

DUO-VSR: Dual-Stream Distillation for One-Step Video Super-Resolution

Zhengyao Lv^{1*} Menghan Xia^{2†} Xintao Wang³ Kwan-Yee K. Wong^{1†}

¹The University of Hong Kong ²Huazhong University of Science and Technology ³Kling Team, Kuaishou Technology

Project webpage: <https://cszy98.github.io/DUO-VSR/>

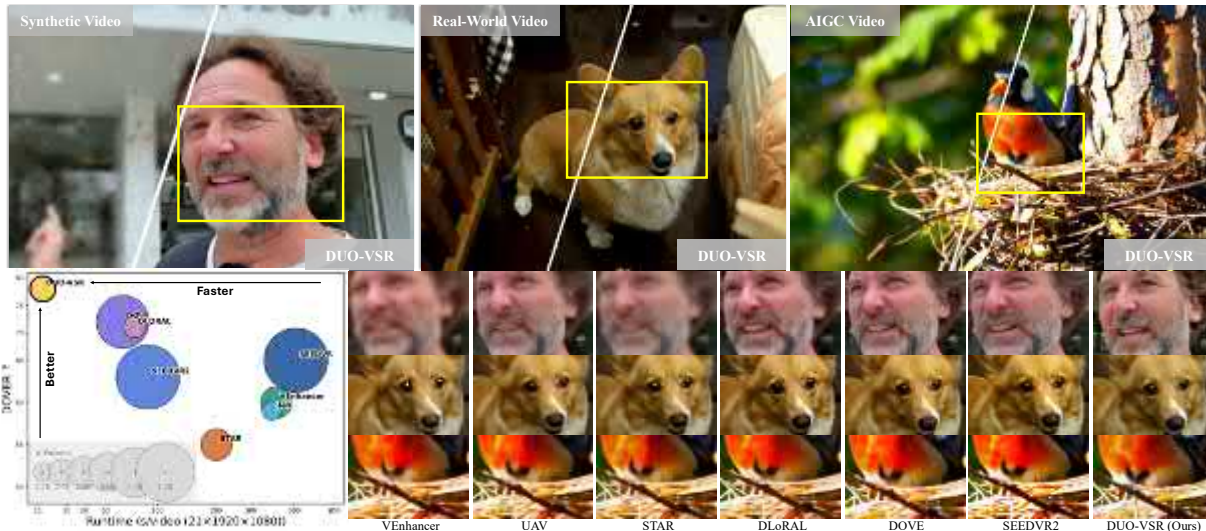


Figure 1. Inference Speed and Performance Comparison. The bubble chart on the left compares model parameter scale, inference time, and DOVER score across methods, with inference speed measured on a single GPU using a 21-frame, 1920×1080 resolution video. The right-side images show super-resolution results for different videos. Our method not only demonstrates remarkable detail generation capabilities but also achieves superior inference efficiency, accelerating inference speed by approximately $50\times$ compared to SeedVR-7B.

Abstract

Diffusion-based video super-resolution (VSR) has recently achieved remarkable fidelity but still suffers from prohibitive sampling costs. While distribution matching distillation (DMD) can accelerate diffusion models toward one-step generation, directly applying it to VSR often results in training instability alongside degraded and insufficient supervision. To address these issues, we propose **DUO-VSR**, a three-stage framework built upon a **D**ual-Stream Distillation strategy that unifies distribution matching and adversarial supervision for **O**ne-step VSR. Firstly, a Progressive Guided Distillation Initialization is employed to stabilize subsequent training through trajectory-preserving distillation. Next, the Dual-Stream Distillation jointly optimizes the DMD and Real-Fake Score Feature GAN (RFS-GAN) streams, with the latter providing complementary adversarial supervision leveraging discriminative features from both real and fake score models. Finally, a Preference-Guided Refinement stage further aligns the student with perceptual quality preferences. Extensive experiments demon-

strate that **DUO-VSR** achieves superior visual quality and efficiency over previous one-step VSR approaches.

1. Introduction

Video super-resolution (VSR) aims to recover high-resolution (HR) videos from low-resolution (LR) inputs [14, 63], serving as a fundamental technique for video quality enhancement. Beyond reconstruction-based methods [2, 3, 63], recent studies have increasingly turned to generative paradigms [64], particularly diffusion models [11, 52], which offer superior visual quality and realism. By leveraging large-scale pretrained priors [46, 80], these models achieve remarkable detail restoration even under challenging degradations [67, 73]. Despite their impressive performance, these methods rely on dozens of iterative sampling steps [61, 96], which incur substantial inference computational overhead and latency, making them impractical for real-world deployment.

A common strategy to accelerate diffusion models is to reduce the number of sampling steps [53, 83], which has been widely explored in image super-resolution (ISR) [7, 69, 87]. One line of work extends this idea to VSR by adapting one-step ISR models with temporal alignment mod-

*Work done during an internship at Kling Team, Kuaishou Tech.

†Corresponding Author.

ules [30, 54], which requires additional fine-tuning to maintain temporal consistency. Another line of research distills pretrained multi-step text-to-video (T2V) [80] or VSR [61] models into one-step generators for VSR. DOVE [6] stabilizes training with a regression loss, but it tends to compromise fine details. SeedVR2 [60] improves perceptual fidelity via adversarial post-training [26], but it often suffers from instability due to the large discriminator which may dominate the optimization dynamics and introduce unnatural artifacts. Despite these advances, one-step VSR methods still face trade-offs among stability, temporal consistency, and perceptual quality, thereby motivating the exploration of alternative distillation strategies.

Recently, Distribution Matching Distillation (DMD) [83, 84] has proven effective for accelerating video diffusion models, outperforming GAN-based counterparts [12]. It trains a student model to directly match the distribution of a pretrained teacher, thereby enabling one-step generation. However, applying DMD to VSR reveals three key limitations. **(1) Training instability.** Directly initializing the student from a pretrained multi-step VSR model produces one-step outputs whose distribution deviates substantially from real HR videos, leading to instability in subsequent training. **(2) Degraded supervision.** The frozen real score model (i.e., the teacher model), never exposed to the noised versions of the student outputs, may produce biased or spatially shifted guidance relative to the given LR anchor, causing artifacts or temporal inconsistencies. **(3) Insufficient supervision.** Although the real score model generates visually high-quality results, it still falls short of real HR videos, which fundamentally limits the achievable performance of the student when relying solely on DMD.

To address these issues, we introduce a three-stage distillation framework, featuring a novel **DUal-Stream Distillation** strategy that unifies distribution matching and adversarial supervision for **One-step VSR**, termed **DUO-VSR**. We first perform Progressive Guided Distillation to obtain a one-step initialization that stabilizes subsequent training. In the second stage, we introduce the Dual-Stream Distillation, where the distribution matching distillation stream ensures stable alignment with the teacher distribution, while the Real-Fake Score Feature GAN (RFS-GAN) stream provides supervision from high-quality real videos. Unlike DMD2 [82], which applies GAN loss only during a late fine-tuning stage and computes it solely from features of the fake score model, we jointly optimize both streams and incorporate features from both real and fake score models. The adversarial supervision from real videos serves as a regularizing signal, mitigating the adverse influence of degraded supervision from the real score model and enabling the student to achieve higher visual quality. Finally, we apply Preference-Guided Refinement to further boost perceptual quality through preference alignment optimization.

Extensive experiments demonstrate that DUO-VSR achieves superior perceptual quality over prior one-step VSR methods. Our main contributions are as follows:

- We identify the optimization challenges in applying DMD alone to one-step VSR training, namely instability and inherent degraded and insufficient supervision.
- We propose a Dual-Stream Distillation Strategy that jointly optimizes DMD and RFS-GAN losses, alleviating the adverse effects of degraded supervision and breaking the quality bound of the teacher model.
- We develop a three-stage pipeline with Progressive Guided Distillation, Dual-Stream Distillation, and Preference-Guided Refinement, enabling stable optimization and high-quality one-step video super-resolution.

2. Related Work

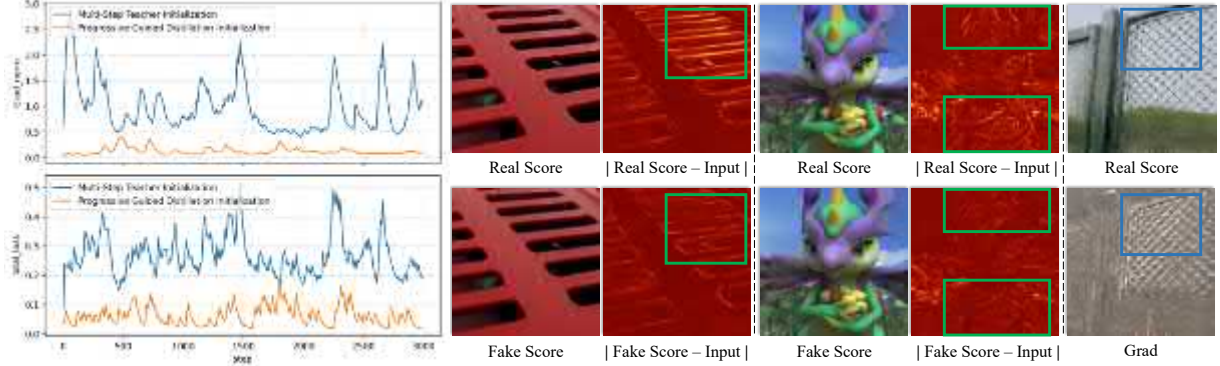
2.1. Video Super-Resolution

Video Super-Resolution (VSR) aims to recover high-quality videos from degraded inputs by leveraging spatial and temporal information. Early sliding-window-based [22, 81], recurrent-based [2, 3, 13, 24, 51], as well as other VSR methods [5, 14, 20, 25, 32, 56, 63, 75, 76, 86], mainly rely on synthetic degradation, which limits their applicability in real-world scenarios. More recent efforts [62, 71, 78, 92] have increasingly focused on addressing VSR in real-world scenarios. These works have explored various architectural designs [43, 70] and degradation pipelines [4, 64], yet they still struggle to synthesize realistic textures and fine details.

