CtrlVSR: Controllable Video Super-Resolution via Motion-Aware Latent World Modeling

Anonymous CVPR submission

Paper ID *****

Abstract

*Video super-resolution (VSR) aims to recover high-fidelity high-resolution videos from low-resolution inputs and is central to applications ranging from mobile capture to streaming and archival restoration. Existing approaches trade off among local-detail fidelity, long-range spatio-temporal modeling, perceptual realism and efficiency: convolutional alignment techniques preserve local structure but suffer when motion is large or degradations are complex; transformer-based methods capture long-range dependencies yet require architectural or algorithmic adaptations to be computationally feasible; and recent latent or diffusion-based generators synthesize rich texture but demand specialized temporal constraints to maintain coherence. In this work, we present **CtrlVSR**, a controllable VSR framework that casts restoration as motion-aware latent world modeling and integrates adaptive sparse attention with an explicit, user-accessible control interface. CtrlVSR combines robust motion fusion, a Latent World Transformer that balances locality and targeted non-local interactions, and a compact conditional decoder to deliver temporally consistent, high-quality reconstructions under streaming constraints. Empirical evaluations demonstrate that our design attains strong reconstruction and perceptual performance while enabling predictable trade-offs between temporal smoothness and fidelity.*

Keywords: video super-resolution, latent diffusion, motion-aware modeling, sparse attention

1. Introduction

Video super-resolution recovers high-resolution sequences from low-resolution inputs and remains a fundamental challenge in computer vision. Over the last decade the field has progressed from spatio-temporal convolutional pipelines toward models that exploit long-range correlations and powerful generative priors. Survey papers and recent benchmarks highlight a rapidly expanding literature and a wide set of practical use cases for VSR [2, 24].

Our contributions are presented up front. First, we propose CtrlVSR, a motion-aware latent world modeling framework that explicitly measures per-location motion reliability and corrects biased external motion cues. Second, we design an adaptive sparse attention mechanism inside a latent transformer that preserves local context while selectively retrieving remote blocks, thereby bounding compute for large frames. Third, we introduce a compact and interpretable controllability interface that exposes a global scalar and a learned spatial gate to trade off temporal smoothness and reconstruction fidelity. Finally, we provide a practical training and distillation recipe that yields a streaming-capable system with strong empirical performance on synthetic and real-world benchmarks.

Classical convolutional and alignment-based methods remain valuable because they excel at preserving local structure and sharp edges when motion is modest and degradations are moderate [19, 21]. These approaches typically fuse nearby frames after motion compensation and achieve reliable per-pixel fidelity in many practical scenarios. However, when motion becomes large or degradations become heterogeneous, errors in explicit alignment propagate through the temporal pipeline and harm both visual quality and temporal consistency.

Transformer-based solutions extend the effective receptive field and provide a natural mechanism for modeling long-range spatio-temporal dependencies [18, 33]. Such models can capture complex cross-frame correlations that convolutional kernels cannot, but naive full attention is computationally expensive for high-resolution and long-duration videos and therefore requires sparse or block-wise approximations to be practical [7].

Generative latent and diffusion methods have recently shown remarkable ability to synthesize realistic textures and perceptually rich details by leveraging strong priors in a compact latent space [3, 27, 34]. These approaches open new possibilities for perceptual restoration, yet their generative nature introduces stochasticity and temporal discontinuities unless sequence-level guidance, motion condition-

ing or multi-stage distillation are applied [24].

Convolutional alignment remains competitive for local fidelity. Transformer-based solutions provide powerful long-range modeling but demand architectural or algorithmic measures to remain efficient. Generative latent approaches enable high-quality detail synthesis but require carefully designed temporal constraints to avoid discontinuities. CtrlVSR unifies motion-aware latent modeling with adaptive sparse attention and an explicit controllability mechanism to address these complementary challenges.

Despite this rich progress, three practical hurdles remain for deploying high-quality VSR in streaming and real-world systems. First, dense spatio-temporal attention grows quadratically with spatial resolution and becomes infeasible without sparse or block-sparse strategies [1, 7]. Second, many pipelines rely on external motion cues such as optical flow or camera pose; these signals can be unreliable in occlusions, low-texture regions and highly dynamic scenes, and naive fusion of noisy motion leads to temporal artifacts [8, 11]. Third, diffusion and latent generators offer strong perceptual priors but require conditioning and distillation to preserve fine-grained reference details and to limit stochastic inconsistency across frames [6, 27].

This paper makes the following contributions. We introduce CtrlVSR, a motion-aware latent world modeling framework that explicitly measures motion reliability and corrects external motion signals for robust temporal forecasting. We design an adaptive sparse attention scheme within a latent transformer that balances locality with targeted non-local interactions to achieve efficient high-resolution restoration. We propose a simple and effective controllability mechanism that gives users predictable control over the smoothness–fidelity trade-off. Finally, we present a practical training and distillation recipe that yields a streaming-capable system with strong empirical performance on synthetic and real-world benchmarks. Throughout the paper we compare and contrast our design with contemporary convolutional, transformer, and diffusion-based VSR approaches to highlight complementary strengths and remaining limitations in prior work.

2. Related Work

We position CtrlVSR within four interrelated strands of prior research, including convolutional alignment methods, transformer-based spatio-temporal modeling and efficient attention, generative latent and diffusion approaches, and motion estimation combined with controllability mechanisms. The following subsections summarize representative work and highlight the gaps that CtrlVSR addresses.

