Appendix

A. More Ablation Studies

A.1. Runtime Analysis for ILT and its alternatives

Here we examine the computational efficiency of our Interweaved Latent Transition method compared to alternatives. It is well known that incorporating temporal consistency inevitably increases inference time, regardless of the method chosen. As shown in Tab. 1, achieving an acceptable balance between quality and efficiency is crucial.

Our ILT approach increases runtime compared to the baseline without temporal modeling, but delivers substantial improvements in both quality metrics and temporal consistency. Notably, when compared to flow-based alignment methods commonly used in super-resolution, including those in UAV [10], our approach offers several advantages despite similar computational costs: ILT requires no additional training, introduces no extra parameters, and avoids the error accumulation problems inherent in flow-based methods.

The results demonstrate that ILT achieves superior temporal consistency (lower warping error) while maintaining competitive or better performance across other metrics. This favorable trade-off between computational cost and quality improvement aligns with our overall framework philosophy: thoughtful design choices that complement our learning strategy can provide significant benefits without excessive computational burden.

Datasets	Metrics	w/o ILT	ILT $\rightarrow$ FA	w/ ILT
YouHQ40	PSNR↑	23.467	23.529	23.713
	LPIPS↓	0.293	0.287	0.288
	MUSIQ↑	69.257	66.195	68.040
	$E^*_{warp}\downarrow$	1.870	1.589	1.492
	Runtime (s)	424.81	728.22	727.85

Table 1. Computational cost and performance comparison of ILT. ILT $\rightarrow$ FA: replacing ILT with flow alignment in UAV. Runtime: processing video 000, calculated with **official time computation code**.

A.2. Effectiveness of Temporal-Enhanced 3DVAE.

While our primary contribution is the Progressive Learning Strategy, architectural components still provide complementary benefits. We evaluate VAE variants trained on OpenVid-1M using our collected VAE-VAL5 dataset for texture and motion assessment. As shown in Tab. 2, compared to 2D VAE, the 3D VAE improves PSNR by 0.4dB and reduces warping error by 24%. Our TE-3DVAE further achieves incremental improvements across all metrics, particularly in temporal consistency. These results demonstrate that architectural enhancements, though secondary to learning strategy, contribute meaningful refinements to the overall framework through better temporal information preservation in the latent space.

	PSNR↑	SSIM↑	$E^*_{warp}\downarrow$	TF↑
2DVAE	30.251	0.917	1.049	0.965
3DVAE	30.645	0.926	0.800	0.967
TE-3DVAE	30.874	0.927	0.761	0.968

Table 2. Quantitative comparison of VAE variants on VAE-VAL5 dataset. While 3D architecture provides substantial improvement over 2D, temporal enhancement offers further refinement, consistent with our thesis that architectural improvements provide complementary benefits to our learning strategy.

A.3. Effectiveness of Multi-Scale Temporal Attention.

While our Progressive Learning Strategy is the key to handling complex degradations, architectural components like MSTA provide modest but consistent improvements. As shown in Tab. 3, when we remove MSTA while keeping our core learning

Exp.	MSTA	PLS	ILT	PSNR↑	LPIPS↓	MUSIQ↑	E_{warp}↓
(1)	✓	✓	✓	24.172	0.285	68.982	1.707
(2)		✓	✓	23.671	0.299	68.357	2.330

Table 3. Ablation study on MSTA. While maintaining our core Progressive Learning Strategy, removing MSTA causes performance degradation across all metrics, particularly in temporal consistency.

strategy (Exp. 2), we observe performance decreases across all metrics, with temporal consistency (warping error) showing particularly notable decline (36% increase in error).

The multi-scale design in MSTA applies 2× downsampling to features at intermediate stages, enabling efficient processing of motion information at various scales. This architectural enhancement helps capture temporal dependencies when combined with our learning-focused approach. These results align with our central thesis that while learning strategy is the primary factor for robust performance, well-designed architectural components like MSTA can still provide complementary benefits by improving the model's ability to understand motion dynamics and maintain temporal coherence.

