
DOVE: Efficient One-Step Diffusion Model for Real-World Video Super-Resolution

Supplementary Material

Zheng Chen^{1*}, Zichen Zou^{2*}, Kewei Zhang¹, Xiongfei Su³,
Xin Yuan⁴, Yong Guo⁵, Yulun Zhang^{1†}

¹School of Computer Science, Shanghai Jiao Tong University,
²Zhiyuan College, Shanghai Jiao Tong University, ³China Mobile Research Institute,
⁴Westlake University, ⁵Huawei Consumer Business Group

Overview

In the supplementary material, we provide additional analysis and results, including:

- Sec. A: A variant of our DOVE, which is trained with a smaller model, CogVideoX-2B [18].
- Sec. B: Evaluation on more NR-IQA metrics.
- Sec. C: Comparison with recent one-step (video/image) super-resolution methods.
- Sec. D: A more comprehensive comparison of model complexity.
- Sec. E: Additional ablation study on the setting of the inference step.
- Sec. F: Pipeline details, including specific configurations and statistics of the HQ-VSR.
- Sec. G: More quantitative results, including temporal consistency and visual comparison.
- Sec. H: Explanations for the checklist, covering limitations and broader impacts.

A Model Variant: DOVE-2B

Method	UDM10					RealVSR				
	PSNR \uparrow	LPIPS \downarrow	CLIP-IQA \uparrow	DOVER \uparrow	E_{warp}^* \downarrow	PSNR \uparrow	LPIPS \downarrow	CLIP-IQA \uparrow	DOVER \uparrow	E_{warp}^* \downarrow
Upscale-A-Video [22]	21.72	0.4116	0.4697	0.7291	3.97	20.29	0.2671	0.4855	0.7114	6.25
MGLD-VSR [16]	24.23	0.3272	0.4557	0.7264	3.59	22.02	0.2182	0.4510	0.7508	3.16
VEncoder [2]	21.32	0.4344	0.2852	0.4576	1.03	15.75	0.3784	0.3880	0.7637	5.15
STAR [14]	23.47	0.4242	0.2417	0.4830	2.08	17.43	0.2943	0.3641	0.7051	9.88
DOVE-2B (ours)	26.27	0.2786	0.4810	0.7782	1.62	22.11	0.1998	0.5340	0.7822	4.15
DOVE (ours)	26.48	0.2696	0.5107	0.7809	1.77	22.32	0.1851	0.5207	0.7867	3.52

Table 1: Quantitative comparison on both synthetic (*i.e.*, UDM10 [6]) and real-world (*i.e.*, RealVSR [17]) datasets. DOVE-2B is based on CogVideoX-2B [18], while DOVE is built on CogVideoX1.5-5B [18]. The best and second best results are colored with red and blue.

Our DOVE (main paper) is based on CogVideoX1.5-5B [18], which contains approximately 5B parameters. Benefits to the model-agnostic design, we can construct DOVE variants by using different pretrained backbones. Therefore, we develop a lighter version, DOVE-2B, built on the smaller model, CogVideoX-2B [18]. To train DOVE-2B, we use the HQ-VSR dataset along with the two-stage latent-pixel training strategy. Considering DOVE-2B has fewer parameters than DOVE, we train 5,000 iterations in stage-1, and 500 iterations in stage-2. All other training settings are consistent with DOVE. Details are provided in Sec. 4.1 of the main paper.

*Equal contribution.

†Corresponding author: Yulun Zhang, yulun100@gmail.com

We compare DOVE-2B and DOVE with some diffusion-based video super-resolution (VSR) methods on both synthetic and real-world datasets. Results are shown in Tab. 1. Both DOVE-2B and DOVE outperform recent VSR methods regarding fidelity, perceptual quality, and consistency. These results further demonstrate the effectiveness and generalizability of our one-step diffusion framework.