2.1. Convolutional methods and explicit alignment

Early VSR methods rely on explicit motion estimation and compensation to aggregate multi-frame evidence. Caballero et al.[4] proposed a spatio-temporal sub-pixel convolution pipeline for real-time VSR. Later work introduced implicit motion handling via dynamic filters and refinement to reduce flow sensitivity[14]. RealBasicVSR[5] addressed realistic degradations with pre-cleaning and stochastic strategies. Despite improvements, pixel-level alignment errors still degrade temporal coherence under large motion or severe artifacts[10].

2.2. Transformer-based spatio-temporal modeling and attention efficiency

Transformers offer a natural mechanism for long-range spatio-temporal modeling. Trajectory-aware Transformer (TTVSR) restricts attention to pre-aligned token trajectories, enabling long-range modeling with reduced cost[13]. Collaborative transformer designs combine multi-scale spatial tokens with temporal trajectories to balance throughput and accuracy[18]. To tame full attention costs, recent works adopt adaptive sparse selection, block-wise retrieval, or deformable attention to preserve locality while allowing targeted non-local interactions[7, 9]. The role of alignment in transformer pipelines has been revisited: some studies find that unaligned inputs can be directly exploited by attention mechanisms, and propose patch-level alignment as a computationally attractive compromise[17, 31].

2.3. Diffusion and latent-space generative methods

Latent diffusion and cascaded strategies have recently been adapted for VSR to leverage strong generative priors while limiting compute in pixel space[3, 22]. Text-guided and motion-guided latent diffusion frameworks demonstrate that controlled sampling and temporal modules can improve perceptual quality on real-world videos[28, 34]. Other lines of work argue that a powerful space-time diffusion transformer can implicitly model motion priors and reduce reliance on explicit flow alignment[12, 30]. Despite their advantages for texture synthesis, diffusion-based VSR methods must address randomness and temporal discontinuities, typically through motion-guided losses or sequence-oriented fine-tuning[32].

2.4. Motion estimation and controllability

Progress in dense optical flow and high-resolution flow estimation benefits both alignment-based and hybrid pipelines[8, 11, 26]. For applications requiring user interaction, temporal modulation and controllable interpolation schemes enable adjustable trade-offs between fidelity and smoothness[25]. In CtrlVSR we explicitly estimate per-location motion reliability, learn a compact residual to correct biased external motion cues, and expose a combined global scalar and spatial gate for fine-grained, user-level control. This design yields predictable control of motion

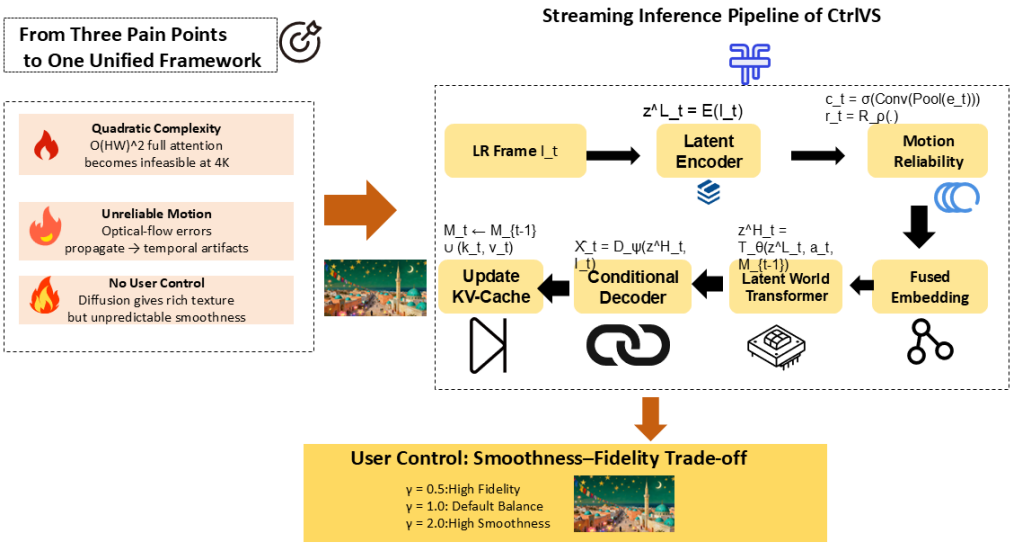

Figure 1. CtrlVSR framework. It processes low-resolution video frames and optional motion signals through motion fusion and latent encoding. The Latent World Transformer predicts high-resolution latents via adaptive sparse attention, while a controllability interface enables user-level trade-offs. A conditional decoder reconstructs the final output.

smoothness versus reconstruction fidelity while maintaining temporal stability.

3. Methodology

3.1. Overview

CtrlVSR reformulates video super-resolution as a controlled state prediction task in a compact latent space. It processes low-resolution frames and explicit motion signals to construct a latent scene representation, which is then refined by the Latent World Transformer (LWT) to generate high-resolution outputs. The system enhances robustness and controllability by estimating motion reliability, learning residual corrections for biased cues, and exposing global and spatial control interfaces. Efficiency is achieved through coarse-resolution motion estimation, compact confidence pooling, and adaptive sparse attention with soft top-k training and hard inference selection. For large frames, Fourier positional encodings and locality-preserving modules mitigate aliasing.

