One-Step Diffusion for Detail-Rich and Temporally Consistent Video Super-Resolution

Yujing Sun[1,2,⋆], **Lingchen Sun**[1,2,⋆], **Shuaizheng Liu**[1,2],
Rongyuan Wu[1,2], **Zhengqiang Zhang**[1,2], **Lei Zhang**[1,2,†]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute
{yukki.sun, ling-chen.sun, shuaizheng.liu, rong-yuan.wu, zhengqiang.zhang}@connect.polyu.hk,
cslzhang@comp.polyu.edu.hk
[⋆]Equal contribution [†]Corresponding author

Abstract

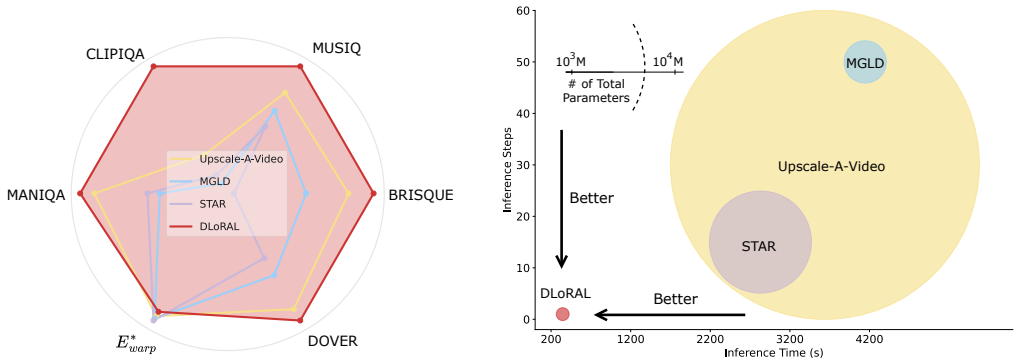
It is a challenging problem to reproduce rich spatial details while maintaining temporal consistency in real-world video super-resolution (Real-VSR), especially when we leverage pre-trained generative models such as stable diffusion (SD) for realistic details synthesis. Existing SD-based Real-VSR methods often compromise spatial details for temporal coherence, resulting in suboptimal visual quality. We argue that the key lies in how to effectively extract the degradation-robust temporal consistency priors from the low-quality (LQ) input video and enhance the video details while maintaining the extracted consistency priors. To achieve this, we propose a Dual LoRA Learning (DLoRAL) paradigm to train an effective SD-based one-step diffusion model, achieving realistic frame details and temporal consistency simultaneously. Specifically, we introduce a Cross-Frame Retrieval (CFR) module to aggregate complementary information across frames, and train a Consistency-LoRA (C-LoRA) to learn robust temporal representations from degraded inputs. After consistency learning, we fix the CFR and C-LoRA modules and train a Detail-LoRA (D-LoRA) to enhance spatial details while aligning with the temporal space defined by C-LoRA to keep temporal coherence. The two phases alternate iteratively for optimization, collaboratively delivering consistent and detail-rich outputs. During inference, the two LoRA branches are merged into the SD model, allowing efficient and high-quality video restoration in a single diffusion step. Experiments show that DLoRAL achieves strong performance in both accuracy and speed. Code and models are available at `https://github.com/yjsunnn/DLoRAL`.

1 Introduction

Video super-resolution (VSR) aims to reconstruct high-quality (HQ) videos from low-quality (LQ) inputs. Traditional VSR methods typically rely on convolutional neural network (CNN)-based [32, 7] and Transformer-based designs [5, 19], trained with pixel-wise L_2 or L_1 losses. While effective in some metrics (*e.g.,* PSNR), these methods often produce over-smoothed results without fine details. To improve perceptual quality, generative adversarial network (GAN)-based VSR methods incorporate the adversarial loss [15] during training to encourage sharper details restoration [3, 21, 4, 41]. However, many VSR models [42, 11, 53] are trained under simplified degradation assumptions (*e.g.*, bicubic downsampling), limiting their performance on real-world LQ videos with complex and unknown degradations. Additionally, GAN-based methods can produce unnatural

[†]This work is supported by the PolyU-OPPO Joint Innovative Research Center.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Quality comparison

(b) Efficiency Comparison

Figure 1: Quality and efficiency comparison among SD-based Real-VSR methods. (a) Quality comparison on the VideoLQ benchmark [8]. (b) Efficiency comparison tested on an A100 GPU (512×512 input with 50 frames for $\times 4$ VSR). DLoRAL achieves the best perceptual quality with only one diffusion step, about $10\times$ faster than Upscale-A-Video [54], MGLD [44], and STAR [40].

artifacts and generalize poorly to diverse video content. Recently, pre-trained diffusion-based text-to-image (T2I) models such as Stable Diffusion (SD) [25, 2] have shown impressive results in real-world image super-resolution (Real-ISR) [31, 38, 48, 26, 37, 27, 34] with realistic textures. One line of research treats the LQ image as a control signal and employs ControlNet-like structures [51] to guide generation [31, 48, 38, 26], and another line of research directly fine-tunes the SD model with LoRA [13] for efficient one-step restoration [37, 27].

The success of SD in Real-ISR inspired exploration of diffusion models for real-world video super-resolution (Real-VSR). Although the powerful generative priors of SD can enhance details, they can introduce inconsistencies among frames when the generated textures sometimes deviate from the content of the LQ inputs [31, 27]. To alleviate this issue, existing SD-based Real-VSR methods typically suppress such fluctuations at the cost of perceptual quality. These methods, such as Upscale-A-Video [54] and MGLD-VSR [44], incorporate temporal modules into pre-trained SD models and adopt frame-wise losses to balance spatial detail and temporal consistency. Despite the significant progress achieved, these methods have two major limitations. First, these approaches optimize detail and consistency jointly in a single model, resulting in suboptimal trade-offs. Improving one objective usually harms the other due to their conflicting nature. Second, the temporal consistency existing in real-world LQ videos is ignored, which can be effectively leveraged to help anchor detail generation on a consistent temporal basis.

