PS-SR: Pseudo-Single-Step Video Super-Resolution via Speculative Diffusion*

Aiqiu Wu[1], Zhaofan Qiu[2], Ting Yao[2], and Tao Mei[2]
[1]University of Science and Technology of China [2]HiDream.ai Inc.
wuaiqiu@mail.ustc.edu.cn, {qiuzhaofan, tiyao, tmei}@hidream.ai

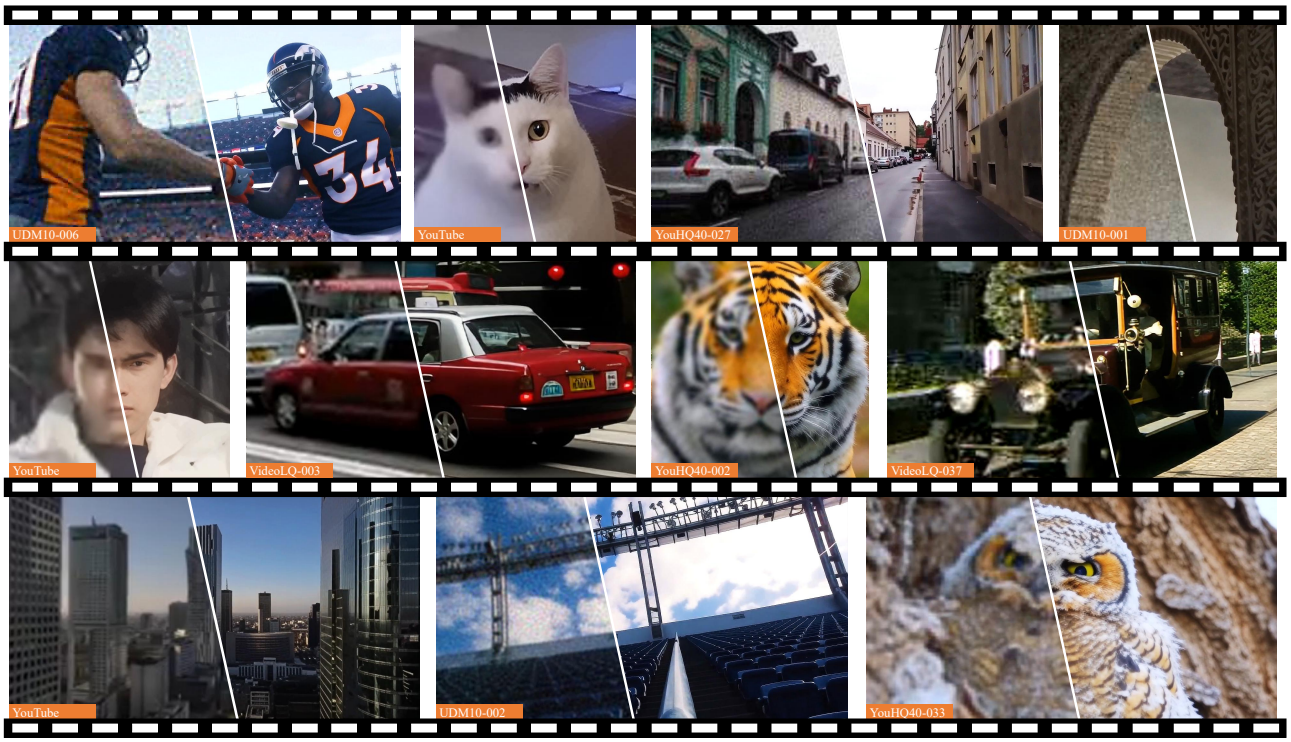

Figure 1. Examples of generated high-quality videos by our proposed PS-SR. Input videos are sampled from synthetic datasets UDM10 [52] and YouHQ40 [60], real-world dataset VideoLQ [7], and downloaded 240p videos from YouTube website.

Abstract

Video Super-Resolution (VSR) fundamentally struggles with a critical trade-off: single-step models offer unmatched efficiency but often lack the high-frequency detail, creativity, and visual quality of their multi-step diffusion counterparts, which are computationally prohibitive for practical use. In this paper, we propose PS-SR, a novel "pseudo" single-step VSR framework that transcends this trade-off through a computationally asymmetric sampling pipeline. The key to PS-SR lies in its speculative diffusion mechanism: a powerful base model performs only a single, comprehensive sampling step, establishing the global structure and content fidelity, after which a lightweight draft model, directly augmented by the base model's features, speculatively performs subsequent refinements. Crucially, we further enforce a frequency-domain update rule that constrains these refinements to exclusively inject high-frequency details, preserving the foundational low-frequency content and preventing semantic drift across sampling steps. By doing so, PS-SR creates the "illusion" of a single-step model—delivering the similar inference speeds and input-output content consistency—while achieving the visual richness and creativity typically reserved for costly multi-step generative models. We demonstrate that our "pseudo-single-step" paradigm achieves state-of-the-art quality with a comparable speed to single-step models, paving the way for real-time, high-fidelity video enhancement. Please refer to our project page for more results: https://waq2001.github.io/PS-SR-page/.

*This work was performed at HiDream.ai.

1. Introduction

Video Super-Resolution (VSR) is fundamentally challenged by a critical trade-off between efficiency and quality. Single-step models, typically based on CNNs or lightweight transformers [3, 19, 20, 24, 27, 47], offer high inference speeds suitable for real-time applications, but often fail to produce the high-frequency details and visual richness required for a compelling high-resolution experience. In contrast, multi-step diffusion models [10, 12, 15, 41, 46, 49, 60] excel at generating photorealistic and detailed videos, yet their iterative nature makes them computationally prohibitive for practical use in real life.

A recently emerged strategy to bridge this gap is model distillation, where a large, multi-step diffusion model is distilled into a single-step student model [28, 33, 45, 54]. While this approach can preserve a significant degree of the teacher model's perceptual quality, the distilled single-step model inherently lacks the creative capacity of its multi-step counterpart. The complex, iterative reasoning process that allows diffusion models to "hallucinate" plausible and diverse high-frequency details is difficult to fully capture in a single, deterministic forward pass. Consequently, distilled models often converge to safer, more averaged predictions, resulting in a noticeable drop in visual creativity and texture richness compared to the original multi-step sampling.