B Additional NR-IQA Metrics

Method	UDM10			SPMCS		
	MUSIQ \uparrow	MANIQA \uparrow	Q-Align \uparrow	MUSIQ \uparrow	MANIQA \uparrow	Q-Align \uparrow
Upscale-A-Video [22]	59.06	0.2868	0.7625	64.19	0.4231	0.7507
MGLD-VSR [16]	60.55	0.2783	0.7794	65.56	0.3099	0.7701
VEnhancer [2]	37.25	0.2120	0.5513	42.71	0.2351	0.5340
STAR [14]	41.98	0.2088	0.5340	36.66	0.2270	0.4168
DOVE (ours)	61.68	0.3284	0.8250	69.06	0.3714	0.8429

Table 2: Quantitative comparison using additional NR-IQA metrics on UDM10 [6] and SPMCS [19]. The best and second best results are colored with red and blue.

To further evaluate the perceptual quality of generated videos, we conduct experiments using additional no-reference image quality assessment (NR-IQA) metrics, including MUSIQ, MANIQA, and Q-Align. We compare our proposed method (DOVE) with recent state-of-the-art approaches: Upscale-A-Video [22], MGLD-VSR [16], VEnhancer [2], and STAR [14].

As shown in Tab 2, our method consistently achieves the highest scores across all three metrics. These results further validate the strong perceptual quality and generalization capability of DOVE, and complement the quantitative results in the main paper.

C Comparison with One-Step VSR Methods

Method	UDM10		SPMCS		RealVSR		MVS4x	
	DOVER \uparrow	$E_{warp}^*\downarrow$	DOVER \uparrow	$E_{warp}^*\downarrow$	DOVER \uparrow	$E_{warp}^*\downarrow$	DOVER \uparrow	$E_{warp}^*\downarrow$
SinSR [10]	0.4091	8.59	0.6052	11.20	0.7192	11.97	0.3974	5.86
OSDiff [13]	0.7240	5.73	0.7380	4.77	0.7650	7.21	0.6729	2.75
DLoRAL [5]	0.7219	3.40	0.6741	2.98	0.7539	4.73	0.6267	1.60
SeedVR2 [8]	0.5568	1.98	0.6320	1.23	0.7209	4.77	0.3098	1.08
DOVE (ours)	0.7809	1.77	0.7828	1.04	0.7867	3.52	0.6984	0.78

Table 3: Quantitative comparison with the existing one-step SR method on four benchmarks. The best and second best results are colored with red and blue.

We compare DOVE with recent one-step super-resolution (SR) models, including OSDiff [13], SinSR [10], DLoRAL [5], and SeedVR2 [8]. Most methods (except SeedVR2) rely on pretrained **image** models and introduce handcrafted modules to adapt them to video data. In contrast, our method is built upon a pretrained **video** model and employs a dedicated training strategy.

We evaluate DOVE and the baselines across four benchmarks: UDM10 [6], SPMCS [19], RealVSR [17], and MVS4x [9]. As observed in Tab 3, image-based SR methods (OSDiff and SinSR) show inferior performance on video-specific metrics such as DOVER and E_{warp}^* . While DLoRAL and SeedVR2 [8] achieve competitive results, our method consistently outperforms all baselines across datasets. These results validate the advantage of leveraging a pretrained video model and our tailored training strategy for one-step video super-resolution.

D More Complexity Comparison

We compare various diffusion-based VSR methods in terms of inference step, parameters, computational complexity (*i.e.*, MACs), and running time. Running time is measured on an A100 GPU using a 33-frame 720×1280 video. Please note that VEnhancer uses the new version v2, and DOVE is optimized for faster inference without affecting performance. The results are provided in Tab. 4.

Although DOVE has a larger number of parameters, its computational complexity and running time are much lower than other multi-step diffusion-based VSR methods. Meanwhile, DOVE-2B has fewer parameters than VEnhancer [2] and STAR [14], with faster speed and better performance.

Method	Upscale-A-Video [22]	MGLD-VSR [16]	VEnhancer [2]	STAR [14]	DOVE-2B (ours)	DOVE (ours)
Inference Step	30	50	15	15	1	1
Parameters (M)	1,086.75	1,564.66	2,496.59	2,492.90	1,910.28	5,787.19
MACs (T)	9,084.73	8,528.7	3,056.16	4,281.67	461.38	504.81
Running Time (s)	279.32	425.23	121.27	173.07	14.88	14.90

Table 4: More comprehensive complexity comparison. We compare several diffusion-based VSR methods in terms of inference step, parameters, MACs, and running time. Running time is measured on one A100 GPU by processing a 33-frame 720×1280 video.