With the rapid advancement of diffusion models [11, 44, 52], several diffusion-based VSR methods [9, 23] have demonstrated remarkable performance. Some methods incorporate additional temporal modules into pretrained T2I models [46, 91] to leverage their rich priors while ensuring temporal consistency. Upscale-A-Video [96] enhances a pretrained diffusion model by integrating temporal layers and a flow-guided recurrent latent propagation module. MGLD-VSR [79] employs a motion-guided loss to guide the diffusion process and embeds a temporal module in the decoder for temporal modeling. Several other works directly leverage pretrained T2V models [1, 57, 80] for VSR. In STAR [73], fine details are recovered through a local enhancement module integrated into the model. Besides, SeedVR [61] adopts a sliding-window strategy to process long video sequences. However, the considerable parameter scale and iterative denoising of diffusion models lead to substantial latency, hindering real-world deployment.

2.2. Diffusion Model Acceleration

Acceleration methods for diffusion models typically include caching-based strategies [37, 94], efficient attention [88, 89], and distillation. Existing distillation methods for accelerating diffusion models generally fall into two main categories, namely trajectory-preserving and distribution-matching. Trajectory-preserving distillation



(a) Impact of initialization on distillation optimization

(b) Degraded supervision potentially introduced by the real score model

Figure 2. (a) Effect of initialization on the stability of the second-stage training. The proposed progressive guided distillation initialization leads to more stable loss and gradient norm trends during the second-stage distillation. (b) Compared with the fake score model, the real score model occasionally produces outputs that are spatially shifted relative to the inputs (highlighted in green boxes in the first two cases) or contain artifacts (blue boxes in the third case), leading to degraded supervision propagated to the student model.

exploits the ODE trajectory of diffusion models to match teacher outputs with fewer steps, as exemplified by methods such as progressive distillation [40, 47], consistency distillation [18, 31, 33, 38, 45, 53, 58], and rectified flow [17, 28, 29, 77]. Distribution-matching distillation bypasses the ODE trajectory and trains the student to align with the distribution of the teacher model. This can be achieved either through adversarial training [15, 36, 49, 50, 74] or through score distillation [34, 35, 82, 83, 95]. Due to the inherent difficulty of preserving diffusion trajectories in few-step settings, trajectory-preserving methods often produce blurry results, whereas distribution-matching approaches tend to yield better video quality under few-step sampling. Despite their effectiveness, GAN-based distribution matching [26] often suffers from training instability caused by heavy discriminators, whereas DMD [83] has been widely adopted in autoregressive video generation [12, 84] for its efficiency.

2.3. One-Step Video Super-Resolution

Based on these acceleration methods, recent image super-resolution (ISR) studies [7, 10, 21, 42, 48, 65, 69, 72, 85, 87, 97] have investigated efficient few-step diffusion sampling. In VSR, SEEDVR2 [60] explores applying Adversarial Post-Training (APT) [26] to VSR, enabling one-step diffusion. DOVE [6] introduces a latent-pixel training strategy that employs a two-stage scheme to adapt pretrained T2V model to one-step VSR. UltraVSR [30] introduces a degradation-aware reconstruction scheduling that reformulates multi-step denoising into a single-step process. DLoRA [54] extends ISR-based one-step models with temporal alignment. Nevertheless, existing one-step methods still exhibit limited realism and temporal consistency.

3. Methodology

3.1. Preliminary

Base VSR Model. Given a low-resolution video x^{LR} , we first upscale it to the target resolution and then encode

it into the latent space using a VAE \mathcal{E} to obtain its latent representation z^{LR} . We train a video diffusion transformer (DiT) [44] conditioned on z^{LR} and text embedding c , to predict clean HR latents from noisy samples obtained by perturbing HR latents z^{HR} with random noise ϵ

$$z_t^{HR} = (1-t)z_0^{HR} + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (1)$$

where $t \in [0, 1]$. The VSR denoiser v_θ with parameter θ is trained to predict the target velocity $v = \epsilon - z_0^{HR}$

$$\mathcal{L}(\theta) = \mathbb{E}_{t, z_0^{HR}} \|v_\theta(z_t^{HR}, t, z^{LR}, c) - v\|^2. \quad (2)$$

Following previous works [61], we concatenate the noisy HR latents z_t^{HR} and the LR latents z^{LR} as the input to the DiT. Similar to recent advanced video DiT models [57, 80], our DiT layers incorporate a cross-attention module to integrate textual conditioning, as well as 3D full attention to capture long-range spatial and temporal dependencies. Our base VSR model, containing about one billion parameters, requires 50 sampling steps by default to generate clean high-resolution videos.

Distribution Matching Distillation (DMD). DMD [82, 83] distills a multi-step diffusion model into a one-step student generator by minimizing the expected approximate Kullback-Leibler (KL) divergence D_{KL} between the diffused target and student distributions over timesteps t .

Given a pretrained diffusion model, the distribution score can be formulated as $s = -\frac{z_t^{HR} + (1-t)v_\theta}{t}$ [52], allowing the student parameters θ_s to be optimized by directly computing the gradient of the KL divergence

$$\nabla_{\theta} D_{KL} = \mathbb{E}_{\epsilon} \left[-\left(s_{\text{real}}(z_t^{HR}) - s_{\text{fake}}(z_t^{HR}) \right) \frac{dv}{d\theta_s} \right], \quad (3)$$

where s_{real} and s_{fake} are computed by the real and fake score models, respectively. Both models are initialized with the same architecture and weights as the teacher model. The real score model is frozen during training to capture the teacher distribution, while the fake score model is continuously updated to track the student distribution.

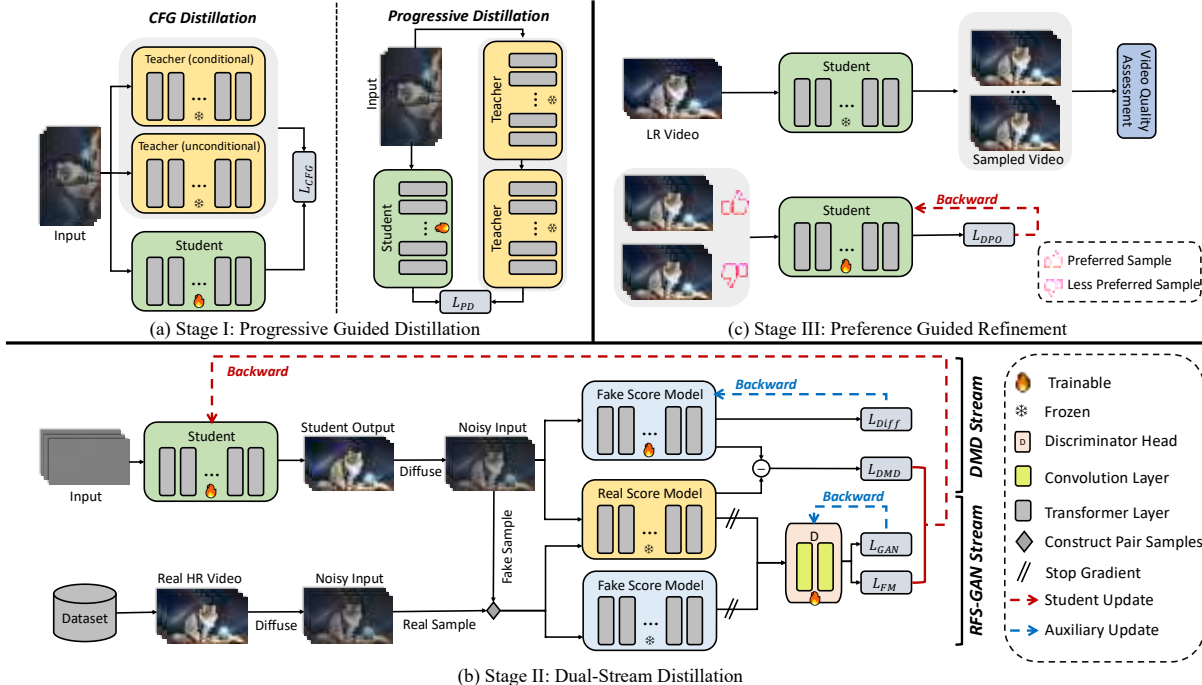


Figure 3. Overview of our three-stage distillation framework. (a) We initialize the student model with trajectory-preserving Progressive Guided Distillation, which consists of CFG Distillation and Progressive Distillation steps. (b) The core of our method, Dual-Stream Distillation, jointly optimizes the DMD and RFS-GAN streams through alternating Student Update and Auxiliary Update, providing reliable and sufficient supervision. (c) In the final stage, we construct a generated preference dataset and apply DPO-based Preference-Guided Refinement to enhance perceptual quality.

