SparkVSR: Interactive Video Super-Resolution via Sparse Keyframe Propagation

Jiongze Yu[1], Xiangbo Gao[1], Pooja Verlani[2], Akshay Gadde[2], Yilin Wang[2], Balu Adsumilli[2], and Zhengzhong Tu[1]*

[1] Texas A&M University, College Station, TX, USA
[2] Google, USA

Abstract. Video Super-Resolution (VSR) aims to restore high-quality video frames from low-resolution (LR) estimates, yet most existing VSR approaches behave like black boxes at inference time: users cannot reliably correct unexpected artifacts, but instead can only accept whatever the model produces. In this paper, we propose a novel interactive VSR framework dubbed SparkVSR that makes sparse keyframes a simple and expressive control signal. Specifically, users can first super-resolve or optionally a small set of keyframes using any off-the-shelf image super-resolution (ISR) model, then SparkVSR propagates the keyframe priors to the entire video sequence while remaining grounded by the original LR video motion. Concretely, we introduce a keyframe-conditioned latent-pixel two-stage training pipeline that fuses LR video latents with sparsely encoded HR keyframe latents to learn robust cross-space propagation and refine perceptual details. At inference time, SparkVSR supports flexible keyframe selection and a reference-free guidance mechanism that continuously balances keyframe adherence and blind restoration, ensuring robust performance even when reference keyframes are absent or imperfect. Experiments on multiple VSR benchmarks demonstrate improved temporal consistency and strong restoration quality, surpassing baselines by up to 24.6%, 21.8%, and 5.6% on CLIP-IQA, DOVER, and MUSIQ, respectively, enabling controllable, keyframe-driven video super-resolution. Moreover, we demonstrate that SparkVSR is a generic interactive, keyframe-conditioned video processing framework as it can be applied out of the box to unseen tasks such as old-film restoration and video style transfer. Code is available at: `https://github.com/taco-group/SparkVSR`.

1 Introduction

Video Super-Resolution (VSR) workflow plays a critical role in modern computational photography and video production, aiming to restore visually pleasing, high-frequency details from degraded, low-resolution (LR) video sequences [18, 31]. Despite rapid progress in learning-based VSR models and impressive gains in benchmark metrics [9, 41, 42, 62], the majority of these approaches still

* Corresponding author.

operate as black boxes. That is, once trained, users have little ability to steer the inference results, while the model fully determines the outputs. Recent attempts [51] to introduce controllability via text prompts describing the video content provide only coarse, high-level guidance and often fall short when users need precise, frame-level controls. In practice, this "hope-for-the-best" inference paradigm (i.e., most existing blind VSR) limits the usability of VSR models in real restoration and content creation workflows where subjective preference and targeted corrections are indispensable.

A central reason why controllability is difficult— but necessary—is that super-resolution is inherently ill-posed [19]. The same LR input can correspond to multiple plausible HR reconstructions that differ in texture, sharpness, and fine appearance, and choosing among these plausible outputs is more of a user-intention problem than a learning problem. This motivates a control interface that is both expressive and lightweight. In video workflows, we deem keyframes an effective interface: editing or validating a small number of anchor frames is far more practical than supervising every frame, yet those anchors can strongly constrain the overall result if the model can propagate them reliably.

In parallel, single-image super-resolution [25,33,37,43,50,64] (or generic editing [63]) has recently made dramatic advances, particularly with strong priors to synthesize photorealistic textures and details. However, applying ISR independently to each frame typically causes severe temporal inconsistency and flicker, because per-frame generation ignores cross-frame dynamics. More broadly, we observed that VSR models are often lagging behind the best frame-level ISR in per-frame visual quality due to the fact that VSR is forced to learn both (i) spatial priors and (ii) complex temporal consistency, simultaneously. This further suggests a principled decomposition that leverages a state-of-the-art (possibly interactive) ISR model to obtain high-quality keyframe anchors, and trains a dedicated VSR model to propagate these anchor priors across time while staying faithful to the LR motion structure.

Based on these observations, we propose **SparkVSR**, an interactive framework for VSR via sparse keyframe propagation. SparkVSR transforms VSR into a human-in-the-loop process: users (or an automatic policy) select a small set of keyframes, generate high-quality HR keyframe references using any off-the-shelf ISR model, and then SparkVSR propagates these reference priors to reconstruct a temporally consistent HR video, as shown in Fig. 1. Specifically, SparkVSR introduces a keyframe-conditioned latent–pixel two-stage training pipeline 2. We explicitly retain the LR video information by encoding it into latents and concatenating it with sparsely encoded HR keyframe latents, enabling robust cross-space propagation while grounding reconstruction in the input video structure. At inference time, SparkVSR supports flexible keyframe selection (manual, codec I-frame, etc.). We adopt classifier-free guidance (CFG) [14] during training to allow the model to adjust the condition strength when reference frames are absent or noisy. This design allows SparkVSR to both improve controllability and stand on top of modern ISR advances, yielding stronger per-frame quality together with temporal consistency. Extensive evaluations across multiple benchmarks

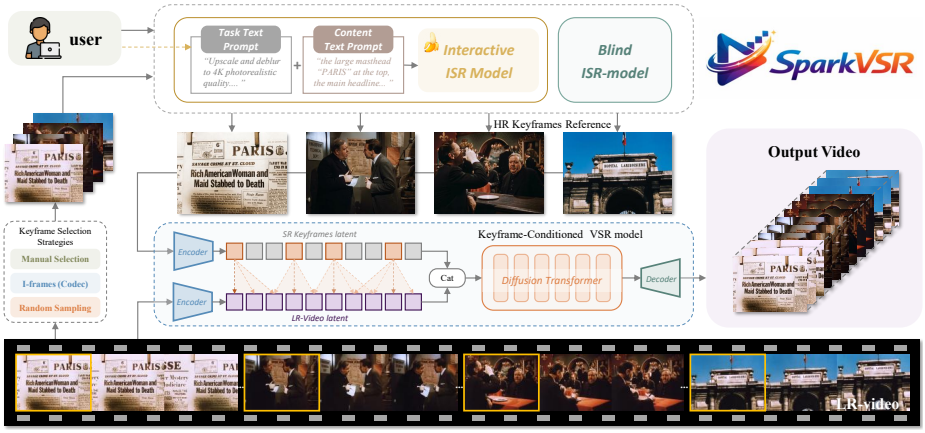

Fig. 1: Overall inference framework of SparkVSR. The pipeline consists of three main stages: (1) Keyframe Selection: LR keyframes are extracted using manual, I-frame, or random sampling strategies; (2) HR Reference Generation: Selected frames are upscaled into HR reference keyframes via an interactive (task/content prompt-guided) or blind ISR model; (3) Conditional Video Reconstruction: A Diffusion Transformer-based VSR model fuses the HR keyframe and LR video latents to guide the generation of the final HR video.