To address these issues, we propose a Dual LoRA Learning (DLoRAL) framework for Real-VSR. Our method is built on a one-step residual diffusion model [37, 27], which significantly reduces inference time while maintaining strong generative capability. Inspired by PiSA-SR [27], which learns two LoRA modules to achieve adjustable Real-ISR results, we design two decoupled LoRA branches within the shared diffusion UNet to resolve the conflict between spatial detail and temporal coherence. Specifically, a Consistency-LoRA (C-LoRA) is designed to learn temporal consistency representation, and a Detail-LoRA (D-LoRA) is designed to restore high-frequency spatial details. To exploit the inherent temporal consistency in LQ videos, we introduce a Cross-Frame Retrieval (CFR) module, which extracts structure-aligned temporal features from adjacent degraded frames, helping the model learn degradation-robust representations. CFR not only provides a stable and informative intermediate representation for C-LoRA to build upon, but also serves as the anchor for the subsequent detail enhancement stage to maintain temporal alignment.

Instead of optimizing both objectives jointly, we adopt a dual-stage training strategy. The training begins from the temporal consistency stage, in which we fine-tune C-LoRA and CFR modules using consistency-related losses. In the detail enhancement stage, we freeze C-LoRA and CFR, and train D-LoRA to refine high-frequency details with the additional classifier score distillation (CSD) [27] loss. These two stages are alternatively trained to allow each branch to specialize in its objective. During inference, the two LoRA modules can be integrated in one-step diffusion. As illustrated in Fig. 1, our DLoRAL method achieves both high temporal consistency and superior visual quality,

outperforming previous Real-VSR methods in overall quality, as well as inference speed (about 10× speedup over current methods [54, 44, 40], as illustrated in Fig. 1(b)).

Our main contributions are summarized as follows. (1) We propose a Dual LoRA Learning (DLoRAL) paradigm for Real-VSR, which decouples the learning of temporal consistency and spatial details into two dedicated LoRA modules under a unified one-step diffusion framework. (2) We introduce a Cross-Frame Retrieval (CFR) module to extract degradation-robust temporal priors for Consistency-LoRA (C-LoRA) training, providing structure-aligned intermediate representations that guide the subsequent training of Detail-LoRA (D-LoRA) for high-fidelity restoration. (3) Our DLoRAL model achieves state-of-the-art performance on Real-VSR benchmarks, producing visually realistic frame details and stable temporal consistency.

2 Related Work

Real-World VSR. Conventional VSR methods [32, 11, 16] typically rely on simply synthesized data (*e.g.*, bicubic downsampling), leading to a significant performance gap when applied to real-world videos. Early works [45, 35] addressed this by collecting real-world LQ-HQ video pairs, such as the iPhone-captured dataset [45]. However, these datasets are limited by device bias and scalability. The following works [33, 8] simulated realistic degradations by combining blur, noise, and compression, while others enhanced robustness through architectural design. For instance, RealVSR [45] introduces a domain adaptation mechanism that aligns feature distributions between synthetic and real domains through adversarial learning. RealBasicVSR [8] proposes a degradation modeling framework that refines the restoration process through iterative correction modules. Despite these advances, existing methods still struggle to recover fine details and generalize across diverse real-world scenarios, often producing over-smoothed outputs.

Diffusion Based Real-VSR. Recent advances in diffusion models for image restoration [1, 10, 23, 49, 50] have inspired the extension to Real-VSR tasks [54, 44, 17, 40]. A common approach is to adapt pre-trained T2I models by injecting temporal modules to ensure both perceptual quality and temporal consistency. For example, Upscale-A-Video [54] integrates temporal layers into the pre-trained diffusion model and proposes a flow-guided recurrent latent propagation module. MGLD-VSR [44] guides the diffusion process with a motion-guided loss and inserts a temporal module into the diffusion decoder. The other directions include decomposing the complex learning burden into staged training phases [17] and reformulating attention mechanisms in diffusion transformers [30] to process videos of arbitrary length. Rather than leveraging the pre-trained T2I model, STAR [40] leverages compressed temporal representations from text-to-video (T2V) models [2].Despite these efforts, balancing spatial detail and temporal consistency remains a key challenge. Most existing methods enforce frame-level constraints to improve consistency by sacrificing visual fidelity. In this work, we propose a decoupled learning strategy: first learning degradation-robust temporal priors from LQ inputs then guiding HQ generation with these features. This design ensures both high-quality detail restoration and stable temporal coherence.

Real-VSR Paradigms. Recent VSR methods follow two main paradigms: sliding-window-based [32, 11, 42, 16] and recurrent-based [6, 7, 5, 20, 54, 44, 40]. Sliding-window-based methods reconstruct each output frame using a set of neighboring frames, capturing fine-grained local details and short-term temporal dependencies. In contrast, recurrent-based methods propagate features across frames sequentially, offering higher efficiency, but are prone to error accumulation and detail degradation. Most diffusion-based Real-VSR methods [54, 44, 40] adopt the recurrent design for its inference efficiency. In this work, we build on a sliding-window framework to better preserve spatial and temporal details. To mitigate the computational overhead, we adopt a one-step diffusion strategy that eliminates redundancy while maintaining high reconstruction quality.

3 Methodology

3.1 Preliminary

Diffusion models [25] simulate a forward process where a clean latent code z_0 is gradually noised into z_t using Gaussian noise: $z_t = \sqrt{\bar{\alpha}_t} \cdot z_0 + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t$ following a predefined schedule. During training, a model $\epsilon_\theta(t, z_t)$ is trained to predict the added noise at each timestep t. During inference, z_0 is recovered from pure noise $z_T \sim \mathcal{N}(0, I)$ by iterative denoising. However,

this multi-step process is slow and stochastic, limiting its efficiency and stability in super-resolution (SR) tasks that demand fast and reliable reconstruction. To address this issue, recent works [37, 27] propose one-step diffusion that skips iterative sampling by directly refining an LQ latent into its HQ counterpart. To further improve controllability, PiSA-SR [27] introduces a residual learning formulation that allows the model to focuses on high-frequency corrections:

$$z^{HQ} = z^{LQ} - \epsilon_\theta(z^{LQ}) \tag{1}$$

where z^{LQ} and z^{HQ} represent the latent codes of LQ and HQ respectively.