In this paper, we propose **PS-SR**, a novel framework that transcends this trade-off by introducing a **Speculative Diffusion** process. Instead of a full multi-step cascade or a single function call, our method distills a pre-trained diffusion model into an asymmetric, collaborative system. The core generative process of PS-SR is defined as:

$$\hat{\mathbf{x}}_H = \left(\prod^{T-1} (\mathbf{I} + \mathcal{H} \circ \phi_{\text{draft}}) \right) \circ \phi_{\text{base}}(\mathbf{x}_L). \quad (1)$$

Here, $\mathbf{x}_L$ is the low-resolution input and $\hat{\mathbf{x}}_H$ is the high-resolution output. This formulation encapsulates our two key innovations:

(1) **Computational Asymmetry**: A powerful base model ϕ_{base} performs only the first and most critical denoising step, establishing the global structure and semantic content. Subsequently, a lightweight draft model ϕ_{draft}, augmented with features from the base model, speculatively executes the subsequent $T-1$ refinement steps.

(2) **Frequency-Domain Constraint**: A high-pass filter $\mathcal{H}$ is applied to the draft model's output. This frequency-domain update rule ensures that all refinements after the first step exclusively inject high-frequency details, while the identity operator ($\mathbf{I}$) preserves the foundational low-frequency content. This prevents semantic drift and guarantees strong input-output consistency.

By materializing this asymmetric multi-step process, PS-SR creates a "pseudo-single-step" experience. It delivers inference speeds and content consistency comparable to a single-step model, while achieving the visual richness and high-frequency creativity previously reserved for costly multi-step models. We demonstrate that our paradigm achieves state-of-the-art quality, effectively bridging the efficiency-quality gap and paving the way for high-fidelity video enhancement.

Our main contributions are summarized as follows: (1) We propose PS-SR and its core Speculative Diffusion process, a novel VSR framework that uses a computationally asymmetric pipeline to break the efficiency-quality trade-off. (2) A frequency-domain update rule is introduced to confine iterative refinements to high-frequency details, ensuring strong content consistency. (3) We demonstrate that our method achieves state-of-the-art performance, matching the speed of single-step models while rivaling the visual quality of multi-step diffusion models.

2. Related Work

Early methods for video super-resolution fall into two main categories: sliding-window and recurrent approaches. Sliding-window methods [3, 24, 27, 47, 48, 52] use CNNs or Transformers to fuse a local frame sequence for reconstructing the target high-resolution frame. Recurrent methods [5, 6, 19, 26, 34, 53], by contrast, leverage hidden states to propagate temporal information sequentially across frames, enabling effective long-range modeling. While achieving an efficiency-fidelity trade-off, these methods typically lack the ability to generate the rich high-frequency details found in real high-resolution videos.

Diffusion models have recently achieved significant success [4, 30, 38, 50, 51, 54, 56, 58, 59], which has promoted their adoption in VSR tasks [11, 15, 41, 46, 49, 60]. For example, Upscale-A-Video [60] adapts an image diffusion model with temporal modules to boost inter-frame consistency and texture detail. Similarly, STAR [46] fine-tunes a video foundation model with specialized losses to enhance both spatial and temporal quality. SeedVR [41] trains a dedicated video diffusion model from scratch, showing strong visual fidelity. However, these approaches inherently rely on iterative denoising, making them computationally prohibitive for practical use.

To mitigate the inefficiency of iterative sampling, a prominent line of research focuses on **distilling multi-step diffusion models** into single-step generators. In the image domain, OSEDiff [45] leverages variational score distillation [44, 54] to align its single-step output with the distribution of a multi-step teacher model. This paradigm has been extended to video: SeedVR2 [40] employs adversarial fine-tuning [28] and progressive distillation [32] to transfer knowledge from its multi-step predecessor, SeedVR, enabling high-quality video generation in a single pass. DOVE [13] further fine-tunes a pre-trained video generation

model for single-step inference and introduces a refinement stage trained on mixed video and image data to enhance perceptual sharpness. While these methods achieve significant speed-up, the complex, iterative reasoning process of the original diffusion model is difficult to fully encapsulate in a single step, often leading to a noticeable drop in visual creativity and texture richness.

In summary, existing diffusion-based VSR methods are caught in a dichotomy: they either rely on slow, multi-step sampling to achieve high visual quality or sacrifice generative richness for the sake of efficiency via single-step distillation. Our work, PS-SR, departs from this trade-off by introducing a pseudo-single-step framework. Through a computationally asymmetric sampling pipeline and a frequency-domain constrained refinement process, our method preserves the visual richness associated with multi-step diffusion models while attaining an inference efficiency comparable to single-step models.

3. PS-SR

3.1. Preliminary

Video diffusion models. Video Diffusion Models [2, 9, 16, 17, 29, 38, 50] learn to transform a noise distribution into a video distribution through iterative refinement. A typical framework uses a 3D VAE for video compression, where an encoder $\mathcal{E}$ maps input video $\mathbf{x}_0$ to latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, and a decoder reconstructs the video. Denoising is handled by a 3D U-Net or Diffusion Transformer (DiT). In the forward process, Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$ is added to $\mathbf{z}_0$ over steps $t \in [1, T]$, yielding:

$$\mathbf{z}_t = (1 - \sigma_t)\mathbf{z}_0 + \sigma_t \varepsilon, \tag{2}$$

where $\sigma_t \in [0, 1]$ is a noise schedule. The reverse process iteratively denoises $\mathbf{z}_t$ to recover $\mathbf{z}_0$:

$$\mathbf{z}_{t-1} = \mathbf{z}_t - (\sigma_t - \sigma_{t-1})\phi(\mathbf{z}_t; t), \tag{3}$$

with $\phi(\mathbf{z}_t; t)$ as the velocity field estimated by the denoising network ϕ. Through this iterative refinement, the model reconstructs a clean video latent $\mathbf{z}_0$ from noise ε.

Flow Matching with Paired Data. For the paired data $(\mathbf{x}_L, \mathbf{x}_H)$, where $\mathbf{x}_L$ is a low-quality source and $\mathbf{x}_H$ the high-quality target, diffusion can be reformulated as conditional flow matching for improved quality and stability [1]. Given latents $(\mathbf{z}_L, \mathbf{z}_H)$, the intermediate state is:

$$\mathbf{z}_t = (1 - \sigma_t)\mathbf{z}_H + \sigma_t \mathbf{z}_L, \tag{4}$$

forming a straight-line flow from $\mathbf{z}_L$ to $\mathbf{z}_H$. The model $\phi(\mathbf{z}_t; t)$ is trained to regress the vector field driving $\mathbf{z}_L$ toward $\mathbf{z}_H$. This objective is simpler than denoising from Gaussian noise, as it learns a deterministic mapping for each pair, often leading to faster convergence and higher fidelity in reconstruction tasks like super-resolution.