Besides, it is interesting that while DOVE has around 5B parameters, its running time is comparable to that of DOVE-2B. To better understand this, we analyze the components of the model. Both DOVE and DOVE-2B consist of three parts: the VAE, the Transformer, and the empty text embedding. The empty prompt is pre-encoded to reduce encoding overhead.

DOVE-2B and DOVE use the same VAE with small parameters but high complexity. Its complexity is 418.06T, accounting for approximately 90% of the total computational cost, yet contains only 215.58M parameters. In contrast, the Transformer module has 1,693.68M parameters in DOVE-2B and 5,570.68M in DOVE, but contributes only 43.32T and 86.74T to the total cost, respectively. Thus, despite the large parameter difference between DOVE and DOVE-2B, their overall complexity and running time remain comparable. This inspired us to reduce overhead by optimizing VAE.

E Additional Ablation Study

In our one-step diffusion framework, denoising is performed in a single step, which requires selecting an appropriate starting timestep t . We conduct an ablation study on step t . The results are listed in Tab. 5. In the diffusion model, a smaller t corresponds to a state closer to the clean target.

Start Step (t)	9	99	399	599	999
PSNR \uparrow	27.24	27.22	27.20	27.28	22.10
LPIPS \downarrow	0.3226	0.3115	0.3037	0.3138	0.3947
CLIP-IQA \uparrow	0.3050	0.3116	0.3236	0.3036	0.2645
DOVER \uparrow	0.5647	0.5862	0.6154	0.5543	0.4298

Table 5: Ablation study on start step (t) in stage-1. The smaller t denotes the closer state to the target.

We observe that when t is large (*e.g.*, 999), overall performance is poor. This may be because large t values bias the model toward generating global structure, whereas VSR emphasizes fine-detail reconstruction, creating a mismatch that hinders learning effectiveness. Meanwhile, very small values of t (*e.g.*, 9) result in better fidelity but lower perceptual quality, due to reduced capacity for generating details. Conversely, a relatively larger t , *e.g.*, 399, achieves a better trade-off between fidelity and perceptual quality. This is consistent with our analysis in Sec. 3.1 of the main paper.

F Pipeline Details

In this section, we provide a detailed description of the video processing pipeline configuration and some statistics of the constructed dataset, HQ-VSR.

F.1 Pipeline Configuration

Our pipeline consists of four steps, detailed are as follows:

Step 1: Metadata Filtering. We first read the metadata of each video to perform preliminary filtering. This step can reduce the number of videos that proceed to the more time-consuming stages, thereby improving overall efficiency. We set the pixel and frame thresholds to 720 and 50, respectively.

Step 2: Scene Filtering. Videos containing multiple scenes (*e.g.*, edited clips) are unsuitable for training, as they hinder consistent semantic learning. Thus, we perform scene filtering using PySceneCut [21], which detects scene boundaries and provides timestamps. We then segment the videos based on these timestamps and discard clips with fewer than 50 frames.

Step 3: Quality Filtering. We apply diverse quality metrics tailored for video super-resolution, covering aesthetics, single-frame quality, and overall video quality. For aesthetic quality, we use the LAION aesthetic model [4] with a threshold of 6.5. For single-frame quality, we adopt the no-reference IQA metric CLIP-IQA [7]. To reduce computation, we uniformly sample 10 frames per video and use the average score, setting a threshold of 0.4. For video-level quality, we employ FasterVQA [11] and DOVER [12], with thresholds set to 0.6 and 0.7, respectively.

Step 4: Motion Processing. We compute the optical flow using UniMatch [15]. From the flow map, we generate a motion intensity map, followed by a motion mask and bounding box for cropping. We set the motion threshold τ to 1.0, and set the bounding box padding $p=200$.