3.2. DMD in VSR: On Stability and Supervision

Despite the impressive performance of DMD in image and video generation, we observe that directly applying it to one-step VSR training faces several challenges. First, DMD initializes the student, real score, and fake score models from the pretrained multi-step VSR model. Since the pre-trained model yields low-quality results under the one-step setting, its distribution differs notably from that of the real score model, causing unstable optimization and degraded results. As shown in Fig. 2 (a), directly initializing from the teacher model results in unstable gradients and training dynamics. Second, the real score model has never been exposed to the noisy outputs of the student model. Compared with the fake score model that continuously tracks the outputs of the student model, the real score model generates results with richer high-frequency details and textures, but often exhibit undesired spatial shifts relative to the inputs. Moreover, it occasionally produces artifact-contaminated outputs, which can be further propagated to the student through gradient updates, as illustrated in Fig. 2 (b). These issues are particularly evident in VSR, where the LR video serves as a strong spatial-temporal anchor, making the system more sensitive to degraded supervision than in text-conditioned image or video generation. Finally, while the real score model represents a high-quality distribution, it remains inferior to real HR videos. Consequently, relying solely on the DMD loss restricts the student to limited rep-

resentational capacity of the teacher model.

To mitigate instability during training, we propose the Progressive Guided Distillation Initialization in Sec. 3.3. To alleviate the adverse effects of degraded and insufficient supervision, we introduce the Dual-Stream Distillation Strategy in Sec. 3.4. Finally, to further enhance the perceptual quality of the generated videos, we incorporate a Preference-Guided Refinement stage in Sec. 3.5.

3.3. Progressive Guided Distillation Initialization

Having identified the instability caused by direct one-step distillation, we adopt a trajectory-preserving Progressive Guided Distillation Initialization to provide a stable foundation for subsequent dual-stream optimization.

Specifically, following [40], we first train a single model θ_S to match the combined output of the conditional and unconditional diffusion branches (CFG-Distillation in Fig. 3 (a)). This can be formulated as

$$\mathbf{v}_{\text{cfg}} = (1 + w)\mathbf{v}_\theta(z_t^{\text{HR}}, t, z^{\text{LR}}, \mathbf{c}) - \mathbf{v}_\theta(z_t^{\text{HR}}, t, z^{\text{LR}}, \emptyset)$$

$$\mathcal{L}_{\text{CFG}}(\theta_S) = \mathbb{E}_{t, z_0^{\text{HR}}} \|\mathbf{v}_{\theta_S}(z_t^{\text{HR}}, t, z^{\text{LR}}, \mathbf{c}) - \mathbf{v}_{\text{cfg}}\|^2. \quad (4)$$

We then treat the CFG-Distilled \mathbf{v}_{θ_S} as the teacher model and progressively distill it into a one-step student (Progressive Distillation in Fig. 3 (a)):

$$\mathcal{L}_{\text{PD}}(\theta_S) = \mathbb{E}_{t, z_0^{\text{HR}}} \left\| \underbrace{z_t^{\text{HR}} - (t - t')\mathbf{v}_{\theta_S}(z_t^{\text{HR}})}_{\text{student}} - \underbrace{z_{t'}^{\text{HR}}(\theta)}_{\text{teacher}} \right\|^2, \quad (5)$$

where $\mathbf{v}_{\theta_s}(\mathbf{z}_t^{HR})$ is the predicted velocity of student model at timestep t , and $\hat{\mathbf{z}}_{t''}^{HR}(\theta)$ denotes the two-step prediction at timestep t'' obtained by integrating the teacher model over timesteps (t, t', t'') . For simplicity, the conditions like text embeddings and timesteps are omitted from the notations.

3.4. Dual-Stream Distillation Strategy

Building upon a stable initialization, we further address the degraded and insufficient supervision in DMD by introducing a Dual-Stream Distillation Strategy that unifies distribution matching (DMD Stream) and adversarial supervision (RFS-GAN Stream), as shown in Fig. 3 (b).

DMD Stream. For the distribution matching distillation stream, we follow the setting in [83] and initialize both the real score model θ_R and fake score model θ_F from the pre-trained teacher model. The real score model remains frozen to capture the distribution of high-quality videos, while the fake score model is updated to track the evolving distribution of the one-step student. During training, we optimize the fake score model θ_F with diffusion loss \mathcal{L}_{Diff} :

$$\hat{\mathbf{z}}_0^S = \epsilon - \mathbf{v}_{\theta_s}(\epsilon, t, \mathbf{z}^{LR}, \mathbf{c})$$

$$\mathcal{L}_{Diff}(\theta_F) = \mathbb{E}_{t, \hat{\mathbf{z}}_0^S} \|\mathbf{v}_{\theta_F}(\hat{\mathbf{z}}_t^S, t, \mathbf{z}^{LR}, \mathbf{c}) - \mathbf{v}\|^2, \quad (6)$$

where $\hat{\mathbf{z}}_0^S$ represents the latent of the HR video predicted by one-step student model, and $\hat{\mathbf{z}}_t^S$ is obtained by diffusing it. Meanwhile, we alternately optimize the student model θ_s using the DMD loss \mathcal{L}_{DMD} :

$$\text{Grad} = \frac{\hat{\mathbf{z}}_0^F(\hat{\mathbf{z}}_t^S; \theta_F) - \hat{\mathbf{z}}_0^R(\hat{\mathbf{z}}_t^S; \theta_R)}{\text{mean}(\text{abs}(\hat{\mathbf{z}}_0^S - \hat{\mathbf{z}}_0^R(\hat{\mathbf{z}}_t^S; \theta_R)))}$$

$$\mathcal{L}_{DMD}(\theta_s) = \mathbb{E}_{t, \hat{\mathbf{z}}_0^S} \|\hat{\mathbf{z}}_0^S - \text{sg}[\hat{\mathbf{z}}_0^S - \text{Grad}]\|^2, \quad (7)$$

where $\hat{\mathbf{z}}_0^R$ and $\hat{\mathbf{z}}_0^F$ are the outputs of the real and fake score models, corresponding to the real and fake scores respectively, and $\text{sg}(\cdot)$ is the stop-gradient operator.

RFS-GAN Stream. In the Real-Fake Score Feature (RFS-GAN) stream, we employ both the frozen Real Score and the Fake Score models as discriminator backbone to extract features. The backbone takes the diffused output of the one-step student $\hat{\mathbf{z}}_t^S$ as fake samples and the diffused HR video \mathbf{z}_t^{HR} as real samples, sharing the same conditioning inputs as the DMD stream, including LR video and corresponding timestep. The intermediate features from transformer layers are concatenated and fed into additional convolutional discriminator heads to compute the RFS-GAN loss, adopting a hinge GAN objective for stable training:

$$\mathcal{L}_D = \mathbb{E}[\max(0, 1 - D(\mathbf{z}_t^{HR}))] + \mathbb{E}[\max(0, 1 + D(\hat{\mathbf{z}}_t^S))]$$

$$\mathcal{L}_G = -\mathbb{E}[D(\hat{\mathbf{z}}_t^S)]. \quad (8)$$

To further stabilize training, we introduce a feature matching loss \mathcal{L}_{FM} computed as the mean squared error between intermediate features extracted from the score models.

Dual-Stream Joint Optimization. To exploit the complementary strengths of DMD and adversarial supervision, we

perform dual-stream joint optimization over the student, fake score model, and convolutional discriminator heads, as illustrated in Fig. 3 (b). We alternate between two interleaved optimization phases: **(a) Student update**, where the one-step student is updated jointly by the \mathcal{L}_{DMD} , \mathcal{L}_G , and \mathcal{L}_{FM} losses; and **(b) Auxiliary update**, where the fake score model and discriminator heads are separately updated with diffusion loss \mathcal{L}_{Diff} and GAN objective \mathcal{L}_D . We apply a stop-gradient between backbone features and discriminator heads to prevent GAN gradients from affecting the score models during discriminator head updates. The detailed algorithm is provided in the supplementary material.

This joint formulation constitutes the core of our framework and delivers two interrelated benefits. **(1) Reliable and comprehensive supervision.** The RFS-GAN stream regularizes and complements the degraded and insufficient DMD supervision. It suppresses the biased gradients induced when the frozen real score model encounters unseen noisy student outputs, and introduces real-video adversarial signals that enrich and extend the guidance beyond the teacher distribution. By leveraging features from both real and fake score models, the adversarial supervision becomes more complete and balanced. **(2) Stability and efficiency.** Operating on diffused samples, RFS-GAN naturally benefits from shared partial forward passes with the DMD stream, improving computational efficiency while the injected noise stabilizes adversarial dynamics. In addition, the stop-gradient between the score-model backbones and discriminator heads decouples their optimization, ensuring that adversarial gradients do not interfere with the distribution tracking of score models. Together, it enables a steady, efficient, and well-regularized joint training process that integrates the strengths of both streams.

