

APPENDIX: TOWARDS REDUNDANCY REDUCTION IN DIFFUSION MODELS FOR EFFICIENT VIDEO SUPER-RESOLUTION

Anonymous authors

Paper under double-blind review

A SUPPLEMENTARY EXPERIMENTS

A.1 ATTENTION SPECIALIZATION ROUTINE

The attention specialization routine (ASR) is the core of our method. We further compare it with single-pattern attention baselines (global, intra-frame, and window) on UDM10 (Tao et al., 2017), SPMCS (Yi et al., 2019), and RealVSR (Yang et al., 2021) datasets. All models are trained with stage 2 only. For ASR, we set ρ to 0.4. As shown in Tab. 1, ASR consistently achieves the best results across almost all metrics. On synthetic datasets UDM10 and SPMCS, it surpasses all baselines in PSNR, LPIPS (Zhang et al., 2018), DOVER (Wu et al., 2023), and E_{warp}^* (Lai et al., 2018), showing clear advantages in both fidelity and temporal consistency. On the challenging RealVSR dataset, ASR maintains the best DOVER and E_{warp}^* scores while keeping perceptual quality competitive. Overall, these results highlight that combining global and localized heads through ASR yields stronger performance than relying on any single attention pattern.

Type	UDM10				SPMCS				RealVSR			
	PSNR \uparrow	LPIPS \downarrow	DOVER \uparrow	E_{warp}^* \downarrow	PSNR \uparrow	LPIPS \downarrow	DOVER \uparrow	E_{warp}^* \downarrow	PSNR \uparrow	LPIPS \downarrow	DOVER \uparrow	E_{warp}^* \downarrow
Global	25.13	0.2651	0.7845	2.20	22.36	0.2790	0.7337	1.34	20.64	0.2014	0.7492	3.39
Intra-Frame	24.47	0.2840	0.7614	2.80	21.97	0.2825	0.7218	1.70	20.14	0.2309	0.7371	3.76
Window	24.69	0.3008	0.7828	2.21	22.00	0.2974	0.7280	1.31	20.27	0.2297	0.7460	4.16
ASR (ours)	25.53	0.2513	0.7914	1.82	22.62	0.2727	0.7440	1.08	20.89	0.2045	0.7587	2.66

Table 1: Additional results of attention specialization routine. The best results are colored with red. All models are trained on the HQ-VSR (Chen et al., 2025) dataset.

A.2 GLOBAL-HEAD RATIO

We provide additional results to further examine the performance of OASIS under different global-head ratios ρ . As shown in Fig. 1, the curves of PSNR and E_{warp}^* exhibit similar trends across both the synthetic dataset (e.g., UDM10) and the real-world dataset (e.g., RealVSR). In most of the settings, the results obtained with ASR surpass those of the baseline that employs global attention only, indicating clear benefits from introducing localized specialization. Moreover, the consistency of the trends across datasets demonstrates that the effect of ρ is stable, and the model retains strong performance under a wide range of settings. These observations highlight the robustness of our ASR design and its ability to effectively balance global and localized patterns.

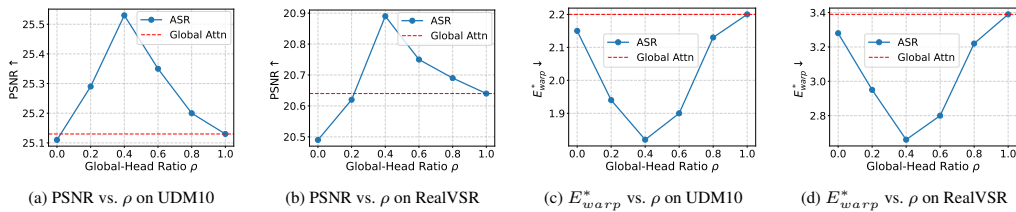


Figure 1: Additional results on the impact of the global-head ratio ρ on PSNR and E_{warp}^* . Global Attn refers to the baseline using the global attention only.

A.3 PROGRESSIVE TRAINING STRATEGY

In this section, we present additional evaluations of our progressive training strategy on three representative datasets: UDM10, SPMCS, and RealVSR. We include two standard training settings, where the model is trained for 30,000 iterations using either only stage 1 degradations (S1) or only stage 2 degradations (S2). In contrast, our progressive scheme (S1+S2) splits the training into two phases of 15,000 iterations each, beginning with temporally consistent degradations in stage 1 and then introducing temporally inconsistent degradations in stage 2. This design allows the model to gradually adapt from easier to more challenging degradations, leading to more effective learning.