F.2 Dataset Statistics

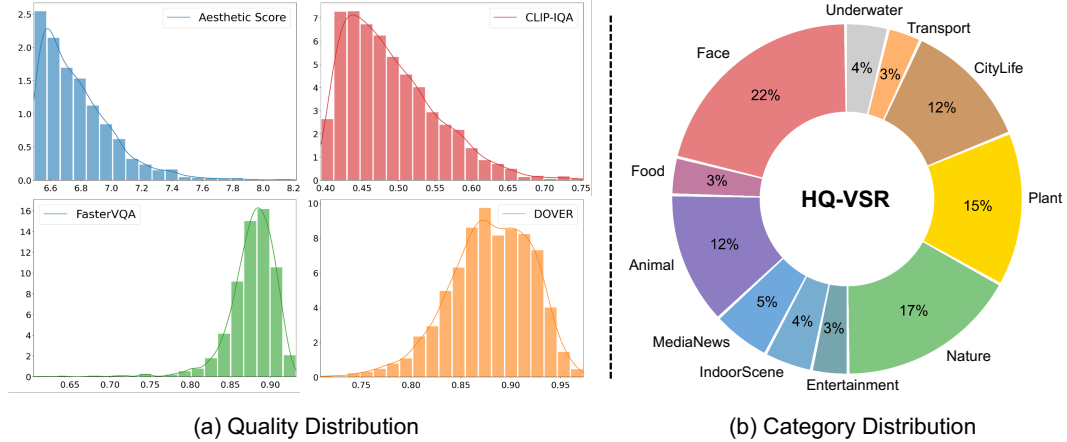


Figure 1: Data distribution of HQ-VSR. (a) Quality distribution, including Aesthetic Score, CLIP-IQA, DOVER, and FasterVQA. (b) Category distribution, showing the proportion of 11 categories.

We curate a high-quality dataset, HQ-VSR, from OpenVid-1M [3] using the proposed pipeline. The HQ-VSR contains 2,055 videos. We visualize its quality and category distributions in Fig. 1.

For quality distribution, we observe that HQ-VSR samples exhibit high scores across various metrics while maintaining a broad range. This diversity ensures the dataset is suitable for training models across different quality levels. Additionally, HQ-VSR covers 11 distinct scene categories (*e.g.*, face, food, and animal). This broad coverage helps enhance the generalization of VSR models.

G More Qualitative Results

We present more visual results, including temporal consistency and qualitative comparison.

G.1 Temporal Consistency

We visualize more temporal profiles across various datasets, as shown in Fig. 2. Compared methods exhibit misalignment or excessive smoothing in some challenging cases. While the latter may yield higher temporal consistency scores (*i.e.*, E_{warp}^*), it often comes at the cost of reduced perceptual quality. In contrast, our method produces smoother temporal profiles with richer details. These results demonstrate that our DOVE can maintain temporal consistency while preserving details.

G.2 Visual Comparison

We provide more visual comparisons. First, we compare the performance of different methods across multiple frames, as shown in Figs. 3 and 4. Our method consistently maintains temporal coherence and clear detail across frames. For instance, in Fig. 3, while MGLD-VSR [16] reconstructs a relatively sharp face in the first frame, its performance quickly degrades as the degradation intensifies in subsequent frames. In contrast, our method achieves more robust information propagation, successfully restoring fine details even under severe degradation. In the second example (Fig. 4), our method also preserves text consistency across frames under real-world conditions.

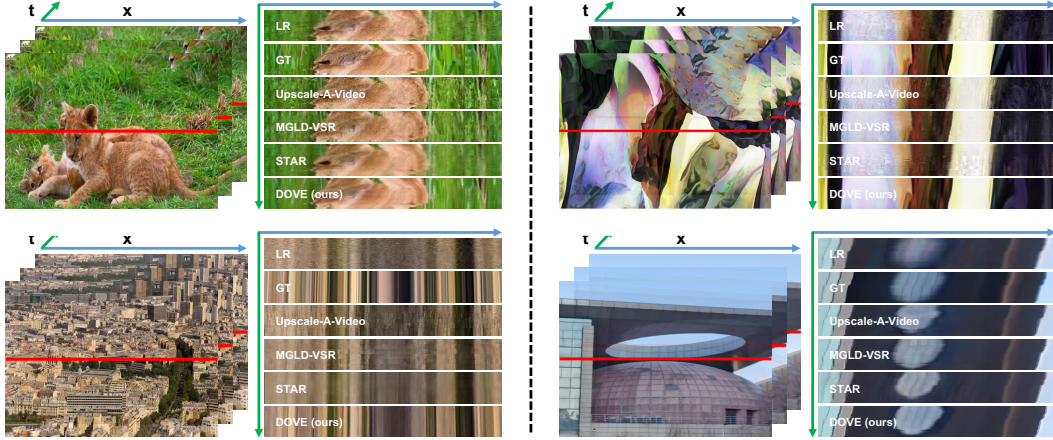


Figure 2: Comparison of temporal consistency (stacking the red line across frames).



