Rethinking Diffusion Model-Based Video Super-Resolution: Leveraging Dense Guidance from Aligned Features

Supplementary Material

1. Limitations of Existing Diffusion Model-based VSR

Instead of training from scratch, existing diffusion model (DM)-based VSR methods typically employ pre-trained text-to-image diffusion models [1, 16, 18] as priors. However, the direct application of image-oriented methods introduces flickering artifacts to the VSR results, and the inherent randomness of the diffusion process aggravates this issue. Recent approaches including UAV [26], MGLD-VSR [23] and SATeCo [5] address temporal consistency by processing multiple frames using 3D CNNs or temporal attentions. Although these methods improve temporal stability, they require substantial computational resources, e.g., UAV [26] utilizes 32 NVIDIA A100-80G GPUs and 37K HD videos for training. Following the framework represented by BasicVSR[2], StableVSR [17] makes new attempts to incorporate bidirectional propagation with ControlNet-based [24] guidance during each diffusion step. Although the bidirectional guiding mechanism helps improve temporal consistency and mitigates error accumulation during diffusion process, StableVSR still suffers from the following limitations: (a) Inaccurate alignment in the pixel domain generates spatial artifacts that degrade fidelity, and the frame reconstruction process introduces unnecessary computational redundancy. (b) Limited guidance from U-Net encoder fails to provide sufficient inter-frame information compensation in pixel-to-pixel VSR task, leading to a perceptual-fidelity trade-off dilemma.

2. Comparison of Observation 1 and BasicVSR

The differences between our Observation 1 and BasicVSR are evident in the application architecture, empirical analysis, and conclusions regarding contributory factors, as shown in Table 1.

(1) **Application Architecture.** BasicVSR investigates feature alignment in a non-generative pipeline (non-diffusion-based), where features are directly extracted from a convolutional neural network. In contrast, our work explores in a diffusion-based pipeline, where features are progressively generated and embedded into stochastic noise during the diffusion process. To our knowledge, bservation 1 constitutes the first comprehensive exploration of the interplay between feature and pixel domains within the context of diffusion-based video super-resolution (VSR) pipeline.

(2) **Empirical Analysis.** BasicVSR derives its analysis of feature and image alignment solely from VSR performance, noting a PSNR drop of 0.17 dB due to image alignment. In contrast, our analysis begins at the representation level: we systematically compare intra- and inter-frame correlations between feature and pixel domains across four quantitative metrics, offering direct evidence of their structural differences.

(3) **Conclusions of Contributory Factor.** BasicVSR claims that the reason for the poorer performance of image alignment 'resulting from the inaccuracy of optical flow estimation'. However, since both feature and pixel warping rely on the same optical flow (computed from the input frames), any flow errors would degrade both warped features and warped frames similarly. Therefore, the inaccuracy of optical flow estimation cannot explain why feature-domain alignment consistently outperforms pixel-domain alignment. In contrast, our work provides numerical evidence that the intra- and inter-frame correlations are significantly stronger in the feature domain. This stronger correlation could be a fundamental reason for the superior performance of feature alignment in diffusion-based video super-resolution (VSR).

3. Edge Strength Based on Different Operators

Figure 1 compares edge strength across different alignment processes using three standard edge detectors: Canny, Sobel, and Laplacian. As demonstrated, despite variations in absolute numerical values across operators, all three operators produce consistent rankings in quantifying edge strength. This pattern strongly supports the first three key points presented in bservation 2 of Section 3.2 of the manuscript.

Focusing on the second key point in bservation 2, Figure 1 (b) presents a comprehensive comparison of edge strength degradation across two distinct scenarios: (1) low-resolution features $\{\tilde{\boldsymbol{z}}^{i}_{t\to 0}\}^{N,T}_{i=1,t=1}$ and warped low-resolution features $\{\tilde{\boldsymbol{z}}^{i,warp}_{t\to 0}\}^{N,T}_{i=1,t=1}$, (2) high-resolution features $\{\tilde{\boldsymbol{z}}^{i,s\times \mathrm{Ne}}_{1\to 0}\}^{N,T}_{i=1,t=1}$ and warped high-resolution features $\{\tilde{\boldsymbol{z}}^{i,s\times warp}_{1\to 0}\}^{N,T}_{i=1,t=1}$. Specifically, warping on low-resolution features induces substantial edge strength degradation under all edge detection operators by 12.68% for Canny, 9.23% for Sobel, and 65.08% for Laplacian. In contrast, high-resolution feature warping shows markedly better preservation of edge information, reducing the degradation to just 4.38%, 5.52% and 40.69% for Canny, Sobel and Laplacian operators, respectively.

Regarding the third key point in bservation 2, Figure 1 (c) compares edge strength measurements across three feature representations: original low-

Table 1. Comparion of bservation 1 and BasicVSR.