As reported in Tab. 2, the progressive strategy consistently surpasses both S1 and S2 across all datasets and evaluation metrics. In particular, it delivers higher pixel fidelity measured by PSNR, better perceptual quality reflected in LPIPS and DOVER, and stronger temporal consistency indicated by lower E_{warp}^* . These results confirm that progressive training provides a more balanced and powerful optimization scheme, yielding clear advantages over standard single-stage training.

Type	UDM10				SPMCS				RealVSR			
	PSNR \uparrow	LPIPS \downarrow	DOVER \uparrow	$E_{warp}^*\downarrow$	PSNR \uparrow	LPIPS \downarrow	DOVER \uparrow	$E_{warp}^*\downarrow$	PSNR \uparrow	LPIPS \downarrow	DOVER \uparrow	$E_{warp}^*\downarrow$
S1	24.89	0.2544	0.7736	2.49	22.34	0.2845	0.7279	1.49	20.46	0.2151	0.7360	3.63
S2	25.13	0.2651	0.7845	2.20	22.36	0.2790	0.7337	1.34	20.64	0.2014	0.7492	3.39
S1+S2 (ours)	25.49	0.2485	0.7959	2.03	22.52	0.2726	0.7433	1.12	20.89	0.1938	0.7527	3.11

Table 2: Additional results of progressive training. The best results are colored with red. All models are trained on the HQ-VSR dataset.

B ADDITIONAL METHOD DETAILS

B.1 ATTENTION PATTERN PROPORTIONS UNDER DIFFERENT GLOBAL-HEAD RATIOS

Figure 2 reports the proportions of global, intra-frame, and window attention patterns under different target global-head ratios ρ . As ρ increases, the fraction of global heads rises accordingly, while both intra-frame and window heads decrease. When $\rho=0.4$, the proportions of global and window heads become roughly comparable, and this setting corresponds to the model’s best performance. Notably, both window and intra-frame patterns are consistently present across all settings, showing that the model maintains diverse localized specializations. At the same time, window attention is more frequent, with its number being $1.7\text{--}3.5\times$ larger than intra-frame, reflecting a stronger tendency toward spatio-temporal localized patterns.

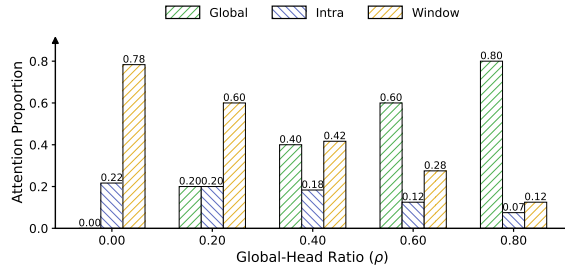
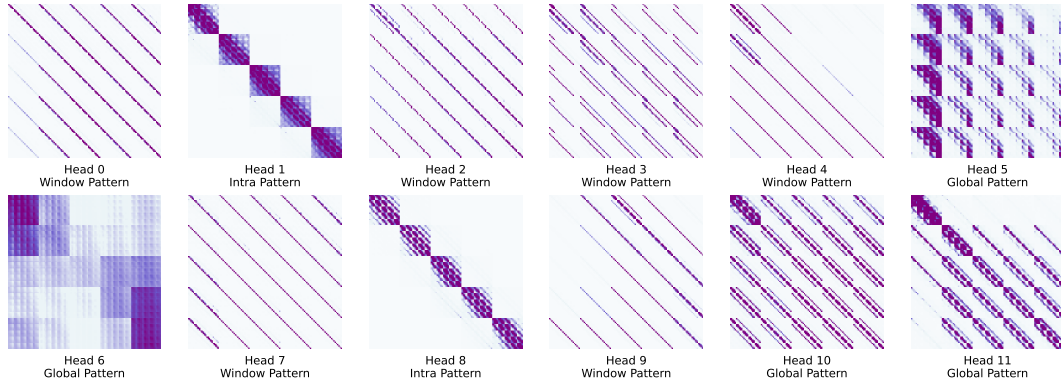


