EFFICIENT VIDEO SUPER-RESOLUTION VIA HYBRID ATTENTION AND REWARD-YIELD DISTRIBUTION MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent video super-resolution (VSR) models achieve efficient one-step generation through diffusion distillation. However, high-resolution inference remains costly, and distilled models are difficult to align with human preferences. These challenges call for improvements in both model architecture and training strategy. We present **HyR-VSR** to address both limitations. On the architecture side, we introduce hybrid attention based on the strong locality observed in pretrained VSR models, while retaining sparse global interaction for long-range consistency. It achieves up to 2× DiT speedup and 45% lower peak memory, enabling spatially untiled 4K VSR on a single 24GB RTX 4090. On the training side, we propose Reward-Yield Distribution Matching Distillation (Ry-DMD) for preference-aligned one-step generation. Ry-DMD uses reward-weighted teacher rollouts to incorporate black-box rewards into distribution matching, without backpropagating through the reward model. We further show that Ry-DMD generalizes beyond VSR to T2I generation. Together, HyR-VSR delivers stronger restoration quality and substantially improves high-resolution efficiency, with consistent gains in user preference.

1 INTRODUCTION

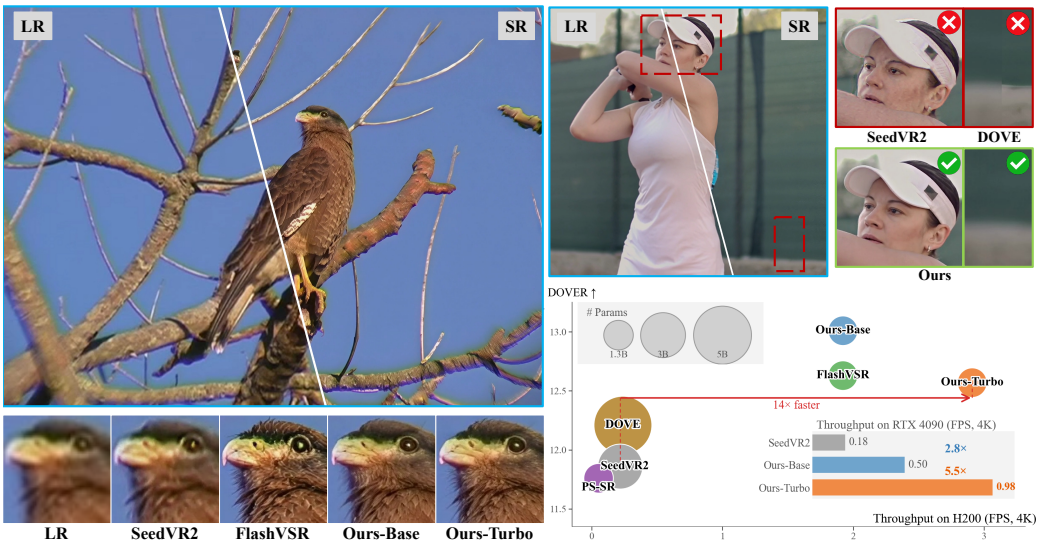


Figure 1: **HyR-VSR achieves high-quality and efficient video restoration. Left:** We recover fine visual details while maintaining natural textures. **Right:** At 4K resolution, we suppress visible artifacts and enable substantially faster inference on both H200 and a single RTX 4090.

Video super-resolution (VSR) reconstructs high-resolution videos from low-resolution inputs. It supports applications such as smartphone photography, social media, live streaming, and cost-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

effective upscaling of AI-generated videos to 2K or 4K resolution. Reconstruction-based VSR aggregates complementary information across frames to improve distortion oriented fidelity (Chan et al., 2022a;b; Shi et al., 2022). These methods benefit from stronger propagation, alignment, and spatiotemporal modeling, but the perception–distortion trade-off often leads to over-smoothed textures when high-frequency details are ambiguous (Blau & Michaeli, 2018). Generative VSR addresses this limitation by utilizing learned image or video priors to synthesize perceptually realistic details (Zhou et al., 2024; Chen et al., 2024; Xu et al., 2025c; Li et al., 2025), and has achieved substantial progress in visual quality.