Method	BasicVSR	Ours
Application Architecture	CNN-based Network	Diffusion-based Network
Empirical Analysis	VSR performance	Data Analysis & VSR performance
Contributory Factors	Stronger optical flow estimation	Stronger intra- and inter-frame correlations

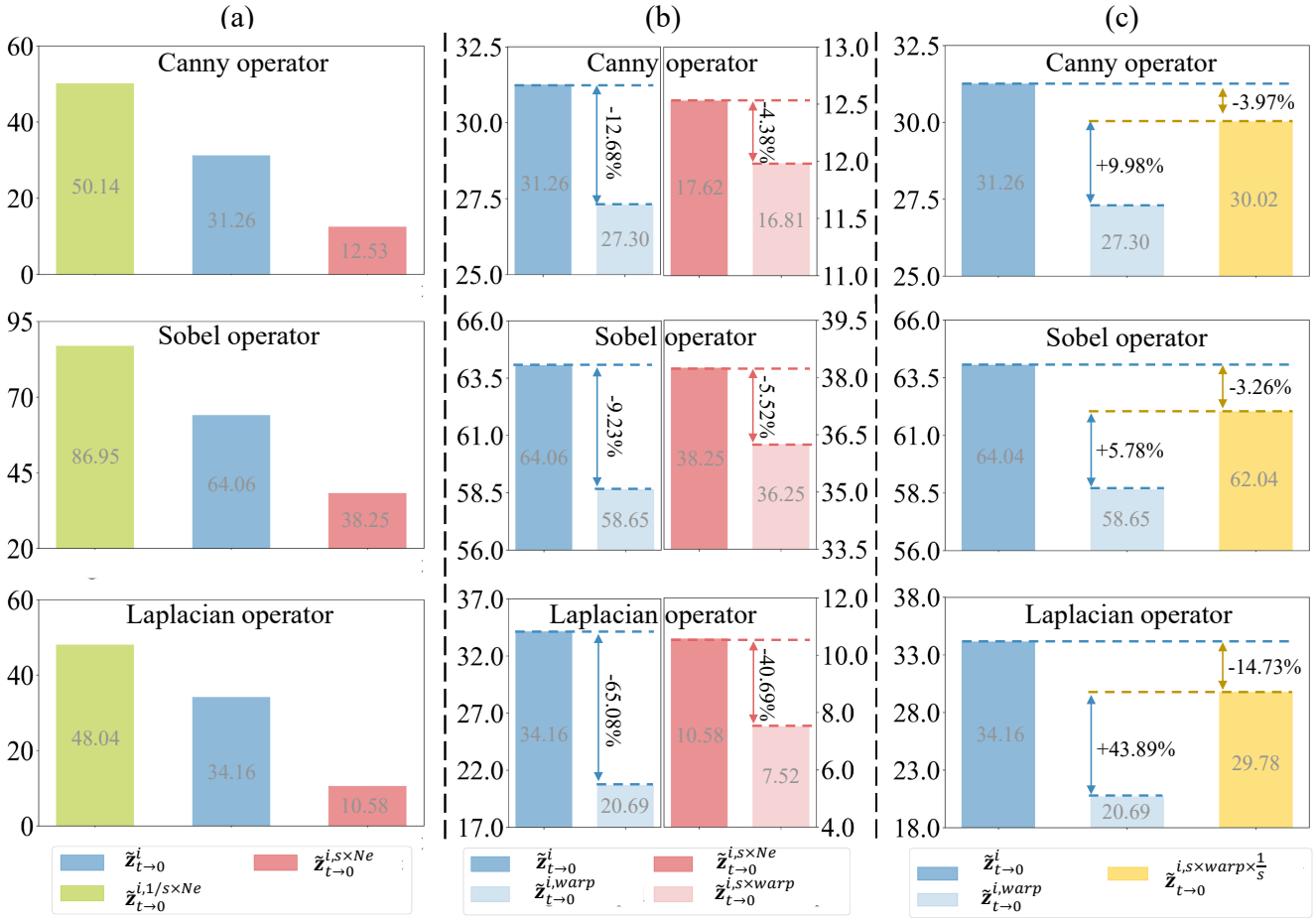


Figure 1. Illustration of edge strength across features aligned through different strategies, using different edge detection operators including the Canny, Sobel, and Laplacian, from top to bottom. (a) the upscaling and downscaling operations, (b) the warping operation on features of different resolutions, and (c) the rescaling-based warping strategy.

resolution features $\{\tilde{\boldsymbol{z}}^i_{t\to 0}\}^{N,T}_{i=1,t=1}$, warped low-resolution features $\{\tilde{\boldsymbol{z}}^{i,warp}_{t\to 0}\}^{N,T}_{i=1,t=1}$, and downscaled features $\{\tilde{\boldsymbol{z}}^{i,s\times warp\times \frac{1}{s}}_{t\to 0}\}^{N,T}_{i=1,t=1}$ after warping on high-resolution features. The results demonstrate that performing the rescaling-based warping strategy significantly enhances edge preservation by 9.98%, 5.78% and 43.89% under Canny, Sobel and Laplacian based edge strength measurement, respectively.