validate these advantages, where SparkVSR outperforms state-of-the-art baselines by up to 24.6% in CLIP-IQA, 21.8% in DOVER, and 5.6% in MUSIQ. We also demonstrate that our SparkVSR framework exhibits strong task generalization abilities as it can be directly applied to old-film restoration and video style transfer. Our contributions are summarized as follows:

- **A Novel Interactive VSR Paradigm.** We formulate VSR as an interactive reconstruction process and propose SparkVSR, where sparse, editable keyframes act as controllable anchors, enabling fine-grained correction and customization beyond black-box inference.
- **Robust Keyframe-Conditioned Latent-Pixel Training.** We propose a two-stage training strategy that fuses LR video latents with sparse HR keyframe latents and refines results in pixel space, equipping the model with robust propagation while maintaining structural fidelity.
- **Flexible inference with controllable guidance.** We provide practical keyframe selection strategies and a method to trade off keyframe adherence and blind restoration, ensuring robustness across diverse scenarios.
- **State-of-the-art Performance.** We demonstrate that SparkVSR achieves state-of-the-art performance in terms of both full-reference and no-reference metrics (using different modes), and reaches Pareto optimality in the perception-distortion tradeoff diagram.

2 Related Work

2.1 Video Super-Resolution

The evolution of Video Super-Resolution (VSR) has been fundamentally driven by architectures designed to capture complex spatio-temporal correlations. Early methods relied on implicit temporal aggregation [17, 18, 22, 31], while subsequent frameworks introduced explicit alignment mechanisms, such as optical flow and deformable convolutions [6, 7, 39, 44], to improve structural reconstruction. More recently, Transformer-based [23,24,28] and Diffusion-based VSR models [9, 15, 41, 42, 51, 52, 62] have achieved state-of-the-art visual quality by synthesizing highly realistic textures from complex degradations. However, despite these impressive quantitative gains, contemporary models predominantly operate as deterministic, end-to-end mapping functions. They essentially function as black boxes during inference, lacking the fine-grained, interactive mechanisms necessary for users to actively guide the reconstruction process, correct specific artifacts, or inject customized visual intentions.

2.2 Controllable Image Super-Resolution

To overcome the limitations of deterministic restoration, recent Image Super-Resolution (ISR) approaches heavily leverage the generative priors of large-scale diffusion models [25,33–37,43,50,59]. These robust spatial priors enable the synthesis of high-fidelity details from severely degraded inputs—a feat VSR struggles to match due to complex cross-frame dynamics and motion blur [56]. Crucially, modern generative paradigms have introduced unprecedented user controllability. Interactive models, such as Nano-Banana-Pro [13, 63], along with text- or spatially-guided frameworks [4,30,57], allow users to actively shape the restored output. While these high-quality, customized single frames serve as ideal reference anchors for reconstruction, applying such ISR models independently across video sequences inevitably disrupts the underlying motion dynamics, resulting in severe temporal flickering and structural inconsistency [21].

2.3 Keyframe-Conditioned Video Processing

In the broader domain of video generation and editing, utilizing sparse keyframes to guide temporal synthesis has emerged as a highly effective paradigm [10–12, 16, 26, 27, 32, 49]. These methods demonstrate that powerful visual priors from individual anchor frames can be robustly propagated across the temporal dimension, often by distributing globally selected keyframes to guide specific local sequence chunks [2]. However, directly adopting these techniques for VSR presents a critical challenge: VSR demands absolute structural fidelity. Existing generative methods often hallucinate content that deviates from original motion constraints, causing severe distortions in the strict LR-to-HR mapping required for restoration [1, 60]. To bridge this gap, SparkVSR introduces a novel keyframe-conditioned latent-pixel training strategy. By seamlessly integrating the high-quality priors of modern ISR models (via sparse keyframes) with the

retained LR latents, our framework explicitly equips the model with robust temporal propagation capabilities while strictly anchoring the output to the video's original structural dynamics.

3 Methodology

This section details the proposed SparkVSR framework, beginning with its overall architecture (Sec. 3.1). We then introduce the keyframe-conditioned latent-pixel training for robust prior propagation (Sec. 3.2), and conclude with the interactive inference phase featuring customizable keyframe selection and flexible reference guidance (Sec. 3.3).

3.1 Overall Architecture

The core of SparkVSR lies in breaking the deterministic black-box mapping of traditional Video Super-Resolution (VSR) by introducing high-quality external reference frames to explicitly guide video generation. To achieve this, we build upon the pretrained weights of the CogVideoX1.5-5B Image-to-Video (I2V) model and design a dual-encoding mechanism to process continuous video sequences and sparse keyframes independently.

Dual-Encoding and Sparse Reference Generation: As illustrated in the inference pipeline (Fig. 1), given a degraded low-resolution (LR) video sequence $x_{lr} \in \mathbb{R}^{T\times H\times W\times 3}$, we first encode it into the latent space using the 3D causal VAE inherited from the pretrained model. Due to the temporal downsampling rate of 4 inherent to the 3D VAE, we obtain a 16-channel video latent representation, denoted as $Z_{LR} \in \mathbb{R}^{\frac{T}{4}\times 16\times H'\times W'}$.