Figure 3: Multi-frame comparison on the synthetic dataset (*i.e.*, UDM10 [6]: 008).

Furthermore, we present more visual results in Figs. 5, 6, and 7. Our method delivers impressive results across diverse scenarios, including architecture, portraits, text, animals, and nature. Compared to other methods, our DOVE produces clearer and more realistic reconstructions. These results further demonstrate the effectiveness of our proposed method. Besides, on the website (<https://zheng-chen.cn/DOVE>), we provide video comparisons across different datasets.

H Explanations for Checklist

H.1 Limitations

In this work, we propose an efficient one-step diffusion-based video super-resolution method. While it achieves good restoration performance and running efficiency, the acceleration primarily benefits the Transformer denoiser. The VAE remains a major bottleneck, which is not addressed. Additionally, although our method is based on a text-to-video model, we do not utilize textual information. Future work could explore multi-modal extensions in VSR tasks.

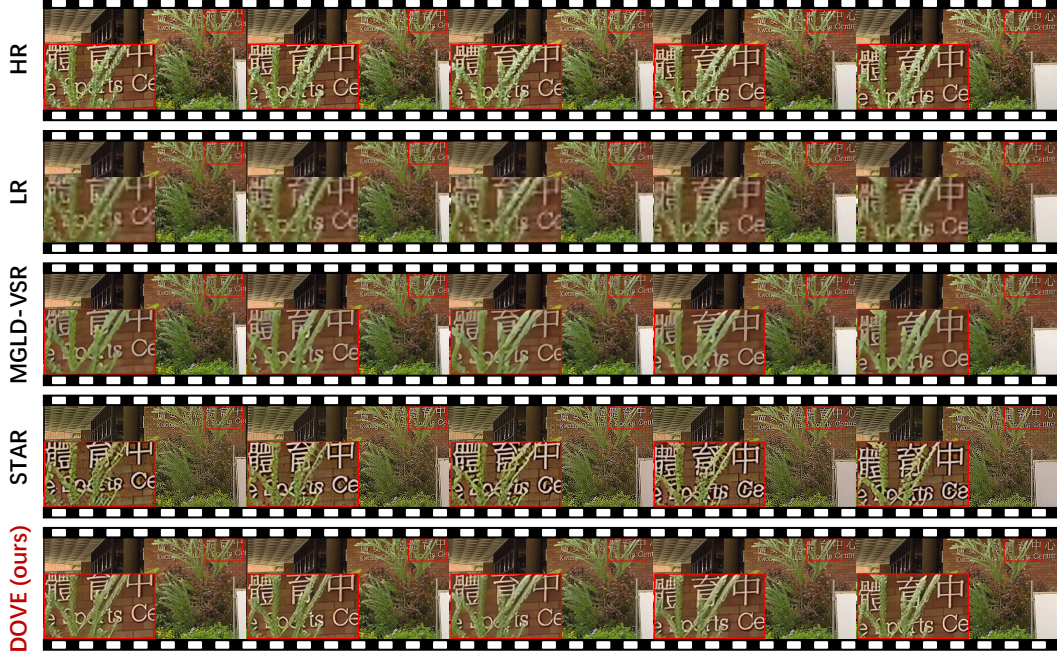


Figure 4: Multi-frame comparison on the real-world dataset (*i.e.*, RealVSR [17]: 135).

H.2 Broader Impacts

Our proposed method, DOVE, offers superior VSR performance and efficiency, benefiting academia and the industry. We believe the method poses no foreseeable negative societal impacts.

References

- [1] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022.
- [2] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024.
- [3] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *ICLR*, 2024.
- [4] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.
- [5] 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.
- [6] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017.
- [7] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.
- [8] 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.
- [9] Ruohao Wang, Xiaohui Liu, Zhilu Zhang, Xiaohe Wu, Chun-Mei Feng, Lei Zhang, and Wangmeng Zuo. Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In *CVPRW*, 2023.