4. High-pass Strength and Edge Strength across Different Diffusion Steps

Figure 2 illustrates the results of the high-pass strength and edge strength averaged on the features $\{\tilde{\boldsymbol{z}}^i_{t\to 0}\}^{N}_{i=1}$, $\{\tilde{\boldsymbol{z}}^{i,warp}_{t\to 0}\}^{N}_{i=1}$, and $\{\tilde{\boldsymbol{z}}^{i,s\times warp\times \frac{1}{s}}_{t\to 0}\}^{N}_{i=1}$ at different diffusion steps t. As shown, at all diffusion steps, the rescaling-based warping strategy (denoted as $s \times warp \times \frac{1}{s}$) preserves more high-frequency information compared to directly warping on the original resolution($warp$). As the reverse process

progresses (i.e., as t decreases), the high-frequency density measured by multiple indicators steadily increases. This implies that high-frequency information is gradually introduced into the low-resolution with the denoising and guiding process, which is in line with the principle of the diffusion model.

5. The Optimal Rescaling Factor of the Rescaling-based Warping Strategy

Figure 3 shows the impact of the rescaling factor in the rescaling-based warping strategy on VSR tasks at different scales. As shown in the top row, the reduction ratio of edge strength during warping decreases as the rescaling factor increases, and this reduction gradually saturates. Moreover, the bottom row reveals that the rescaling factor not only affects the degree of high-frequency loss during warping, but also influences the absolute edge strength value of the noise-free approximation $\{\tilde{\boldsymbol{z}}_{t\to 0}^{i}\}_{i=1}^{N}$ and the aligned feature $\{\tilde{\boldsymbol{z}}_{t\to 0}^{i,s\times warp\times\frac{1}{s},}\}_{i=1}^{N}$. Importantly, this relationship is non-monotonic and exhibits a maximum. We attribute this behavior to the fact that different upscaling and downscaling factors alter feature distributions in distinct ways, thereby affecting high-frequency preservation during the denoising process within DM-based pipeline. Furthermore, the location of this optimal point shifts to higher rescaling factors as the super-resolution scale increases. Therefore, to better preserve high-frequency information during DM-based VSR diffusion steps while avoiding excessive computational overhead, we select the rescaling factor s in the rescaling-based warping strategy as $4, 8$ and 16 for the $4\times, 8\times$ and $16\times$ VSR tasks, respectively.

6. Comparison of DGAF-VSR and MGLD-VSR

MGLD-VSR leverages motion constraints to guide the diffusion process, ensuring temporal consistency. While both methods exploit inter-frame information, our approach is grounded in a quantitative analysis of alignment and compensation between adjacent frames. By enabling effective feature alignment and dense guidance, DGAF-VSR achieves strong temporal consistency while balancing fidelity and perceptual quality without requiring additional temporal modules, unlike MGLD-VSR. Key differences are summarized in Table 2.

(1) **Warping Procedure.** MGLD-VSR performs warping on low-resolution (LR) features, whereas DGAF-VSR operates on upscaled high-resolution (HR) features, better preserving high-frequency content.

(2) **Adjacent Information Utilization.** MGLD-VSR incorporates information from adjacent frames only through a motion-guided loss applied after each denoising step. In contrast, DGAF-VSR feeds aligned features directly into both the denoising and guidance U-Nets, enabling more comprehensive exploitation of temporal context.

(3) **Guiding Source and Strategy.** MGLD-VSR derives guidance exclusively from LR inputs via a U-Net encoder. In contrast, DGAF-VSR receives dense guidance from both LR images and aligned adjacent features through a full U-Net architecture, facilitating richer information integration.

(4) **Bidirectional Strategy.** MGLD-VSR enforces temporal consistency by constraining forward and backward warping errors during sampling. DGAF-VSR alternates temporal guidance between the previous and next frames at each diffusion step, enabling dynamic and adaptive feature propagation.

7. Comparison of FTCM and BrushNet

The FTCM module in DGAF-VSR is fundamentally distinct from BrushNet in several aspects: Different tasks (video super-resolution vs. image inpainting), different motivations (Providing dense temporal compensation vs. preventing mask corruption), different inputs (Aligned latents, LR frames vs. masks, text prompts, encoded images), different pipelines (w/o vs. w/ classifier-free guidance, reducing inference memory and time by $\sim$50%).

8. Experimental Settings

Datasets. We conduct series of experiments on two benchmark synthetic datasets including REDS [14] and Vid4 [12], and a real-world dataset, VideoLQ [4]. For a fair comparison, we adopt the same training set as [11, 21]. The training dataset consists of 296 video sequences from the REDS dataset, with each sequence containing 100 frames and having a standard resolution of 1280×720. As for synthetic testing, we utilize four sequences reserved in the REDS dataset [1], which is referred to as the REDS4 dataset, along with the Vid4 dataset. The low-resolution videos in the REDS and Vid4 datasets both undergo $4\times$ bicubic downscaling. For real-world VSR testing, we adopt the testing dataset, VideoLQ, which consists of 50 low-resolution video sequences.