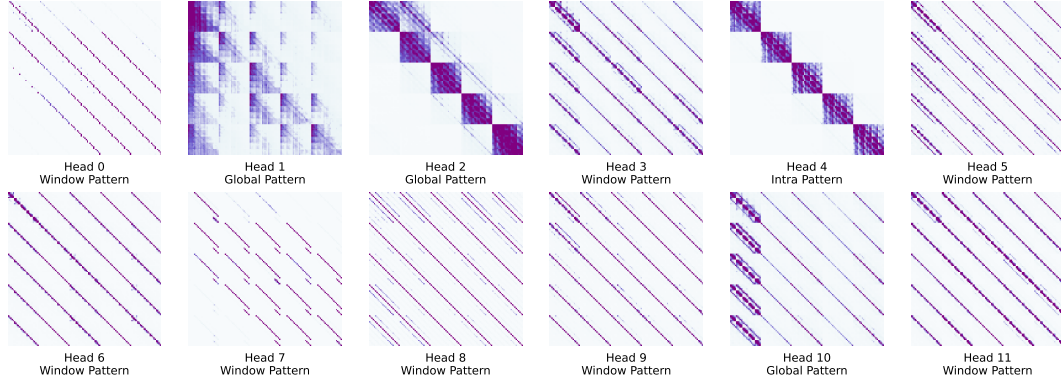
Figure 2: Proportions of global, intra-frame, and window attention under different global-head ratios ρ .

B.2 ADDITIONAL VISUALIZATIONS OF ATTENTION HEATMAPS

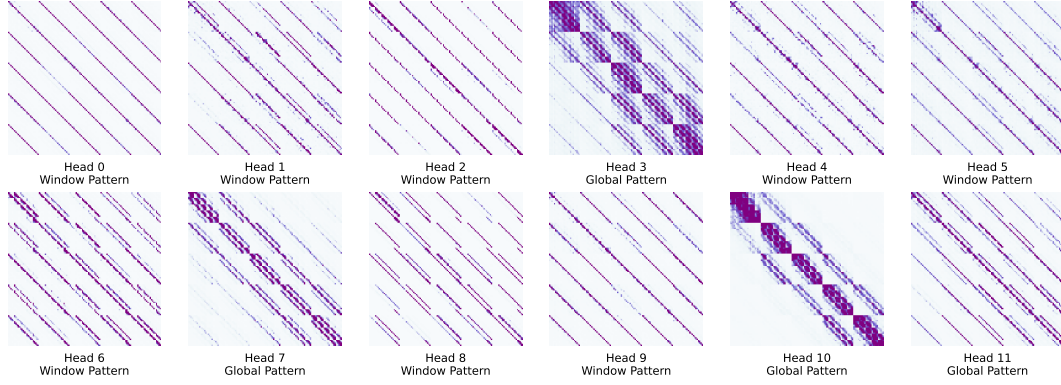
To further illustrate the redundancy of relying solely on global attention in the DiT, we provide additional visualizations of head attention maps from multiple layers, as shown in Fig. 3. Across different layers, we observe a clear diversity of specialization patterns: a substantial number of heads consistently exhibit localized behaviors, either concentrating within intra-frame regions or focusing on compact spatio-temporal neighborhoods, while other heads retain broad global receptive fields. This mixture of localized and global patterns recurs throughout the network, suggesting that the attention heads do not all serve the same role but instead divide into complementary functions. Importantly, the coexistence of these distinct behaviors across layers indicates that treating all heads as global introduces redundancy, since many of them naturally specialize in more constrained contexts. These visualizations highlight the intrinsic specialization tendencies of DiTs and further motivate our strategy of explicitly assigning heads to distinct attention patterns.



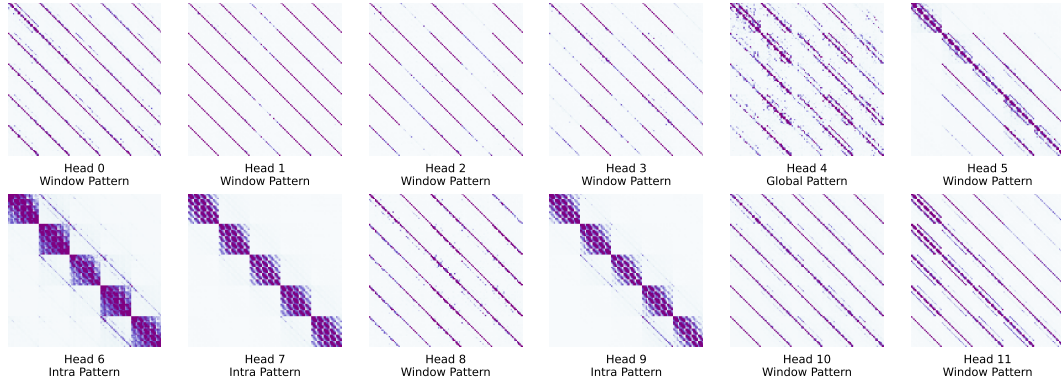
(a) Attention heatmap visualizations of Layer 1