3.2. Speculative Diffusion

Based on the flow matching formulation for paired data $(\mathbf{x}_L, \mathbf{x}_H)$, our proposed **Speculative Diffusion** framework elegantly extends this paradigm into an asymmetric, multi-stage refinement process. Given a paired instance $(\mathbf{z}_L, \mathbf{z}_H)$, our method decomposes the flow from $\mathbf{z}_L$ to $\mathbf{z}_H$ into a collaborative multi-model sequence. In this flow sequence, $\mathbf{z}_t$ denotes the intermediate latent in conventional flow matching, representing the gradual transition from $\mathbf{z}_L$ to $\mathbf{z}_H$ over T steps. Meanwhile, $\mathbf{x}_t$ represents the estimate of the target high-quality video at each step, which is updated through a frequency-domain update rule based on the intermediate $\mathbf{z}_t$. By synchronously updating both sequences, the final $\mathbf{x}_0$ is selected as the high-quality video output.

The process is initiated by the powerful **base model**, which performs a comprehensive initial step. This can be viewed as making a major stride along the flow, transforming the source latent significantly towards the target:

$$\begin{cases} \mathbf{z}_{T-1} = \mathbf{z}_L - (1 - \sigma_{T-1})\phi_{\text{base}}(\mathbf{z}_L; T), \\ \mathbf{x}_{T-1} = \mathcal{E}^{-1}(\mathbf{z}_{T-1} - \sigma_{T-1}\phi_{\text{base}}(\mathbf{z}_L; T)), \end{cases} \tag{5}$$

where the base model ϕ_{base} predicts the velocity field that drives $\mathbf{z}_L$ towards $\mathbf{z}_H$ for this critical first step.

Subsequently, the lightweight **draft model** takes over to perform $T-1$ speculative refinement steps. The draft model ϕ_{draft} updates both the intermediate latent and target high-quality video as:

$$\begin{cases} \mathbf{z}_{t-1} = \mathbf{z}_t - (\sigma_t - \sigma_{t-1})\phi_{\text{draft}}(\mathbf{z}_t; t), \\ \mathbf{x}_{t-1} = \mathbf{x}_t + \mathcal{H} \circ \mathcal{E}^{-1}(\mathbf{z}_{t-1} - \sigma_{t-1}\phi_{\text{draft}}(\mathbf{z}_t; t)), \end{cases} \tag{6}$$

where $\mathcal{E}^{-1}$ is VAE decoder, and $\mathcal{H}$ is a high-pass filter to ensure that only high-frequency details are added while preserving the foundational low-frequency content established by the base model.

The remainder of this section is structured as follows: we first detail the architectural designs and training objectives for both the base model (Sec. 3.3) and draft model (Sec. 3.4), then introduce the specific frequency-domain update rule to refine the high-frequency content (Sec. 3.5).

3.3. Base Model in PS-SR

Our base model ϕ_{base} is designed to restore the global structure and semantic content of a high-quality video from a low-quality input in a single diffusion step, as shown in Figure 2. To inherit rich generative and motion priors, we initialize the model from the Wan2.1 [38] foundation model and adapt it to the VSR task via Low-Rank Adaptation (LoRA) [18] fine-tuning of all DiT blocks. The model is trained on video pairs using a two-stage strategy that first optimizes the model in the latent space, followed by a fine grain stage in the pixel space to ensure high-fidelity output.

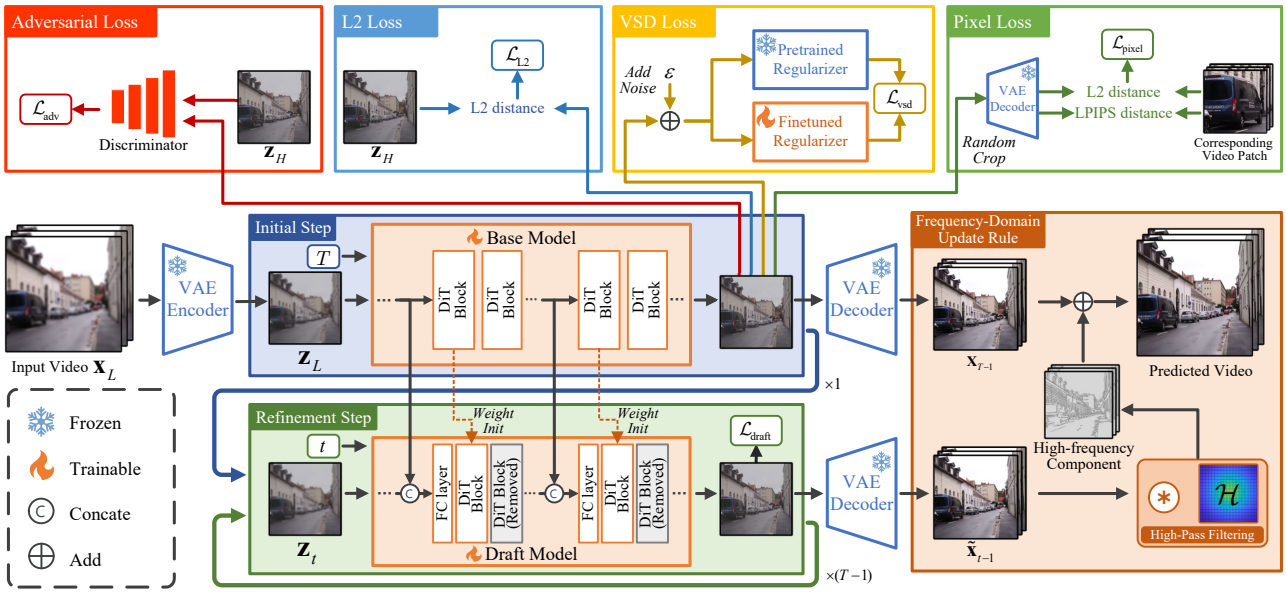

Figure 2. An overview of our proposed PS-SR framework. Given a low-quality video input $\mathbf{x}_L$, it is first encoded into a latent representation $\mathbf{z}_L$. Our computationally asymmetric sampling pipeline concludes: (1) Base Model Execution: A powerful base model ϕ_{base} performs a single, comprehensive denoising step, transforming $\mathbf{z}_L$ into an intermediate latent $\mathbf{z}_{T-1}$ that establishes the global structure and content. (2) Draft Model Refinement: The latent $\mathbf{z}_{T-1}$ is then iteratively refined over multiple steps by a lightweight draft model ϕ_{draft}. Crucially, each draft model prediction is guided by features inherited from the base model. (3) Frequency-Domain Update: After each draft step, the prediction is converted to pixel space, and our frequency-domain update rule is applied: it preserves the low-frequency content from the previous step while adaptively blending only the high-frequency components from the new prediction.