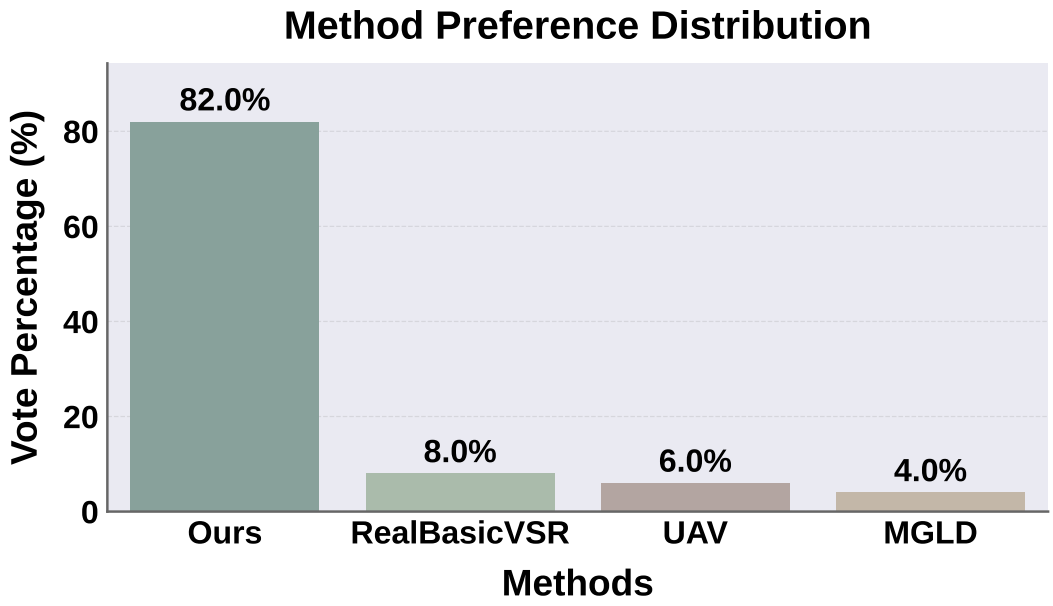

Figure 1. Quantitative comparison of user preferences among different video restoration methods. Our method achieves the highest preference rate (82.0%) in the user study, significantly outperforming RealBasicVSR [3], MGLD [9], and UAV [10].

A.4. Effectiveness of Text Prompts

Our DiffVSR model supports both text-guided and text-free restoration approaches. As shown in Fig. 2, incorporating classifier-free guidance [5] with meaningful text prompts significantly enhances visual quality compared to using empty prompts. This improvement is particularly evident in fine-grained details, such as the enhanced fur textures of the piglet and owl, sharper vegetation in the zebra scene, and more distinct zebra patterns in distant regions. These results demonstrate that appropriate text prompts can effectively guide the restoration process, enhancing intricate textures while maintaining a natural and realistic appearance.

The ability of text prompts to improve results highlights their role in activating the pretrained generative priors within our framework, enabling the model to leverage its full potential when dealing with severely degraded videos. Unlike previous approaches that may struggle with complex degradations regardless of text guidance, our Progressive Learning Strategy creates a foundation where text prompts can more effectively steer the restoration process. These findings align with prior work [10], further validating that text guidance combined with an effective learning strategy can significantly boost performance for complex video restoration tasks.

B. User Study

We conducted a user study to evaluate perceptual quality through blind comparison. Twenty participants with a computer vision background were asked to compare restoration results from RealBasicVSR [3], MGLD [9], UAV [10], and our DiffVSR. The study included 20 test sets, containing both real-world and synthetic degraded videos. Participants were instructed to select the most visually appealing result based on three criteria: overall fidelity, detail preservation, and temporal consistency. To ensure unbiased evaluations, method names were hidden, and display order was randomized. As shown in Fig. 1, our method achieved the highest preference rate across various degradation scenarios, particularly on severely degraded videos where competing methods struggle. This validates our approach of addressing the fundamental learning burden through Progressive Learning Strategy rather than solely focusing on architectural complexity.