(b) Attention heatmap visualizations of Layer 3



(c) Attention heatmap visualizations of Layer 12



(d) Attention heatmap visualizations of Layer 25

Figure 3: Illustrations of head-level specializations. Each heatmap shows the mean attention map computed over 50 videos from the training set, revealing global, intra-frame, and window patterns.

B.3 PSEUDOCODE FOR OASIS TRAINING AND INFERENCE

This section provides the pseudocode for OASIS, covering both training and inference stages, which are provided in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 OASIS Progressive Training

```

1: Input: Training set  $\mathcal{S}$ , DiT  $\mathcal{DN}_\theta$ , VAE decoder  $\mathcal{D}_\phi$ , iterations  $\{T_1, T_2\}$ , learning rate  $\eta$ 
2: Output: Finetuned DiT  $\mathcal{DN}_\theta$ 
3: for stage  $s = 1, 2$  do
4:   for iteration  $t = 1, \dots, T_s$  do
5:     Sample HQ video  $\mathbf{V}_H \sim \mathcal{S}$ 
6:     if  $s = 1$  then
7:       Apply temporally consistent degradations to  $\mathbf{V}_H$  to obtain  $\mathbf{V}_L$ 
8:     else
9:       Apply temporally inconsistent degradations to  $\mathbf{V}_H$  to obtain  $\mathbf{V}_L$ 
10:     $\tilde{\mathbf{z}}_L = \text{UnShuffle}(\mathbf{V}_L)$ ;  $\mathbf{z}_L = \mathbf{W}_{\text{proj}}\tilde{\mathbf{z}}_L + \mathbf{b}_{\text{proj}}$ 
11:     $\tilde{\mathbf{z}}_H = \text{UnShuffle}(\mathbf{V}_H)$ ;  $\mathbf{z}_H = \mathbf{W}_{\text{proj}}\tilde{\mathbf{z}}_H + \mathbf{b}_{\text{proj}}$ 
12:     $\hat{\mathbf{z}}_H = \mathbf{z}_L - \sigma_{T_L} \mathcal{DN}_\theta(\mathbf{z}_L, T_L)$ ;  $\hat{\mathbf{V}}_H = \mathcal{D}_\phi(\hat{\mathbf{z}}_H)$ 
13:    Compute losses:
14:     $\mathcal{L}_{\text{latent}} = \|\hat{\mathbf{z}}_H - \mathbf{z}_H\|_2^2$ ;  $\mathcal{L}_{\text{per}} = \|\hat{\mathbf{V}}_H - \mathbf{V}_H\|_2^2 + \mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{V}}_H, \mathbf{V}_H)$ 
15:     $\mathbf{O}^{\text{fw},i}, \mathbf{O}^{\text{bw},i} = \text{RAFT}(\mathbf{V}_H)$ 
16:     $\mathcal{L}_{\text{warp}} = \sum_i \|\text{Warp}(\hat{\mathbf{V}}_H^i, \mathbf{O}^{\text{bw},i}) - \hat{\mathbf{V}}_H^{i+1}\|_1 + \|\text{Warp}(\hat{\mathbf{V}}_H^i, \mathbf{O}^{\text{fw},i}) - \hat{\mathbf{V}}_H^{i-1}\|_1$ 
17:     $\mathcal{L} = \mathcal{L}_{\text{latent}} + \mathcal{L}_{\text{per}} + \lambda_{\text{warp}} \mathcal{L}_{\text{warp}}$ 
18:    Update  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ 
19: return  $\mathcal{DN}_\theta$ 