3.5. Preference-Guided Refinement

To further enhance the perceptual quality of the one-step VSR student, we introduce a Preference-Guided Refinement, as illustrated in Fig. 3 (c). The second-stage student model generates multiple HR candidates for each LR video, which are ranked by video quality assessment models to form a synthetic preference dataset $\mathcal{D} = \{\mathbf{z}^{LR}, \hat{\mathbf{z}}_0^{S^w}, \hat{\mathbf{z}}_0^{S^l}\}$, with $\hat{\mathbf{z}}_0^{S^w}$ preferred over $\hat{\mathbf{z}}_0^{S^l}$. The student model is then fine-tuned with Direct Preference Optimization (DPO) [27] loss \mathcal{L}_{DPO} to better align with perceptual preferences:

$$-\mathbb{E}[\log\sigma(-\frac{\beta_t}{2}(\|\mathbf{v}^w - \mathbf{v}_{\theta_s}(\hat{\mathbf{z}}_t^{S^w})\|^2 - \|\mathbf{v}^w - \mathbf{v}_{\theta_{\text{ref}}}(\hat{\mathbf{z}}_t^{S^w})\|^2 - (\|\mathbf{v}^l - \mathbf{v}_{\theta_s}(\hat{\mathbf{z}}_t^{S^l})\|^2 - \|\mathbf{v}^l - \mathbf{v}_{\theta_{\text{ref}}}(\hat{\mathbf{z}}_t^{S^l})\|^2)))]), \quad (9)$$

where β_t is a hyperparameter and $\mathbf{v}_{\theta_{\text{ref}}}$ is the reference model. This loss encourages \mathbf{v}_{θ_s} to approach the target velocity \mathbf{v}^w of the preferred data, while repelling it from \mathbf{v}^l associated with the less preferred data. This refinement further aligns the one-step generator with perceptual preferences, yielding high-fidelity video results.



Figure 4. Visual comparison on synthetic (YouHQ40), real-world (VideoLQ) and AIGC (AIGC60) datasets. Zoom in for details.

4. Experiments

4.1. Experimental Settings

Implementation Details. Our base VSR model is built upon an internal 1.3B-parameter text-to-video model, which is adapted through 10k iterations of training on 830k paired samples synthesized by RealBasicVSR [4] degradation pipeline, with a batch size of 64. In the Progressive Guided Distillation stage, we first perform CFG Distillation for 500 iterations. Next, starting from a 64-step teacher, we progressively halve number of denoising steps of student, using a learning rate of 5×10^{-5} and a batch size of 32. Meanwhile, teacher is updated with the latest student every 500 iterations, until obtaining a single-step model. In the Dual-Stream Distillation stage, we perform one student update after every three auxiliary updates, iterating for 2,000 steps in total. The DMD loss, RFS-GAN loss, and feature matching loss are weighted by 1.0, 0.1, and 0.05 respectively. The learning rate and batch size are set to 5×10^{-6} and 32 respectively. In the Preference-Guided Refinement stage, we construct 2,000 preference pairs and fine-tune the model for 1,000 iterations with a learning rate of 1×10^{-6} .

Evaluation Settings. Following previous work [61], we conduct evaluations on synthetic benchmarks including SPMCS [81], UDM10 [55], and YouHQ40 [96] under the same degradation settings as in training. Furthermore, we evaluate on a real-world dataset VideoLQ [4] and a self-constructed AIGC60 dataset comprising 60 AI-generated videos covering a wide range of visual scenes.

For synthetic datasets, we evaluate the fidelity using full-reference metrics including PSNR, SSIM [66], and LPIPS [90]. To further assess perceptual quality, we report no-reference metrics such as NIQE [41], CLIP-IQA [59], MUSIQ [16], and DOVER [68]. We also employ the flow warping error E_{warp}^* (scaled by 10^{-3}) [19] to evaluate temporal consistency. For real-world (VideoLQ) and AIGC (AIGC60) datasets, where ground-truth HR videos are unavailable, we rely solely on no-reference metrics and E_{warp}^* for evaluation.

4.2. Comparison with Prior Works

We compare our DUO-VSR with several recent state-of-the-art video super-resolution (VSR) models, including RealVformer [92], VEnhancer [9], MGLD [79], UAV [96], STAR [73], DLoRAL [54], DOVE [6], and SEEDVR2 [60].

Qualitative Comparison. Fig. 1 and Fig. 4 present qualitative comparisons with various methods on synthetic, real-world, and AIGC video datasets. DUO-VSR demonstrates strong capability in reconstructing realistic textures and structures under diverse and challenging degradations. For example, in Fig. 4, the first row shows that DUO-VSR successfully restores a visually convincing brick-wall pattern; in the second row, it reconstructs a clear human face even under severe degradation; and in the last row, it produces fine-grained, natural fur. The temporal profiles visualized in Fig. 5 illustrate the comparison of temporal consistency. Under severely degraded LR inputs, existing methods tend to produce noticeable misalignment or blurring, whereas

Table 1. Quantitative comparisons on benchmarks, including synthetic (SPMCS [55], UDM10 [81], YouHQ40 [96]), real-world (VideoLQ [4]), and AIGC (AIGC60) videos. The best and second performances are marked in red and blue respectively.

Datasets	Metrics	RealViformer	VENhancer	MGLD	UAV	STAR	DLoRAL	DOVE	SeedVR2-7B	DUO-VSR
SPMCS	PSNR \uparrow	21.34	19.92	21.89	19.67	21.56	22.87	22.69	23.08	22.90
	SSIM \uparrow	0.601	0.523	0.642	0.538	0.613	0.674	0.694	0.685	0.691
	LPIPS \downarrow	0.394	0.417	0.348	0.424	0.386	0.331	0.318	0.302	0.315
	NIQE \downarrow	4.67	4.36	3.79	3.87	6.17	3.96	4.79	4.68	3.59
	MUSIQ \uparrow	62.13	61.10	63.43	58.52	29.08	61.42	65.27	64.74	66.91
	CLIP-IQA \uparrow	0.3443	0.3398	0.4333	0.4582	0.3827	0.5218	0.4922	0.5073	0.5459
	DOVER \uparrow	61.32	58.25	75.54	69.84	35.17	72.30	79.94	75.40	81.47
E_{warp}^* \downarrow	4.45	3.92	4.04	4.99	6.53	3.45	2.10	2.96	1.67	
UDM10	PSNR \uparrow	23.75	23.38	23.89	23.16	23.97	24.83	24.32	24.56	24.94
	SSIM \uparrow	0.638	0.612	0.667	0.607	0.659	0.739	0.723	0.745	0.726
	LPIPS \downarrow	0.364	0.398	0.385	0.401	0.324	0.272	0.284	0.267	0.259
	NIQE \downarrow	5.17	4.99	4.93	4.64	5.79	4.65	4.51	4.46	4.07
	MUSIQ \uparrow	59.11	54.51	58.71	59.53	48.21	56.96	54.51	51.89	62.25
	CLIP-IQA \uparrow	0.4134	0.3859	0.4047	0.4002	0.2636	0.4163	0.4412	0.4346	0.4898
	DOVER \uparrow	70.51	73.48	71.35	65.72	56.72	62.03	74.87	69.09	75.32
E_{warp}^* \downarrow	3.43	3.65	3.81	3.76	2.96	3.89	2.89	3.12	2.44	
YouHQ40	PSNR \uparrow	20.98	19.23	21.35	18.97	21.76	22.57	23.12	22.87	22.96
	SSIM \uparrow	0.621	0.543	0.661	0.564	0.642	0.658	0.682	0.691	0.674
	LPIPS \downarrow	0.379	0.426	0.355	0.433	0.364	0.311	0.299	0.291	0.289
	NIQE \downarrow	5.30	4.97	4.52	4.11	5.69	4.18	4.91	4.86	3.92
	MUSIQ \uparrow	54.86	56.83	57.30	57.52	49.19	59.64	61.36	58.44	65.24
	CLIP-IQA \uparrow	0.3185	0.3203	0.4106	0.4028	0.2871	0.4188	0.4167	0.3736	0.4222
	DOVER \uparrow	71.13	75.36	81.23	83.68	50.96	68.32	84.43	73.60	87.28
E_{warp}^* \downarrow	3.63	3.54	3.32	2.68	2.09	2.71	2.37	2.54	1.98	
VideoLQ	NIQE \downarrow	5.52	5.38	4.78	4.77	5.16	5.17	4.43	4.63	4.08
	MUSIQ \uparrow	49.20	46.21	50.87	51.47	45.90	59.08	51.25	55.45	59.24
	CLIP-IQA \uparrow	0.3221	0.3106	0.3633	0.3460	0.2753	0.4068	0.3209	0.3387	0.3925
	DOVER \uparrow	61.09	58.87	65.36	62.57	63.43	69.29	69.36	59.56	69.71
	E_{warp}^* \downarrow	5.03	4.93	4.10	4.82	4.29	4.46	3.91	4.08	3.67
AIGC60	NIQE \downarrow	6.61	5.86	5.56	5.73	5.28	5.44	5.47	4.99	4.42
	MUSIQ \uparrow	50.99	50.85	53.82	52.34	57.52	58.65	57.89	62.30	63.68
	CLIP-IQA \uparrow	0.3464	0.3933	0.4203	0.4209	0.3509	0.4668	0.4061	0.4376	0.4886
	DOVER \uparrow	84.55	83.65	83.52	83.50	87.32	87.89	87.61	86.79	88.15
E_{warp}^* \downarrow	3.89	3.76	3.40	4.17	1.22	1.67	0.95	1.48	1.08	

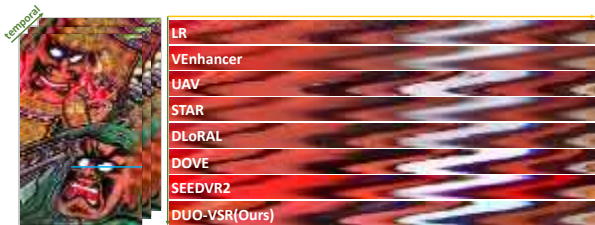


Figure 5. Comparison of temporal consistency. Extracted and stacked along the blue line in the width-temporal plane.

our DUO-VSR achieves a good balance between detail enhancement and temporal coherence. More results are provided in the supplementary materials.