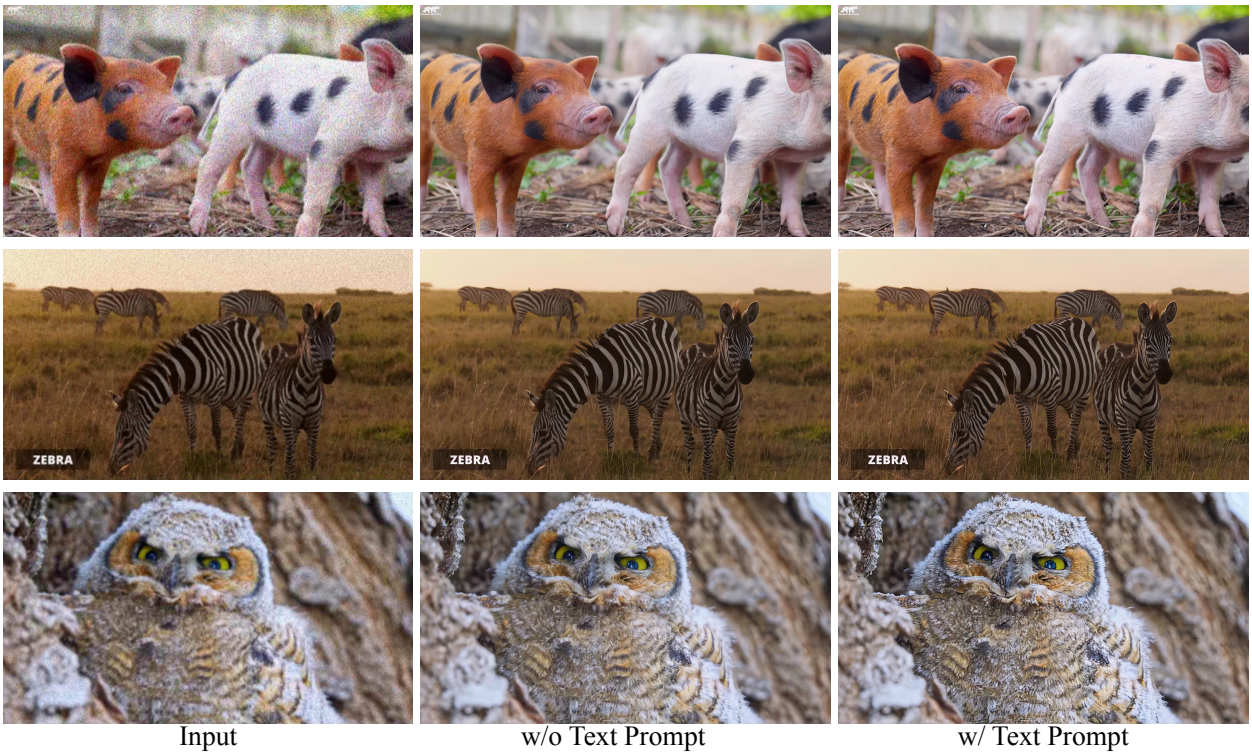

Figure 2. Visual comparison of restoration results with and without text prompts. Incorporating meaningful text prompts as guidance significantly enhances visual quality, improving texture sharpness, fine-grained details, and overall realism. **(Zoom-in for best view)**

Datasets	Metrics	Real-ESRGAN[8]	SD ×4 Upscaler[1]	DiffBIR[6]	RealBasicVSR[3]	MGLD[9]	VEnhancer[4]	UAV[10]	Ours
VideoLQ	NRQM↑	5.772	4.956	6.430	6.175	5.990	4.351	5.090	6.257
	CLIP-IQA↑	0.633	0.682	0.611	0.669	0.698	0.659	0.639	0.682
	MUSIQ↑	49.837	34.864	53.575	55.975	47.958	43.555	36.369	57.812
	DOVER↑	0.728	0.518	0.679	0.742	0.730	0.657	0.632	0.747

Table 4. Quantitative comparison with state-of-the-art methods on VideoLQ dataset.