- [10] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024.
- [11] 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. *TPAMI*, 2023.
- [12] 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.
- [13] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024.
- [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. *arXiv preprint arXiv:2501.02976*, 2025.
- [15] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthslat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024.
- [16] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *ECCV*, 2024.
- [17] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *CVPR*, 2021.
- [18] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.
- [19] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019.
- [20] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2023.
- [21] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.
- [22] 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.

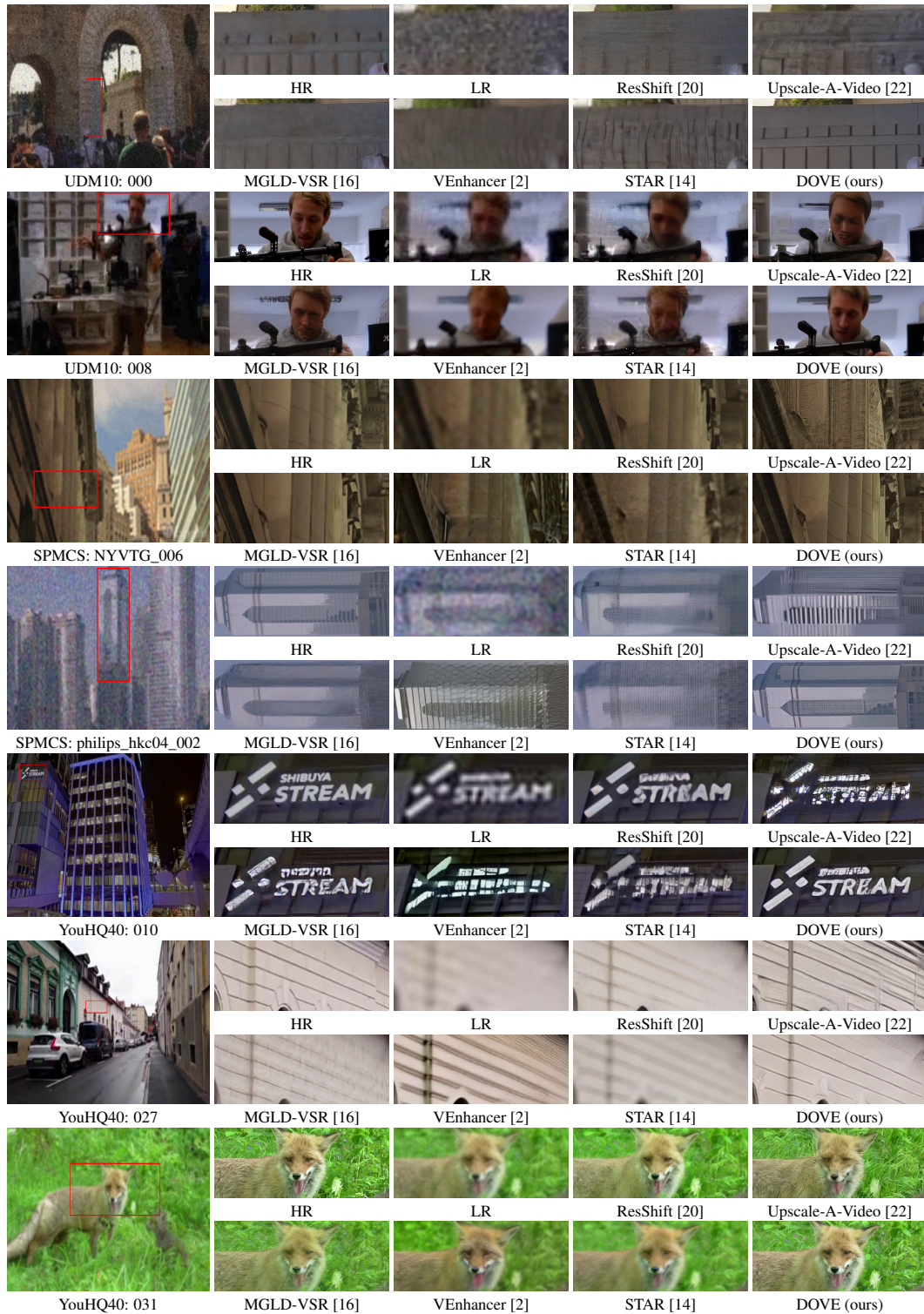


Figure 5: Visual comparison on synthetic (UDM10 [6], SPMCS [19], and YouHQ40 [22]) datasets.