Most existing VSR methods [54, 44, 40] rely on multi-step diffusion, resulting in high computational cost. In this work, we make the first attempt to apply a one-step diffusion framework to VSR, improving efficiency while preserving restoration quality by adapting the residual learning formulation in Eq. (1) to accelerate convergence. To this end, we introduce VSR-specific modules and learning strategies to produce detail-rich and temporally consistent results.

3.2 Dual LoRA Learning Network for Real-VSR

Motivation. There is a fundamental challenge in Real-VSR: how to balance the preservation of spatial details and the enforcement of temporal consistency. To simultaneously achieve both objectives, we begin by analyzing the characteristics of real-world LQ videos and the limitations of current SD-based VSR methods, which motivate the design of our proposed framework.

- ***Temporal Consistency in Degraded Videos.*** Despite degradations, such as noise, blur, and compression, real-world LQ videos retain stable information across frames, preserving inherent structural and semantic consistency. Leveraging these consistent representations provides a strong foundation for reconstructing HQ videos with realistic details and temporal coherence. To exploit this, we propose a Cross-Frame Retrieval (CFR) module to aggregate complementary information across frames to enhance consistency. In addition, we design a Consistency-LoRA (C-LoRA) to further improve reconstruction by reinforcing temporal alignment and structural integrity. This stage lays the groundwork for more accurate guidance in the subsequent detail enhancement phase.
- ***Conflict in Optimizing Details and Consistency.*** To adapt pre-trained diffusion models for VSR and balance spatial detail and temporal coherence, existing methods [25, 46] typically introduce trainable layers optimized jointly with diffusion and temporal losses. However, detail generation and consistency preservation are inherently conflicting objectives, and joint optimization often results in suboptimal trade-offs. To address this, we propose two decoupled weight spaces: one for temporal consistency modeling and another for detail enhancement. Rather than training two networks, which is costly, we adopt a decoupled scheme inspired by PiSA-SR [27], embedding two specialized LoRA branches into a shared SD UNet. This lightweight design enables alternative refinement, allowing each branch to focus on its objective.

Framework Overview. Building on the above insights, we design a Dual LoRA Learning (DLoRAL) framework to generate HQ video outputs from degraded inputs. Given an LQ sequence of N frames, $\mathbf{I}^{LQ} = \{I_n^{LQ} \mid n = 1, \ldots, N\}$, our model G_θ generates a corresponding HQ sequence $\mathbf{I}^{HQ} = \{I_n^{HQ} \mid n = 1, \ldots, N\}$. To utilize information from neighboring frames, we adopt a sliding-window strategy [32, 11, 16], where each HQ frame I_n^{HQ} is generated from two adjacent LQ frames: the current n-th frame I_n^{LQ} and its preceding frame I_{n-1}^{LQ}. For the first frame I_1^{LQ}, which lacks a previous frame, we adopt a self-replication approach to generate I_1^{LQ}.

As illustrated in Fig. 2, our generator G_θ leverages the pre-trained SD model, which consists of a VAE encoder E_θ, an SD UNet ϵ_θ, and a VAE decoder D_θ. Our DLoRAL framework employs two specialized training stages, *i.e.,* temporal consistency stage and detail enhancement stage. In the temporal consistency stage, the CFR module retrieves inter-frame relevant information from degraded inputs, then the UNet is finetuned by C-LoRA for further reinforcement of temporal alignment. In the detail enhancement stage, D-LoRA is optimized to improve spatial visual quality. These two stages are trained alternately in an iterative manner to progressively refine both temporal consistency and spatial quality, ultimately leading to coherent and detail-preserved video restoration. During inference, the C-LoRA and D-LoRA are merged into the SD UNet to ensure efficient deployment.

Temporal Consistency Stage. This stage is to establish a temporally coherent and robust representation from the LQ video sequence $\mathbf{I}^{LQ}$ before enhancing details. This stage involves two main steps: temporal feature fusion using a CFR module and fine-tuning the SD UNet to improve consistency.

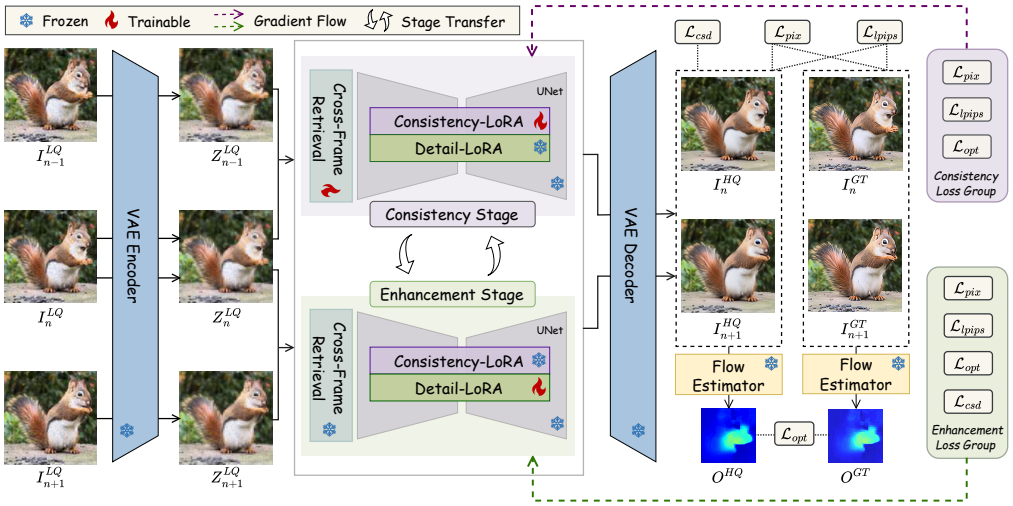


Figure 2: The training pipeline of our proposed DLoRAL. The Cross-Frame Retrieval (CFR) and Consistency-LoRA (C-LoRA) modules are optimized in the consistency stage, while the Detail-LoRA (D-LoRA) is optimized in the enhancement stage. Both stages are alternately trained to ensure temporal coherence and visual quality.