Implementation details. We construct the DGAF-VSR on the pre-trained Stable Diffusion x4 Upscaler model [16, 18], and use pre-trained RAFT [19] to initialize the optical flow estimator introduced in our OGWM module. We utilize DDPM sampling with $T = 1000$ during training and sampling 50 steps during inference. Implemented in PyTorch, the DGAF-VSR is trained on 4 NVIDIA V100 GPUs for around 24 epochs using the Adam optimizer with default parameters (ρ_1=0.9, ρ_2=0.999, δ=1×10^{-8}). The training patch size of low-resolution frames is set to 64×64 with a 50% probability horizontal flip as data augmentation, and the training batch size is set to 32. The learning rate is fixed to $5e-5$.

[1] Clips 000, 011, 015, 020 of REDS training set.

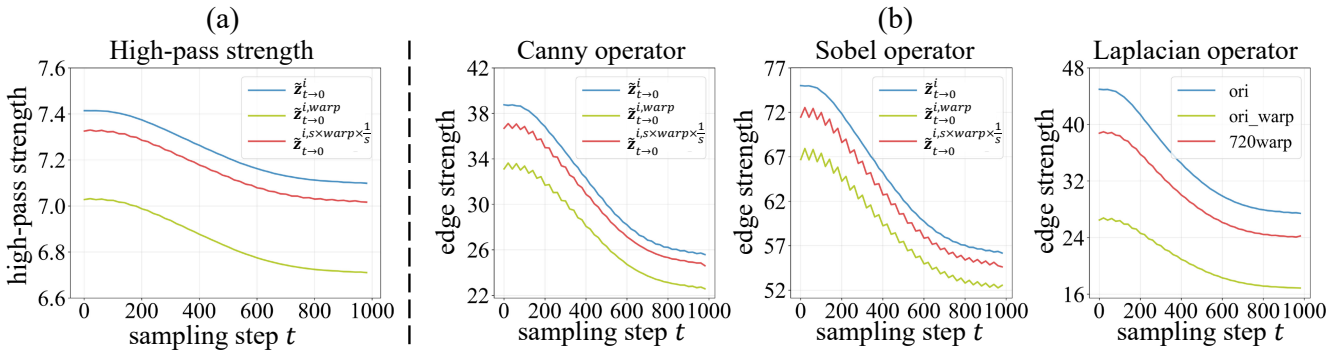


Figure 2. Illustration of high-pass strength and edge strength across different diffusion steps.