Latent-Space Training. In the first training phase, we optimize the base model ϕ_{base} directly in the latent space. Given paired latents $(\mathbf{z}_L, \mathbf{z}_H)$, the model is supervised using an L2 loss on the predicted velocity:

$$\mathcal{L}_{\text{L2}} = \mathbb{E}\left\|\phi_{\text{base}}(\mathbf{z}_L) - (\mathbf{z}_L - \mathbf{z}_H)\right\|^2. \tag{7}$$

While minimizing L2 loss provides stable training, it often leads to over-smoothed results that lack perceptual quality. To overcome this limitation, we incorporate Variational Score Distillation (VSD) [44, 45, 54] as a regularization term. This aligns the distribution of our single-step output with that of a multi-step teacher model by harmonizing predictions between a LoRA-fine-tuned version ϕ'_{reg} and a fixed pre-trained DiT regularizer ϕ_{reg}:

$$\nabla_\theta \mathcal{L}_{\text{vsd}} = \mathbb{E}_{t,\varepsilon}\left[\omega(t)\left(\phi_{\text{reg}}(\hat{\mathbf{z}}_t; t) - \phi'_{\text{reg}}(\hat{\mathbf{z}}_t; t)\right)\frac{\partial \hat{\mathbf{z}}_H}{\partial \theta}\right], \tag{8}$$

where $\hat{\mathbf{z}}_H$ is the predicted high-quality latents, $\hat{\mathbf{z}}_t = \sigma_t \hat{\mathbf{z}}_H + (1-\sigma_t)\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$, $t \sim \mathcal{U}(T_{\min}, T_{\max})$, and $\omega(t)$ is a time-dependent weighting factor.

Additionally, we introduce a latent-space adversarial loss to enhance visual realism. A VGG-16-based discriminator D is trained to distinguish between generated $\hat{\mathbf{z}}_H$ and ground-truth $\mathbf{z}_H$ latents via:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}\left[\log D(\mathbf{z}_H)\right] + \mathbb{E}\left[\log(1 - D(\hat{\mathbf{z}}_H))\right], \tag{9}$$

while the base model is simultaneously optimized to fool the discriminator, thereby improving output quality.

The overall training objective function of base model in latent-space training is comprised of the L2 loss in Eq. (7), VSD loss in Eq. (8) and adversarial loss in Eq. (9):

$$\mathcal{L}_{\text{latent}} = \lambda_{\text{L2}}\mathcal{L}_{\text{L2}} + \lambda_{\text{vsd}}\mathcal{L}_{\text{vsd}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{10}$$

where λ_{L2}, λ_{vsd} and λ_{adv} are tradeoff parameters.

Pixel-Space Training. Following latent-space convergence, we refine the base model in the pixel space to enhance visual fidelity. To maintain memory efficiency, we remove the VSD and adversarial objectives, adopting instead a patch-wise training strategy [57]. Specifically, we spatially crop the predicted high-quality latent $\hat{\mathbf{z}}_H$ into patches $\hat{\mathbf{z}}_H^{\text{crop}}$, decode them to pixel patches $\hat{\mathbf{x}}_H^{\text{crop}} = \mathcal{E}^{-1}(\hat{\mathbf{z}}_H^{\text{crop}})$, and align them with corresponding ground-truth patches $\mathbf{x}_H^{\text{crop}}$. The model is optimized using a composite objective combining pixel-wise L2 and perceptual LPIPS loss [55]:

$$\mathcal{L}_{\text{pixel}} = \lambda_{\text{L2}}\mathbb{E}\left\|\hat{\mathbf{x}}_H^{\text{crop}} - \mathbf{x}_H^{\text{crop}}\right\|^2 + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}}(\hat{\mathbf{x}}_H^{\text{crop}}, \mathbf{x}_H^{\text{crop}}). \tag{11}$$

This focused refinement significantly improves the visual quality while keeping manageable memory consumption.

3.4. Draft Model in PS-SR

While multi-step sampling in diffusion models effectively refines details, its computational demands hinder practical

deployment. Inspired by speculative sampling [8, 23, 25] in large language models, we introduce a lightweight draft model ϕ_{draft} that leverages informative features from base model to enable efficient multi-step refinement. As shown in Figure 2, the draft model adopts a simplified version of the base model architecture—initialized from ϕ_{base} and pruned by uniformly removing DiT blocks. To enhance representation capacity, features from corresponding blocks in ϕ_{base} are concatenated channel-wise with those in ϕ_{draft}, followed by a fully-connected layer to restore the original hidden dimension. Unlike the base model, the draft model is fully fine-tuned to adapt to its more complex target.

During training, the draft model takes as input an interpolated latent $\mathbf{z}_t = \sigma_t \mathbf{z}_L + (1 - \sigma_t)\mathbf{z}_H$ and predicts the velocity $\phi_{\text{draft}}(\mathbf{z}_t; t)$. It is supervised using a combined objective of L2 loss and pixel loss:

$$\mathcal{L}_{\text{draft}} = \lambda_{\text{L2}}\mathcal{L}_{\text{L2}} + \lambda_{\text{pixel}}\mathcal{L}_{\text{pixel}}. \tag{12}$$

where $\mathcal{L}_{\text{L2}}$ and $\mathcal{L}_{\text{pixel}}$ are defined in Eq. (7) and Eq. (11), respectively. Notably, we omit VSD and adversarial losses to focus the draft model on recovering high-frequency details rather than distribution-level alignment, ensuring both efficiency and enhanced detail synthesis.

3.5. Frequency-Domain Update Rule

To effectively integrate the powerful base model with the lightweight draft model, we design a frequency-domain update rule that preserves global structure and low-frequency content from the base model while progressively enhancing high-frequency details through subsequent refinement steps. This approach maintains content consistency while leveraging the creative potential of multi-step diffusion.