C. Trade-off Between Visual Quality and Temporal Consistency

In video restoration tasks, achieving a balance between visual quality and temporal consistency is a well-known challenge [7, 9, 10], akin to the perception-distortion trade-off in image restoration [2]. Higher restoration quality typically enhances texture detail but can often lead to reduced temporal consistency across frames.

To illustrate this trade-off, we analyze MUSIQ (video quality) and Warping Error (temporal consistency) scores on the YouHQ40 dataset, as shown in Fig. 3. The ideal performance lies in the lower-right region, reflecting high video quality and

low warping error. As observed, image-based methods achieve higher quality scores but poor temporal consistency, while video-based approaches maintain better temporal coherence at the cost of reduced quality. Our DiffVSR achieves a favorable balance between these competing objectives, demonstrating how our Interweaved Latent Transition technique effectively maintains temporal consistency without additional training overhead. This further supports our thesis that addressing learning strategy and making intelligent design choices can lead to superior performance in complex degradation scenarios, without requiring overly complex architectural modifications.

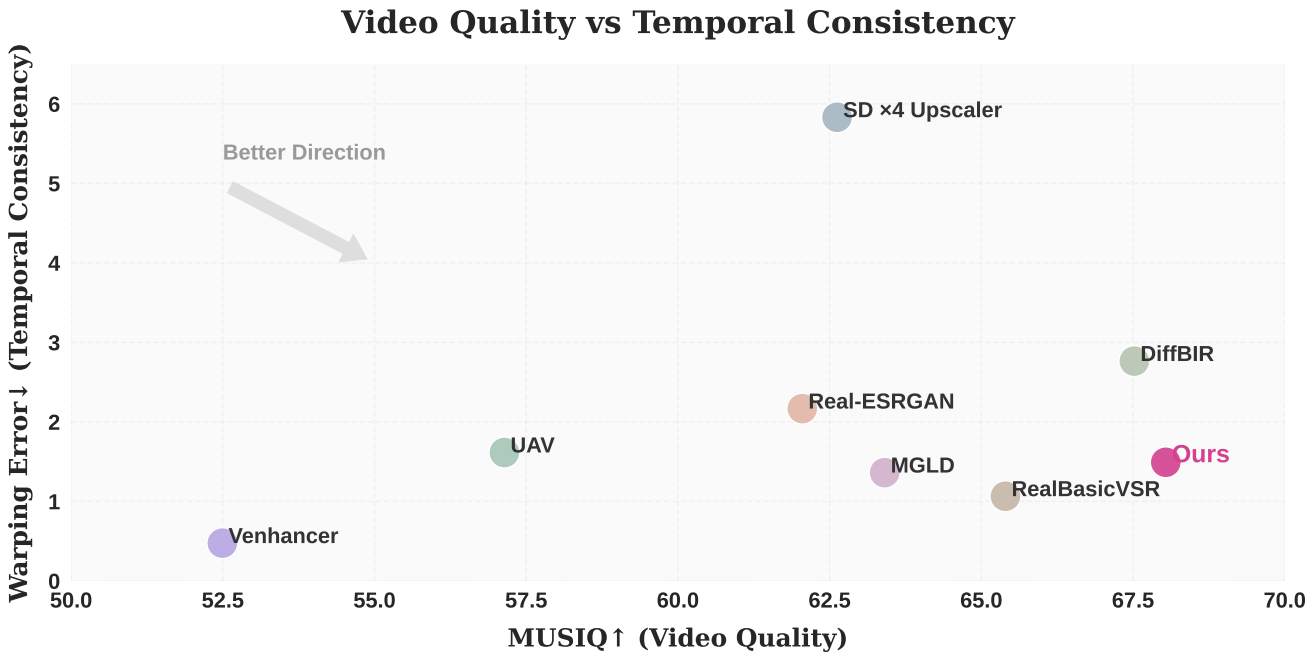

Figure 3. Analysis of the trade-off between video quality (MUSIQ) and temporal consistency (Warping Error) on YouHQ40 dataset. Methods towards the lower-right corner achieve better overall performance, with higher MUSIQ and lower warping error.