To unlock the inherent consistency among degraded inputs, CFR improves the current latent representation z_n^{LQ} by employing a specialized attention mechanism that integrates complementary information from the previous latent feature z_{n-1}^{LQ}. Specifically, with encoded features z_n^{LQ} and z_{n-1}^{LQ}, CFR first warps them into the same coordinate space with SpyNet [24] following a common frame alignment procedure [32, 44] (denoted as F_{wp}). The current latent features z_n^{LQ} and aligned latent features $F_{wp}(z_{n-1}^{LQ})$ are then projected into query (Q_n), key (K_{n-1}), and value (V_{n-1}) embeddings through 1×1 convolutions (denoted as $\circ$) parameterized by W_Q, W_K, and W_V, as shown below:

$$Q_n = W_Q \circ z_n^{LQ}, \quad K_{n-1} = W_K \circ F_{wp}(z_{n-1}^{LQ}), \quad V_{n-1} = W_V \circ F_{wp}(z_{n-1}^{LQ}). \tag{2}$$

With obtained embeddings extracted from adjacent frames, CFR employs two mechanisms to enhance fusion quality. First, for each query position p, it selectively attends to only the top-k most similar positions (denoted as $F_{topk}[p]$) in the aligned previous frame, avoiding perturbations from uncorrelated noises. Second, for each query position p, a learnable threshold $\tau_n[p]$ is predicted via a lightweight MLP. It dynamically adapts to regional characteristics - enforcing stricter filtering in detail-rich areas while being more permissive in flat regions, ensuring that only confident matches could contribute to the final fusion. The fused feature $\bar{z}_n^{LQ}[p]$ is computed as:

$$\bar{z}_n^{LQ}[p] = z_n^{LQ}[p] + \sum_{q \in F_{topk}[p]} \phi\left(\frac{\langle Q_n[p], K_{n-1}[q] \rangle}{\sqrt{d}} - \tau_n[p]\right) \cdot V_{n-1}[q], \tag{3}$$

where $\phi(\cdot)$ is a non-negative gating function (*e.g.*, ReLU [12]), and d is the channel dimension.

The latent feature $\bar{z}_n^{LQ}$ is then processed by the UNet to generate the HQ latent z_n^{HQ}. In this stage, only the C-LoRA is trainable, while D-LoRA remains frozen. The final HQ frame is reconstructed via the VAE decoder by $I_n^{HQ} = D_\theta(z_n^{HQ})$. All trainable components in this stage, including the CFR module and C-LoRA, are optimized using the consistency loss $\mathcal{L}_{\text{cons}}$, which is designed to ensure both the quality of individual frames and the temporal consistency across the sequence. It combines the pixel-level loss ($\mathcal{L}_{\text{pix}}$), LPIPS loss ($\mathcal{L}_{\text{lpips}}$), and optical flow loss ($\mathcal{L}_{\text{opt}}$), as shown below:

$$\begin{aligned} \mathcal{L}_{\text{cons}} &= \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{opt}}\mathcal{L}_{\text{opt}}, \\ \mathcal{L}_{\text{opt}} &= \left\| O_n^{HQ} - O_n^{\text{GT}} \right\|_1 = \left\| F(I_n^{HQ}, I_{n+1}^{HQ}) - F(I_n^{\text{GT}}, I_{n+1}^{\text{GT}}) \right\|_1. \end{aligned} \tag{4}$$

Here, the ℓ_2 loss is adopted as the $\mathcal{L}_{\text{pix}}$, and $\mathcal{L}_{\text{opt}}$ measures the L_1 distance between optical flow maps estimated from generated and ground-truth frame pairs, promoting motion alignment and smooth

transitions. The loss weights λ_{pix}, λ_{lpips}, and λ_{opt} are empirically set to balance spatial accuracy, perceptual quality, and temporal consistency.

Detail Enhancement Stage. Different from the temporal consistency stage, which yields aligned and coherent latent representations, the detail enhancement stage focuses on restoring high-frequency visual details. In this stage, adjacent latent features z_{n-1}^{LQ} and z_n^{LQ} are processed by the frozen CFR module to reapply the learned alignment and fusion, thus the temporal consistency learned in the consistency stage is maintained without introducing new variations.

The resulting temporally enriched latent $\bar{z}_n^{LQ}$ is then fed into the diffusion UNet ϵ_θ. We employ a decoupled finetuning strategy: only the D-LoRA parameters, responsible for detail synthesis, are trainable, while the C-LoRA parameters, associated with consistency, remain frozen. This setting allows the D-LoRA to focus solely on detail synthesis without compromising the temporal structure previously established. The output HQ latent z_n^{HQ} is then decoded using the frozen decoder D_θ to produce the final super-resolved frame I_n^{HQ}.

To guide this detail enhancement while preserving the structure learned previously, the loss function $\mathcal{L}_{\text{enh}}$ combines several components as follows:

$$\mathcal{L}_{\text{enh}} = \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{opt}}\mathcal{L}_{\text{opt}} + \lambda_{\text{csd}}\mathcal{L}_{\text{csd}}. \tag{5}$$

We retain $\mathcal{L}_{\text{pix}}$, $\mathcal{L}_{\text{lpips}}$, and $\mathcal{L}_{\text{opt}}$ used in the consistency stage (as Eq. (4)), serving as anchors to maintain spatial fidelity and motion coherence. Furthermore, we introduce the Classifier Score Distillation (CSD) loss [27], $\mathcal{L}_{\text{csd}}$, which encourages the generation of richer and finer details.

3.3 Training and Inference

Dynamic Dual-Stage Training. We adopt a dynamic dual-stage training scheme. The training begins with the consistency stage, aiming at learning degradation-robust features and establishing strong temporal coherence among frames. In this stage, only the CFR and C-LoRA modules are trainable, while the D-LoRA is fixed. Once the model converges in the consistency stage, the training switches to refine high-frequency spatial details, guided by $\mathcal{L}_{\text{enh}}$ with the additional CSD loss. In this stage, only the D-LoRA parameters are trainable, while the CFR module and C-LoRA are fixed. Such an alternative training is iterated, allowing the model to dynamically converge toward a solution that balances temporal coherence and visual fidelity.