Simultaneously, we generate high-resolution (HR) reference images for the selected sparse keyframes. For interactive restoration, we employ the powerful Nano-Banana-Pro [13, 63], which offers exceptional single-frame generative capabilities. For blind ISR without user intervention, we adopt the current state-of-the-art model, PiSA-SR [37]. Once the HR keyframes, denoted as X_{ref}, are obtained, we design a specific sparse keyframes encoding branch to map these images into the latent space. Crucially, the HR keyframes are sparsely encoded into their corresponding latent indices based on their temporal positions in the video. Let $\mathcal{K}$ denote the set of latent indices corresponding to the selected keyframes. For each latent frame index $i \in \{1, 2, \ldots, \frac{T}{4}\}$, the reference latent $Z_{ref}^{(i)}$ is mathematically formulated as:

$$Z_{ref}^{(i)} = \begin{cases} \mathcal{E}_{sparse}(X_{ref}^{(i)}) & \text{if } i \in \mathcal{K}, \\ \mathbf{0} & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{E}_{sparse}$ denotes the spatial encoder and $\mathbf{0}$ represents a zero tensor of identical spatial and channel dimensions. Consequently, we construct a 16-channel sparse reference latent representation, $Z_{ref} \in \mathbb{R}^{\frac{T}{4}\times 16\times H'\times W'}$.

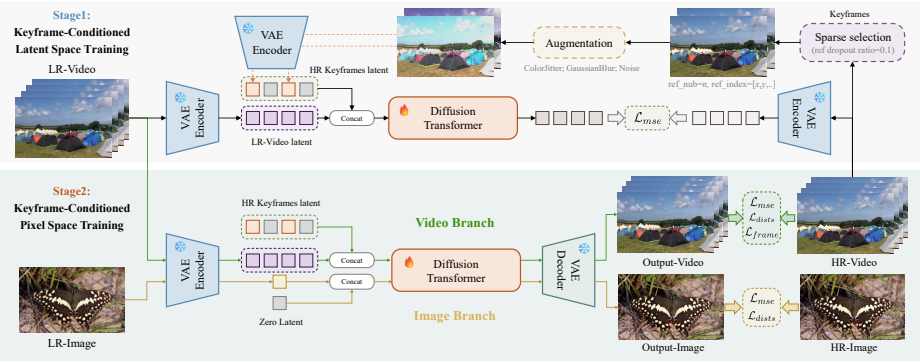

Fig. 2: Keyframe-conditioned two-stage training pipeline of SparkVSR. (1) Stage 1 (Latent Space Training): Augmented HR keyframe latents are concatenated with LR video latents to optimize the Diffusion Transformer using $\mathcal{L}_{mse}$. (2) Stage 2 (Pixel Space Training): A joint video-image training mechanism is employed. The video branch is conditioned on HR keyframe latents, while the image branch uses a zero latent. Finally, outputs are decoded by the VAE and refined in the pixel space using mixed losses.

Feature Fusion and One-Step Denoising: After obtaining the dual-space features, we concatenate the 16-channel LR video latent Z_{LR} and the 16-channel reference latent Z_{ref} along the channel dimension to form the joint conditional input: $Z_{in} = \text{Concat}(Z_{LR}, Z_{ref}) \in \mathbb{R}^{\frac{T}{4} \times 32 \times H' \times W'}$.

Following recent diffusion-based VSR paradigms, we initialize the denoising process directly with the encoded LR video latent Z_{LR}, treating it as the noisy latent Z_t. To minimize computational overhead, we adopt a one-step diffusion strategy inspired by DOVE [9], specifically setting the timestep to $t = 399$. This intermediate step strikes an optimal balance: it preserves sufficient global structure from the LR video while allowing the Diffusion Transformer v_θ to focus exclusively on hallucinating high-frequency details conditioned on Z_{in}. Finally, the denoised latent Z_{sr} is decoded by the VAE decoder $\mathcal{D}$ to reconstruct the HR video $x_{sr} = \mathcal{D}(Z_{sr})$.

3.2 Keyframe-Conditioned Latent-Pixel Training

To achieve precise keyframe control and effectively propagate the high-quality spatial priors generated by the external ISR model throughout the entire video sequence, we formulate a two-stage Keyframe-Conditioned Latent-Pixel Training strategy, as depicted in the training pipeline (Fig. 2), inspired by the highly efficient two-stage paradigm [9].

Stage 1 (Latent-Space): In the first stage, we fix the VAE decoder and train the Transformer v_θ entirely in the latent space, which significantly improves training efficiency. During the extraction of HR keyframes from the ground-truth video x_{hr}, we implement a sparse random selection strategy. Specifically, the total number of selected keyframes is randomly determined with an upper

bound of $T/4$, and their temporal indices are randomly sampled with the strict constraint that the interval between any two selected keyframes must be greater than the inherent temporal downsampling rate (> 4). We then apply severe augmentations (e.g., Color Jitter, Gaussian Blur, Noise) to these selected HR frames to simulate the distribution of external ISR outputs. Guided by the 32-channel concatenated condition Z_{in}, the model learns to align and absorb reference features. The optimization objective minimizes the Mean Squared Error (MSE) between the predicted latent Z_{sr} and the HR ground-truth latent Z_{hr}. To preserve robust performance when reference frames are unavailable, we introduce a reference-dropout strategy. During training, the reference latent Z_{ref} is omitted and replaced with zero tensors with a predefined probability p_{drop}, compelling the model to perform reference-free blind restoration.

Stage 2 (Pixel-Space): While latent space training effectively captures the semantic layout, pixel-level constraints are necessary to eliminate temporal flickering and refine perceptual textures. Therefore, we transition to the pixel space and introduce a joint image-video training scheme.

For the **Video Branch**, the sparse keyframe selection and encoding strategies remain strictly identical to those utilized in Stage 1. The sequence is processed by the network and decoded into the pixel space to generate the output video $\hat{x}_{sr}$. To enforce temporal coherence and exceptional perceptual quality, this branch is supervised by a combination of pixel-wise MSE ($\mathcal{L}_{mse}$), perceptual DISTS loss ($\mathcal{L}_{dists}$), and a frame consistency loss ($\mathcal{L}_{frame}$):

$$\mathcal{L}_{s2-video} = \mathcal{L}_{mse}(\hat{x}_{sr}, x_{hr}) + \lambda_1 \mathcal{L}_{dists}(\hat{x}_{sr}, x_{hr}) + \lambda_2 \mathcal{L}_{frame}(\hat{x}_{sr}, x_{hr}) \quad (2)$$

Simultaneously, for the **Image Branch**, we process single LR images and explicitly concatenate the encoded image latent with a Zero Latent. This concatenation of the Zero Latent is purposefully designed not only to align the channel dimensions (maintaining the 32-channel input format required by the DiT) but also to substantially solidify the model's robust generative priors in scenarios where reference frames are entirely absent. This branch is optimized using only $\mathcal{L}_{mse}$ and $\mathcal{L}_{dists}$.