Despite these advances, efficient and flexible generative VSR remains hindered by two bottlenecks. (1) **Computational efficiency**: Diffusion-based approaches rely on costly multi-step denoising trajectories (Zhou et al., 2024; Chen et al., 2024). Although recent one-step VSR methods greatly reduce this latency by distilling the multi-step generation trajectory into a single step (Song et al., 2023; Yin et al., 2024a;b; Sun et al., 2026; Zhuang et al., 2026; Lv et al., 2026), the computation and memory costs of spatiotemporal attention remain prohibitive for high-resolution videos (*e.g.*, 2K and 4K resolution). As a result, inference on consumer-grade GPUs (*e.g.*, RTX 4090) often requires spatial tiling, which limits global context and may introduce boundary artifacts, as illustrated in Fig. 1. (2) **Preference-aligned distillation**: Standard one-step distillation method DMD (Yin et al., 2024a;b) matches the teacher distribution but cannot explicitly favor outputs that better align with human preferences, such as more accurate text rendering. Existing reward-guided methods typically rely on multi-step stochastic trajectories (Black et al., 2024; Fan et al., 2023) or differentiable reward signals (Clark et al., 2024; Jia et al., 2025), making them ill-suited to deterministic one-step generators with black-box rewards.

To address these bottlenecks, we present HyR-VSR, a one-step VSR framework with hybrid attention for efficient high-resolution restoration and Ry-DMD for preference-aligned distillation with black-box rewards. Together, they make one-step generative VSR more efficient, **enabling spatially untiled 4K VSR inference on a single 24GB RTX 4090**, while better aligning outputs with human preferences.

For high-resolution efficiency, we find that pretrained VSR models such as FlashVSR (Zhuang et al., 2026) exhibit strong spatiotemporal attention locality, but with substantial variation across layers. Attention is highly concentrated within compact neighborhoods in many blocks, suggesting that dense attention is unnecessary throughout the backbone, while fully local attention would be overly restrictive. We therefore introduce a **hybrid attention architecture** that replaces most original attention blocks with efficient local spatiotemporal attention while retaining a small number of sparse global blocks. This design achieves up to $2\times$ DiT acceleration and nearly halves peak memory, enabling spatially untiled 4K VSR on a single 24GB RTX 4090 while maintaining comparable restoration quality.

For preference-aligned distillation, one-step generators are particularly challenging because stochasticity is limited to the input seed, while black-box rewards provide no gradients for direct optimization. We propose **Reward-Yield Distribution Matching Distillation (Ry-DMD)**, which uses multiple teacher rollouts to expose the stochasticity available to a one-step generator and reweights them with black-box rewards. The difference between the reward-weighted and original teacher predictions provides a preference direction, which is added to standard DMD to favor higher-reward teacher-supported outputs. We validate Ry-DMD on OCR-guided VSR and further demonstrate its applicability to text-to-image generation.

To further scale up one-step VSR distillation, a natural strategy is to increase the training budget and data size. However, we find that longer DMD training does not yield consistent gains, revealing non-monotonic behavior. Instead, student initialization and teacher capacity play a more important role in final performance.

2 Related Work

2.1 Video Super-Resolution

Conventional VSR methods recover missing details through alignment, propagation, and spatiotemporal aggregation (Wang et al., 2019; Chan et al., 2021; 2022a; Shi et al., 2022; Liang et al., 2022),

