Fig. 5 provides the visual comparisons of the temporal profile, illustrating that as propagation occurs later, the videos exhibit improved temporal coherence.

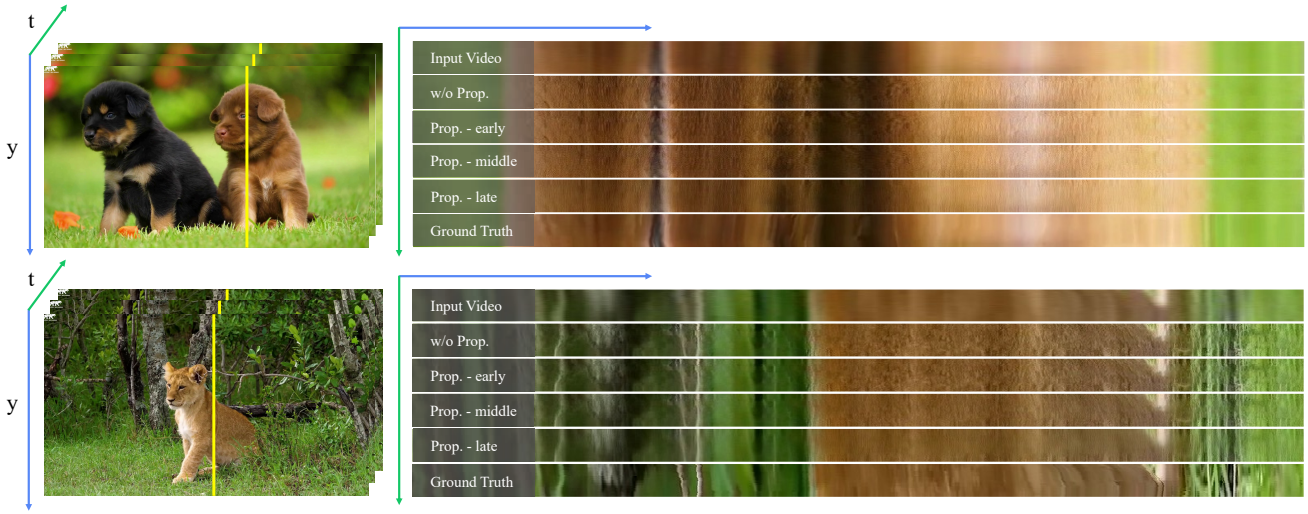


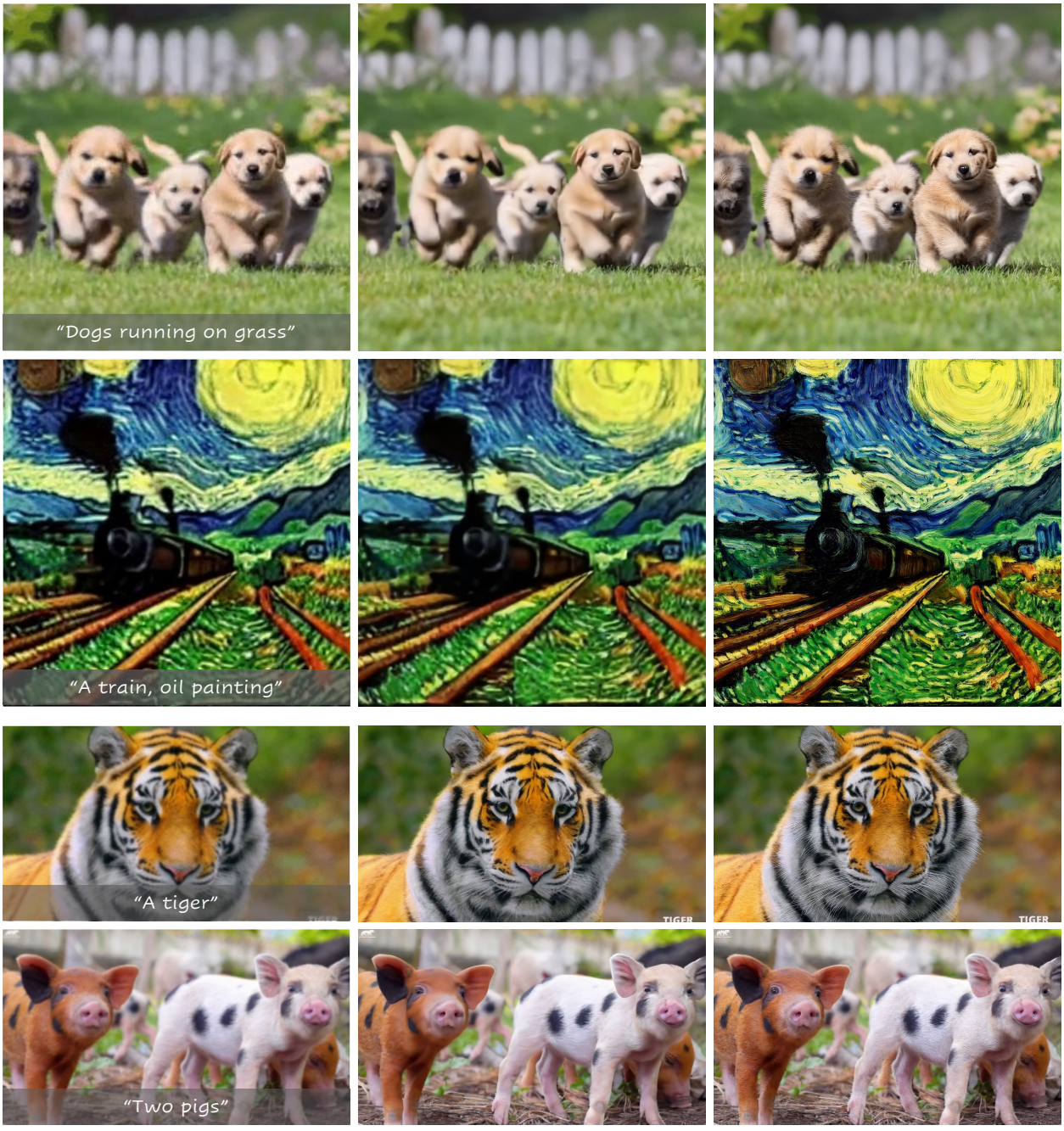
Figure 5. Visual comparison on temporal profile with different positions of recurrent latent propagation module.

Table 3. Ablation study of different positions of recurrent latent propagation module on the YouHQ40 test set.

Metrics	w/o prop. -	early prop. {4, 5, 6, 7}	middle prop. {14, 15, 16, 17}	late prop. {24, 25, 26, 27}
PSNR ↑	23.82	24.18	24.53	24.10
SSIM ↑	0.639	0.646	0.671	0.670
E^*_{warp} ↓	2.398	1.931	0.638	0.618

4.4. Effectiveness of Text Prompt

Upscale-A-Video is trained using video data that includes labeled prompts or no prompts, allowing it to work effectively in both situations. However, when employing the classifier-free guidance approach [5], utilizing proper text prompts as guidance can noticeably enhance the visual quality. As illustrated in Fig. 6, the use of appropriate text prompts leads to significantly improved results with finer and more faithful details compared to using empty prompts.



Input w/o Text Prompt w/ Text Prompt

Figure 6. Visual comparison of using proper text prompts and empty prompts. When employing the classifier-free guidance [5], using proper text prompts as guidance can significantly improve the visual quality and realism, resulting in finer details. This improvement is observed in both real-world scene videos (the last two rows) and AIGC videos (the first two rows).

4.5. More Qualitative Comparisons

In this section, we provide additional visual comparisons of our method with the state-of-the-art methods, including Real-ESRGAN [11], SD ×4 Upscaler [1], ResShift [12], StableSR [10], DBVSR [7], and RealBasicVSR [3]. Fig. 7, Fig. 8, and Fig. 9 present the visual results on synthetic, real-world, and AIGC videos, respectively.