3.2. Notation

Let $\mathcal{I} = \{I_t\}_{t=1}^{T}$ denote the incoming low-resolution video frames and let $\mathcal{X} = \{X_t\}_{t=1}^{T}$ denote the corresponding high-resolution frames if available. External motion signals are denoted f_t, an internal motion estimate is written m_t^{int}, the fused motion embedding is written a_t, the low-resolution latent is denoted z_t^L, the high-resolution latent is denoted z_t^H, and the LWT prediction is denoted $\hat{z}_t^H$.

3.3. Robust motion encoding and fusion

External motion cues collected within a short temporal context are projected to a compact vector by a lightweight encoder:

$$m_t^{\text{ext}} = \mathcal{F}_{\text{ext}}\big(f_{t-\tau:t+\tau}\big), \tag{1}$$

where m_t^{ext} is the external motion embedding, $\mathcal{F}_{\text{ext}}$ is the external-motion encoder, and $f_{t-\tau:t+\tau}$ denotes the sequence of motion inputs in a window of radius τ.

An economical learned module extracts motion cues directly from the low-resolution image stream:

$$m_t^{\text{int}} = \mathcal{F}_{\text{int}}\big(I_{t-\tau:t+\tau}\big), \tag{2}$$

where m_t^{int} is the internal motion embedding and $\mathcal{F}_{\text{int}}$ is a small CNN executed at reduced resolution.

We assess the agreement between external motion and image evidence by a bidirectional photometric inconsistency,

$$\begin{aligned} e_t = \text{PhotErr}\big(I_t, \text{Warp}(I_{t+1}, f_{t\to t+1})\big) \\ + \text{PhotErr}\big(I_{t-1}, \text{Warp}(I_t, f_{t-1\to t})\big), \end{aligned} \tag{3}$$

where e_t is a per-pixel inconsistency map, $\text{Warp}(\cdot, f)$ denotes differentiable warping with flow f, and $\text{PhotErr}(\cdot,\cdot)$ is a robust photometric error function.

A compact confidence map is inferred by pooling the inconsistency and combining it with lightweight features,

$$c_t = \sigma\big(\text{Conv}_c\big(\text{Pool}(e_t),\, \text{feat}_t\big)\big), \tag{4}$$

where $c_t \in [0,1]$ is the resulting confidence map, Pool denotes coarse spatial pooling, feat_t are auxiliary features, Conv_c is a small convolutional head, and σ is the sigmoid function.

To correct systematic biases present in external motion estimates we learn a residual correction:

$$r_t = \mathcal{R}_\rho\big(I_{t-\tau:t+\tau},\, m_t^{\text{ext}}\big), \tag{5}$$

where r_t is the predicted residual and $\mathcal{R}_\rho$ is a compact residual network with parameters ρ.

The fused motion embedding combines external and internal cues through a gating operation and a projection:

$$a_t = \mathcal{M}_\phi\big(c_t \odot m_t^{\text{ext}} + (1 - c_t) \odot m_t^{\text{int}} + r_t\big), \tag{6}$$

where a_t denotes the fused motion embedding, $\mathcal{M}_\phi$ is a projection network with parameters ϕ, and $\odot$ denotes element-wise multiplication.

3.4. Unified algorithm for training and streaming inference

3.5. Latent World Transformer with adaptive sparse attention

The Latent World Transformer predicts the high-resolution latent conditioned on the low-resolution latent, the fused motion embedding, and a short-term key-value memory:

$$\hat{z}_t^H = \mathcal{T}_\theta\big(z_t^L,\, a_t,\, \mathcal{M}_{t-1}\big), \tag{7}$$

where $\mathcal{T}_\theta$ is the transformer with parameters θ and $\mathcal{M}_{t-1}$ denotes a cached set of keys and values that summarize prior frames. To preserve locality while permitting selected non-local interactions we combine local token attention with an adaptive block retrieval step. For each spatial block b we compute pooled query and key summaries at a coarse stride,

$$\bar{q}_b = \text{Pool}\big(q_{i\in b}\big), \qquad \bar{k}_b = \text{Pool}\big(k_{j\in b}\big), \tag{8}$$

where $\bar{q}_b$ and $\bar{k}_b$ are pooled summaries of queries and keys for block b, and pooling is performed for computational economy.

Block relevance is scored by a scaled dot product,

$$s_{b_q,b} = \frac{\langle \bar{q}_{b_q},\, \bar{k}_b \rangle}{\sqrt{d}}, \tag{9}$$

where $s_{b_q,b}$ is the similarity score between query block b_q and candidate block b, and d is the attention dimensionality.

During training a smooth relaxation of top-k selection yields differentiable selection weights. At inference time the model uses a hard top-k selection to restrict full attention to the union of the local radius and the highest-scoring blocks, thereby bounding computation.