Smooth Transition Between Training Stages. Compared to the consistency stage, the enhancement stage introduces an additional loss function $\mathcal{L}_{\text{csd}}$ for enriching semantic details. Directly switching between the full loss functions $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{enh}}$ can lead to instability due to the abrupt change in learning targets. To prevent this, we employ a re-weighting strategy that progressively shifts the loss objective, ensuring a smooth transition between stages. Taking the transition from the consistency stage to the enhancement stage as an example, after the consistency stage, the two loss functions are interpolated as the optimization objective for a warm-up phase of s_t steps, as shown below:

$$\mathcal{L}(s) = (1 - \frac{s}{s_t}) \cdot \mathcal{L}_{\text{cons}} + \frac{s}{s_t} \cdot \mathcal{L}_{\text{enh}}, \quad s \in [0, s_t], \tag{6}$$

where s denotes the current step within the transition. Symmetric interpolation is applied when we switch back from the enhancement stage to the consistency stage.

Inference Phase. At test time, both C-LoRA and D-LoRA are activated and merged into the frozen diffusion UNet. A single diffusion step is used to enhance the LQ input to HQ video frames.

4 Experiment

4.1 Experimental Settings

Implementation Details. We adopt the pre-trained Stable Diffusion V2.1 as the backbone of denoising U-Net. Training is carried out with a batch size of 16, a sequence length of 3, and a video resolution of 512×512. All models are trained using the PyTorch framework on 4 NVIDIA A100 GPUs. We use Adam optimizer with an initial learning rate of 5×10^{-5}. For inference, both C-LoRA and D-LoRA are activated simultaneously in a frozen UNet. Videos are processed in sliding sequences to fit GPU memory limits.

Training Datasets. To support the decoupled training design of our DLoRAL framework, we construct two training datasets for the consistency and enhancement stages, respectively.

For the ***consistency stage***, the training data needs to contain realistic motion while maintaining reasonable image quality. To this end, we select 44,162 high-quality frames from the REDS dataset [22], which offers professionally captured sequences with rich dynamics, and a curated set of videos [39] from Pexels[1], chosen based on aesthetic and temporal smoothness criteria. These sequences provide necessary temporal priors for learning degradation-robust representations.

For the ***enhancement stage***, the training data should prioritize visual quality. Thus, we select the LSDIR [18] dataset, known for its rich textures and more fine-grained details than existing public video datasets. To preserve the learned consistency modeling capability and enable the optical flow regularization among frames, we generate simulated video sequences based on LSDIR. Specifically, for each ground-truth image in LSDIR, we apply random pixel-level translations to it to generate multiple shifted images. The resulting pseudo-video sequences inherently support consistency constraints through synthetic motion, while surpassing real video datasets in visual quality.

The data in both stages are degraded using the RealESRGAN [33] degradation pipeline. We apply identical degradation parameters across frames within the same video, while using random parameters for different video sequences.

Testing Datasets. We evaluate our method on both synthetic and real-world datasets, including UDM10 [47], SPMCS [28], RealVSR [45], and VideoLQ [8]. Among them, UDM10 contains 10 sequences, each having 32 frames. SPMCS contains 30 sequences, each having 31 frames. RealVSR contains 50 real-world sequences, each having 50 frames. VideoLQ contains 50 real-world sequences with complex degradations. For the synthetic dataset (UDM10 and SPMCS), we synthesize LQ-HQ pairs following the same degradation pipeline in training. For real-world datasets (RealVSR and VideoLQ), we directly adopt the given LQ-HQ pairs.

Evaluation Metrics. A set of full-reference and no-reference metrics are selected to evaluate different real-world VSR methods. The full-reference metrics include PSNR and SSIM, and perceptual quality with LPIPS [52] and DISTS [9]. No-reference quality assessment involves MUSIQ [14], MANIQA [43], CLIPIQA [29], and the video quality assessment metric DOVER [36]. Compared to Real-ISR, Real-VSR places greater emphasis on temporal consistency. Following prior works [44, 53], we use the average warping error E^*_{warp} to quantitatively assess temporal consistency: $E^*_{warp} = \frac{1}{N-1}\sum_{i=1}^{N-1}||I^{HQ}_{i+1} - F_{wp}(I^{HQ}_i)||_1$. For the test datasets with GT, optical flow in F_{wp} is estimated from GT frames. For real-world datasets without GT (*e.g.*, VideoLQ test set), we use the flow estimated from predicted frames.

4.2 Experimental Results

To demonstrate the effectiveness of our DLoRAL algorithm, we compare it with seven represnetative and state-of-the-art methods, including three Real-ISR models (RealESRGAN [33], StableSR [31], and the one-step model OSEDiff [37]), a discriminative VSR model (RealBasicVSR [8]), and three diffusion-based VSR models (Upscale-A-Video [54], MGLD-VSR [44] and STAR [40]).

Quantitative Comparison. We show the quantitative comparison on both synthetic and real-world video benchmarks (where real-world testing videos were centrally cropped to 128×128 resolution) in Tab 1, from which several key observations can be made. First, non-diffusion-based methods (*e.g.*, RealESRGAN and RealBasicVSR) perform worse than diffusion-based methods on no-reference perceptual quality metrics, such as MUSIQ and CLIPIQA, mainly because they lack the strong image priors provided by pre-trained SD models, leading to over-smoothed results. Second, SD-based Real-ISR methods (StableSR and OSEDiff) can achieve comparable or even better perceptual quality scores than existing Real-VSR methods. In particular, OSEDiff achieves the best DOVER scores on both the UDM10 and SPMCS datasets. However, its warping error evaluated by E^*_{warp} is worse. This is because the Real-ISR methods generate details for each frame without considering the inter-frame consistency. Finally, compared to the existing Real-VSR methods, our DLoRAL consistently ranks first or second across a range of perceptual quality metrics, including LPIPS, DISTS, MUSIQ, CLIPIQA, MANIQA, and DOVER, demonstrating its strong alignment with human perception. At the same time, DLoRAL does not compromise temporal consistency, as evidenced