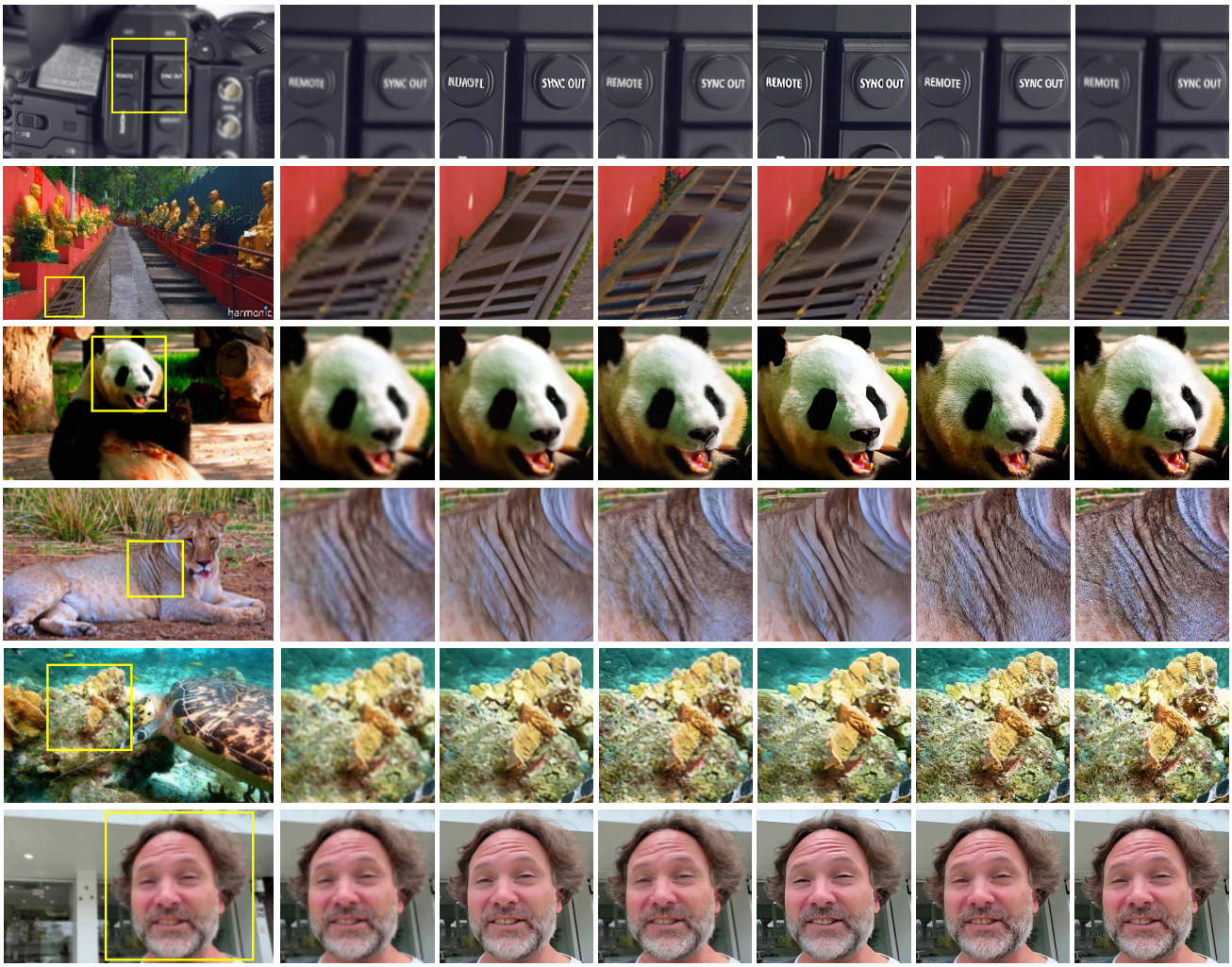


Figure 7. Qualitative comparisons on synthetic datasets. Our Upscale-A-Video exhibits promising enhanced results with more details and heightened realism. (**Zoom in for best view.**)

4.6. Video Demo

We also offer a demo video [Upscale-A-Video-demo.mp4] to showcase more video results and comparisons, which are evaluated on synthetic, real-world, and AIGC videos.