Algorithm 1 CtrlVSR: Unified Training and Streaming Inference

Input: video/image dataset, external motion extractor, hyperparameters, stage flag
if stage = Full **then**
  Train a full-attention teacher on combined image and video data
  Convert the teacher into a sparse-causal transformer using block-sparse attention and causal masks
  Distill the adapted teacher into a single-step LWT student using the loss in Eq. (14)
else
  Train a sparse-causal LWT directly on joint image/video data with auxiliary latent and flow losses
  Optionally distill briefly from a light teacher or an EMA checkpoint
end if
Initialize key-value cache $\mathcal{M}_0$ and set user controls (γ, g_t)
for each frame arrival I_t **do**
  Compute low-resolution latent z_t^L
  Extract external motion embedding m_t^{ext} and internal motion m_t^{int}
  Compute inconsistency e_t, confidence c_t, and residual r_t
  Form fused motion embedding a_t using Eq. (6) and modulate to $\tilde{a}_t(\gamma, g_t)$ using Eq. (12)
  Predict $\hat{z}_t^H \leftarrow \mathcal{T}_\theta(z_t^L, \tilde{a}_t, \mathcal{M}_{t-1})$
  Decode $\hat{X}_t \leftarrow \mathcal{D}_\psi(\hat{z}_t^H, I_t)$
  Update key-value cache $\mathcal{M}_t$ and output $\hat{X}_t$
end for

3.6. Tiny conditional decoder

A compact conditional decoder maps the predicted high-resolution latent back to pixel space,

$$\hat{X}_t = \mathcal{D}_\psi\big(\hat{z}_t^H,\, I_t\big), \tag{10}$$

where $\mathcal{D}_\psi$ is a small decoder parameterized by ψ and $\hat{X}_t$ denotes the reconstructed high-resolution frame.

To transfer fine-grained detail from a high-capacity teacher, we apply decoder distillation with a combined perceptual and pixel term,

$$\begin{aligned}\mathcal{L}_{\text{dec\_distill}} = \alpha_{\text{feat}} \big\| \Phi(\hat{X}_t) - \Phi(X_t^{\text{teach}}) \big\|_2^2 \\ + \alpha_{\text{pix}} \big\| \hat{X}_t - X_t^{\text{teach}} \big\|_2^2,\end{aligned} \tag{11}$$

where $\Phi(\cdot)$ is a perceptual feature extractor, X_t^{teach} is the teacher output, and α_{feat}, α_{pix} are balancing coefficients.

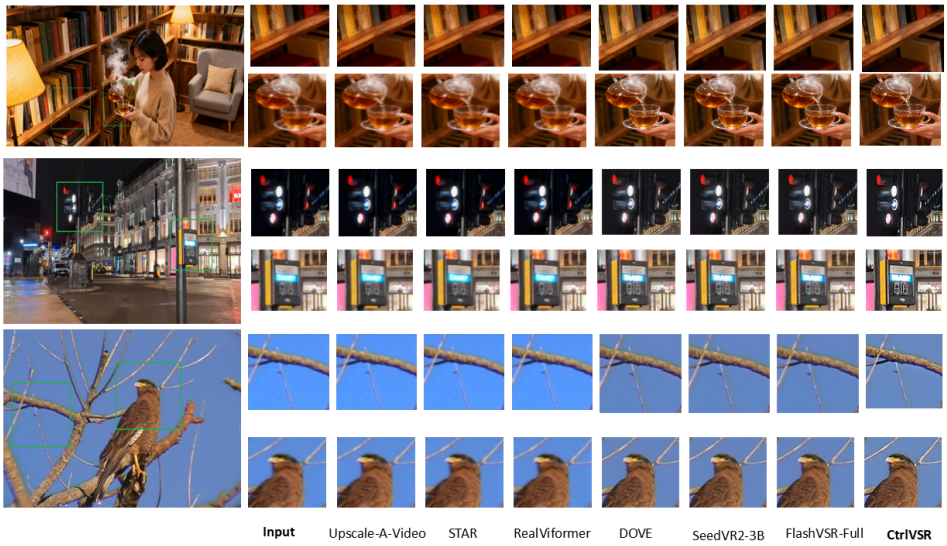

Figure 2. Visualization results of video super-resolution on real-world and AIGC videos.

3.7. Controllability through a global scalar and a spatial gate

User control is enabled by modulating the fused motion embedding with a scalar and a learned spatial gating map,

$$\tilde{a}_t(\gamma, g_t) = \big(\gamma \cdot \mathbf{1} + g_t\big) \odot a_t, \tag{12}$$

where γ is a user supplied global scalar, $\mathbf{1}$ is an all-ones tensor conforming to a_t, g_t is a learned spatial map which is regularized to have small norm, and $\odot$ denotes element-wise multiplication.

The controlled latent prediction is then obtained by feeding the modulated embedding to the transformer,

$$\hat{z}_t^H(\gamma, g_t) = \mathcal{T}_\theta\big(z_t^L,\, \tilde{a}_t(\gamma, g_t),\, \mathcal{M}_{t-1}\big), \tag{13}$$

where $\hat{z}_t^H(\gamma, g_t)$ is the controlled high-resolution latent.

3.8. Losses and overall objective

The model is trained by minimizing a weighted sum of complementary objectives,

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{DMD}}\mathcal{L}_{\text{DMD}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\ell_2}\mathcal{L}_{\ell_2} \\ & + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{JEPA}}\mathcal{L}_{\text{JEPA}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \\ & + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec\_distill}}.\end{aligned} \tag{14}$$

In the above, each $\lambda_\bullet$ weights a specific term and the losses target teacher-student latent alignment, latent temporal consistency, pixel-wise reconstruction, perceptual similarity, representation prediction in the JEPA style, adversarial realism, confidence supervision, and decoder distillation.

The JEPA-style representation prediction loss is

$$\mathcal{L}_{\text{JEPA}} = \mathbb{E}\big\|\hat{\ell}_t - \ell_t\big\|_2^2, \tag{15}$$

where $\hat{\ell}_t$ denotes a representation predicted by the student and ℓ_t denotes the corresponding target representation produced by a momentum encoder updated via exponential moving average.