Quantitative Comparison. We present quantitative comparisons in Tab. 1 and Tab. 2. As can be seen, DUO-VSR consistently achieves the highest or near-highest scores on non-reference perceptual metrics such as NIQE and MUSIQ across all datasets, demonstrating its superior perceptual quality. In terms of fidelity metrics, our method attains performance comparable to competing approaches. Moreover,

DUO-VSR exhibits highly stable and consistent results in temporal coherence (E_{warp}^*). In terms of efficiency, Tab. 2 shows that DUO-VSR maintains low inference latency with a relatively small parameter scale. Compared with previous multi-step methods such as MGLD [79], it achieves near $90\times$ faster inference, and even compared with recent one-step approaches, its speed is generally more than $5\times$ higher. Overall, these comprehensive evaluations verify the effectiveness and superiority of our approach.

Table 2. Inference efficiency comparison. Measured on a single GPU using a 21-frame 1920×1080 video. The model parameters are counted only for the generator part.

Metric	UAV	MGLD	VENh.	STAR	DOVE	DLoRAL	SeedVR2	DUO-VSR
Step	30	50	15	15	1	1	1	1
Time (s)	382.1	956.7	404.5	200.4	66.7	76.6	89.7	11.3
Params (B)	0.7	1.4	2.0	2.0	5.6	0.9	8.2	1.3

4.3. Ablation Study

We conduct ablation studies to evaluate the contribution of each component and design choice, following the training configurations in Sec. 4.1 and using the AIGC60 dataset.

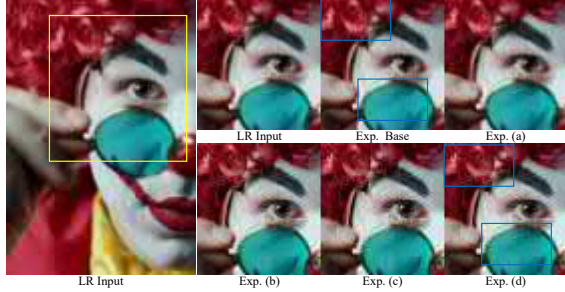


Figure 6. Visual comparison of ablation on three stage distillation. Experiment indices refer to Tab. 3. Zoom in for details.

Further analysis is provided in the supplementary material.

Ablation on Three Stage Distillation. We present the ablation on the impact of our three-stage fine-tuning pipeline in Tab. 3 and Fig. 6. We report the performance of the base model with 50 inference steps (the first row highlighted in gray) and compare it with variants equipped with different fine-tuning stages (Exps. (a)–(d), where ✓ indicates the inclusion of that stage). Comparing (a) and (b) shows that incorporating Dual-Stream Distillation (Stage II) notably improves the distilled model, benefiting from the RFS-GAN supervision derived from real-world videos and even surpassing the base model on perceptual metrics such as CLIPQA and DOVER. Comparing (b) and (d) reveals that the Preference-Guided Refinement (Stage III) further enhances perceptual quality, demonstrating the effectiveness of preference-based alignment for human-perceived realism. Finally, the comparison between (c) and (d) highlights that the Trajectory-Preserving Distillation (Stage I) provides a strong initialization that stabilizes subsequent training and contributes to consistent quality improvements.

Table 3. Ablation on Three Stage Distillation.

Exp.	Variants			Metrics			
	I	II	III	NIQE	MUSIQ	CLIPQA	DOVER
Base				4.31	63.46	0.471	87.98
(a)	✓			5.45	58.97	0.408	86.49
(b)	✓	✓		4.64	63.36	0.487	88.01
(c)		✓	✓	5.11	60.22	0.423	87.63
(d)	✓	✓	✓	4.42	63.68	0.489	88.15

Ablation on Dual-Stream Distillation Strategy. We further analyze the Dual-Stream Distillation strategy in Tab. 4 and Fig. 7, where we separately examine its effectiveness from the component and optimization perspectives. At the component level, we compare the effects of using DMD or RFS-GAN alone against their combination. While each individual branch provides moderate improvements compared to using only Stage I, the Dual-Stream configuration enables them to play complementary roles, achieving notably better performance across all perceptual metrics. As shown in Fig. 7, while RFS-GAN alone does not enhance textures as effectively as DMD (e.g., plants in the red box), its inclusion mitigates quality degradation or insufficient super-

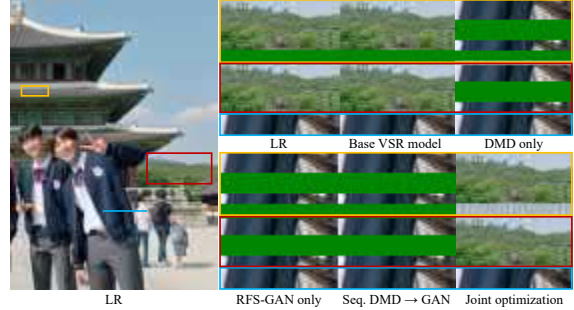


Figure 7. Visual comparison of ablation on Dual-Stream Distillation Strategy. The orange and red boxes show spatial comparison in the LR. The blue box shows the temporal profile along the blue line in the LR. Zoom in for details.

vision from DMD alone (e.g., tiles in the orange box, temporal profile in the blue box), improving artifacts and temporal consistency. At the optimization level, we compare our joint optimization scheme with the sequential strategy adopted in DMD2, which first distills with DMD and then fine-tunes with GAN supervision. The results show that joint optimization allows the two objectives to interact more effectively during training, leading to stronger mutual reinforcement and the best overall performance.

Table 4. Ablation on Dual-Stream Distillation Strategy. “Joint” and “Seq.” denote different optimization schemes.

Setting	NIQE	MUSIQ	CLIPQA	DOVER
<i>Component-level</i>				
DMD only	4.99	61.46	0.432	87.38
RFS-GAN only	5.32	62.64	0.427	87.53
Dual-Stream (Joint)	4.42	63.68	0.489	88.15
<i>Optimization-level</i>				
Sequential DMD→GAN (Seq.)	5.17	62.76	0.419	87.67
Joint optimization (ours)	4.42	63.68	0.489	88.15

5. Conclusion

In this paper, we identified that directly applying distribution matching distillation (DMD) to one-step video super-resolution suffers from training instability, degraded supervision from the real score model, and insufficient guidance toward real HR videos. To address these issues, we proposed DUO-VSR, a three-stage framework built upon a Dual-Stream Distillation Strategy that integrates DMD with Real-Fake Score Feature GAN for stable and comprehensive supervision. Through Progressive Guided Distillation Initialization, Dual-Stream Distillation, and Preference-Guided Refinement, DUO-VSR effectively stabilizes optimization, enhances supervision, and aligns perceptual quality preferences. Our findings reveal that combining distribution matching and adversarial supervision provides an effective path toward efficient, high-fidelity one-step VSR.

6. Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (Grant No. 62502169)

References

- [1] Haoran Bai, Xiaoxu Chen, Canqian Yang, Zongyao He, Sibin Deng, and Ying Chen. Vivid-vr: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. *arXiv preprint arXiv:2508.14483*, 2025. [2](#)
- [2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. [1](#), [2](#)
- [3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. [1](#), [2](#)
- [4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5962–5971, 2022. [2](#), [6](#), [7](#)
- [5] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9232–9241, 2024. [2](#)
- [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](#), [3](#), [6](#)
- [7] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23174–23184, 2025. [1](#), [3](#)
- [8] Jinpei Guo, Yifei Ji, Zheng Chen, Yufei Wang, Sizhuo Ma, Yong Guo, Yulun Zhang, and Jian Wang. Towards redundancy reduction in diffusion models for efficient video super-resolution. *arXiv preprint arXiv:2509.23980*, 2025. [2](#)
- [9] 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 preprint arXiv:2407.07667*, 2024. [2](#), [6](#)
- [10] Xiao He, Huaao Tang, Zhijun Tu, Junchao Zhang, Kun Cheng, Hanting Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, et al. One step diffusion-based super-resolution with time-aware distillation. *arXiv preprint arXiv:2408.07476*, 2024. [3](#)
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [1](#), [2](#)
- [12] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. [2](#), [3](#)
- [13] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European conference on computer vision*, pages 645–660. Springer, 2020. [2](#)
- [14] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018. [1](#), [2](#)
- [15] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, pages 428–447. Springer, 2024. [3](#)
- [16] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. [6](#), [1](#)
- [17] Lei Ke, Hubery Yin, Gongye Liu, Zhengyao Lv, Jingcai Guo, Chen Li, Wenhan Luo, Yujiu Yang, and Jing Lyu. Flowsteer: Guiding few-step image synthesis with authentic trajectories. *arXiv preprint arXiv:2511.18834*, 2025. [3](#)
- [18] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023. [3](#)
- [19] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. [6](#)
- [20] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023. [2](#)
- [21] Jianze Li, Jiezhong Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. *arXiv preprint arXiv:2502.01993*, 2025. [3](#)
- [22] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *European conference on computer vision*, pages 335–351. Springer, 2020. [2](#)
- [23] Xiaohui Li, Yihao Liu, Shuo Cao, Ziyang Chen, Shaobin Zhuang, Xiangyu Chen, Yanan 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](#)
- [24] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhong Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. [2](#)
- [25] Jingyun Liang, Jiezhong Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool.

- Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024. 2
- [26] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025. 2, 3
- [27] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 5
- [28] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3
- [29] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. InstafLOW: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 3
- [30] Yong Liu, Jinshan Pan, Yinchuan Li, Qingji Dong, Chao Zhu, Yu Guo, and Fei Wang. Ultravs: Achieving ultra-realistic video super-resolution with efficient one-step diffusion space. *arXiv preprint arXiv:2505.19958*, 2025. 2, 3
- [31] Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024. 3
- [32] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019. 2
- [33] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 3
- [34] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023. 3
- [35] Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2024. 3
- [36] Yihong Luo, Xiaolong Chen, Xinghua Qu, Tianyang Hu, and Jing Tang. You only sample once: Taming one-step text-to-image synthesis by self-cooperative diffusion gans. *arXiv preprint arXiv:2403.12931*, 2024. 3
- [37] Zhengyao Lv, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K Wong. FasterCache: Training-free video diffusion model acceleration with high quality. *arXiv preprint arXiv:2410.19355*, 2024. 2
- [38] Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K Wong, Yu Qiao, and Ziwei Liu. Dcm: Dual-expert consistency model for efficient and high-quality video generation. *arXiv preprint arXiv:2506.03123*, 2025. 3
- [39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 2
- [40] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14297–14306, 2023. 3, 4
- [41] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6
- [42] Mehdi Noroozi, Isma Hadji, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. You only need one step: Fast super-resolution with stable diffusion via scale distillation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2024. 3
- [43] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4811–4820, 2021. 2
- [44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3
- [45] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *Advances in Neural Information Processing Systems*, 37:117340–117362, 2024. 3
- [46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2
- [47] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3
- [48] Shoab Meraj Sami, Md Mahedi Hasan, Jeremy Dawson, and Nasser Nasrabadi. Hf-diff: High-frequency perceptual loss and distribution matching for one-step diffusion-based image super-resolution. *arXiv preprint arXiv:2411.13548*, 2024. 3
- [49] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3
- [50] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 3
- [51] 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
- [52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2, 3
- [53] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 1, 3

- [54] 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. *arXiv preprint arXiv:2506.15591*, 2025. 2, 3, 6
- [55] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 4472–4480, 2017. 6, 7
- [56] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3360–3369, 2020. 2
- [57] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3
- [58] Fu-Yun Wang, Zhaoyang Huang, Alexander Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024. 3
- [59] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 6
- [60] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv preprint arXiv:2506.05301*, 2025. 2, 3, 6
- [61] 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. 1, 2, 3, 6
- [62] Ruohao Wang, Xiaohui Liu, Zhilu Zhang, Xiaohe Wu, Chun-Mei Feng, Lei Zhang, and Wangmeng Zuo. Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1168–1177, 2023. 2
- [63] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 2
- [64] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1, 2
- [65] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25796–25805, 2024. 3
- [66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [67] Zhongdao Wang, Guodongfang Zhao, Jingjing Ren, Bailan Feng, Shifeng Zhang, and Wenbo Li. Turbovsr: Fantastic video upscalers and where to find them. *arXiv preprint arXiv:2506.23618*, 2025. 1
- [68] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 6, 1
- [69] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024. 1, 3
- [70] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. Animesr: Learning real-world super-resolution models for animation videos. *Advances in Neural Information Processing Systems*, 35:11241–11252, 2022. 2
- [71] Liangbin Xie, Xintao Wang, Shuwei Shi, Jinjin Gu, Chao Dong, and Ying Shan. Mitigating artifacts in real-world video super-resolution models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2956–2964, 2023. 2
- [72] Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, Jian Yang, and Ying Tai. Addr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717*, 2024. 3
- [73] 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. 1, 2, 6
- [74] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8196–8206, 2024. 3
- [75] 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. 2
- [76] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2
- [77] Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *Advances in Neural Information Processing Systems*, 37:78630–78652, 2024. 3
- [78] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a

- decomposition based learning scheme. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4781–4790, 2021. 2
- [79] 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, 6, 7
- [80] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 3
- [81] 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. 2, 6, 7
- [82] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024. 2, 3
- [83] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 1, 2, 3, 5
- [84] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22963–22974, 2025. 2, 3
- [85] Weiyi You, Mingyang Zhang, Leheng Zhang, Xingyu Zhou, Kexuan Shi, and Shuhang Gu. Consistency trajectory matching for one-step generative super-resolution. *arXiv preprint arXiv:2503.20349*, 2025. 3
- [86] Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. Fmanet: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 44–55, 2024. 2
- [87] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 3
- [88] Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. *arXiv preprint arXiv:2505.13389*, 2025. 2
- [89] Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025. 2
- [90] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 1
- [91] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2
- [92] 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. 2, 6
- [93] 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 preprint arXiv:2510.00948*, 2025. 2
- [94] Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588*, 2024. 2
- [95] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [96] 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, 6, 7
- [97] Yuanzhi Zhu, Ruiqing Wang, Shilin Lu, Junnan Li, Han-shu Yan, and Kai Zhang. Ofts: One-step flow for image super-resolution with tunable fidelity-realism trade-offs. *arXiv preprint arXiv:2412.09465*, 2024. 3
- [98] 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. 2, 3

DUO-VSR: Dual-Stream Distillation for One-Step Video Super-Resolution

Supplementary Material

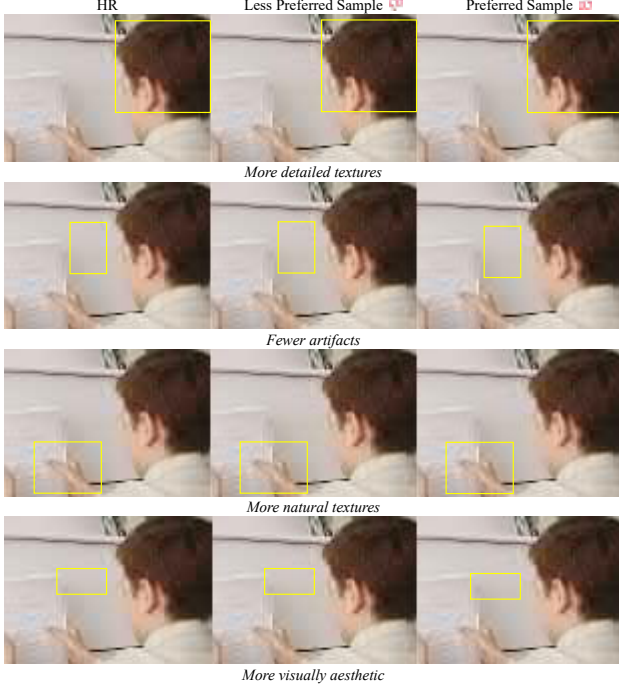


Figure 8. Examples of preferred and less-preferred samples in the constructed preference dataset. Zoom in for details.

7. Further Implementation Details

7.1. Algorithm for Dual-Stream Distillation

The detailed procedure of the dual-stream distillation strategy is outlined in Algorithm 1, comprising interleaved Auxiliary and Student updates. In our implementation, we set the update interval $N = 3$ by default.

7.2. Construction of Preference Dataset

In the preference-guided refinement stage, we construct a preference dataset for Direct Preference Optimization. Specifically, for each LR video, we generate five candidate reconstructions using the second-stage model. We then evaluate these candidates using the LPIPS [90], MUSIQ [16] and DOVER [68] metrics and rank them according to their combined quality scores. The highest-scoring output is selected as the preferred sample, while the lowest-scoring one serves as the less preferred sample. As illustrated in Fig. 8, the preferred samples typically exhibit richer, more natural, and aesthetically pleasing textures. In total, we construct 2000 preference pairs for fine-tuning.