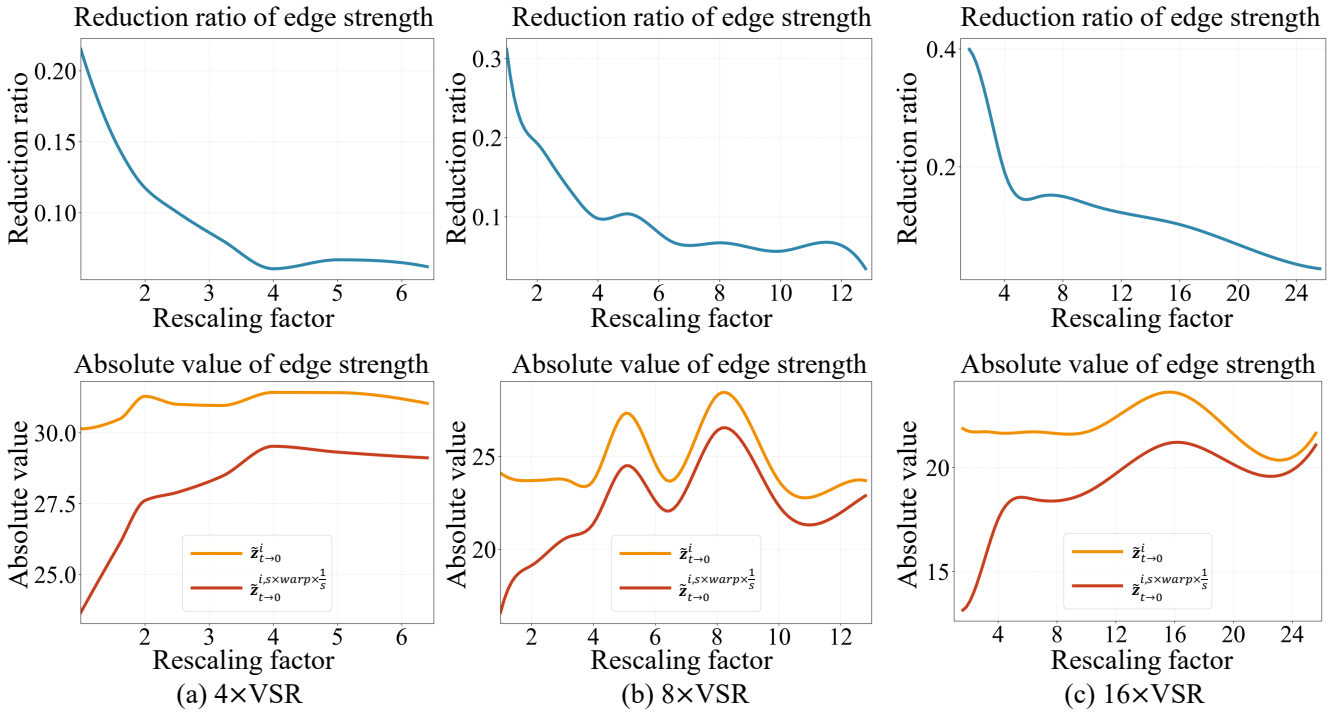


Figure 3. Reduction ratio and absolute edge strength in the rescaling-based warping strategy, as functions of the rescaling factor s. Experiments are conducted on the REDS4 dataset for (a) $4\times$, (b) $8\times$ and (c) $16\times$ VSR task. Reported values are averaged across all diffusion steps and over edge strength computed using three edge detectors: Canny, Sobel, and Laplacian.

Evaluation Metrics. Our assessment employs nine metrics spanning three critical aspects: fidelity, perceptual quality, and temporal consistency. For fidelity evaluation, we utilize standard quantitative measures, including PSNR and SSIM, to assess reconstruction accuracy. The perceptual quality analysis combines two reference-based metrics LPIPS [25] and DISTS [7] with three reference-free metrics MUSIQ [9], CLIP-IQA [20] and NIQE [13] for comprehensive quality assessment. Temporal consistency is evaluated using tLPIPS [6] and tOF [6] across video sequences.

9. Discussion on video-based methods

Unlike video models relying on implicit temporal handling via 3D layers, incorporating optical flow can provide superior motion compensation, as demonstrated in various prior works like BasicVSR++. Our model consumes only 11.94G memory (on 720P) and is trainable on V100 GPUs, whereas CogVideoX2B requires ∼17.8G as reported in its released codes, just for its 24 CausalConv3d layers on 480P. Excessive memory demands force video backbones to rely on tiled inference, which can compromise spatial consistency. Our framework's advantages in temporal and spatial consistency

Table 2. Comparion of DGAF-VSR and MGLD-VSR.

Method	MGLD-VSR	Ours
Warping Procedure	Low-resolution	High-resolution
Adjacent Information	Motion-guide loss	Aligned features
Bidirectional Strategy	Motion-guided sampling	In-pair guidance strategy
Guiding Source	Image	Image & Feature
Dense Guidance	No	Yes
Quantitative Analysis	No	Yes

are validated by the superior metrics shown in the tables in the manuscript.

10. Quantitative Comparisons on Different Scale Factors

To validate the adaptability of DGAF-VSR to different scale factors, we compare the VSR performance of various DM-based methods at $8\times$ and $16\times$ scale factors. The experiments are conducted on the REDS4 dataset. As shown in Table 3, our method improves the MUSIQ value by 8.50 and 9.34 over the second-best algorithm for $8\times$ and $16\times$ VSR, respectively. Besides, our network also enhances the CLIP-IQA metric by 24.73% and 46.45% at $8\times$ and $16\times$ scale factors, respectively. This demonstrates that by utilizing a pre-trained $4\times$ Upscaler model along with sufficient alignment and guidance, our DGAF-VSR achieves excellent results in VSR tasks across various scale factors.

11. Qualitative comparisons on the Vid4 dataset and the VideoLQ dataset

Figure 4 visualizes the $4\times$ super-resolution results generated by different methods on the Vid4 dataset. As illustrated in the figure, our DGAF-VSR achieves superior performance in reconstructing fine texture details and structural clarity, particularly in regions such as the beak and edge of the chick, where our method preserves sharpness and natural texture patterns, while competing approaches exhibit blurred edges[3, 11], as well as in the distant building structure, where our approach maintains distinct architectural features without introducing unrealistic artifacts observed in DM-based methods like StableVSR[17].

Figure 5 visualizes the $4\times$ super-resolution results generated by different methods on the real-world VideoLQ dataset. As illustrated, compared to other DM-based VSR approaches, our DGAF-VSR achieves superior performance in edge reconstruction and the generation of fine-grained, colorful details. This advantage is particularly evident in regions such as the patterns on the carpet and the shape of the letter 'F', where competing methods produce blurred contours and implausible structural artifacts [23, 26].