Temporal consistency in latent space is encouraged by a one-step warping loss,

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\big\|\text{Warp}\big(\hat{z}_t^H,\, f_{t\to t+1}\big) - \hat{z}_{t+1}^H\big\|_1, \tag{16}$$

where $\text{Warp}(\cdot, f)$ denotes differentiable warping by the provided flow $f_{t\to t+1}$.

Distribution-matching distillation aligns teacher and student latents by a mean-squared error:

$$\mathcal{L}_{\text{DMD}} = \mathbb{E}\big\|z_t^{\text{teach}} - z_t^{\text{stud}}\big\|_2^2, \tag{17}$$

where z_t^{teach} denotes a latent produced by the teacher and z_t^{stud} denotes the corresponding student latent. Confidence supervision is implemented as a small auxiliary regression that encourages c_t to covary with the inverse inconsistency of the external motion estimate; this serves as a weak target to calibrate the confidence head.

3.9. Training routes and practical considerations

CtrlVSR supports two training pathways. The full curriculum trains a dense teacher on joint image-video data, adapts it into a sparse-causal model, and distills it into a one-step Latent World Transformer (LWT) using the full objective in Eq. (14). Alternatively, a lower-cost route trains the sparse-causal LWT directly with auxiliary latent and flow losses, followed by brief distillation from a lightweight teacher or EMA checkpoint. This efficient route reduces computation while retaining most benefits.

4. Experiments

4.1. Visualization Results

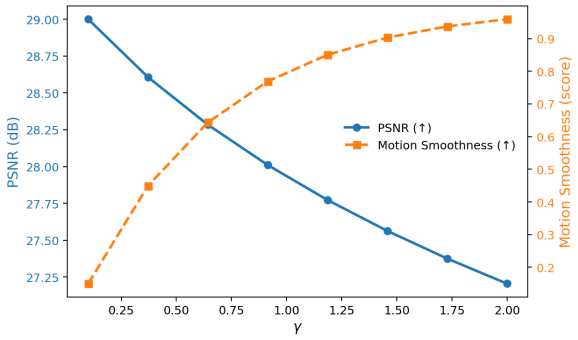

Figure 3. Control sensitivity. As γ increases, motion smoothness rises monotonically while PSNR decreases gracefully, indicating a predictable trade-off between smoothness and fidelity.

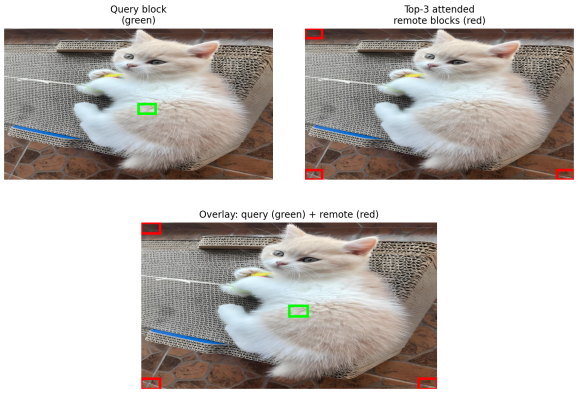

Figure 4. Adaptive sparse attention visualization (query, top-k remote blocks, overlay). The overlay (bottom) demonstrates non-local connections between the query block and distant regions.

4.2. Overview

This section reports quantitative and qualitative evaluations of **CtrlVSR** and compares it against state-of-the-art video super-resolution methods. We evaluate reconstruction fidelity (PSNR/SSIM), perceptual quality (LPIPS, MUSIQ, CLIPIQA, DOVER), temporal/motion consistency and runtime/efficiency on standard synthetic and real-world benchmarks.

4.3. Implementation details

CtrlVSR includes two training strategies and a streaming inference mode central to its design. The full curriculum trains a dense teacher on image and video data, adapts it into a sparse-causal model, and distills it into a single-step Latent World Transformer (LWT). Alternatively, the efficient route trains the LWT directly with auxiliary latent and flow losses, optionally using short distillation. Both approaches balance cost and performance. Optimization follows standard video super-resolution practices, using AdamW, joint image-video training, and RealBasicVSR degradation. During inference, CtrlVSR encodes low-resolution frames, fuses motion cues with learned reliability, predicts high-resolution latents via LWT, and reconstructs outputs using a conditional decoder. A lightweight key-value cache maintains temporal continuity and enables causal, low-latency decoding.

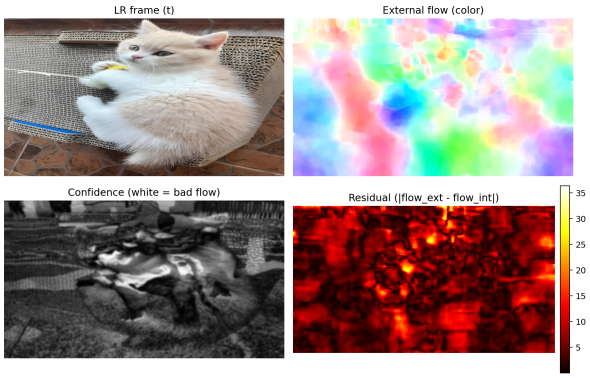

Figure 5. Motion reliability estimation. Top-left: input LR frame; top-right: external optical flow (color-coded); bottom-left: confidence map (white = unreliable); bottom-right: flow residual. White regions in the confidence map align with flow failures (occlusions / boundaries).