[1] `https://www.pexels.com/`

Datasets	Metrics	Real-ISR Methods			Real-VSR Methods				
		RealESRGSN	StableSR	OSEDiff	RealBasicVSR	Upscale-A-Video	MGLD	STAR	**DLoRAL**
UDM10	PSNR ↑	21.345	22.042	23.761	**24.334**	22.364	24.192	**24.451**	23.975
	SSIM ↑	0.565	0.568	0.696	**0.723**	0.584	0.685	**0.714**	0.710
	LPIPS ↓	0.451	0.455	0.367	0.363	0.410	**0.335**	0.417	**0.327**
	DISTS ↓	**0.175**	0.185	**0.175**	0.204	0.198	0.176	0.230	0.179
	BRISQUE ↓	29.843	26.310	20.718	**14.129**	17.607	22.701	36.910	**16.250**
	MUSIQ ↑	49.838	47.805	**63.146**	62.360	61.046	61.309	40.789	**65.620**
	CLIPIQA ↑	0.474	0.445	**0.574**	0.474	0.445	0.453	0.267	**0.652**
	MANIQA ↑	0.330	0.319	**0.334**	0.330	0.318	0.291	0.244	**0.373**
	E^*_{warp} ↓	7.580	8.440	5.220	4.670	5.790	**4.610**	**3.510**	4.720
	DOVER ↑	36.860	30.470	**48.404**	37.572	37.694	40.045	30.384	**42.871**
SPMCS	PSNR ↑	**21.660**	19.260	20.650	**21.580**	19.030	21.260	20.730	21.240
	SSIM ↑	0.569	**0.585**	**0.696**	0.545	0.386	0.515	0.489	0.524
	LPIPS ↓	0.444	0.432	**0.354**	0.404	0.485	0.384	0.606	**0.375**
	DISTS ↓	0.246	0.235	**0.229**	0.237	0.274	0.234	0.342	**0.222**
	BRISQUE ↓	25.240	26.310	19.471	12.048	**19.784**	23.184	27.902	**11.030**
	MUSIQ ↑	53.221	47.805	64.619	66.683	**66.912**	65.079	33.247	**67.390**
	CLIPIQA ↑	0.515	0.445	**0.526**	0.515	0.517	0.437	0.240	**0.581**
	MANIQA ↑	0.308	0.319	0.308	0.308	**0.443**	0.312	0.237	**0.340**
	E^*_{warp} ↓	7.570	8.430	7.500	5.400	7.570	**4.410**	**4.080**	6.250
	DOVER ↑	32.151	30.470	**40.160**	30.953	32.151	31.118	17.220	**34.895**
RealVSR	PSNR ↑	**21.340**	18.950	19.920	**22.270**	20.060	21.120	15.080	20.360
	SSIM ↑	0.565	0.583	0.588	**0.720**	0.591	**0.646**	0.433	0.606
	LPIPS ↓	0.451	0.225	0.282	**0.193**	0.263	**0.219**	0.409	0.242
	DISTS ↓	0.175	0.154	0.164	0.160	0.158	**0.151**	0.279	**0.150**
	BRISQUE ↓	29.843	28.250	31.794	30.362	**25.476**	39.082	62.750	**27.893**
	MUSIQ ↑	49.838	69.962	64.101	**71.413**	67.714	70.734	67.947	**70.908**
	CLIPIQA ↑	0.474	**0.612**	0.546	0.370	0.436	0.530	0.532	**0.617**
	MANIQA ↑	0.330	0.345	0.341	0.384	**0.414**	**0.496**	0.438	0.386
	E^*_{warp} ↓	17.580	25.010	18.300	18.720	**18.200**	19.210	24.600	**17.300**
	DOVER ↑	36.860	**46.846**	42.138	46.439	36.136	42.044	30.214	**49.646**
VideoLQ	BRISQUE ↓	29.605	**22.337**	26.403	24.790	25.101	29.606	42.582	**23.039**
	MUSIQ ↑	53.138	52.975	58.959	**59.475**	57.489	53.092	49.305	**63.846**
	CLIPIQA ↑	0.334	0.478	**0.499**	0.393	0.377	0.315	0.333	**0.567**
	MANIQA ↑	0.232	0.278	0.254	0.312	**0.328**	0.254	0.268	**0.344**
	E^*_{warp} ↓	7.580	8.430	8.406	8.108	7.586	**7.409**	**7.280**	7.897
	DOVER ↑	28.400	30.470	**37.580**	34.772	36.860	31.899	29.400	**38.505**

Table 1: Comparison of various Real-ISR and Real-VSR methods across different datasets. The best and second best results of each metric are highlighted in **red** and **blue**, respectively.

by its superior performance on the E^*_{warp} metric. For example, on the RealVSR dataset, DLoRAL achieves state-of-the-art results in DISTS, CLIPIQA, and DOVER, while ranking among the top in E^*_{warp}, highlighting its ability to produce visually pleasing and temporally coherent outputs.

It should be mentioned that although E^*_{warp} is widely used to assess temporal consistency, it does not correlate well with human perception. For example, blurry Real-VSR outputs can achieve lower warping errors but exhibit poorer visual quality. DLoRAL may report slightly larger E^*_{warp} values than some methods (*e.g.*, STAR), but this is because DLoRAL better preserves fine details that the warping error metric tends to penalize.

Qualitative Comparison. To further demonstrate the effectiveness of DLoRAL, we visualize the Real-VSR results in Fig. 3. One can see that DLoRAL can remove complex spatial-variant degradations and generate realistic details, significantly outperforming other Real-VSR models. Specifically, for the severely degraded facial region (first row), RealBasicVSR fails to reconstruct the facial structural, and Upscale-A-Video and STAR lose facial details. MGLD produces sharper outputs, but suffers from severe structural distortions, particularly around the eye regions. In contrast, DLoRAL successfully recovers fine facial features while maintaining structural integrity. The second row highlights the performance in texture reconstruction, where our method restores sharper and more legible texture patterns compared to the blurry or distorted outputs from other algorithms.