```

Algorithm 2 OASIS Inference with Attention Specialization Routing (ASR)

```

1: Input: LQ video  $\mathbf{V}_L$ , DiT  $\mathcal{DN}_\theta$ , VAE decoder  $\mathcal{D}_\phi$ , assignment map  $\mathcal{A}$  (head  $\rightarrow$  {global, intra-frame, window}), timestep  $T_L$ , window sizes  $(P_t, P_h, P_w)$ 
2: Output: Reconstructed HQ video  $\hat{\mathbf{V}}_H$ 
3: Preprocess:  $\tilde{\mathbf{z}}_L \leftarrow \text{UnShuffle}(\mathbf{V}_L)$ ;  $\mathbf{z}_L \leftarrow \mathbf{W}_{\text{proj}}\tilde{\mathbf{z}}_L + \mathbf{b}_{\text{proj}}$ 
4: Obtain per-head queries/keys/values ( $\mathbf{q}, \mathbf{k}, \mathbf{v}$ ) from  $\mathbf{z}_L$  via linear projections
5: for each attention layer  $\ell$  in  $\mathcal{DN}_\theta$  do
6:   for each head  $h$  in layer  $\ell$  do
7:     Routing via  $\mathcal{A}(h)$  for a query at position  $(t, h, w)$ :
8:     if  $\mathcal{A}(h) = \text{global}$  then
9:        $\mathcal{S}_h \leftarrow \{(t', h', w') : t' \in [1, T], h' \in [1, H], w' \in [1, W]\}$ 
10:    if  $\mathcal{A}(h) = \text{intra-frame}$  then
11:       $\mathcal{S}_h \leftarrow \{(t, h', w') : h' \in [1, H], w' \in [1, W]\}$ 
12:    else  $\triangleright \mathcal{A}(h) = \text{window}$ 
13:       $\mathcal{S}_h \leftarrow \{(t', h', w') : |t' - t| \leq P_t/2, |h' - h| \leq P_h/2, |w' - w| \leq P_w/2\}$ 
14:      Gather  $\mathbf{K}_h \leftarrow \{\mathbf{k}_{t', h', w'} : (t', h', w') \in \mathcal{S}_h\}$ ;  $\mathbf{V}_h \leftarrow \{\mathbf{v}_{t', h', w'} : (t', h', w') \in \mathcal{S}_h\}$ 
15:      Attend to  $(\mathbf{K}_h, \mathbf{V}_h)$  with query  $\mathbf{q}_{t, h, w}$  to produce head output  $\mathbf{o}^{(h)}$ 
16:      Concatenate  $\{\mathbf{o}^{(h)}\}_h$  and continue the layer's feed-forward
17: One-step latent reconstruction:  $\hat{\mathbf{z}}_H \leftarrow \mathbf{z}_L - \sigma_{T_L} \mathcal{DN}_\theta(\mathbf{z}_L, T_L)$ 
18: Decode to pixel space:  $\hat{\mathbf{V}}_H \leftarrow \mathcal{D}_\phi(\hat{\mathbf{z}}_H)$ 
19: return  $\hat{\mathbf{V}}_H$ 

```

C ADDITIONAL QUALITY COMPARISONS

C.1 TEMPORAL CONSISTENCY

We provide additional temporal profile visualizations in Fig. 4. Compared with competing methods, which still exhibit noticeable gaps from the ground truth, our approach achieves smoother frame-to-frame transitions and better preserves fine details, highlighting its stronger temporal consistency.

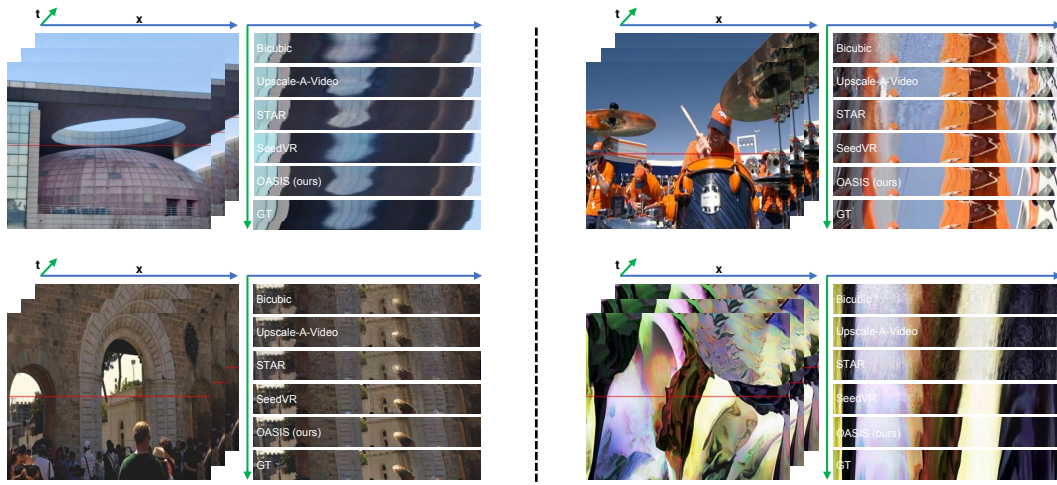


Figure 4: Additional comparisons of temporal consistency (stacking the red line across frames).