D. More Quantitative and Qualitative Comparisons

D.1. More Quantitative Evaluation

We provide additional quantitative evaluations on the VideoLQ [3] dataset, which contains real-world low-quality videos with diverse degradation types. As shown in Tab. 4, our DiffVSR consistently outperforms existing approaches across all metrics. Notably, we achieve significant improvements in perceptual quality metrics, with the highest scores in MUSIQ (57.812). The superior DOVER score (0.747) further validates our approach's effectiveness in maintaining both visual quality and temporal consistency. These results align with our findings in the main paper, demonstrating that addressing the learning burden through our Progressive Learning Strategy leads to robust performance across different real-world degradation scenarios.

D.2. More Qualitative Results

We conduct comprehensive visual comparisons against state-of-the-art approaches, including both image-based methods (ESRGAN [8], SD ×4 Upscaler [1], DiffBIR [6]) and video-based methods (RealBasicVSR [3], MGLD [9], VEnhancer [4], UAV [10]). Figs.5 and4 showcase the qualitative results on synthetic and real-world test videos, respectively. Our method demonstrates superior capability in recovering diverse textures and structures across various severely degraded scenarios, including architectural elements (wall textures, brick patterns), organic details (facial features, hair strands), natural scenes (vegetation, marine life), and high-frequency patterns (text, dental structures). The restored results exhibit both high fidelity to reference images and rich textural details, while avoiding common artifacts like over-smoothing or false pattern generation that often plague methods struggling with complex degradation modeling.

E. Video Demo

We also provide a demonstration video [DiffVSR.mp4] to showcase the capabilities of our method on both synthetic and real-world videos. The video highlights temporal coherence and dynamic detail preservation enabled by our Interweaved Latent Transition technique, which are better appreciated in motion than through static comparisons. **The demonstration video is included in the supplementary materials. Note that due to file size limitations, the video has been compressed; the original results exhibit even higher visual quality.**

F. Limitations

While DiffVSR achieves significant improvements over existing methods, it has several limitations: (1) As a diffusion-based model, DiffVSR requires repetitive iterations for inference, resulting in slower processing times. This makes it challenging to deploy in real-time applications, though our ILT approach helps maintain competitive computational efficiency compared to alternative temporal consistency methods. (2) DiffVSR struggles with certain challenging scenarios, such as small faces, small human bodies, and complex street scenes, due to inherited limitations of current diffusion-based generative models. Addressing these issues may require further refinements to our Progressive Learning Strategy and additional task-specific adaptations. (3) Due to limited computational resources, we have not conducted larger-scale experiments with increased input sizes or batch sizes. Scaling up the training process with more GPUs could potentially further improve our model's ability to handle even more complex degradation distributions through our staged learning approach.

References

[1] Stability AI. Stable diffusion x4 upscaler. https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler, 2022. 3, 4

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 2, 3, 4

[4] 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, 4

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[6] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 3, 4

[7] Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *Computational Visual Media*, 10(2):375–395, 2024. 3

[8] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 3, 4

[9] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European Conference on Computer Vision*, pages 224–242. Springer, 2025. 2, 3, 4