8. Further Discussions and Ablation Analyses.

In the ablation study presented in Sec. 4.3 of the main text, we analyzed the effectiveness of the three-stage distillation

Algorithm 1: Dual-Stream Distillation Strategy

Input: Frozen Real Score model θ_R ; trainable Fake Score model θ_F ; student θ_S ; discriminator heads H_ϕ ; loss weights $\lambda_{\text{DMD}}, \lambda_{\text{GAN}}, \lambda_{\text{FM}}$; interval N .

while not converged do

for $i \leftarrow 1$ **to** N **do**

/* Auxiliary update */

Sample $(z^{LR}, z^{HR}, c), t, \epsilon$;

$\hat{z}_0^S \leftarrow \epsilon - v_{\theta_S}(\epsilon, t, z^{LR}, c)$;

$\hat{z}_t^S \leftarrow q_t(\hat{z}_0^S), z_t^{HR} \leftarrow q_t(z^{HR})$;

// Diffusion loss for θ_F

Compute target v at $(\hat{z}_t^S, t, z^{LR}, c)$;

$\mathcal{L}_{\text{Diff}} \leftarrow \|v_{\theta_F}(\hat{z}_t^S, t, z^{LR}, c) - v\|^2$;

// GAN discriminator loss for ϕ
with stop_grad backbones

$\mathbf{h}^S \leftarrow \text{concat}(\text{Feat}_{\theta_R}(\hat{z}_t^S), \text{Feat}_{\theta_F}(\hat{z}_t^S))$;

$\mathbf{h}^{HR} \leftarrow \text{concat}(\text{Feat}_{\theta_R}(z_t^{HR}), \text{Feat}_{\theta_F}(z_t^{HR}))$;

$D_S \leftarrow H_\phi(\text{sg}[\mathbf{h}^S])$;

$D_{HR} \leftarrow H_\phi(\text{sg}[\mathbf{h}^{HR}])$;

$\mathcal{L}_D \leftarrow \mathbb{E}[\max(0, 1 - D_{HR})] + \mathbb{E}[\max(0, 1 + D_S)]$;

Update θ_F by descending $\nabla_{\theta_F} \mathcal{L}_{\text{Diff}}$;

Update ϕ by descending $\nabla_{\phi} \mathcal{L}_D$;

/* Student update (after every N Auxiliary steps) */

Sample $(z^{LR}, z^{HR}, c), t, \epsilon$;

$\hat{z}_0^S \leftarrow \epsilon - v_{\theta_S}(\epsilon, t, z^{LR}, c)$;

$\hat{z}_t^S \leftarrow q_t(\hat{z}_0^S), z_t^{HR} \leftarrow q_t(z^{HR})$;

// DMD loss

$\hat{z}_0^R \leftarrow \hat{z}_0^R(\hat{z}_t^S; \theta_R), \hat{z}_0^F \leftarrow \hat{z}_0^F(\hat{z}_t^S; \theta_F)$;

$\hat{z}_t^R \leftarrow \frac{\hat{z}_0^F - \hat{z}_0^R}{\text{mean}(\text{abs}(\hat{z}_0^S - \hat{z}_0^R))}$;

$\mathcal{L}_{\text{DMD}} \leftarrow \|\hat{z}_t^S - \text{sg}[\hat{z}_t^R - \text{Grad}]\|^2$;

// GAN generator loss

$\mathbf{h}^S \leftarrow \text{concat}(\text{Feat}_{\theta_R}(\hat{z}_t^S), \text{Feat}_{\theta_F}(\hat{z}_t^S))$;

$D(\hat{z}_t^S) \leftarrow H_\phi(\text{sg}[\mathbf{h}^S])$;

$\mathcal{L}_G \leftarrow -\mathbb{E}[D(\hat{z}_t^S)]$;

// Feature matching loss

$\mathbf{h}^{HR} \leftarrow \text{concat}(\text{Feat}_{\theta_R}(z_t^{HR}), \text{Feat}_{\theta_F}(z_t^{HR}))$;

$\mathcal{L}_{\text{FM}} \leftarrow \|\mathbf{h}^S - \mathbf{h}^{HR}\|^2$;

$\mathcal{L}_S \leftarrow \lambda_{\text{DMD}} \mathcal{L}_{\text{DMD}} + \lambda_{\text{GAN}} \mathcal{L}_G + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}}$;

Update θ_S by descending $\nabla_{\theta_S} \mathcal{L}_S$;

framework and the two branches in the Dual-Stream Distillation, namely the DMD stream and the RFS-GAN stream, along with the exploration of different optimization strategies. In this section, we provide additional discussions on the design and training of the RFS-GAN.



Figure 9. Noise-perturbed samples stabilize adversarial training and suppress artifacts. Zoom in for details.

Noise-Perturbed Sample Input in RFS-GAN. Different from SeedVR2 [60], which directly feeds the clean outputs of the student into the discriminator, we observe that such a design often leads to training instability and occasionally produces grid-like artifacts, as shown in Fig. 9. We hypothesize that this instability stems from a discriminator-generator imbalance, where an overly strong discriminator can easily distinguish real samples from fake ones. Inspired by the perturbation strategy in DMD [83], which intentionally blurs the boundary between real and fake data distributions, we similarly add random noise with varying intensity to both real and fake inputs of the discriminator. This modification effectively stabilizes the adversarial learning while preserving its enhancement effect.

Furthermore, using noisy real and fake samples enables sharing the intermediate features from real and fake score computation for the GAN loss calculation, requiring only an additional extraction of features from real samples and thus reducing the number of forward passes.

Cross-Model and Multi-Layer Feature in RFS-GAN In RFS-GAN, both the real score model and the fake score model are employed as the backbones of the discriminator. As illustrated in Fig. 10, intermediate representations are extracted from the 9th, 18th, and 27th layers of the DiT architecture (consisting of 30 layers in total). RFS-GAN effectively integrates shallow features that capture structural and semantic information with deeper representations that encode richer and more fine-grained details. Furthermore, the two score models are optimized over distinct data distributions: the real score model is intrinsically aligned with the real (teacher) distribution, providing high-quality discriminative guidance, whereas the fake score model dynamically reflects the evolving distribution of the student. The complementarity between these two models substantially enhances the representational capacity of the discriminator, thereby delivering stronger and more reliable gradient feedback to the student model.

Ablation studies are performed on the second-stage model to assess the effectiveness of discriminator features extracted from the real and fake score models. As shown in Tab. 5, the discriminator that combines the real and fake score models achieves the best performance in perceptual metrics, demonstrating the effectiveness of RFS-GAN.

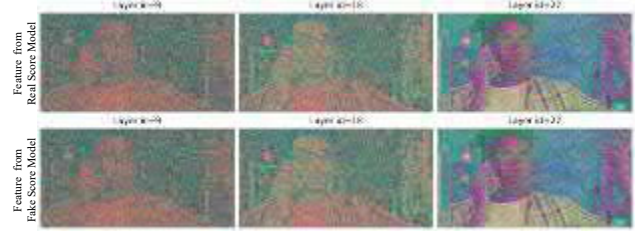


Figure 10. Discriminator features from the real and fake score models used for the RFS-GAN loss computation, reduced to three dimensions via t-SNE [39] for visualization.

Table 5. Ablation study on the discriminator design of RFS-GAN.

Method	NIQE↓	MUSIQ↑	CLIPQA↑	DOVER↑
Real Score Model only	4.71	62.79	0.456	87.95
Fake Score Model only	4.98	62.98	0.475	87.76
RFS-GAN	4.64	63.36	0.487	88.01

9. Additional Evaluation Results

9.1. Additional Visual Comparisons

Comparison with the base model. We first compare DUO-VSR with its base model to examine the effectiveness of the distillation framework, as shown in the Fig. 11. The results indicate that our method achieves a comparable ability to generate textures (first row), while producing more natural and visually coherent details (third and fourth rows).

Comparison with other methods. We present additional visual quality comparisons with VEnhancer [9], UAV [96], STAR [73], DLoRAL [54], DOVE [6], and SEEDVR2 [60] in Fig. 12. These results further demonstrate the advantages of our method when dealing with challenging regions that involve fine textures.

9.2. Discussion of Concurrent Works

We note that several concurrent works [8, 93, 98] have explored efficient video super-resolution, some of which also employ DMD for one-step inference. Both InfVSR [93] and FlashVSR [98] adopt DMD and causal DiT architectures to achieve one-step streaming VSR, focusing primarily on reformulating full-sequence diffusion into a causal structure, where DMD mainly serves as a step-distillation mechanism. Earlier, UltraVSR [30] also employs distribution matching distillation to facilitate one-step VSR, but focuses on degradation-aware scheduling and leverages an image diffusion backbone (extended Stable Diffusion [46] for VSR). In contrast, our DUO-VSR takes an orthogonal perspective by revisiting the intrinsic limitations of DMD in VSR and introducing an effective dual-stream distillation strategy to mitigate them. This design offers a complementary pathway that could potentially be integrated with existing DMD-based frameworks to further enhance their robustness and visual quality.

Recently, both FlashVSR [98] and UltraVSR [30] have made their official implementations publicly available, and