Given the refined video $\mathbf{x}_t$ from step t and the current step's predicted high-quality video $\widetilde{\mathbf{x}}_{t-1} = \mathcal{E}^{-1}(\mathbf{z}_{t-1} - \sigma_{t-1}\phi_{draft}(\mathbf{z}_t; t))$, we first convert both to YUV color space and extract their luminance channels $\mathbf{Y}^t$ and $\widetilde{\mathbf{Y}}^{t-1}$. High-frequency components are obtained via a high-pass filter $\mathcal{H}$: $\mathbf{Y}_H = \mathcal{H}(\mathbf{Y})$. An adaptive weighting scheme balances contributions from previous and current refinements:

$$w_t = \frac{|\widetilde{\mathbf{Y}}_H^{t-1}|}{|\mathbf{Y}_H^t| + |\widetilde{\mathbf{Y}}_H^{t-1}|}. \tag{13}$$

The updated high-frequency component is computed as:

$$\mathbf{Y}_H^{t-1} = \alpha \left(w_t \widetilde{\mathbf{Y}}_H^{t-1} + (1 - w_t)\mathbf{Y}_H^t \right), \tag{14}$$

where α controls the refinement strength. The final result $\mathbf{x}_{t-1}$ is obtained by combining this high-frequency luminance with the low-frequency components and chrominance channels from $\mathbf{x}_t$, then converting back to RGB space. This iterative process ensures both structural preservation and enhanced visual detail through multi-step refinement.

4. Experiments

4.1. Experimental Settings

Datasets. We train our PS-SR on the **YouHQ** [60] dataset, which contains approximately 37K high-quality video clips. The corresponding low-quality input videos are synthesized using the RealESRGAN degradation pipeline [42]. For evaluation, we use both synthetic and real-world datasets. The synthetic datasets include **UDM10** [52] (10 clips, 32 frames each), **SPMCS** [37] (30 clips, 31 frames each), and **YouHQ40** [60] (40 clips, around 30 frames each). Low-quality versions of these videos are generated with the same degradation pipeline as in training. For real-world evaluation, we use the **VideoLQ** [7] dataset, which consists of 50 low-quality Internet-sourced clips with 100 frames each.

Implementation Details. We implement PS-SR on the PyTorch platform. The VAE and base model are initialized from the pre-trained Wan2.1-T2V-1.3B video diffusion model [38], while the draft model is obtained by pruning 20 out of 30 DiT blocks from the fine-tuned base model. The speculative diffusion step T is set to 4, with a refinement strength factor $\alpha = 0.6$. We adopt the following loss weights from prior work [42, 45]: $\lambda_{\text{L2}} = 1$, $\lambda_{\text{vsd}} = 1$, $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{pixel}} = 1$, and $\lambda_{\text{lpips}} = 2$. For pixel loss computation, we randomly crop 160×160 patches to balance memory and performance. The model is trained on 8 NVIDIA A800 GPUs with a total batch size of 8, using AdamW optimizer with a learning rate of 5×10^{-5}, and LoRA rank set to 32.

Evaluation Metrics. We employ a range of quantitative metrics to evaluate video quality comprehensively. PSNR and SSIM [43] assess pixel-wise similarity with ground-truth video. LPIPS [55] and DISTS [14] measure perceptual similarity using VGG-based features [35]. Additionally, we include no-reference metrics: CLIP-IQA [39], MUSIQ [21], and NIQE [31], which predict quality scores without ground-truth reference videos.

4.2. Comparisons with State-of-the-Art Methods

We compare PS-SR against state-of-the-art VSR approaches, including multi-step diffusion models (STAR [46], SeedVR [41]) and single-step diffusion-based methods (DLoRAL [36], SeedVR2 [40], DOVE [13]).

Quantitative Comparisons. As summarized in Table 1, PS-SR consistently achieves top-tier performance across four datasets under diverse evaluation metrics. Notably, on UDM10, PS-SR attains an SSIM of 0.7547, an LPIPS of 0.2444 and a DISTS of 0.1277—the best among all competitors—reflecting its superior ability to reconstruct ground-truth videos. This advantage stems from our integrated training strategy: the powerful base model is optimized using VSD loss to align the output distribution with that of multi-step teacher, complemented

Table 1. Performance comparisons with state-of-the-art methods on UDM10, SPMCS, YouHQ40 and VideoLQ datasets. The best and second best results are **bolded** and underlined, respectively.

Datasets	Metric	STAR [46]	SeedVR [41]	DLoRAL [36]	SeedVR2 [40]	DOVE [13]	PS-SR (Ours)
UDM10	PSNR ↑	23.635	22.860	23.559	22.871	**24.039**	23.913
	SSIM ↑	0.7334	0.7211	0.7323	0.7349	0.7434	**0.7547**
	LPIPS ↓	0.3433	0.2796	0.2839	0.2587	0.2672	**0.2444**
	DISTS ↓	0.1730	0.1301	0.1620	0.1340	0.1569	**0.1277**
	CLIP-IQA ↑	0.2557	0.3086	0.4618	0.2907	**0.5259**	0.3716
	MUSIQ ↑	45.848	52.309	61.605	50.594	**63.140**	57.373
	NIQE ↓	5.7470	5.2708	5.1255	5.2579	5.2837	**4.9902**
SPMCS	PSNR ↑	21.437	20.738	21.175	20.378	21.281	**22.092**
	SSIM ↑	0.5653	0.5901	0.5710	0.5950	0.5802	**0.6287**
	LPIPS ↓	0.4220	0.3313	0.3797	0.3232	0.3727	**0.2940**
	DISTS ↓	0.2179	0.1461	0.1965	0.1526	0.1856	**0.1454**
	CLIP-IQA ↑	0.2749	0.3121	0.4747	0.3431	**0.5762**	0.3686
	MUSIQ ↑	45.549	52.864	65.451	58.613	**69.634**	61.004
	NIQE ↓	5.2590	4.7232	3.9897	4.3218	4.7511	**3.9542**
YouHQ40	PSNR ↑	21.076	20.508	21.440	20.250	21.589	**21.772**
	SSIM ↑	0.5525	0.5596	0.5598	0.5706	0.5741	**0.5873**
	LPIPS ↓	0.4149	0.3542	0.3156	0.3100	0.3192	**0.3011**
	DISTS ↓	0.1852	0.1435	0.1641	**0.1379**	0.1707	0.1390
	CLIP-IQA ↑	0.2871	0.3665	0.4764	0.4007	**0.5314**	0.4189
	MUSIQ ↑	49.418	57.863	66.302	61.462	**68.324**	63.001
	NIQE ↓	5.1299	4.2937	3.9482	4.2196	5.1521	**3.7508**
VideoLQ	CLIP-IQA ↑	0.2919	0.2470	**0.3910**	0.3711	0.2446	0.3155
	MUSIQ ↑	60.411	49.451	**65.119**	59.407	51.213	62.091
	NIQE ↓	4.8153	4.7222	**4.2362**	4.9334	5.0053	4.6975