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024. 1, 2, 3, 4

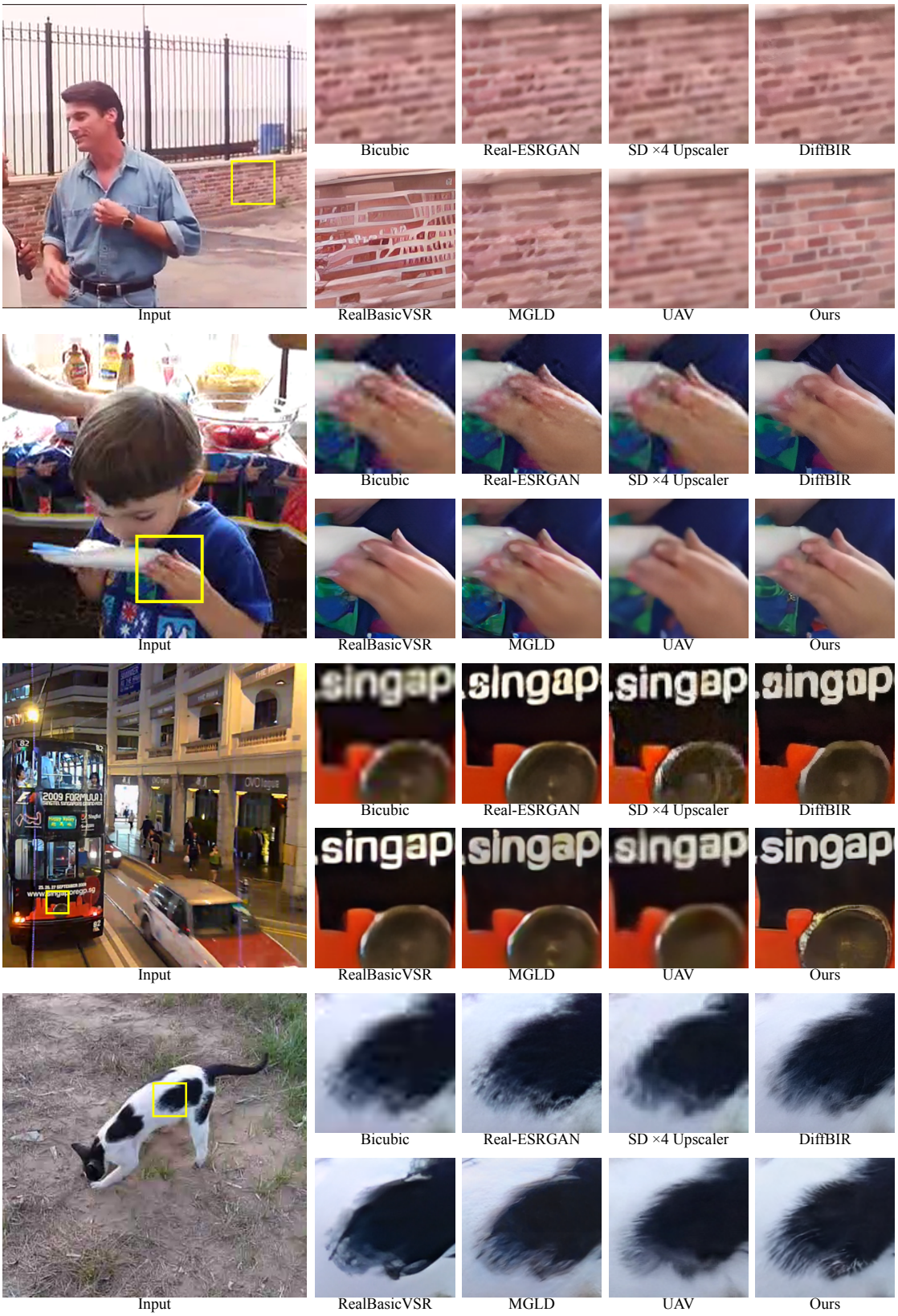

Figure 4. Qualitative comparisons on real-world videos. Our method effectively recovers fine details while maintaining natural textures. **(Zoom-in for best view)**