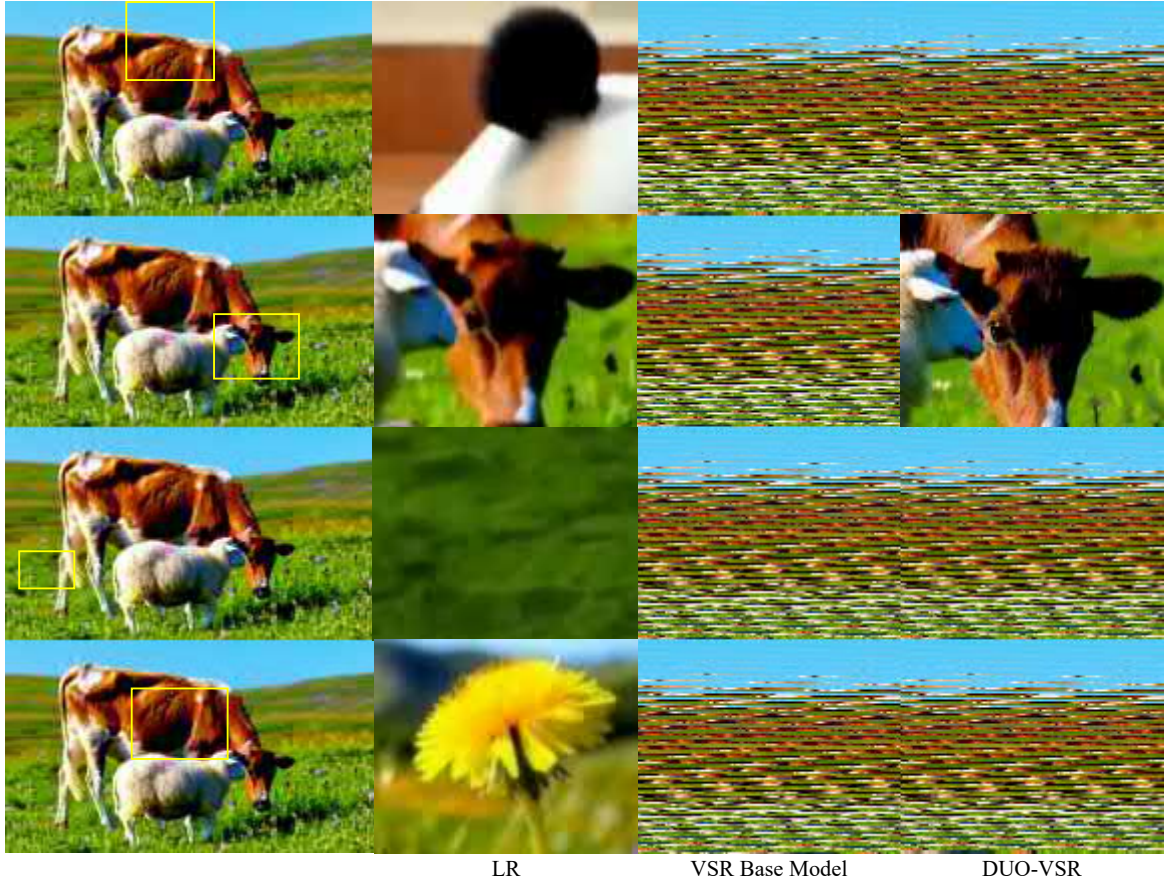


Figure 11. Visual comparison with base VSR model. Zoom in for details.

we include comparative results in this supplementary material. To ensure a fair comparison in terms of performance and quality, we use FlashVSR-Full for evaluation. As shown in Fig. 12, our method produces more realistic and natural details than FlashVSR and UltraVSR. Specifically, in the first case, DUO-VSR reconstructs finer and smoother fur textures on the fox; in the second case, the woman’s eyebrows and eyes appear more natural; and in the fourth case, the wheat spikes exhibit more faithful and visually convincing structures. Tab. 6 presents a quantitative comparison between DUO-VSR and these two methods on the AIGC60 dataset. It can be seen that DUO-VSR achieves superior performance in perceptual metrics while exhibiting comparable inference efficiency to FlashVSR-Full.

9.3. User Study

Following APT [26] and SeedVR2 [60], we conducted a blind user study using the GSB test to more comprehensively assess the subjective visual quality of our method. Specifically, the preference score is computed as $\frac{G-B}{(G+B+S)}$, where G denotes the number of samples judged as good, B as bad, and S as similar. The score ranges from -100% to 100%, with 0% indicating equal performance. We ran-

Table 6. Quantitative comparison on the AIGC60 dataset.

Metric	UltraVSR	FlashVSR	DUO-VSR
NIQE ↓	5.58	<u>4.67</u>	4.42
MUSIQ ↑	58.23	<u>63.11</u>	63.68
CLIP-IQA ↑	0.4434	<u>0.4690</u>	0.4886
DOVER ↑	86.45	<u>87.49</u>	88.15
E_{warp}^* ↓	<u>1.54</u>	1.76	1.08
Time (s)	126.5	10.7	<u>11.3</u>
Params (B)	<u>1.9</u>	1.3	1.3

domly selected 30 samples from the VideoLQ and AIGC60 datasets. The evaluation primarily compared our approach with recent one-step video super-resolution methods, including SeedVR2-7B [60], DOVE [6], DLoRAL [54], UltraVSR [30], and FlashVSR-Full [98]. Participants rated three aspects: visual fidelity, visual quality, and overall quality. Twenty researchers with computer vision backgrounds took part in the evaluation. As shown in Tab. 7, DUO-VSR achieves higher subjective preference scores than previous methods.

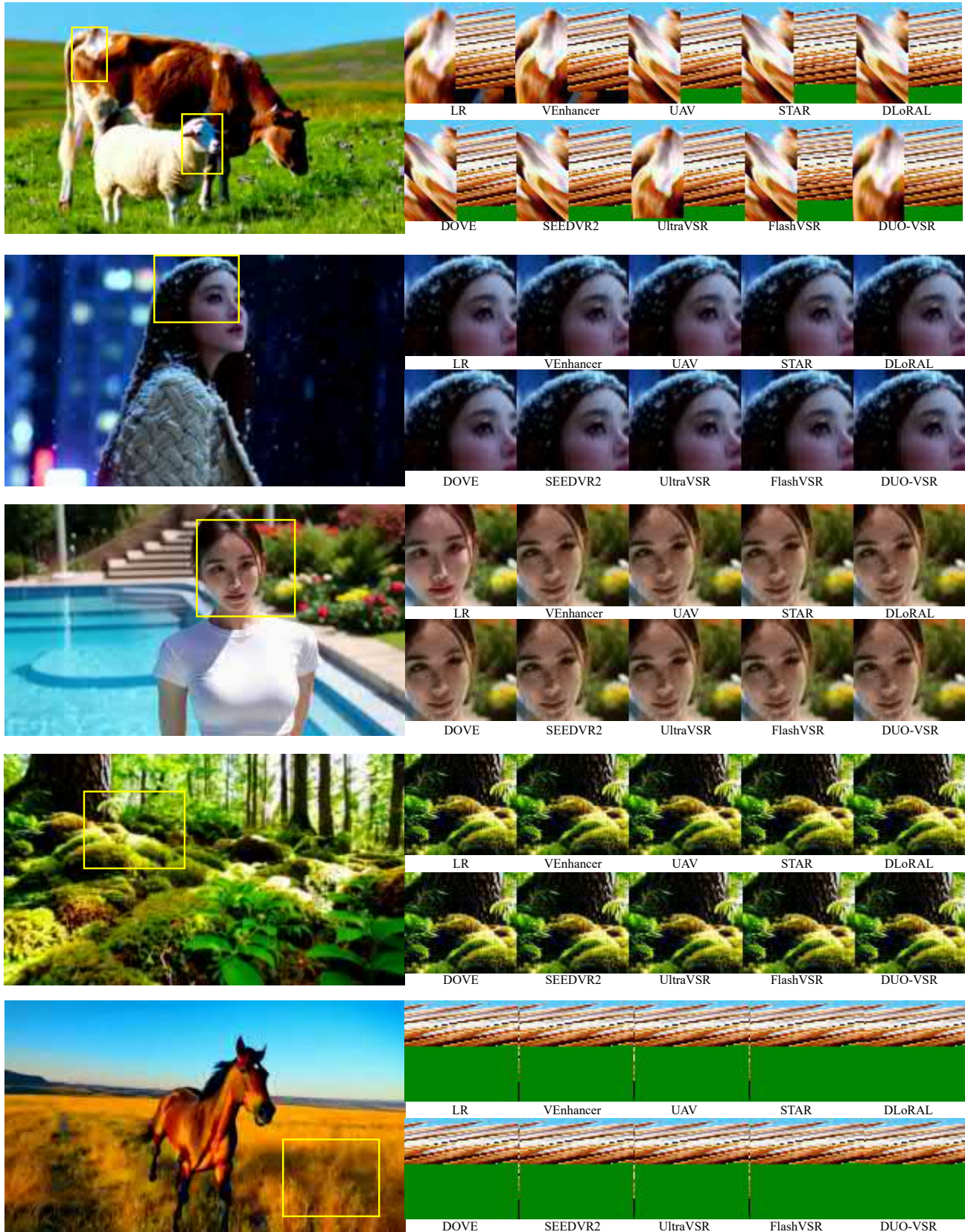


Figure 12. Visual comparison of different VSR methods. DUO-VSR consistently reconstructs finer textures. Zoom in for details.

Table 7. Blind user study results based on GSB test.

Method	Overall Quality	Visual Fidelity	Visual Quality
DUO-VSR	0%	0%	0%
Our Base VSR model	-1.3%	-3.7%	+2.3%
SeedVR2-7B	-32.7%	-13.3%	-39.2%
DOVE	-29.3%	-8.0%	-36.2%
DLoRAL	-34.0%	-10.8%	-37.8%
UltraVSR	-39.8%	-16.7%	-43.3%
FlashVSR-Full	-25.5%	-6.7%	-28.2%

10. Limitations and Future Work

Limitations. Despite the strong efficiency and perceptual quality achieved by our one-step framework, several limitations remain. Since our method is trained in the latent space, the underlying VAE applies an aggressive spatiotemporal compression ($8\times$ spatial and $4\times$ temporal), which can hinder the reconstruction of extremely fine-grained details such as tiny text. In addition, the video VAE becomes the dominant computational bottleneck during inference, accounting for more than 90% of the total runtime.

Future Work. In future work, we plan to explore more efficient or task-specific video VAEs that not only preserve high-frequency details and temporal coherence but also significantly accelerate the decoding process, thereby reducing the overall inference latency of our one-step framework.