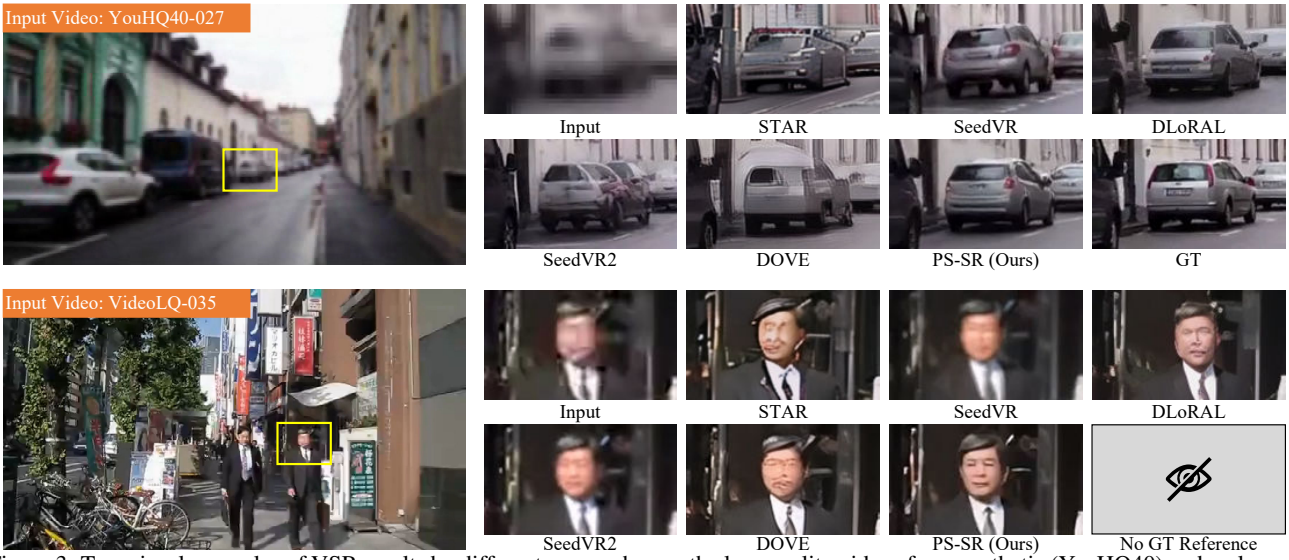

Figure 3. Two visual examples of VSR results by different approaches on the low-quality videos from synthetic (YouHQ40) and real-world (VideoLQ) datasets. The videos in VideoLQ are low-quality videos crawled from the Internet without high-quality ground-truth references.

by pixel-level supervision on randomly cropped local regions to enforce fine-grained spatial accuracy. In terms of no-reference sharpness metrics—which assess perceptual quality without ground-truth references—PS-SR achieves competitively low NIQE values. While some competing methods obtain higher values on CLIP-IQA and MUSIQ, their pursuit of extreme sharpness often comes at the cost of excessive deviation from the low-resolution input, leading to semantic drift and degraded performance on reconstruction-oriented metrics. In contrast, our frequency-constrained refinement mechanism enhances visual creativity through multi-step detail generation while preventing over-modification of low-frequency content. This enables PS-SR to maintain an optimal balance between input-output consistency and visual richness, outperforming other approaches in overall video reconstruction quality.

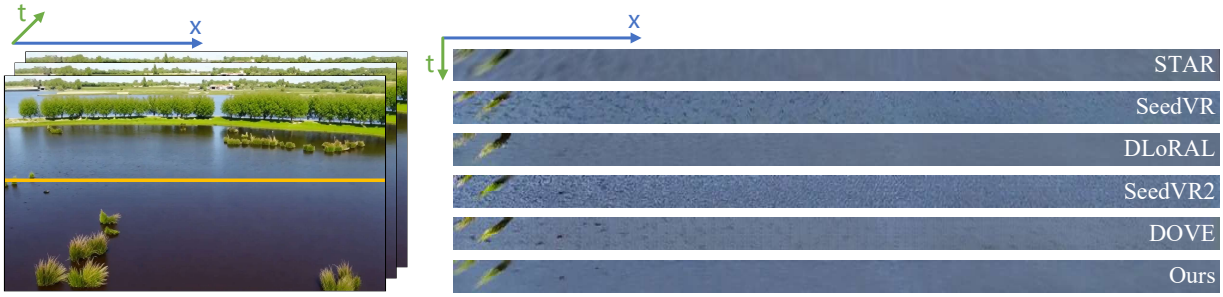

Figure 4. Comparison of temporal profile that tracks variation of the pixels in a spatial row (highlighted by the yellow line). The width of the temporal profile equals to the video width, and the height is the number of frames of generated high-quality video.

Table 2. Temporal consistency comparisons in terms of flow warping error ($E^*_{warp} \downarrow$) on different datasets.

Datasets	STAR	SeedVR	DLoRAL	SeedVR2	DOVE	Ours
UDM10	1.66	4.19	3.72	4.78	1.79	**1.43**
SPMCS	1.41	2.00	2.82	2.60	1.16	**0.82**
YouHQ40	2.98	4.22	4.79	4.55	1.84	**1.56**
VideoLQ	9.28	9.96	7.14	11.09	6.74	**6.46**

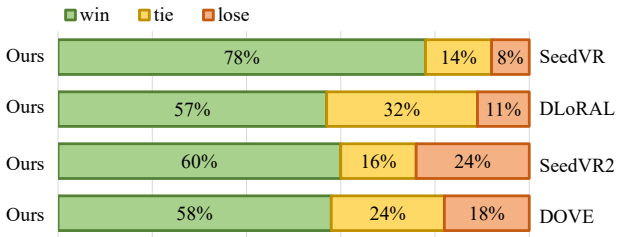

Figure 5. Human evaluation of different methods on sampled 20 input low-quality videos. Each pair is evaluated by 20 people.

Qualitative Comparisons. Figure 3 presents qualitative comparisons on representative samples from the synthetic YouHQ40 and real-world VideoLQ datasets. PS-SR consistently produces visually plausible details, particularly in semantically sensitive regions such as human faces and vehicle structures. Unlike DOVE, which generates oversharpened color patches lacking structural accuracy, or SeedVR2, which tends to yield oversmoothed results with missing textures, our method achieves an optimal balance: it preserves low-frequency content alignment with the input video while fully leveraging the creativity of multi-step refinement to reconstruct semantically faithful and detailed visual elements.