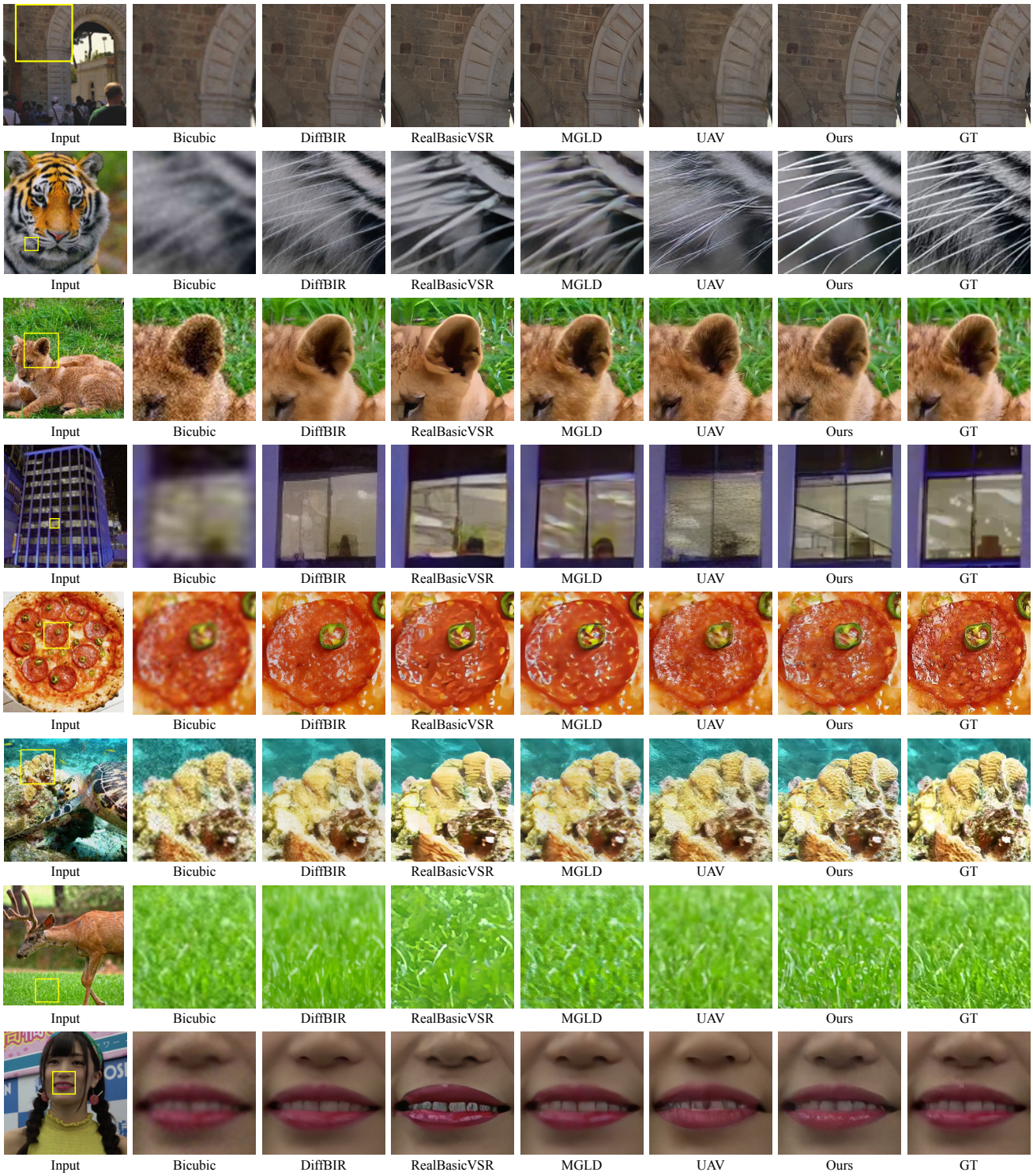

Figure 5. Qualitative comparisons on synthetic datasets. Our method demonstrates superior capability in recovering accurate facial details and textual information, while other methods struggle with either over-smoothing or detail distortion. **(Zoom-in for best view)**