To better compare the consistency, we plot the temporal profiles of the VSR results produced by competing methods Fig. 4. Real-ISR approaches such as StableSR and OSEDiff restore sharper details but suffer from severe temporal instability, as shown by the erratic fluctuations in their profiles,

resulting in unpleasant flickering that harms the video quality. On the other hand, while existing Real-VSR methods can offer better temporal consistency than Real-ISR methods, this comes at the cost of blurred details (see the results of Upscale-A-Video, MGLD and STAR in the left case of Fig. 4) or intra-frame artifacts (see the results of RealBasicVSR, Upscale-A-Video and STAR in the right case of Fig. 4). In comparison, our DLoRAL produces smooth and stable transitions across frames, as reflected by its consistent temporal profiles. This qualitative evidence aligns with our quantitative results, demonstrating DLoRAL's ability to preserve fine visual details while ensuring natural temporal consistency. More visual comparisons can be found in the **Appendix**.

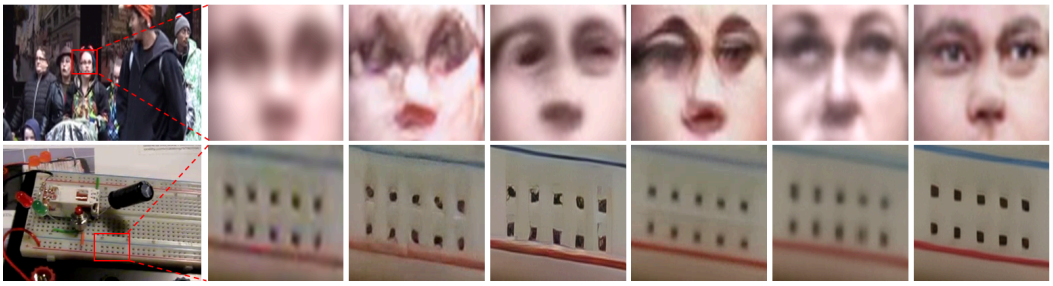


Figure 3: Qualitative comparison of VSR models on real-world VideoLQ dataset.

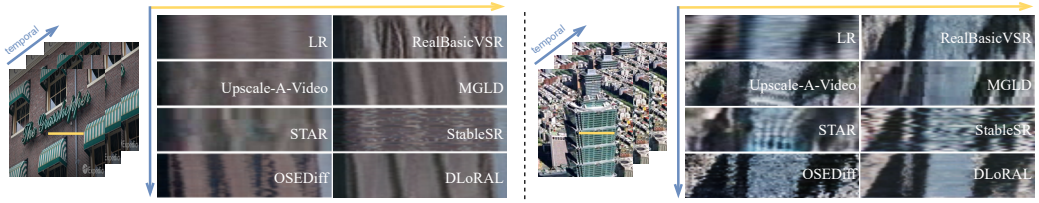


Figure 4: Temporal profiles of competing Real-ISR and Real-VSR methods.

	Real-ISR Methods		Real-VSR Methods			
	StableSR	OSEDiff	Upscale-A-Video	MGLD	STAR	**DLoRAL**
Inference Step	200	1	30	50	15	1
Inference Time (s/50 frames)	32800	340	3640	4146	2830	346
# Total Param (M)	1150	1294	14442	1430	2492	1300

Table 2: Complexity comparison among different methods. All methods are evaluated using 50 512×512 frames for the $\times 4$ VSR task. Inference time is measured on an A100 GPU and includes the entire pipeline: data loading, processing, and result storage.

Complexity Comparison. We compare the inference steps, model size, and inference time of competing diffusion-based models in Tab. 2. The inference time of the whole pipeline (including data loading, data processing, and result storage) is reported, which is measured on the $\times 4$ VSR task with 50 frames of 512×512 LQ images on a single NVIDIA A100 80G GPU. Compared with Real-ISR methods, DLoRAL (346s) achieves a strong balance between quality and complexity, delivering superior visual quality and temporal consistency while maintaining a similar speed to OSEDiff (340s). Among the Real-VSR methods, DLoRAL achieves the fastest inference time and the lowest parameter count, benefiting from its efficient one-step design. Specifically, DLoRAL is more **10**$\times$ **faster** than Upscale-A-Video, MGLD, and **8**$\times$ **faster** than STAR, while maintaining superior visual quality.

Ablation Study. To validate the effectiveness of the proposed components in our model, we conduct ablation studies by selectively removing each of the three key modules: (i) CFR, (ii) C-LoRA, and (iii) D-LoRA, while keeping all other settings identical. For this analysis, we adopt VideoLQ4[2], a subset of four representative sequences with diverse scenes and motions from the VideoLQ dataset. As summarized in Tab. 3, removing either CFR or C-LoRA leads to weaker temporal consistency (*i.e.*, higher warping error), indicating their complementary roles in maintaining temporal coherence. In contrast, removing D-LoRA significantly impairs all perceptual metrics, confirming its core contribution to fine-grained detail enhancement. Further ablations are provided in the **Appendix**.

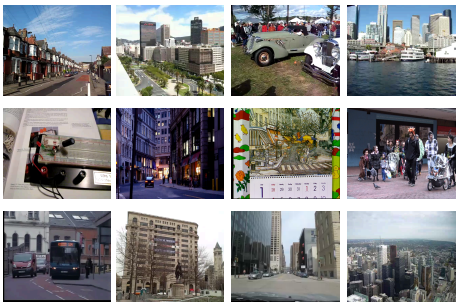
[2]Specifically, VideoLQ4 contains the 013, 015, 020, and 041 clips, each consisting of 100 frames.