Evaluation on Temporal Consistency. We evaluate temporal consistency using the flow warping error E^*_{warp} [22]. As shown in Table 2, our method achieves the lowest warping error, confirming superior temporal coherence. Visualizations of pixel intensity transitions in Figure 4 further reveal that PS-SR maintains smoother and more continuous inter-frame evolution, while competing methods exhibit noticeable jitter. These results validate that PS-SR effectively preserves the motion priors from the initial video diffusion model, ensuring high temporal stability.

Human Evaluation. We conducted a user study with 20 participants evaluating VSR results on 20 input videos randomly sampled from four datasets. As shown in Figure 5, PS-SR is consistently preferred in pairwise comparisons for its superior visual quality and temporal smoothness. This results support the quantitative and qualitative findings, confirming that our method performs well in both objective evaluation and subjective perception.

Table 3. Performance comparisons between different variants of PS-SR on the SPMCS dataset.

Metric	w/o $\mathcal{L}_{\text{vsd}}$	w/o $\mathcal{L}_{\text{adv}}$	w/o $\mathcal{L}_{\text{pixel}}$	w/o FDU	Ours
PSNR ↑	22.097	22.165	**22.266**	18.661	22.092
SSIM ↑	0.6333	**0.6355**	0.6340	0.5299	0.6287
LPIPS ↓	0.3361	0.3448	0.3046	0.3293	**0.2940**
DISTS ↓	0.1718	0.1745	0.1483	0.1665	**0.1454**
CLIP-IQA ↑	0.3300	0.3318	0.3508	**0.4196**	0.3686
MUSIQ ↑	56.369	56.573	56.820	**67.066**	61.004
NIQE ↓	4.6710	4.5826	4.2561	**3.9313**	3.9542

Table 4. Inference speed evaluations (29 frames, 720 × 1280).

	STAR	SeedVR	DLoRAL	SeedVR2	DOVE	Ours
Step	15	50	1	1	1	1+3
Time (s/sample)	98.61	188.93	45.48	22.36	**20.43**	21.11

4.3. Model Analysis

Ablation Study. Ablation results on the SPMCS dataset (Table 3) validate the contribution of each component. Removing VSD loss ($\mathcal{L}_{\text{vsd}}$) degrades perceptual metrics (e.g., CLIP-IQA, MUSIQ), underscoring its role in aligning single-step outputs with the teacher distribution. Omitting adversarial loss ($\mathcal{L}_{\text{adv}}$) or pixel-space supervision ($\mathcal{L}_{\text{pixel}}$) also reduces visual realism. Disabling the frequency-domain update rule (FDU) improves no-reference perceptual scores but lowers PSNR/SSIM, confirming its effectiveness in preserving structural fidelity during refinement.

Inference Efficiency. As shown in Table 4, PS-SR achieves a favorable balance between quality and speed. Evaluated on an NVIDIA A800 GPU for 29-frame 720 × 1280 videos, our method introduces only minimal overhead over single-step models—thanks to the lightweight draft model—while significantly enhancing visual detail through speculative refinements.

Impact of Sampling Steps. We analyze the effect of varying speculative steps T in Table 5. While $T = 1$ yields the highest PSNR/SSIM, perceptual quality improves with more steps. We empirically set $T = 4$ for an optimal trade-

Table 5. Performance comparisons with different diffusion steps.

Metric	Speculative Diffusion (Ours)				Baseline
	$T=1$	$T=2$	$T=4$	$T=8$	$T=50$
PSNR ↑	**22.337**	22.201	22.092	21.983	20.572
SSIM ↑	**0.6418**	0.6352	0.6287	0.6210	0.5332
LPIPS ↓	0.2798	**0.2783**	0.2940	0.3111	0.3909
DISTS ↓	0.1452	**0.1430**	0.1454	0.1510	0.1702
CLIP-IQA ↑	0.2964	0.3319	0.3686	0.4005	**0.4059**
MUSIQ ↑	50.632	56.927	61.004	63.716	**71.088**
NIQE ↓	5.4465	4.6670	3.9542	3.7801	**3.2668**

Table 6. Performance comparisons between PS-SR variants with different numbers of pruned blocks in draft model.

Metric	Pruned Blocks / Total Blocks			
	0/30	10/30	20/30	25/30
CLIP-IQA ↑	**0.3767**	0.3699	0.3686	0.3353
MUSIQ ↑	**61.484**	61.347	61.004	56.936
NIQE ↓	3.9023	**3.8885**	3.9542	4.2615

Table 7. Performance comparisons between PS-SR variants with different refinement strengths.

Metric	$\alpha=0.2$	$\alpha=0.4$	$\alpha=0.6$	$\alpha=0.8$	$\alpha=1.0$
PSNR ↑	**22.272**	22.183	22.092	21.931	21.774
SSIM ↑	**0.6385**	0.6335	0.6287	0.6197	0.6118
LPIPS ↓	**0.2759**	0.2812	0.2940	0.3106	0.3266
DISTS ↓	**0.1427**	0.1428	0.1454	0.1503	0.1546
CLIP-IQA ↑	0.3125	0.3397	0.3686	0.4021	**0.4208**
MUSIQ ↑	53.899	57.583	61.004	64.358	**66.535**
NIQE ↓	4.9888	4.3543	3.9542	3.7524	**3.6929**

off. A 50-step baseline that built upon the same architecture as base model achieves top no-reference metric scores, confirming the perceptual advantage of multi-step sampling. In contrast, our approach preserves better reconstruction fidelity with far fewer steps.

Draft Model Efficiency. The number of pruned DiT blocks critically determines the efficiency-quality trade-off in our draft model. As validated in Table 6, removing 20 out of 30 blocks yields the optimal balance: insufficient pruning limits speed improvements, while excessive removal (e.g., 25 blocks) severely compromises high-frequency detail synthesis. This configuration ensures efficient refinement while preserving essential enhancement capabilities.

Frequency-Domain Update Analysis. Figure 6 demonstrates that while standard multi-step refinement leads to progressive semantic drift, our frequency-domain update rule (+FDU) effectively enhances textural details while preserving global structural integrity. In addition, we systematically evaluate the refinement strength parameter α (Table 7). As α increases from 0.2 to 1.0, we observe a systematic trade-off: perceptual metrics (CLIP-IQA, MUSIQ) improve consistently, indicating enhanced visual richness, while pixel-level similarity metrics (PSNR, SSIM) experience moderate degradation. Based on this analysis, we select $\alpha = 0.6$ as the balanced operating point that maintains strong reconstruction fidelity while achieving substantial perceptual gains.