C.2 VIDSUAL COMPARISON

We provide visual comparisons on both synthetic and real-world datasets in Figs. 5 and 6. Across multiple frames, OASIS restores sharper details and more stable structures than competing methods. Moreover, patch-level results in Figs. 7 and 8 demonstrate its effectiveness across diverse scenes, where fine textures, subtle edges, and structural details are faithfully recovered.



Figure 5: Comparison across multiple frames (UDM10 (Tao et al., 2017): 003).

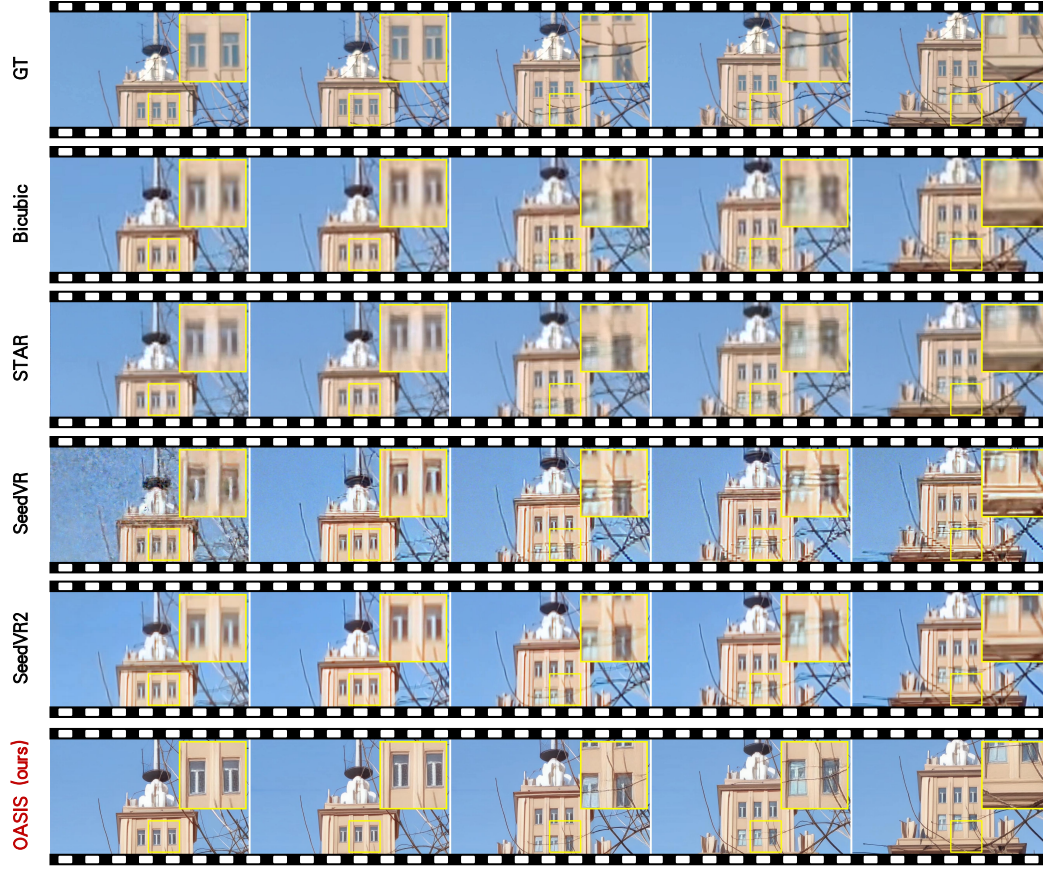


Figure 6: Comparison across multiple frames (MVSR4x (Wang et al., 2023): 065).

D LIMITATIONS AND BROADER IMPACTS

Limitations. Our method primarily reduces redundancy within the diffusion transformer, leading to notable improvements in both performance and efficiency. However, the VAE decoder remains an important bottleneck, as its potential redundancy may introduce computational overhead and affect reconstruction fidelity, particularly for fine temporal and spatial details. Exploring how to further reduce redundancy in the VAE decoder will be an important direction for future work.