3.3 Flexible Interactive Inference

During inference (Fig. 1), SparkVSR introduces a highly customizable interactive paradigm that returns control to the user. This flexibility is achieved through versatile keyframe selection strategies, the integration of prompt-guided Image Super-Resolution (ISR) models, and a tunable reference-free guidance mechanism that modulates the influence of spatial priors.

Flexible Keyframe Selection: SparKVSR supports three distinct strategies to accommodate diverse application scenarios:

1. **Manual Selection:** Users can pinpoint specific frames based on aesthetic intentions or target those suffering from the most severe degradation.

2. **Codec I-frames:** SparKVSR natively extracts intra-coded I-frames directly from the video stream. As these frames retain maximal spatial information with minimal compression artifacts, they serve as optimal, high-quality anchors for global restoration.
3. **Random Sampling:** Designed for automated, large-scale batch processing where human intervention is not required.

Prompt-Guided Interactive Restoration: When employing an interactive ISR model (e.g., Nano-Banana-Pro [63]), users can exert fine-grained control over keyframe restoration via decoupled textual conditioning. As illustrated in Fig. 1, the prompt is systematically divided into two components: a **Task Text Prompt** (e.g., *"Upscale and deblur to 4K photorealistic quality"*) to specify the core restoration objective, and a **Content Text Prompt** (e.g., *"the large masthead 'PARIS' at the top"*) to explicitly describe desired semantic or structural details. This human-in-the-loop dual-prompting ensures the generation of pristine, semantically accurate keyframes, especially when tackling extreme degradations where blind models fail. These customized keyframes are subsequently propagated through the SparkVSR architecture to drive the high-fidelity reconstruction of the entire sequence.

Reference-Free Guidance: To grant precise control over feature propagation and balance external keyframe priors with the model's internal generative capacity, we introduce a Reference-Free Guidance mechanism inspired by Classifier-Free Guidance (CFG) [14]. Benefiting from the reference-dropout strategy and the Zero Latent conditioning, our model inherently supports both referenced and reference-free (blind SR) predictions. During the denoising step, the final prediction $\hat{v}$ is formulated as:

$$\hat{v} = v_\theta(Z_{in}^{\text{uncond}}) + s \cdot \big(v_\theta(Z_{in}^{\text{cond}}) - v_\theta(Z_{in}^{\text{uncond}})\big) \tag{3}$$

where $Z_{in}^{\text{cond}} = \text{Concat}(Z_{LR}, Z_{ref})$ serves as the conditional input, $Z_{in}^{\text{uncond}} = \text{Concat}(Z_{LR}, \mathbf{0})$ replaces reference features with Zero Latents, and s dictates the user-adjustable guidance scale. Specifically, setting $s = 1$ yields standard keyframe-guided generation, whereas $s > 1$ amplifies the high-frequency textures and spatial features from the keyframes, enforcing stronger prior propagation. Conversely, if the external ISR outputs contain slight artifacts, or if the user prefers to rely more heavily on the model's internal blind SR priors, s can be reduced ($s < 1$) or completely disabled ($s = 0$).

4 Experiments

4.1 Experimental Settings

Datasets. Following the training data configuration of DOVE [9], our training corpus combines 2,055 high-resolution video clips from HQ-VSR (degraded via RealBasicVSR [8]) and 900 images from DIV2K [5] (degraded via Real-ESRGAN [45]). For evaluation, we employ synthetic benchmarks (UDM10 [38],

SPMCS [55], YouHQ40 [61]) with training-matched degradations, alongside real-world datasets like RealVSR [53], which contains smartphone-captured LQ-HQ pairs. Furthermore, we propose **MovieLQ**, a novel dataset featuring ten 360p (360×480) vintage film clips from the 1940s to 1950s that cover diverse scenes, including text, human subjects, and landscapes. Each clip lasts 8 seconds at 24 frames per second (fps), totaling 192 frames. Unlike existing benchmarks, it offers longer sequences with complex, authentic degradations inherently originating from historical imaging devices and legacy video compression.

Evaluation Metrics. We comprehensively assess model performance using both image (IQA) and video quality assessment (VQA) metrics. The IQA suite includes standard fidelity metrics (PSNR, SSIM [46]) and perceptual indicators (LPIPS [58], CLIP-IQA [40], MUSIQ [20]). For VQA, FasterVQA [47] and DOVER [48] are utilized to accurately measure the overall spatial-temporal video quality.

Implementation Details. SparkVSR is built upon the CogVideoX1.5-5B I2V [54] foundation model and optimized via a two-stage fine-tuning strategy on four NVIDIA A100-80GB GPUs (total batch size 8) using the AdamW optimizer [29] ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\beta_3 = 0.98$). Throughout training, reference frames are sampled with random quantities and temporal intervals (strictly exceeding the VAE's temporal downsampling rate) and augmented via ColorJitter, GaussianBlur, and noise injection. Additionally, we set the reference-dropout probability to $p_{drop} = 0.1$ to robustly maintain zero-reference super-resolution capability. Stage-1 trains exclusively on video sequences (320×640 resolution, 33 frames) for 10,000 iterations at a learning rate of 2×10^{-5}, while Stage-2 shifts to a joint video-image paradigm ($\varphi = 0.5$) for 500 iterations at 5×10^{-6} with loss weights $\lambda_1 = \lambda_2 = 1$.

4.2 Comparisons

To comprehensively evaluate the effectiveness of our proposed SparkVSR, we compare it against several recent state-of-the-art video super-resolution methods. The evaluated baselines include STAR [51], DOVE [9], SeedVR2 (3B/7B) [41], and FlashVSR (Tiny/Full) [62]. Regarding our keyframe selection strategy, we select only the initial frame as the reference for short sequences (i.e., UDM10, SPMCS, YouHQ40, and RealVSR), while utilizing codec I-frames as reference anchors for the MovieLQ dataset.