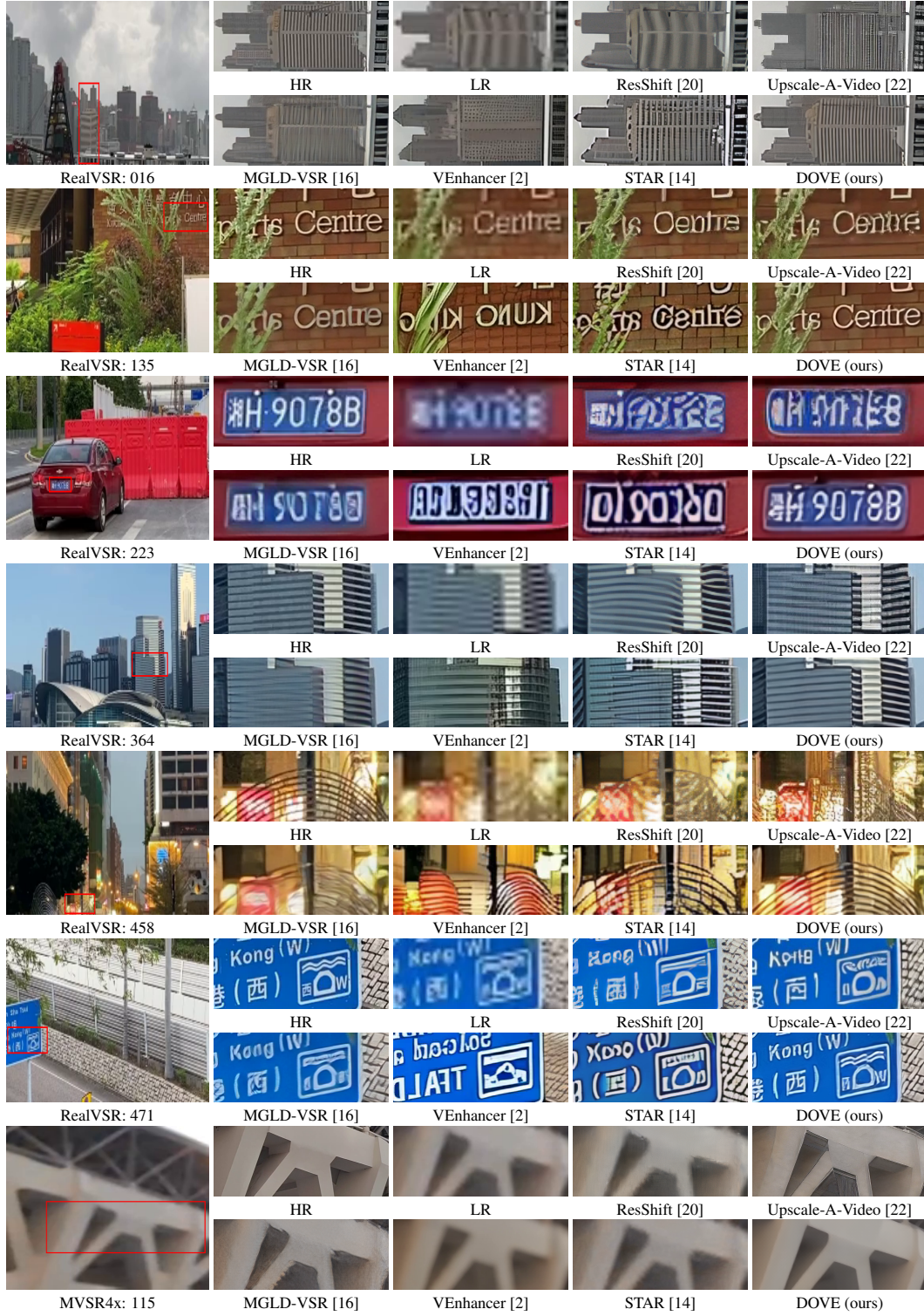


Figure 6: Visual comparison on real-world (RealVSR [17] and MVSR4x [9]) datasets.