12. High-frequency preservation of DGAF-VSR

According to observation 2, performing warping operations on upscaled features has the potential to mitigate the high-frequency loss commonly observed during the warping process. Additionally, the subsequent downscaling of these warped features is capable of counteracting the high-frequency loss introduced by the upscaling operation. This dual mechanism serves as the cornerstone of the design of our Optical Guided Warping Module (OGWM). To validate this concept experimentally, we visualize the edge and high-pass components of three feature representations $\tilde{\boldsymbol{z}}^i_{1\to0}$, $\tilde{\boldsymbol{z}}^{i,warp}_{1\to0}$ and $\tilde{\boldsymbol{z}}^{i,s\times warp\times\frac{1}{s}}_{1\to0}$ in Figure. 6. As is evident from the figure, the high-frequency components of $\tilde{\boldsymbol{z}}^{i,warp}_{1\to0}$ exhibit significant detail loss and edge blurring compared to the original feature $\tilde{\boldsymbol{z}}^i_{1\to0}$. In contrast, the high-frequency components of $\tilde{\boldsymbol{z}}^{i,s\times warp\times\frac{1}{s}}_{1\to0}$ are nearly identical to those of $\tilde{\boldsymbol{z}}^i_{1\to0}$. These experimental results provide compelling evidence supporting the effectiveness of the proposed OGWM module, demonstrating that DGAF-VSR significantly improves the retention of high-frequency information during the warping process.

13. Efficiency of DGAF-VSR

To evaluate the efficiency of our network, we compare the peak GPU memory of DGAF-VSR with that of SOTA DM-based VSR methods. The peak GPU memory consumption is recorded on an NVIDIA V100-32G GPU at an output resolution of 720p (720×1280). As shown in Table 4, the peak GPU memory of our DGAF-VSR is 14.96% and 51.11% lower than that of StableVSR and STAR, respectively, indicating the efficiency of our network.

In terms of runtime, we evaluate the performance of different modules using an NVIDIA V100-32G GPU on the REDS4 dataset. As shown in Table 5, the upscaling and downscaling processes contribute only 2.51% and 1.73% to the total runtime of the OGWM, respectively. Notably, for each diffusion step, the average runtime of the OGWM module and the FTCM module are 2.43 e-5 and 0.4505 seconds, respectively. The FM module incurs an overhead of 0.3178

(a) Bicubic (b) BasicVSR++ (c) RVRT (d) StableVSR (e) Ours GT

Figure 4. Visualization of the VSR results from different methods on the Vid4 dataset. (a) Bicubic, (b) BasicVSR++, (c) RVRT, (d) StableVSR, (e) Ours.

(a) Bicubic (b) UAV (c) MGLD-VSR (d) Ours

Figure 5. Visualization of the VSR results from different methods on the VideoLQ dataset. (a) Bicubic, (b) UAV, (c) MGLD-VSR, (d) Ours.

seconds for each frame, accounting for only 1.32% of the total inference runtime.

Furthermore, we provide the number of parameters in Table 6. As shown, the proposed DGAF-VSR contains 983.3 MB of network parameters, with 473 MB being trainable. The number of parameters in DGAF-VSR is significantly smaller than that of MGLD-VSR (1465 MB) and STAR (2139 MB).

14. Ablation study

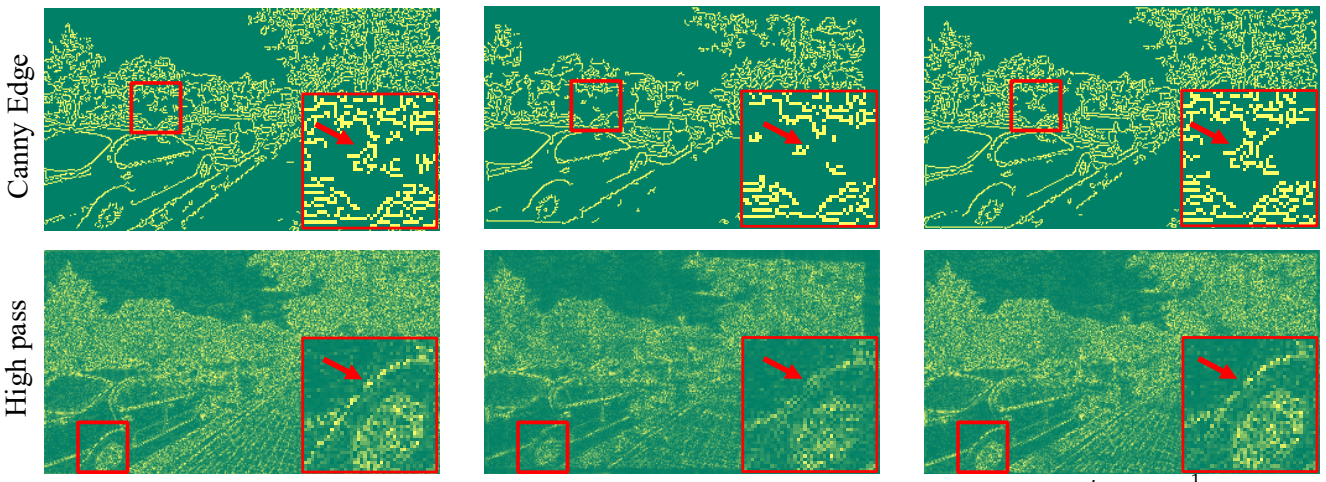
Effect of the in-pair guidance strategy. The purpose of the in-pair forward and backward guidance strategy is to alternately integrate the aligned information from the forward and backward frames with the information from the current frame. Therefore, validating the effectiveness of this strategy is essential. As shown in Table 7, a comparison between Case 1 and Case 5 evaluates the $4\times$ VSR performance with and without the in-pair forward and backward guidance strategy. As shown, incorporating this strategy improves PSNR by 0.29 dB and reduces the tLPIPS value by 20.65%, indicating significant improvements in fidelity, perceptual quality, and temporal consistency. These results substantiate the effectiveness of the in-pair guidance strategy.