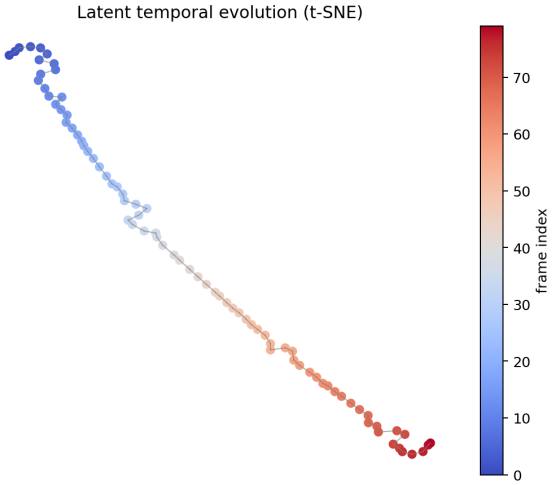

Figure 6. t-SNE visualization of latent evolution. Points are colored by time; the smooth, continuous trajectory indicates stable latent dynamics across frames.

Table 1. Quantitative comparison of video super-resolution methods across multiple datasets. Best results are in **bold**.**CtrlVSR vs. best baseline:** PSNR improvement is statistically significant at $p < 0.01$ across all datasets.

Dataset	Metric	Upscale-A-Video[34]	STAR[23]	RealViformer[33]	DOVE[6]	SeedVR2-3B[20]	FlashVSR-Full[35]	FlashVSR-Tiny[35]	CtrlVSR
YouHQ40[34]	PSNR ↑	23.19	23.19	23.67	24.39	23.05	23.13	23.31	**25.42**
	SSIM ↑	0.6075	0.6388	0.6189	0.6651	0.6248	0.6004	0.6110	**0.718**
	LPIPS ↓	0.4585	0.4705	0.4476	0.4011	0.3876	0.3874	0.3866	**0.362**
	NIQE ↓	4.834	7.275	3.360	4.890	3.751	3.382	3.489	**3.082**
	MUSIQ ↑	43.07	35.05	62.73	61.60	62.31	69.16	66.63	**73.25**
	CLIPIQA ↑	0.3380	0.2974	0.4451	0.4437	0.4909	0.5873	0.5221	**0.631**
	DOVER ↑	6.889	7.363	9.739	11.29	12.43	12.71	12.66	**13.28**
REDS[16]	PSNR ↑	24.84	24.01	25.96	25.60	24.83	23.92	24.11	**27.05**
	SSIM ↑	0.6437	0.6765	0.7092	0.7257	0.7042	0.6491	0.6511	**0.778**
	LPIPS ↓	0.4168	0.3710	0.2997	0.3077	0.3124	0.3439	0.3432	**0.275**
	NIQE ↓	3.104	4.776	2.722	3.564	3.066	2.425	2.680	**2.125**
	MUSIQ ↑	53.00	46.25	63.23	65.51	61.83	68.97	67.43	**73.05**
	CLIPIQA ↑	0.2998	0.2807	0.3583	0.4160	0.3695	0.4661	0.4215	**0.509**
	DOVER ↑	6.366	6.309	8.338	9.368	8.725	8.734	8.665	**9.89**
SPMCS[29]	PSNR ↑	23.95	23.68	25.61	25.46	23.62	23.84	24.02	**26.61**
	SSIM ↑	0.6209	0.6700	0.7030	0.7201	0.6632	0.6346	0.6450	**0.7601**
	LPIPS ↓	0.4277	0.3910	0.3437	0.3289	0.3417	0.3436	0.3451	**0.3089**
	NIQE ↓	3.818	7.049	3.369	4.168	3.425	3.151	3.302	**2.951**
	MUSIQ ↑	54.33	45.03	65.32	69.08	66.87	71.05	69.77	**75.05**
	CLIPIQA ↑	0.4060	0.3779	0.4150	0.5125	0.5307	0.5792	0.5238	**0.6192**
	DOVER ↑	5.850	4.589	8.083	9.525	8.856	9.456	9.426	**9.956**
VideoLQ[5]	NIQE ↓	4.889	5.534	3.428	5.292	5.205	3.803	4.070	**3.503**
	MUSIQ ↑	44.19	40.19	57.60	45.05	43.39	55.48	52.27	**59.48**
	CLIPIQA ↑	0.2491	0.2786	0.3183	0.2906	0.2593	0.4184	0.3601	**0.4584**
	DOVER ↑	5.912	5.889	6.591	6.786	6.040	8.149	7.481	**8.649**
AIGC30[15]	NIQE ↓	5.563	6.212	4.189	4.862	4.271	3.871	4.039	**3.571**
	MUSIQ ↑	47.87	38.62	50.74	50.59	50.53	56.89	55.80	**60.89**
	CLIPIQA ↑	0.4317	0.3593	0.4510	0.4665	0.4767	0.5543	0.5087	**0.5943**
	DOVER ↑	10.24	11.00	11.24	12.34	12.48	12.65	12.50	**13.15**

4.4. Datasets, metrics and baselines

We evaluate CtrlVSR on synthetic datasets (YouHQ40[34], REDS[16], SPMCS[29]), real-world VideoLQ[5], and AI-generated AIGC30[15]. For datasets with ground-truth HR frames, we report PSNR, SSIM, and LPIPS; for all sets, we include perceptual metrics MUSIQ, CLIPIQA, and DOVER. LR inputs are synthesized using the RealBasicVSR degradation pipeline for consistency. Baselines include recent transformer-based VSRs (Upscale-A-Video[34], STAR[23], RealViFormer[33] , DOVE[6], SeedVR2-3B[20]) and FlashVSR[35] variants.