Quantitative Evaluation. As reported in Table 1, SparkVSR consistently achieves superior performance across five diverse benchmark datasets. Our baseline model without reference (SparkVSR*) exhibits strong fidelity, achieving the highest PSNR (26.62) and SSIM (0.7756) on the UDM10 dataset. Furthermore, by incorporating reference priors, our reference-guided variants (SparkVSR† with Nano-Banana-Pro reference and SparkVSR‡ with PiSA-SR reference) establish new state-of-the-art results in perceptual and video quality assessment (VQA) metrics. Notably, on the challenging real-world MovieLQ dataset, SparkVSR‡ dominates perceptual evaluations, yielding the best scores in MUSIQ (68.88), CLIP-IQA (0.6361), FasterVQA (0.8028), and DOVER (0.6212). This demon-

Table 1: Quantitative comparison of our method against state-of-the-art methods across multiple datasets. The best and second best results are highlighted. Ours*, Ours†, and Ours‡ denote our method without reference, with Nano-Banana-Pro reference, and with PiSA-SR reference, respectively.

Dataset	Metric	STAR	DOVE	SeedVR2-3B	SeedVR2-7B	FlashVSR-Tiny	FlashVSR-Full	Ours*	Ours†	Ours‡
UDM10	PSNR↑	24.11	26.52	25.29	25.44	23.84	23.58	26.62	23.70	23.43
	SSIM↑	0.7052	0.7697	0.7573	0.7583	0.7095	0.6993	0.7756	0.6807	0.6710
	LPIPS↓	0.3987	0.2709	0.2667	0.2526	0.2800	0.2915	0.2830	0.3376	0.3548
	MUSIQ↑	35.58	61.11	47.13	49.06	63.42	65.84	55.79	66.16	67.52
	CLIP-IQA↑	0.2505	0.5011	0.2797	0.2882	0.4587	0.5016	0.4303	0.5501	0.6252
	FasterVQA↑	0.6046	0.8155	0.5894	0.6256	0.7707	0.7899	0.7535	0.8357	0.8202
	DOVER↑	0.3331	0.5664	0.3716	0.4095	0.5178	0.5317	0.5494	0.6902	0.6411
SPMCS	PSNR↑	22.71	23.08	22.2	22.51	21.51	21.39	23.04	19.68	20.12
	SSIM↑	0.6379	0.6190	0.6067	0.6183	0.5463	0.5403	0.6237	0.4704	0.4908
	LPIPS↓	0.4558	0.2887	0.2888	0.2782	0.2907	0.3015	0.3139	0.3356	0.3387
	MUSIQ↑	32.32	69.11	63.73	64.06	68.2	69.48	65.53	72.97	73.38
	CLIP-IQA↑	0.2807	0.5708	0.4288	0.4314	0.4441	0.4738	0.4937	0.6830	0.6811
	FasterVQA↑	0.5398	0.7227	0.681	0.6911	0.7291	0.7306	0.6791	0.7474	0.7105
	DOVER↑	0.4882	0.4858	0.4391	0.4297	0.4856	0.4701	0.4300	0.5781	0.5337
YouHQ40	PSNR↑	22.71	24.44	22.13	22.21	22.35	22.09	24.35	22.27	21.75
	SSIM↑	0.6379	0.6807	0.6338	0.6418	0.5998	0.5912	0.6787	0.599	0.5786
	LPIPS↓	0.4558	0.2950	0.2975	0.2802	0.2835	0.2894	0.3272	0.3178	0.3501
	MUSIQ↑	32.32	61.03	60.804	63.18	65.97	68.69	55.41	64.62	69.10
	CLIP-IQA↑	0.2807	0.4867	0.4106	0.4216	0.4644	0.5043	0.4269	0.5411	0.6346
	FasterVQA↑	0.5398	0.8477	0.8554	0.8581	0.8547	0.8655	0.8075	0.8592	0.8420
	DOVER↑	0.4882	0.7020	0.7041	0.7050	0.6732	0.6783	0.6694	0.7393	0.7315
RealVSR	PSNR↑	17.30	22.32	21.21	22.13	19.84	19.62	22.00	21.35	19.72
	SSIM↑	0.5253	0.7301	0.7010	0.7189	0.5613	0.5561	0.7222	0.6982	0.6183
	LPIPS↓	0.3268	0.1850	0.2168	0.1998	0.2367	0.2316	0.1809	0.1678	0.2165
	MUSIQ↑	68.61	71.69	64.71	63.92	68.02	71.12	71.84	71.86	75.44
	CLIP-IQA↑	0.3377	0.5207	0.3013	0.2853	0.3454	0.3927	0.5407	0.5575	0.6216
	FasterVQA↑	0.7184	0.7929	0.7287	0.7284	0.7514	0.7501	0.7861	0.7727	0.7946
	DOVER↑	0.5672	0.6150	0.5486	0.5241	0.5278	0.5270	0.6180	0.5988	0.6399
MovieLQ	MUSIQ↑	59.68	60.71	49.59	45.97	64.79	66.38	56.34	65.48	68.88
	CLIP-IQA↑	0.3731	0.5433	0.3309	0.2894	0.5487	0.5754	0.4622	0.6128	0.6361
	FasterVQA↑	0.7611	0.7647	0.5534	0.4184	0.78	0.7822	0.7065	0.797	0.8028
	DOVER↑	0.5696	0.5101	0.3619	0.313	0.5485	0.5544	0.5121	0.6194	0.6212

strates the robustness of our reference-guided approach in handling complex real-world degradations.

Qualitative Evaluation. Compared to existing approaches, SparkVSR consistently produces sharper, more realistic restorations while avoiding undesired artifacts. Specifically, on the real-world MovieLQ dataset (Figure 3), our method successfully reconstructs highly legible text and delicate facial details (e.g., skin texture and facial hair), effectively mitigating the severe blurring and over-smoothing prevalent in baselines like DOVE, FlashVSR, and STAR. Furthermore, when evaluated against Ground Truth references on the SPMCS and YouHQ40 datasets (Figure 4), SparkVSR exhibits exceptional high-frequency detail regeneration, precisely restoring complex structural edges and fine natural textures (e.g., animal fur).