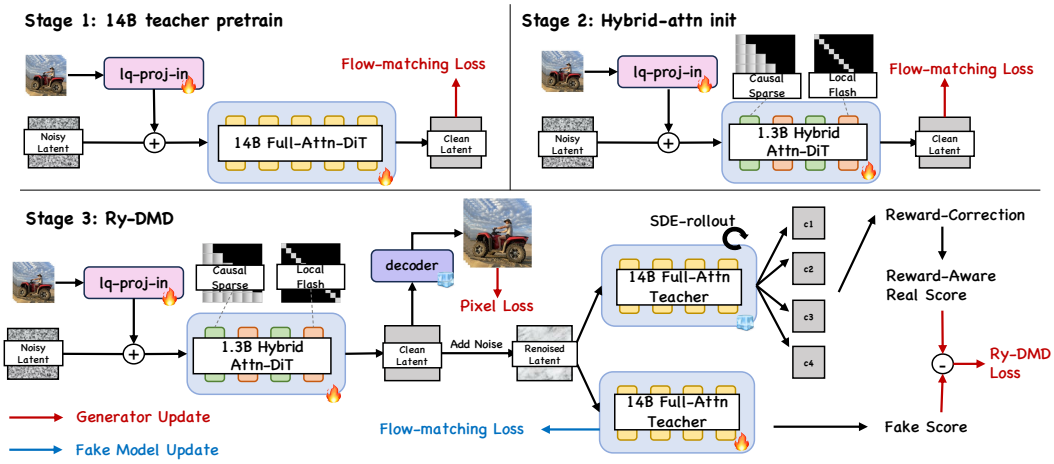


Figure 2: **Overview of the HyR-VSR training pipeline.** Stage 1 trains a 14B full-attention multi-step teacher on image and video data. Stage 2 trains the hybrid-attention student to obtain a strong initialization. Stage 3 distills the student into a one-step generator using Ry-DMD, where reward feedback modifies the teacher term estimation without requiring reward gradients.

but often produce over-smoothed textures in perceptually ambiguous regions (Blau & Michaeli, 2018). Generative VSR instead leverages learned priors to synthesize realistic details, including adversarial (Chu et al., 2020; Xu et al., 2025c) and diffusion-based approaches (Rota et al., 2024; Zhou et al., 2024; Chen et al., 2024; Li et al., 2025; Chen et al., 2026c). To reduce the cost of iterative sampling, recent methods increasingly target one-step restoration through direct fine-tuning, adversarial post-training, or distillation (Liu et al., 2025; Chen et al., 2026d; Wang et al., 2026; Sun et al., 2026; Zhuang et al., 2026; Lv et al., 2026). Recent image SR and restoration methods have also explored reward or preference guided post-training (Yi et al., 2026; Zhao et al., 2026; Xu et al., 2025a). Nevertheless, one-step VSR still faces challenges in both high-resolution inference efficiency and reward-guided post-training. HyR-VSR targets these two aspects with hybrid attention and Ry-DMD, respectively.

2.2 Efficient Attention for Visual Generation

The quadratic cost of dense attention has motivated sparse and sub-quadratic alternatives for visual generation. Video diffusion methods exploit structured spatiotemporal sparsity or locality-aware computation to reduce attention cost (Xi et al., 2025; Yang et al., 2026; Shen et al., 2025; Xu et al., 2025b; Li et al., 2026). Another line of work adopts efficient attention backbones, including linear attention and hybrid architectures that periodically retain full softmax attention for global information exchange (Xie et al., 2025a;b; Chen et al., 2026b;a). In contrast, HyR-VSR exploits the locality observed in a pretrained VSR model, replacing most global-attention blocks with spatiotemporal local attention while retaining a small number of global blocks.

2.3 Reward Optimization for One/Few-Step Generative Models

Reward-guided post-training has been extensively studied for multi-step diffusion models, where denoising trajectories enable policy optimization, differentiable reward backpropagation, and preference learning (Fan et al., 2023; Black et al., 2024; Clark et al., 2024; Wallace et al., 2024; Liu et al., 2026; Zheng et al., 2026). Extending these approaches to few-step and one-step generators remains challenging. While few-step models still preserve short denoising trajectories, one-step generators directly map noise to samples and lack such intermediate optimization signals. Existing few-step approaches explore reward-guided distribution matching and related optimization strategies (Jiang et al., 2026; Huang et al., 2026; Dong et al., 2026), while one-step methods typically rely on inference-time optimization or differentiable reward surrogates (Eyring et al., 2024; Jia et al., 2025; Zhu et al., 2026). Ry-DMD instead enables black-box reward optimization for deterministic one-step generators by introducing a reward-tilted teacher target within the DMD framework.