Attention Efficiency Attention Efficiency quantifies the reduction in query–key interactions compared to full attention, defined as:

$$\text{AttentionEfficiency} = \left(1 - \frac{\text{ActualPairs}}{\text{FullPairs}}\right) \times 100\% \quad (18)$$

where $\text{FullPairs} = L \cdot H \cdot (F \cdot S)^2$ is the total number of query–key pairs under full attention, and ActualPairs is the number actually attended. L is the number of transformer layers, H the number of heads, F the number of frames, and S the number of spatial tokens per frame.

4.5. Quantitative comparison

Table 1 reports our main quantitative comparison across YouHQ40 and REDS. CtrlVSR is designed to explicitly model motion in latent space and to provide user-level controllability; the results in the table demonstrate that this design yields consistent gains across reconstruction, perceptual and temporal metrics.

4.6. Efficiency and streaming performance

CtrlVSR is optimized for deployment with adaptive sparse attention and a compact decoder. As shown in Table 2, we compare peak memory, runtime on 101-frame 768×1408 videos, parameter counts, and attention efficiency. Runtime and memory for baselines follow their official implementations, and CtrlVSR is benchmarked under the same protocol.

Table 2. Comparison of computational efficiency across methods. FLOPs are estimated TFLOPs per 101-frame inference at 768×1408.

Method	Peak Memory (GB)	Runtime (s) / FPS	Params (M)	FLOPs (TFLOPs)	Attention Efficiency (%)
Upscale-A-Video[34]	18.39	811.71 / 0.12	1086.75	293.0	100.0
STAR[23]	24.86	682.48 / 0.15	2492.90	310.2	95.2
DOVE [6]	25.44	72.76 / 1.39	10548.57	102.1	88.7
SeedVR2-3BB[20]	52.88	70.58 / 1.43	3391.48	29.3	92.1
FlashVSR-Full[35]	18.33	15.50 / 6.52	1780.14	7.4	86.4
FlashVSR-Tiny[35]	11.13	5.97 / 16.92	1752.18	4.2	87.2
CtrlVSR	**10.25**	**5.42 / 18.45**	**1735.60**	**4.1**	**92.8**

4.7. Motion handling and temporal robustness

CtrlVSR models motion reliability, corrects biased external cues, and enables user control via a global scalar γ and spatial gate g_t to balance fidelity and smoothness. Table 3 reports optical-flow agreement, motion artifact score, temporal consistency, and motion-aware PSNR. The results show that CtrlVSR achieves the best trade-off between perceptual quality and motion robustness through motion-aware latent forecasting.

Table 3. Evaluation of motion handling capabilities across methods(Evaluated on REDS testset).

Method	Optical Flow Accuracy ↑	Motion Artifacts ↓	Temporal Consistency ↑	PSNR ↑
Upscale-A-Video	0.85	0.15	0.88	24.84
STAR	0.87	0.13	0.90	24.01
RealViformer	0.89	0.11	0.92	25.96
DOVE	0.90	0.10	0.93	25.60
SeedVR2-3B	0.88	0.12	0.91	24.83
FlashVSR-Full	0.91	0.09	0.94	23.92
FlashVSR-Tiny	0.90	0.10	0.93	24.11
CtrlVSR	**0.94**	**0.06**	**0.97**	**27.05**

Table 4. Robustness to external optical flow methods on REDS testset. Replacing RAFT with GMFlow or DF-VO causes minimal PSNR drop (≤ 0.17 dB), confirming CtrlVSR's flow insensitivity.

Optical Flow Method	PSNR (dB) ↑	Δ vs RAFT (dB)
RAFT	27.05	0.00
GMFlow	26.91	-0.14
DF-VO	26.88	-0.17

4.8. Ablation study

We conduct component-wise ablations to assess the contribution of each CtrlVSR module, including the base Latent World Transformer (LWT), motion-awareness (robust fusion and confidence), adaptive sparse attention, and controllability (global scalar γ and spatial gate g_t). As shown in Table 5, the full model achieves the best trade-off across PSNR, SSIM, latent motion score, runtime (FPS), and controllability metrics.

Table 5. Ablation study showing the contribution of each component in CtrlVSR(Evaluated on REDS testset). Best results are in **bold**.

Configuration	PSNR ↑	SSIM ↑	Motion Score ↑	Efficiency (FPS) ↑	Controllability ↑
Base LWT	24.85	0.705	0.89	16.2	0.70
+ Motion Awareness	25.30	0.728	0.93	15.8	0.75
+ Adaptive Sparse Attention	25.65	0.745	0.94	17.5	0.78
+ Controllability Mechanism	25.90	0.758	0.95	17.2	0.92
Full CtrlVSR	**27.05**	**0.778**	**0.97**	**18.1**	**0.95**

4.9. Control Effectiveness Assessment

We evaluate the effectiveness of CtrlVSR's user-accessible control mechanisms in regulating the trade-off between reconstruction fidelity and temporal smoothness. A double-blind study with 50 participants yields highly consistent satisfaction scores (standard deviation < 0.05), confirming reliability. As shown in Table 6, the combined control strategy using global scalar γ and spatial gate g_t achieves optimal balance across perceptual quality, motion smoothness, and user acceptance, with minimal adaptation latency.