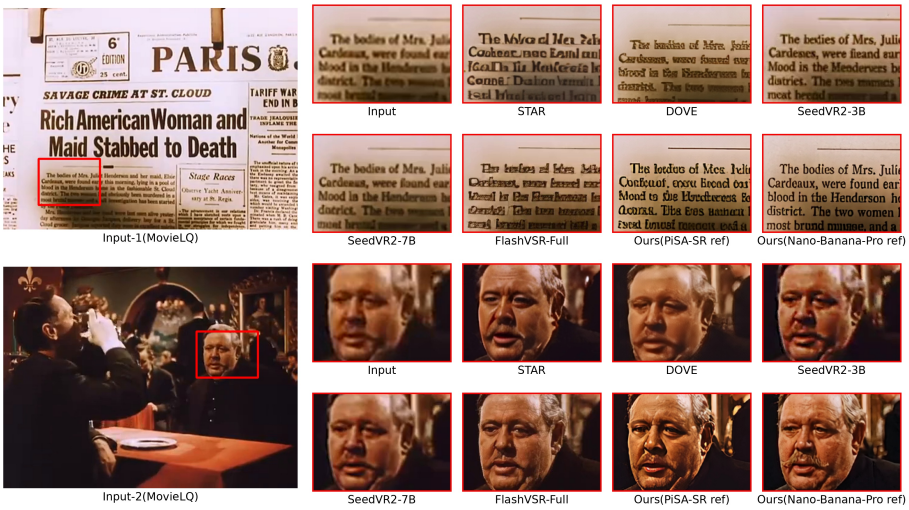

Fig. 3: Qualitative visual comparisons on the MovieLQ dataset. Compared to state-of-the-art VSR methods, SparkVSR demonstrates superior recovery of fine textures and structural details, particularly in restoring highly degraded text and facial features.

4.3 Ablation Study

To validate the effectiveness of the core components in SparkVSR, we conduct comprehensive ablation studies analyzing our training strategies, perception-distortion control via the Reference-Free Guidance (RFG) scale, and reference frame selection.

Effectiveness of Training Strategies. We investigate the impact of our two-stage training paradigm in Table 2. While utilizing only the first stage (S1) achieves high fidelity, evidenced by a PSNR of 26.73 under zero-reference conditions, it yields sub-optimal perceptual quality. Introducing the second stage (S1+S2) significantly improves all perceptual evaluations, regardless of the reference source. This demonstrates that our refinement stage is crucial for enhancing visual realism while maintaining competitive structural integrity.

Perception-Distortion Tradeoff and RFG. We explore the influence of the Reference-Free Guidance (RFG) scale on restoration results, addressing the inherent mathematical tradeoff between distortion and perceptual quality in image restoration [3]. As shown in Table 4 and Figure 5, increasing the RFG scale from 0 to 1.5 gradually decreases distortion-based metrics, such as PSNR and SSIM, but substantially boosts perceptual indicators like MUSIQ and CLIP-IQA. Notably, when compared against recent baselines including DOVE, STAR, SeedVR2, and FlashVSR, varying the RFG scale allows SparkVSR to establish a superior Pareto front, as illustrated in Figure 5. Our method consistently achieves better perceptual quality at equivalent distortion levels, effectively pushing the boundaries of

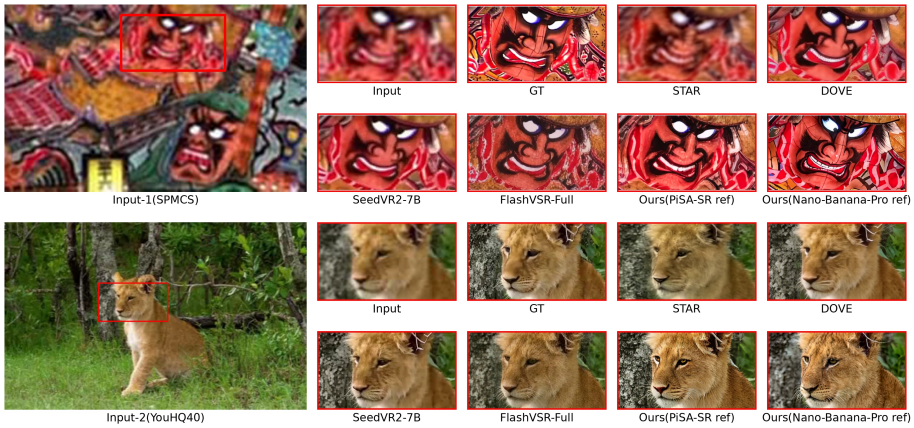

Fig. 4: Qualitative visual comparisons on the SPMCS [55] and YouHQ40 [61] datasets. We compare our method against recent state-of-the-art VSR models. Guided by high-resolution references, SparkVSR excels in reconstructing sharp edges in animation scenes (top) and fine, realistic textures in natural scenes (bottom).

Table 4: Ablation study of the impact of various reference sources and varying RFG scales on the UDM10 [38] dataset.

Ref-source	-	Nano-Banana-pro					PiSA-SR				
RFG	0	0.5	0.8	1.0	1.2	1.5	0.5	0.8	1.0	1.2	1.5
PSNR↑	**26.62**	26.04	24.78	23.70	22.56	20.89	25.66	24.41	23.43	22.42	20.94
SSIM↑	**0.776**	0.757	0.718	0.681	0.641	0.582	0.741	0.703	0.671	0.639	0.589
LPIPS↓	**0.283**	0.292	0.312	0.338	0.367	0.412	0.312	0.337	0.355	0.372	0.399
MUSIQ↑	55.79	58.76	63.93	66.16	67.52	68.36	60.17	65.23	67.52	69.01	**70.39**
CLIP-IQA↑	0.430	0.447	0.515	0.550	0.570	0.576	0.496	0.584	0.625	0.647	**0.652**
FasterVQA↑	0.754	0.794	0.827	0.836	0.850	**0.860**	0.769	0.796	0.820	0.824	0.832
DOVER↑	0.549	0.571	0.645	0.690	0.714	**0.746**	0.562	0.604	0.641	0.676	0.706

this tradeoff. Visually (Figure 6), although the zero-reference variant successfully recovers fundamental structures with relatively smooth textures, higher RFG scales progressively inject richer high-frequency details, specifically the distinct structural grids of stadium seats and complex natural textures, which closely approximate the ground truth.