Broader Impacts. Most existing approaches adapt diffusion models to video super-resolution by incorporating additional modules, which inevitably increase both model complexity and runtime overhead. In contrast, our work takes a redundancy-reduction perspective, which mitigates redundant computation so that the model not only runs more efficiently but also focuses on learning meaningful information rather than filtering out noise. This provides a novel and promising direction for adapting diffusion models to VSR, combining efficiency with stronger task-specific performance.

REFERENCES

- Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. Dove: Efficient one-step diffusion model for real-world video super-resolution. *arXiv preprint arXiv:2505.16239*, 2025.
- Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018.
- Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017.

- 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 *CVPR*, 2023.
- 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.
- Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *ICCV*, 2021.
- 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.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

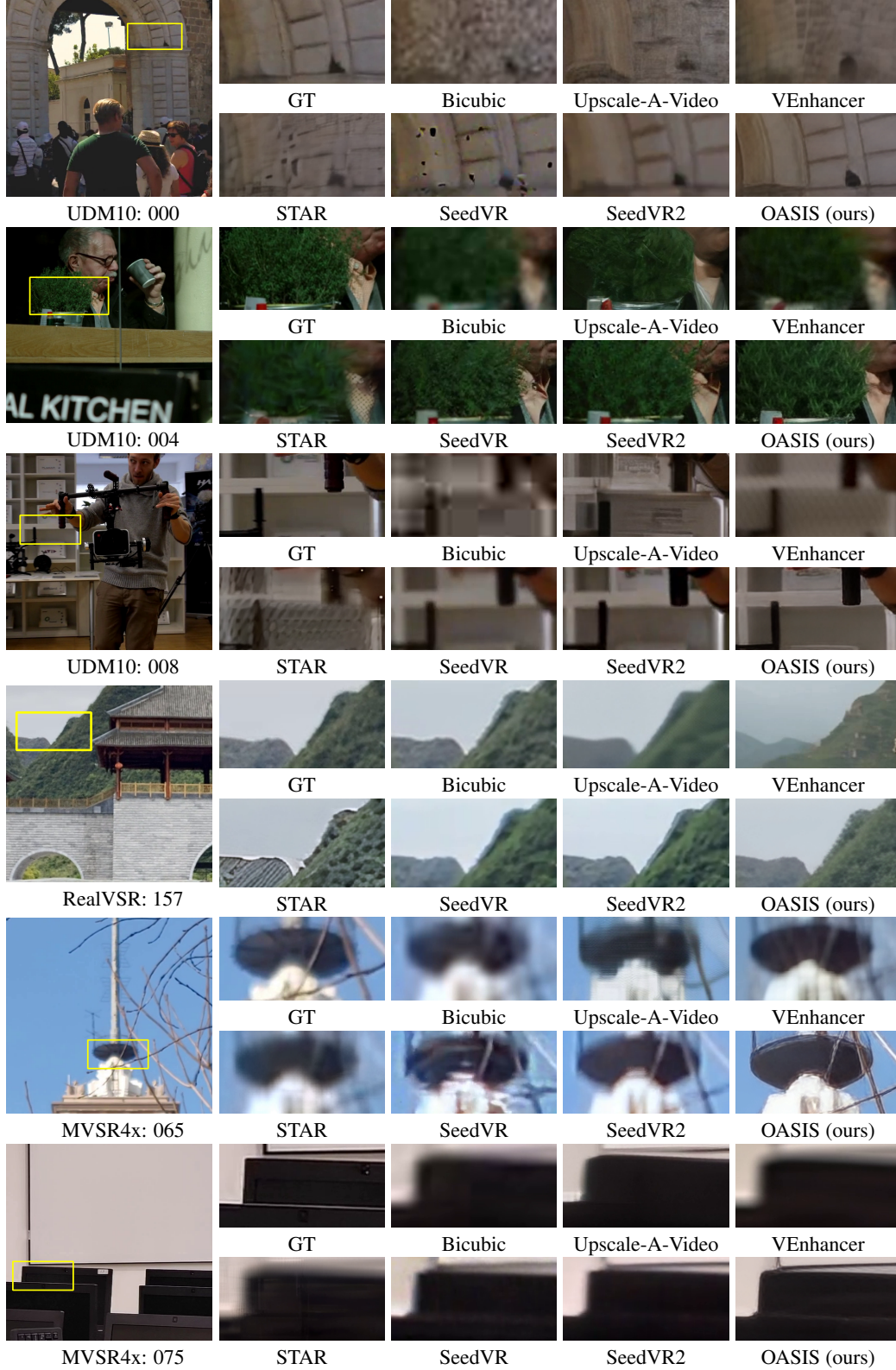


Figure 7: Additional visual comparisons for $\times 4$ VSR. OASIS yields clean and sharp reconstructions while faithfully preserving contours, edges, and fine-scale surface patterns.