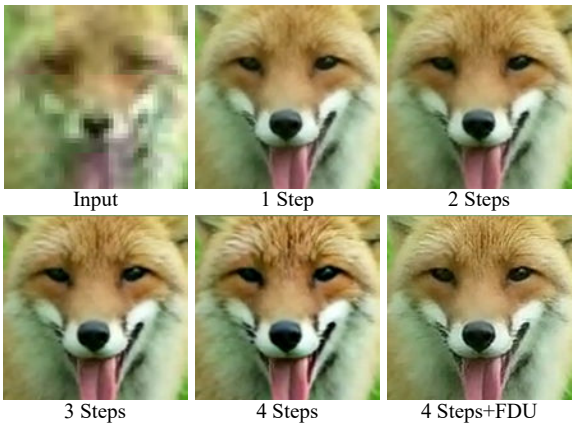

Figure 6. A visualization example of PS-SR variants by using different numbers of steps.

5. Discussion

Towards Long Video Super-Resolution. To process long videos under memory constraints, we employ an overlapping clip splitting and merging strategy. This approach divides the input into overlapping segments, enhances each independently with PS-SR, and then seamlessly integrates them using temporal position-aware averaging. More details are given in the supplementary material.

Balance Reconstruction and Creativity. Achieving optimal VSR requires balancing reconstruction and creativity. We explore this trade-off by adjusting key hyperparameters, as improving one often affects the other. Since our ultimate goal is to optimize the subjective viewing experience rather than any single metric, we base our final configuration on human-centric evaluation. Detailed parameter analyses are provided in the supplementary material.

6. Conclusion

In this paper, we present PS-SR, a novel pseudo-single-step framework that effectively bridges the efficiency-quality gap in video super-resolution. Our approach introduces two key innovations: a computationally asymmetric pipeline where a powerful base model establishes global structure in one step, followed by a lightweight draft model for efficient refinements; and a frequency-domain update rule that confines refinements to high-frequency details while preserving semantic structure. This unique integration enables PS-SR to deliver both the speed of single-step models and the visual richness of multi-step diffusion. Extensive experiments show that PS-SR achieves SOTA performance while maintaining practical efficiency. The speculative diffusion paradigm provides a new architectural blueprint for balancing computational demands with generative quality, with promising implications for various video generation tasks.

Acknowledgments. This work was supported by the Key Science & Technology Project of Anhui Province No. 202523o09050002.

References

[1] Michael S. Albergo, Mark Goldstein, Nicholas M. Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. In *ICML*, 2024. 3

[2] 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 *CVPR*, 2023. 3

[3] 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 *CVPR*, 2017. 2

[4] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv:2505.22705*, 2025. 2

[5] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*, 2021. 2

[6] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022. 2

[7] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022. 1, 5

[8] Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv:2302.01318*, 2023. 5

[9] Jingyuan Chen, Fuchen Long, Jie An, Zhaofan Qiu, Ting Yao, Jiebo Luo, and Tao Mei. Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. In *AAAI*, 2025. 3

[10] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In *CVPR*, 2024. 2

[11] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Aligning global semantics and local textures in generative video enhancement. In *ICCV*, 2025. 2

[12] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Tuning-free high-resolution video diffusion with spatial-temporal latent grouping. *IEEE TMM*, 2025. 2

[13] 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. In *NeurIPS*, 2025. 2, 5, 6

[14] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 2020. 5

[15] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv:2407.07667*, 2024. 2

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022. 3

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 3

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[19] 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 *CVPR*, 2020. 2

[20] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 2018. 2

[21] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 5

[22] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 7

[23] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *ICML*, 2023. 5

[24] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, 2020. 2

[25] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *ICML*, 2024. 5

[26] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. In *NeurIPS*, 2022. 2

[27] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE TIP*, 2024. 2

[28] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. In *ICML*, 2025. 2

[29] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *ECCV*, 2024. 3

[30] Yang Luo, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Zhineng Chen, Yu-Gang Jiang, and Tao Mei. Freeenhance: Tuning-free image enhancement via content-consistent noising-and-denoising process. In *ACM MM*, 2024. 2

[31] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 2012. 5

[32] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 2

[33] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024. 2

[34] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. In *NeurIPS*, 2022. 2

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 5

[36] Yujing Sun, Lingchen Sun, Shuaizheng Liu, Rongyuan Wu, Zhengqiang Zhang, and Lei Zhang. One-step diffusion for detail-rich and temporally consistent video super-resolution. In *NeurIPS*, 2025. 5, 6

[37] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 5

[38] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, et al. Wan: Open and advanced large-scale video generative models. *arXiv:2503.20314*, 2025. 2, 3, 5

[39] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 5

[40] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv:2506.05301*, 2025. 2, 5, 6

[41] 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 *CVPR*, 2025. 2, 5, 6

[42] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 5

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 5

[44] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 2, 4

[45] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024. 2, 4, 5

[46] 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:2501.02976*, 2025. 2, 5, 6

[47] 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 *CVPR*, 2021. 2

[48] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 2019. 2

[49] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *ECCV*, 2024. 2

[50] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv:2408.06072*, 2024. 2, 3

[51] Ting Yao, Yehao Li, Yingwei Pan, Zhaofan Qiu, and Tao Mei. Denoising token prediction in masked autoregressive models. In *ICCV*, 2025. 2

[52] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 1, 2, 5

[53] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, Xin Tian, and Jiayi Ma. Omniscient video super-resolution. In *ICCV*, 2021. 2

[54] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 2, 4

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4, 5

[56] Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. In *CVPR*, 2024. 2

[57] Ziqing Zhang, Kai Liu, Zheng Chen, Xi Li, Yucong Chen, Bingnan Duan, Linghe Kong, and Yulun Zhang. Infvsr: Breaking length limits of generic video super-resolution. *arXiv:2510.00948*, 2025. 4

[58] Zhongwei Zhang, Fuchen Long, Wei Li, Zhaofan Qiu, Wu Liu, Ting Yao, and Tao Mei. Region-Constraint In-Context Generation for Instructional Video Editing. *arXiv:2512.17650*, 2025. 2

[59] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *CVPR*, 2025. 2

[60] 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 *CVPR*, 2024. 1, 2, 5