Influence of Reference Frame Selection. Table 3 and Figure 7 evaluate various reference frame selection strategies. Incorporating a single reference frame significantly boosts perceptual quality (e.g., MUSIQ improves from 56.34 to 61.73). Visually, while the zero-reference baseline yields relatively soft details, injecting a single reference dramatically sharpens the subject, effectively propagating the high-fidelity textures of the keyframe to adjacent frames. Increasing the reference count further enhances performance. Utilizing multiple distributed references (such as three uniformly sampled indices or four extracted I-frames)

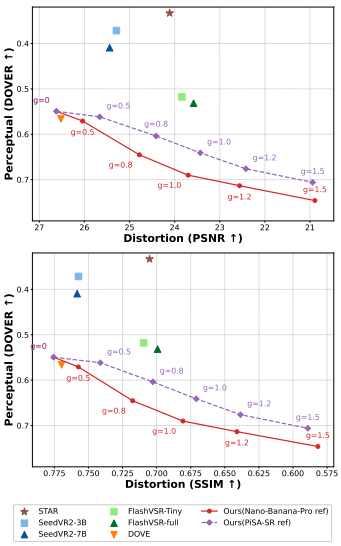

Fig. 5: Perception-distortion trade-off. Comparison on PSNR and SSIM vs. DOVER.

Table 2: Ablation of training strategies on the UDM10 [38] dataset. 'NBP' denotes the NanoBanana-Pro method.

	ours(S1)			ours(S1+S2)		
Metric	None	NBP	PiSA	None	NBP	PiSA
PSNR↑	**26.73**	24.53	24.35	26.62	23.70	23.43
SSIM↑	**0.778**	0.707	0.698	0.776	0.681	0.671
LPIPS↓	0.330	0.338	0.354	**0.283**	0.338	0.355
MUSIQ↑	44.04	62.57	63.67	55.79	66.16	**67.52**
C-IQA↑	0.331	0.474	0.537	0.430	0.550	**0.625**
F-VQA↑	0.656	0.824	0.789	0.754	**0.836**	0.820
DOVER↑	0.439	0.642	0.609	0.549	**0.690**	0.641

Table 3: Ablation of the number and positions of reference frames on the MovieLQ dataset.

Ref. Num	Ref. Idx	MUSIQ↑	C-IQA↑	F-VQA↑	DOVER↑
0	-	56.34	0.462	0.707	0.512
1	[1]	61.73	0.535	0.764	0.562
2	[1,192]	63.76	0.566	0.774	0.580
3	[1,96,192]	64.84	0.594	0.795	0.596
4	[1,64,128,192]	**65.76**	0.610	0.793	0.606
I-frames	[1,48,96,144]	65.48	**0.613**	**0.797**	**0.619**

ensures temporally consistent, delicate textures—specifically hair strands and facial features—across the entire sequence. Achieving optimal and comparable scores across these diverse indices demonstrates that SparkVSR robustly accommodates flexible strategies, encompassing user-defined, random, and codec-aware frame extraction.

4.4 Broader Applications

The reference-guided architecture of SparkVSR extends beyond standard super-resolution, serving as a robust temporal propagation engine for broader low-level video editing tasks. By utilizing sparsely edited keyframes (e.g., colorized, deblurred, or stylized) as references, SparkVSR effectively propagates high-fidelity features across the entire sequence with rigorous temporal consistency. This unlocks several zero-shot applications without task-specific retraining:

- **Old Video Restoration and Colorization:** Given a few manually restored keyframes from severely degraded historical videos, SparkVSR faithfully propagates clean textures and realistic colors to overcome complex real-world degradations.
- **Stylized Video Generation:** Applying artistic edits (e.g., pixel anime art style) to reference frames enables the synthesis of temporally coherent stylized videos while strictly preserving original structural motions.

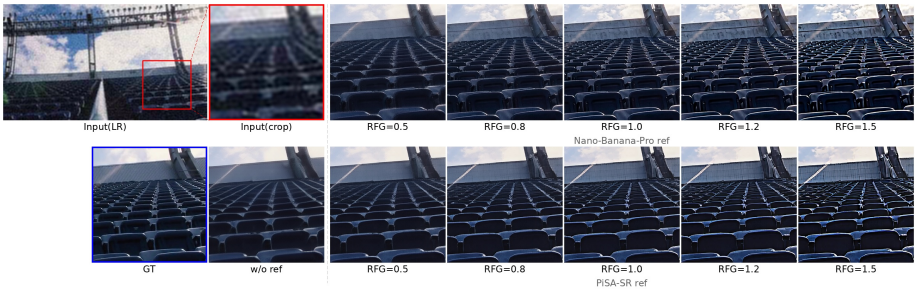

Fig. 6: Visual ablation study of different reference sources and varying RFG scales. We compare the restoration results generated by Nano-Banana-Pro and PiSA-SR across a range of Reference-Free Guidance (RFG) values.

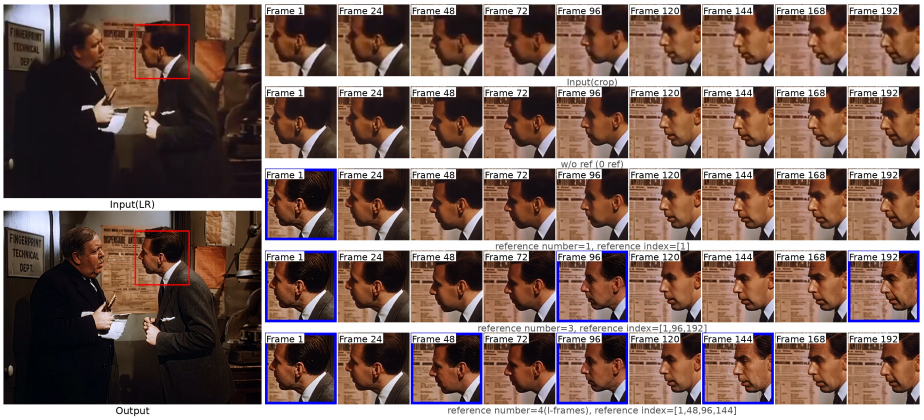

Fig. 7: Visual ablation study of the number and temporal positions of reference frames. Blue boxes indicate the specific temporal indices corresponding to the provided high-resolution reference frames, which are generated using Nano-Banana-Pro.

These capabilities demonstrate that SparkVSR transcends a standard super-resolution baseline, emerging as a highly generalizable framework for consistent video feature propagation.