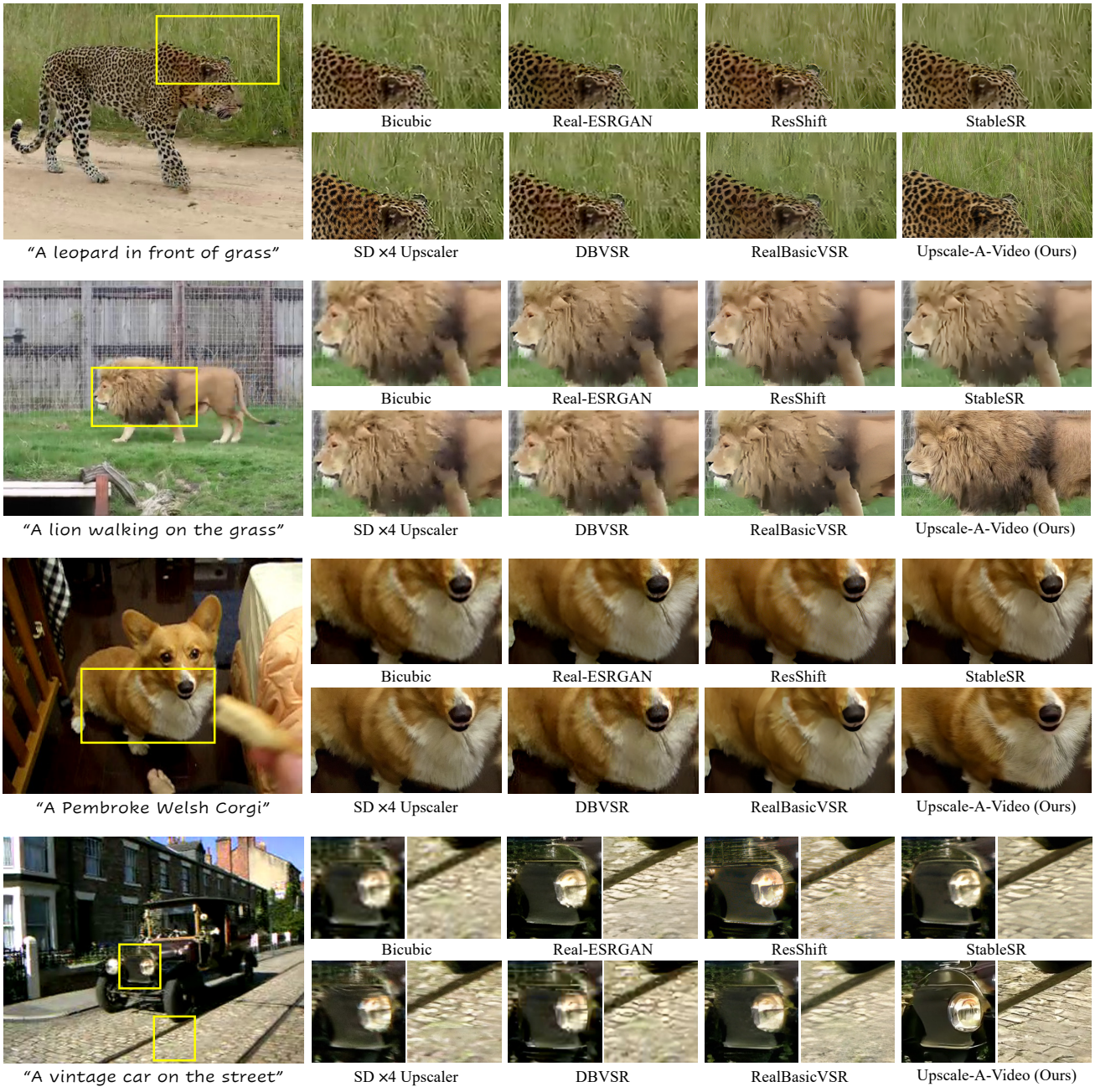


Figure 8. Qualitative comparisons on real-world videos. Our Upscale-A-Video produces promising improvements, delivering increased detail and heightened realism. (**Zoom in for best view.**)