	MUSIQ ↑	CLIP-IQA ↑	MANIQA ↑	E^*_{warp} ↓
Ours (Full)	**66.6174**	0.5475	**0.3791**	1.51×10^{-3}
W/o CFR	64.5732	0.5148	0.3386	1.58×10^{-3}
W/o C-LoRA	64.2623	**0.5492**	0.3520	1.61×10^{-3}
W/o D-LoRA	54.0769	0.3654	0.2471	$\mathbf{1.48\times10^{-3}}$

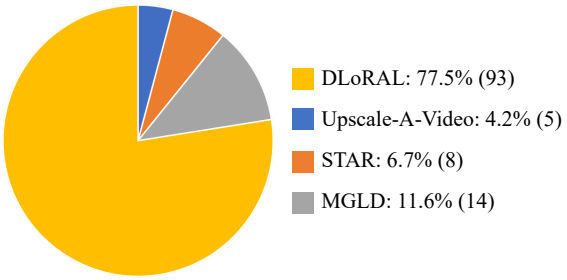
Table 3: Ablation study on key modules on VideoLQ4 dataset.

User Study. We also conduct a user study to further examine the effectiveness of DLoRAL in comparison with existing RealVSR methods. We invited ten volunteers to participate in a user study. Our DLoRAL method was compared with the other three diffusion-based Real-VSR methods: Upscale-A-Video [54], MGLD [44] and STAR [40]. We randomly selected 12 real-world LQ videos with complex degradations and motions from the VideoLQ dataset [8], whose scenes are shown in Fig. 5(a). Each LQ video and its corresponding HQ videos generated by the competing Real-VSR methods were presented to the participants who were asked to select the best HQ result by considering two equally weighted factors: the perceptual quality and temporal consistency of the video.

The results of the user study are shown in Fig. 5(b). DLoRAL received **93** votes, significantly outperforming the other methods, with MGLD, STAR, and Upscale-A-Video receiving 14, 8, and 5 votes, respectively. This overwhelming preference for DLoRAL highlights its effectiveness in addressing the challenges of real-world video restoration. Note that the selected videos include a variety of motion scenarios. In scenarios with complex motion, DLoRAL is able to achieve superior visual quality while maintaining temporal consistency comparable to other methods. In relatively static scenes, DLoRAL demonstrates stable temporal consistency along with equally sharp and clear visual quality.



(a) LQ videos used in user study



(b) Voting results from 10 volunteers

Figure 5: LQ videos used in our user study and the voting results.

5 Conclusion

We proposed DLoRAL to achieve temporally consistent and detail-rich Real-VSR results. To effectively extract degradation-robust temporal priors from low-quality input videos while enhancing details without compromising these priors, we first developed a CFR module and a consistency-LoRA to generate robust temporal representations, and then developed a detail-LoRA to enhance spatial details. We optimized these two objectives alternatively and iteratively, where the results of the previous stage served as an anchor to provide priors for the next stage. The resulting DLoRAL model demonstrated significantly superior performance to previous Real-VSR methods, achieving rich spatial details without compromising the temporal coherence.

Limitations. Despite its strong performance, DLoRAL still has certain limitations. First, since it inherits the $8\times$ downsampling VAE from SD, DLoRAL faces difficulties in restoring very fine-scale details such as small texts. Second, this heavy compression of VAE may disrupt temporal coherence, making it harder to extract robust consistency priors. A VAE specifically designed for Real-VSR tasks could help to address these issues. We leave this challenge for future investigation.

References

[1] Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024.

[2] 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.

[3] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2018.

[4] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pages 185–200, 2018.

[5] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. *arXiv preprint arXiv:2411.18263*, 2024.

[11] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019.

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[14] 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.

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[16] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 335–351. Springer, 2020.

[17] 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 preprint arXiv:2501.10110*, 2025.

[18] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023.

[19] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5687–5696, 2022.

[20] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5687–5696, 2022.

[21] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020.

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

[23] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *European Conference on Computer Vision*, pages 285–303. Springer, 2024.

[24] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[25] 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.

[26] Lingchen Sun, Rongyuan Wu, Jie Liang, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability and efficiency of diffusion models for content consistent super-resolution. *arXiv preprint arXiv:2401.00877*, 2023.

[27] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. *arXiv preprint arXiv:2412.03017*, 2024.

[28] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 4472–4480, 2017.

[29] 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*, volume 37, pages 2555–2563, 2023.

[30] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Fei Xiao, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. *arXiv preprint arXiv:2501.01320*, 2025.

[31] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

[32] 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.

[33] 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.

[34] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25796–25805, 2024.

[35] Pengxu Wei, Yujing Sun, Xingbei Guo, Chang Liu, Guanbin Li, Jie Chen, Xiangyang Ji, and Liang Lin. Towards real-world burst image super-resolution: Benchmark and method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13233–13242, 2023.

[36] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023.

[37] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024.

[38] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024.

[39] Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. *arXiv preprint arXiv:2503.20287*, 2025.

[40] 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.

[41] Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. *arXiv preprint arXiv:2404.12388*, 2024.

[42] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019.

[43] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022.

[44] 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.

[45] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4781–4790, 2021.

[46] 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. *arXiv preprint arXiv:2408.06072*, 2024.

[47] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3106–3115, 2019.

[48] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024.

[49] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024.

[50] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. *arXiv preprint arXiv:2412.09013*, 2024.

[51] 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.

[52] 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.

[53] Zhengqiang Zhang, Ruihuang Li, Shi Guo, Yang Cao, and Lei Zhang. Tmp: Temporal motion propagation for online video super-resolution. *IEEE Transactions on Image Processing*, 2024.

[54] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in the final section (after the conclusion).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Yes, the paper provides a full set of assumptions and a proof for each theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper fully discloses all necessary information to reproduce the main experimental results, supporting the main claims and conclusions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All data are publicly available. We will release the codes and new data if the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper provides all necessary details, including hyperparameters, evaluation metrics and *etc.* to fully understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the type of compute resources in experimental settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research in the paper fully conforms to the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we analyze the potential social impact in appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we clearly indicate the baseline methods and testing data used in the paper. Their licenses permit use with academic scope.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code and dataset will be release if the paper is accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our core method focuses on video super-resolution and does not involve LLMs as part.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.