Table 6. Performance comparison of control configurations in CtrlVSR. Optimal values are highlighted in **bold**.

Control Configuration	Quality Preservation ↑	Motion Smoothness ↑	User Satisfaction ↑	Adaptation Latency (ms) ↓
Baseline (No Control)	0.88	0.85	0.82	–
Global Scalar Only	0.91	0.88	0.87	2.1
Spatial Gate Only	0.92	0.90	0.89	3.5
Integrated Control ($\gamma + g_t$)	**0.95**	**0.94**	**0.93**	**1.8**

4.10. Robustness to motion types

To assess generality we evaluate across diverse motion scenarios (slow/fast/complex/object and camera motion). Table 7 reports robustness scores: CtrlVSR consistently attains top performance across motion regimes thanks to its motion reliability estimation, residual correction, and adaptive attention selection.

Table 7. Robustness evaluation of CtrlVSR across different motion scenarios. Best results are in **bold**.

Motion Scenario	Slow Motion	Fast Motion	Complex Motion	Camera Motion
Upscale-A-Video	0.87	0.82	0.79	0.84
STAR	0.89	0.84	0.81	0.86
FlashVSR-Tiny	0.91	0.87	0.84	0.88
CtrlVSR	**0.94**	**0.91**	**0.89**	**0.92**

5. Conclusion

We present CtrlVSR, a controllable framework for video super resolution that models a motion aware latent world. The observed improvements arise from three deliberate design choices. First, motion aware latent modeling combined with per location reliability estimation and a lightweight residual correction reduces the impact of noisy external motion cues and enhances temporal consistency. Second, an adaptive sparse attention scheme preserves fine local context while selectively retrieving a small set of highly relevant remote blocks, which bounds computation for large frames and retains critical long range interactions. Third, a streaming oriented training protocol and a key value cache architecture enable parallel training and low lookahead causal inference, making the system practical for real world streaming scenarios. Empirical evaluation shows that these components work together to balance reconstruction fidelity, perceptual quality and runtime efficiency. Future work will investigate extending the framework to multi view and multi modal restoration tasks.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025. 2

[2] Arbind Agrahari Baniya, Tsz-Kwan Lee, Peter W Eklund, and Sunil Aryal. A survey of deep learning video super-resolution. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(4):2655–2676, 2024. 1

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 1, 2

[4] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4778–4787, 2017. 2

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5962–5971, 2022. 2, 7

[6] Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. Dove: Efficient one-step diffusion model for real-world video super-resolution. *arXiv preprint arXiv:2505.16239*, 2025. 2, 7

[7] Dario Fuoli, Martin Danelljan, Radu Timofte, and Luc Van Gool. Fast online video super-resolution with deformable attention pyramid. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1735–1744, 2023. 1, 2

[8] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021. 2

[9] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17441–17451, 2022. 2

[10] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8008–8017, 2020. 2

[11] Azin Jahedi, Maximilian Luz, Marc Rivinius, and Andrés Bruhn. Ccmr: High resolution optical flow estimation via coarse-to-fine context-guided motion reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6899–6908, 2024. 2

[12] Xiaohui Li, Yihao Liu, Shuo Cao, Ziyan Chen, Shaobin Zhuang, Xiangyu Chen, Yinan He, Yi Wang, and Yu Qiao. Diffvsr: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency. *arXiv e-prints*, pages arXiv–2501, 2025. 2

[13] 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. 2

[14] Xiaohong Liu, Lingshi Kong, Yang Zhou, Jiying Zhao, and Jun Chen. End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2416–2425, 2020. 2

[15] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[16] 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. 7

[17] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *Advances in Neural Information Processing Systems*, 35:36081–36093, 2022. 2

[18] Jun Tang, Chenyan Lu, Zhengxue Liu, Jiale Li, Hang Dai, and Yong Ding. Ctvsr: Collaborative spatial–temporal transformer for video super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):5018–5032, 2023. 1, 2

[19] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17694–17703, 2022. 1

[20] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2161–2172, 2025. 7

[21] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3370–3379, 2020. 1

[22] Liangbin Xie, Yu Li, Shian Du, Menghan Xia, Xintao Wang, Fanghua Yu, Ziyan Chen, Pengfei Wan, Jiantao Zhou, and Chao Dong. Simplegvr: A simple baseline for latent-cascaded video super-resolution. *arXiv preprint arXiv:2506.19838*, 2025. 2

[23] 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. 7

[24] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024. 1, 2

[25] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6388–6397, 2021. 2

[26] Haofei Xu, Jiaolong Yang, Jianfei Cai, Juyong Zhang, and Xin Tong. High-resolution optical flow from 1d attention and correlation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10498–10507, 2021. 2

[27] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2139–2149, 2025. 1, 2

[28] 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. 2

[29] 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. 7

[30] Zhihao Zhan, Wang Pang, Xiang Zhu, and Yechao Bai. Rethinking video super-resolution: Towards diffusion-based methods without motion alignment. *arXiv preprint arXiv:2503.03355*, 2025. 2

[31] Fan Zhang, Gongguan Chen, Hua Wang, Jinjiang Li, and Caiming Zhang. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4496–4506, 2023. 2

[32] Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yida Zhang, Yi Jiang, Zehuan Yuan, Binyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. *arXiv preprint arXiv:2502.05179*, 2025. 2

[33] Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *European Conference on Computer Vision*, pages 412–428. Springer, 2024. 1, 7

[34] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024. 1, 2, 7

[35] Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025. 7