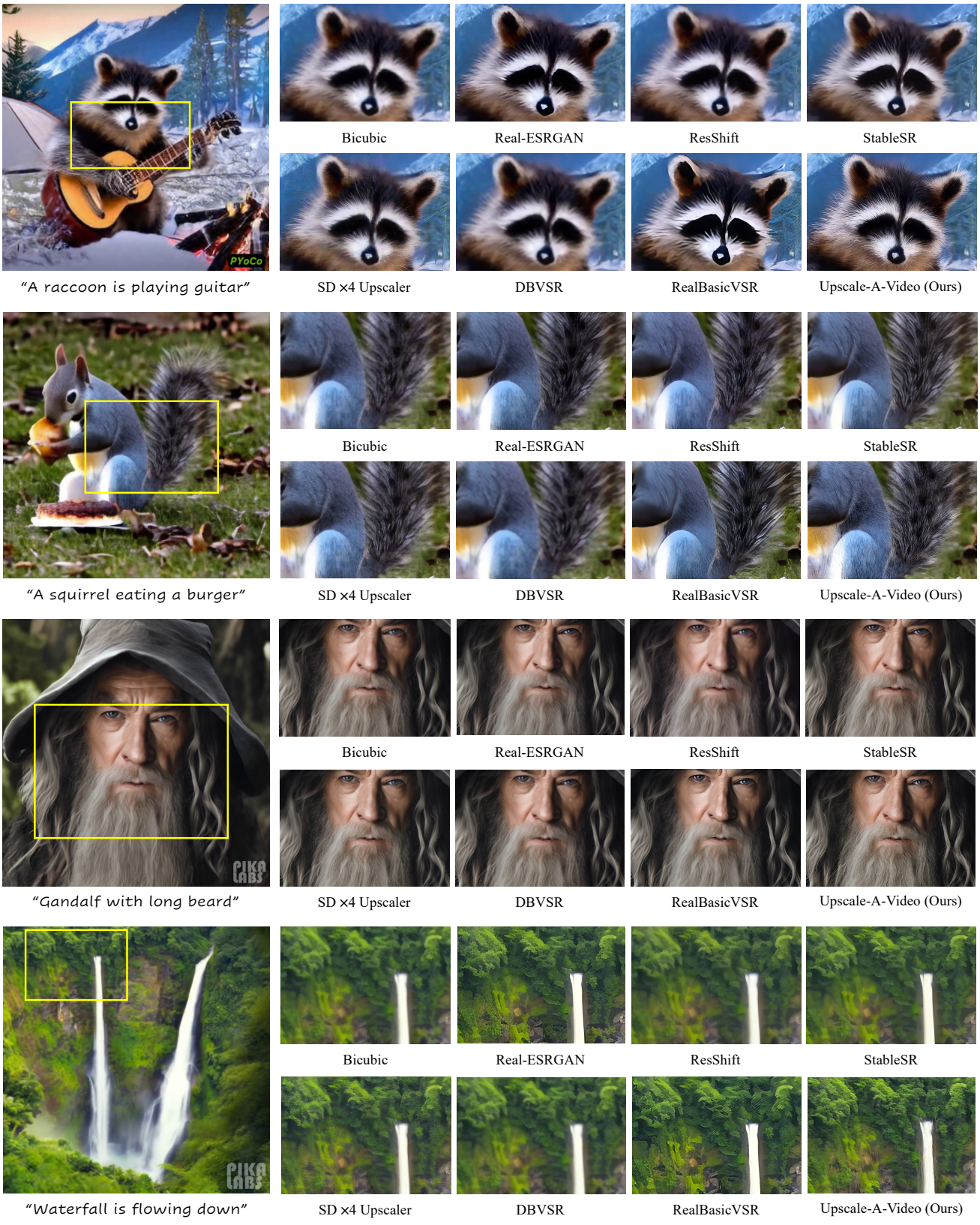


Figure 9. Qualitative comparisons on AIGC videos. When guided by input text prompts, our Upscale-A-Video exhibits promising video results with more details and enhanced realism. (**Zoom in for best view.**)

References

[1] Stable Diffusion x4 Upscaler. https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler, 2023. 2, 3, 4, 7

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022. 3, 7

[4] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, 2022. 3

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPSW*, 2022. 5, 6

[6] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 4

[7] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *ICCV*, 2021. 7

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[10] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3, 4, 7

[11] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 7

[12] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2023. 7