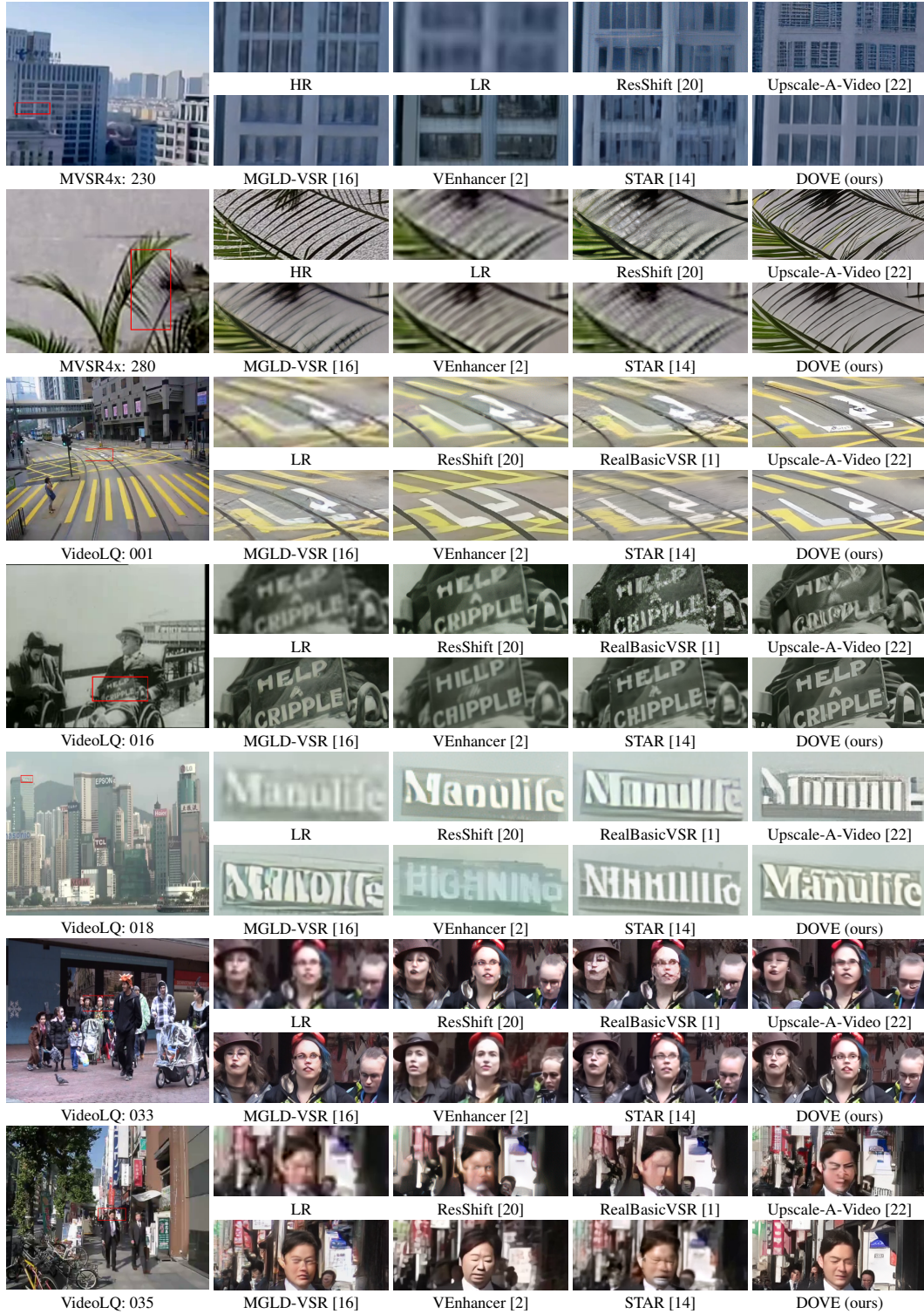


Figure 7: Visual comparison on real-world (MVS4x [9] and VideoLQ [1]) datasets. The videos in VideoLQ are sourced from the Internet without high-resolution (HQ) references.