Effect of the pre-upscaling strategy. In the flow predicting module (FM), input frames are pre-upscaled to serve as inputs for the RAFT-based optical flow estimator. In this ablation study, we validate the effectiveness of this strategy. Comparing the results of Case 2 and Case 5 in Table 7, we observe that the pre-upscaling strategy leads to a 1.26 dB improvement in PSNR, an 18.80% reduction in LPIPS, and a 14.24% reduction in tOF. This indicates that the pre-upscaling strategy can achieve better VSR performance.

Effect of the optical flow estimator. The proposed DGAF-VSR utilizes a RAFT-based optical flow estimator. To verify the effectiveness of the estimator, we compare the VSR results obtained using different optical flow estimators.

Table 3. Quantitative comparison on the REDS4 dataset for 8× and 16× VSR, in terms of five perceptual(⋄) and two fidelity metrics(⋆) metrics, with the best results in bold and second bests underlined.

Types	Method	LPIPS⋄↓	DISTS⋄↓	MUSIQ⋄↑	CLIP-IQA⋄↑	NIQE⋄↓	PSNR⋆↑	SSIM⋆↑
8×	Bicubic	0.688	0.302	21.80	<u>0.372</u>	9.96	21.55	0.588
	MGLD-VSR[23]	<u>0.249</u>	<u>0.104</u>	<u>59.44</u>	0.271	<u>2.86</u>	<u>22.94</u>	<u>0.594</u>
	STAR[22]	0.642	0.329	17.98	0.202	8.08	22.23	0.559
	DGAF-VSR	**0.224**	**0.103**	**67.94**	**0.464**	**2.72**	**23.35**	**0.601**
16×	Bicubic	0.813	0.403	15.62	<u>0.338</u>	13.17	19.27	<u>0.520</u>
	MGLD-VSR[23]	<u>0.381</u>	<u>0.184</u>	<u>58.28</u>	0.255	<u>2.94</u>	<u>20.49</u>	0.494
	STAR[22]	0.753	0.492	14.12	0.242	10.67	20.26	0.505
	DGAF-VSR	**0.343**	**0.140**	**67.62**	**0.495**	**2.62**	**21.16**	**0.537**



(a) $\tilde{\boldsymbol{z}}^{i}_{1\to0}$ (b) $\tilde{\boldsymbol{z}}^{i,warp}_{1\to0}$ (c) $\tilde{\boldsymbol{z}}^{i,s\times warp\times\frac{1}{s}}_{1\to0}$

Figure 6. Visualization of the edge and high-pass component of (a) the original features $\tilde{\boldsymbol{z}}^{i}_{1\to0}$, (b) the features $\tilde{\boldsymbol{z}}^{i,warp}_{1\to0}$ warped on the original size, and (c) the features $\tilde{\boldsymbol{z}}^{i,s\times warp\times\frac{1}{s}}_{1\to0}$ obtained through rescaling-based warping strategy.

Table 4. Comparion of the peak GPU memory overhead, with the lowest consumption in bold.

Method	MGLD-VSR	StableVSR	STAR	Ours
Memory (GB)	25.61	14.04	24.42	**11.94**

Comparing Case 5 with Case 3 and Case 4 in Table 7, it can be observed that adopting the RAFT-based optical flow estimator yields a PSNR improvement of 0.89 dB and 0.13 dB over the Farneback and SpyNet estimators, respectively. Additionally, the temporal consistency metric tLPIPS also shows improvements of 51.66% and 11.31% compared to Farneback and SpyNet. Therefore, we select RAFT as the optical flow estimator for DGAF-VSR.

Effect of the rescaling factor. Table 8 presents the 4× VSR performance of our proposed DGAF-VSR under different rescaling factors in the rescaling-based warping strategy of OGWM. As shown, DGAF-VSR achieves the best perceptual quality, temporal consistency, and fidelity on the REDS4 dataset when the rescaling factor $s = 4$. This result aligns well with Observation 2 in our manuscript.

15. Limitations and Future Work

Our proposed framework introduces an effective alignment mechanism for high-frequency preservation and a novel compensation approach for enhanced information integration, both of which are potentially beneficial for various low-level vision tasks. DGAF-VSR is built upon the pre-trained Stable Diffusion ×4 Upscaler, a backbone commonly adopted by existing diffusion-based VSR methods [10, 17, 26]. Consequently, while our method achieves state-of-the-art performance on 4×, 8×, and 16× VSR tasks, it shares the same

Table 5. Runtime per frame of different modules in our proposed DGAF-VSR.