5 Conclusion

In this paper, we propose SparkVSR, an interactive framework that transforms Video Super-Resolution from a deterministic black box into a controllable, user-guided process. By utilizing sparse, editable keyframes as anchors, SparkVSR efficiently propagates the robust spatial priors of modern image super-resolution models. Our keyframe-conditioned latent-pixel training ensures high-fidelity temporal consistency across the sequence while strictly preserving original motion dynamics. Coupled with flexible keyframe selection and a tunable reference-free guidance mechanism, SparkVSR empowers users to precisely balance perceptual quality and structural fidelity. Extensive experiments

demonstrate that SparkVSR achieves state-of-the-art performance across multiple benchmarks and generalizes seamlessly to zero-shot applications like old-film restoration and stylized video generation, establishing a highly versatile paradigm for video reconstruction.

Acknowledgements

This work was supported in part by Google Inc, and in part by GPU hardware provided to the Texas A&M University through the NVIDIA Academic Grant Program.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
3. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
5. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3086–3095 (2019)
6. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: 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. pp. 4947–4956 (2021)
7. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5972–5981 (2022)
8. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5962–5971 (2022)
9. Chen, Z., Zou, Z., Zhang, K., Su, X., Yuan, X., Guo, Y., Zhang, Y.: Dove: Efficient one-step diffusion model for real-world video super-resolution. arXiv preprint arXiv:2505.16239 (2025)
10. Gao, X., Li, R., Chen, X., Wu, Y., Feng, S., Yin, Q., Tu, Z.: Pisco: Precise video instance insertion with sparse control. arXiv preprint arXiv:2602.08277 (2026)
11. Gao, X., Wu, M., Yang, S., Yu, J., Taghavi, P., Lin, F., Tu, Z.: The pulse of motion: Measuring physical frame rate from visual dynamics. arXiv preprint arXiv:2603.14375 (2026)

12. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
13. Google DeepMind: Gemini 3 pro image – nano banana pro. `https://deepmind.google/models/gemini-image/pro/` (2025)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
16. Huang, X., Xu, C., Luo, D., Hu, X., Tang, P., Peng, X., Zhang, J., Wang, C., Fu, Y.: Ffp-300k: Scaling first-frame propagation for generalizable video editing. arXiv preprint arXiv:2601.01720 (2026)
17. Isobe, T., Li, S., Jia, X., Yuan, S., Slabaugh, G., Xu, C., Li, Y.L., Wang, S., Tian, Q.: Video super-resolution with temporal group attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8008–8017 (2020)
18. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3224–3232 (2018)
19. Jo, Y., Oh, S.W., Vajda, P., Kim, S.J.: Tackling the ill-posedness of super-resolution through adaptive target generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16236–16245 (2021)
20. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
21. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018)
22. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J.: Mucan: Multi-correspondence aggregation network for video super-resolution. In: European conference on computer vision. pp. 335–351. Springer (2020)
23. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. IEEE Transactions on Image Processing **33**, 2171–2182 (2024)
24. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems **35**, 378–393 (2022)
25. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: European conference on computer vision. pp. 430–448. Springer (2024)
26. Liu, J., Li, J., Deng, J., Li, G., Zhou, S., Fang, Z., Lao, S., Deng, Z., Zhu, J., Ma, T., et al.: Dreamontage: Arbitrary frame-guided one-shot video generation. arXiv preprint arXiv:2512.21252 (2025)
27. Liu, S., Wang, T., Wang, J.H., Liu, Q., Zhang, Z., Lee, J.Y., Li, Y., Yu, B., Lin, Z., Kim, S.Y., et al.: Generative video propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17712–17722 (2025)
28. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)

29. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=rk6qdGgCZ
30. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
31. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: 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. pp. 0–0 (2019)
32. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15932–15942 (2023)
33. Qi, C., Tu, Z., Ye, K., Delbracio, M., Milanfar, P., Chen, Q., Talebi, H.: Spire: Semantic prompt-driven image restoration. In: European Conference on Computer Vision. pp. 446–464. Springer (2024)
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
37. Sun, L., Wu, R., Ma, Z., Liu, S., Yi, Q., Zhang, L.: Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2333–2343 (2025)
38. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE international conference on computer vision. pp. 4472–4480 (2017)
39. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3360–3369 (2020)
40. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
41. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., et al.: Seedvr2: One-step video restoration via diffusion adversarial post-training. arXiv preprint arXiv:2506.05301 (2025)
42. Wang, J., Lin, Z., Wei, M., Zhao, Y., Yang, C., Loy, C.C., Jiang, L.: Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2161–2172 (2025)
43. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
44. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
45. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
47. Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., Gu, J., Lin, W.: Neighbourhood representative sampling for efficient end-to-end video quality assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 15185–15202 (2023)
48. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20144–20154 (2023)
49. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)
50. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems **37**, 92529–92553 (2024)
51. Xie, R., Liu, Y., Zhou, P., Zhao, C., Zhou, J., Zhang, K., Zhang, Z., Yang, J., Yang, Z., Tai, Y.: Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17108–17118 (2025)
52. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM computing surveys **56**(4), 1–39 (2023)
53. Yang, X., Xiang, W., Zeng, H., Zhang, L.: 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. pp. 4781–4790 (2021)
54. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations (2025), https://openreview.net/forum?id=LQzN6TRFg9
55. Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J.: Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3106–3115 (2019)
56. Zhang, F., Li, Y., You, S., Fu, Y.: Learning temporal consistency for low light video enhancement from single images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4967–4976 (2021)
57. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

59. Zhang, R., Yu, J., Chen, J., Li, G., Lin, L., Wang, D.: A prior guided wavelet-spatial dual attention transformer framework for heavy rain image restoration. IEEE Transactions on Multimedia **26**, 7043–7057 (2024)
60. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
61. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: 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. pp. 2535–2545 (2024)
62. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)
63. Zuo, J., Deng, H., Zhou, H., Zhu, J., Zhang, Y., Zhang, Y., Yan, Y., Huang, K., Chen, W., Deng, Y., et al.: Is nano banana pro a low-level vision all-rounder? a comprehensive evaluation on 14 tasks and 40 datasets. arXiv preprint arXiv:2512.15110 (2025)
64. Zuo, Y., Zheng, Q., Wu, M., Jiang, X., Li, R., Wang, J., Zhang, Y., Mai, G., Wang, L.V., Zou, J., et al.: 4kagent: agentic any image to 4k super-resolution. arXiv preprint arXiv:2507.07105 (2025)