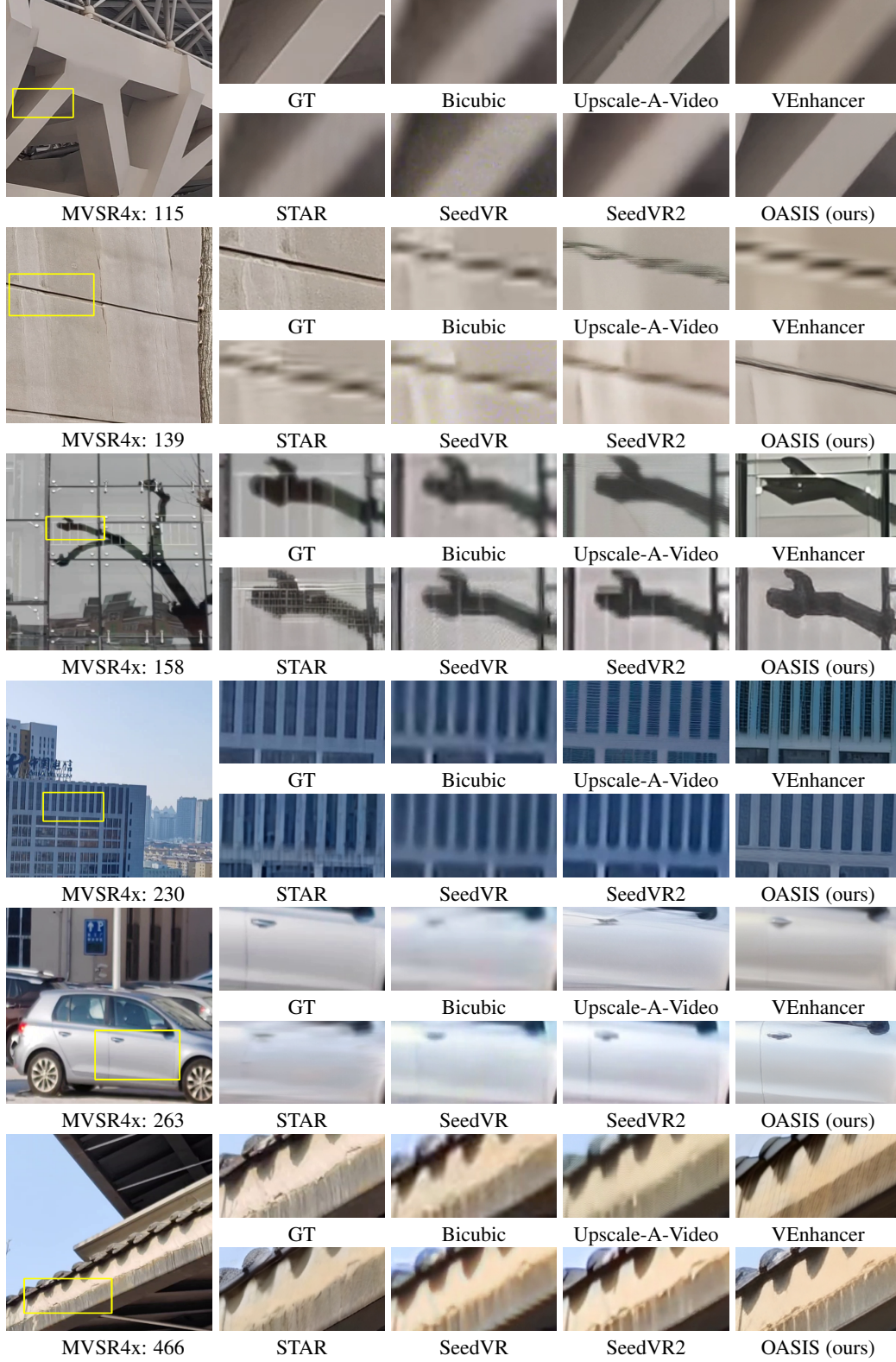


Figure 8: Additional visual comparisons on synthetic and real-world datasets for $\times 4$ VSR. OASIS yields clean reconstructions while preserving contours and fine-scale surface patterns.