	Each diffusion step						FM	VAE decoder	Total(steps=50)
	OGWM				FTCM	Total			
	upscaling	downscaling	Warping	Total					
Runtime(s)	6.11 e-7	4.21 e-7	2.33 e-5	2.43 e-5	0.4504	0.4504	0.3178	1.3137	24.15

Table 6. The number of parammeters of different modules.

	FM	FTCM		VAE	Total	Trainable
		Guiding UNet	Denoising UNet			
Trainable	No	Yes	No	No	-	-
Pamameters(M)	5.3	473	473	32	983.3	473

limitation as prior works: when applied to other low-level video restoration tasks, such as denoising or deblurring, its performance may degrade due to the scale-specific design of the diffusion model. Nevertheless, we argue that by implementing simple, task-oriented modifications like LoRA, any general-purpose T2I diffusion model (e.g., SD-XL [15] and SDv3.5 [8]) can serve as our foundational model. This inherent flexibility implies that the advancements introduced by our DGAF-VSR are readily transferable to diverse video reconstruction tasks, such as video denoising and video deblurring. We leave this promising extension for future research but emphasize its potential significance for advancing general video restoration frameworks.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4947–4956, 2021.

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5972–5981, 2022.

[4] 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.

[5] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9232–9241, 2024.

[6] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. ACM Transactions on Graphics (TOG), 39(4):75–1, 2020.

[7] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence, 44(5):2567–2581, 2020.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first international conference on machine learning, 2024.

[9] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5148–5157, 2021.

[10] Xiaohui Li, Yihao Liu, Shuo Cao, Ziyan Chen, Shaobin Zhuang, Xiangyu Chen, Yinan He, Yi Wang, and Yu Qiao. Diffvsr: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency. arXiv e-prints, pages arXiv–2501, 2025.

[11] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems, 35:378–393, 2022.

[12] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. IEEE transactions on pattern analysis and machine intelligence, 36(2):346–360, 2013.

[13] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. IEEE Signal processing letters, 20(3):209–212, 2012.

[14] 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 Proceedings of the

Table 7. The effects of the in-pair guidance strategy, the pre-upscaling strategy and the choice of optical flow estimator, in terms of two fidelity ($\star$), two perceptual ($\diamond$) and two temporal consistency ($\circ$) metrics, with the best results in bold.

Case	In-pair guidance	Pre/Post Upscaling	Optical flow	PSNR$\star\uparrow$	SSIM$\star\uparrow$	LPIPS$\diamond\downarrow$	DISTS$\diamond\downarrow$	tLPIPS$\circ\downarrow$	tOF$\circ\downarrow$
1		Pre	RAFT	27.88	0.796	0.097	**0.043**	4.94	2.74
2	✓	Post	RAFT	26.91	0.765	0.117	0.051	14.06	3.16
3	✓	Pre	SpyNet	28.04	0.801	0.101	0.045	4.42	2.76
4	✓	Pre	Farneback	27.28	0.774	0.116	0.053	8.11	2.96
5(Ours)	✓	Pre	RAFT	**28.17**	**0.804**	**0.095**	**0.043**	**3.92**	**2.71**

Table 8. Effect of the rescaling factor on the 4$\times$ VSR task over the REDS4 dataset, in terms of two fidelity ($\star$), two perceptual ($\diamond$) and two temporal consistency ($\circ$) metrics, with the best results in bold.

Rescaling factor	PSNR$\star\uparrow$	SSIM$\star\uparrow$	LPIPS$\diamond\downarrow$	DISTS$\diamond\downarrow$	tLPIPS$\circ\downarrow$	tOF$\circ\downarrow$
1.0	27.22	0.768	0.114	0.054	12.03	2.99
1.6	26.89	0.761	0.113	0.051	17.05	3.04
2.0	28.01	0.798	0.098	0.045	4.78	2.82
2.5	27.39	0.780	0.105	0.047	12.0	2.89
3.2	27.50	0.782	0.104	0.047	10.19	2.80
4.0 (Ours)	**28.17**	**0.804**	**0.095**	**0.045**	**3.92**	**2.71**
5.0	28.16	**0.804**	**0.095**	0.046	3.93	2.77
6.4	27.59	0.785	0.103	0.047	9.76	2.84

IEEE/CVF conference on computer vision and pattern recognition workshops, pages 0–0, 2019.

[15] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023.

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[17] Claudio Rota, Marco Buzzelli, and Joost van de Weijer. Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models. In European Conference on Computer Vision, pages 36–53. Springer, 2024.

[18] Stability AI. stabilityai/stable-diffusion-x4-upscaler. `https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler`, 2022.

[19] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 402–419. Springer, 2020.

[20] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In Proceedings of the AAAI conference on artificial intelligence, pages 2555–2563, 2023.

[21] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 0–0, 2019.

[22] 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.

[23] 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, 2024.

[24] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3836–3847, 2023.

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

[26] 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.