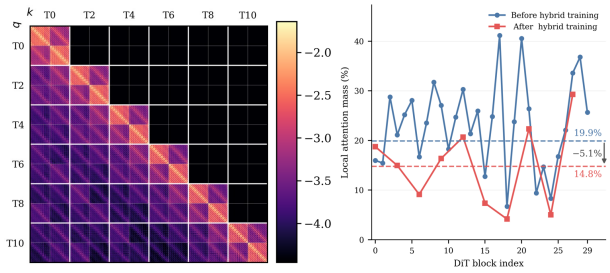
3 METHOD

3.1 OVERVIEW

Following FlashVSR (Zhuang et al., 2026), HyR-VSR adopts the three-stage training framework shown in Fig. 2. In **Stage 1**, we train a 14B full-attention multi-step model with flow matching as the teacher. In **Stage 2**, we replace the 1.3B backbone with our hybrid-attention architecture and train it to provide a strong initialization for one-step generation. In **Stage 3**, we distill the initialized student into a one-step VSR model using distribution matching together with the pixel loss. Ry-DMD further injects scalar reward feedback into the teacher term without reward backpropagation. Training details are provided in Appendix A.

3.2 EFFICIENT ONE-STEP VSR WITH HYBRID ATTENTION

Spatiotemporal attention is a major scalability bottleneck for causal VSR at high resolution. To understand the potential redundancy of full attention, we examine the attention patterns of a pretrained one-step VSR model. As shown in Fig. 3(a), attention is highly concentrated within local neighborhoods, which cover only $\sim$ 1% of valid key blocks yet receive $\sim$ 20% of the total attention mass. We quantify this behavior using the local attention mass, defined as the fraction of attention assigned to valid local key blocks. Higher local attention mass indicates stronger concentration within local neighborhoods, suggesting that many non-local interactions are weakly weighted yet still require computation, leading to unnecessary overhead. As shown in Fig. 3(b), local attention is strongly enriched across DiT blocks but varies substantially across layers, motivating a hybrid rather than purely local design.



(a). Spatiotemporal attention locality (b). Layer-wise attention shift

Figure 3: **Attention locality and adaptation in causal VSR.** (a) Attention concentrates in compact spatiotemporal neighborhoods. (b) Local attention mass varies across DiT blocks, and hybrid training reduces the locality of retained global-attention blocks.

Motivated by this observation, we introduce **hybrid attention**, which replaces most attention blocks with efficient spatiotemporal local attention while retaining sparse global-attention blocks. Specifically, we denote the local attention configuration as (s, t, b), where s and t denote the numbers of spatial and preceding temporal key blocks used in local attention, and b denotes the number of DiT blocks replaced by local attention. We then analyze how attention adapts after hybrid training. For visualization, we use the setting $(s, t, b) = (3, 1, 20)$, with after-training measurements taken only on retained global-blocks. As shown in Fig. 3(b), the retained global-attention blocks exhibit reduced local attention mass after hybrid training, indicating a shift toward non-local interactions. This suggests that hybrid attention allows the retained global-attention blocks to better capture non-local interactions, leading to more efficient attention computation.

In the final model, we adopt a regular interleaving of local and global blocks and use $(s, t, b) = (3, 1, 25)$, retaining one global-attention block every five DiT blocks (Fig. 2, Stages 2&3). This design achieves up to $2\times$ DiT acceleration and nearly halves peak memory usage, with detailed evaluations in Section 4.2.

3.3 REWARD-YIELD DISTRIBUTION MATCHING DISTILLATION

Adapting a deterministic one-step generator to a downstream objective or preferences is not straightforward. Continuing standard DMD on task-specific data can adapt the training distribution, but does not explicitly favor outputs with better human preference. Direct reward optimization within DMD is possible, but non-differentiable rewards require differentiable surrogates for backpropagation. We propose **Reward-Yield Distribution Matching Distillation (Ry-DMD)**. As illustrated in Fig. 4, Ry-DMD transforms the teacher distribution toward higher-reward outputs, such as clearer and more recognizable text, by constructing a reward-tilted target. The resulting target shift is then

converted into a correction to the DMD teacher term. This preserves the overall DMD training framework while enabling the use of black-box objectives such as OCR, which we validate on VSR in Sec. 4.3; we further demonstrate its generality on text-to-image generation in Sec. 4.4.

3.3.1 Reward-Tilted Distribution Matching

Our goal is to modify the teacher distribution toward higher-reward regions while preserving its original support. As a distribution-level motivation, we consider an exponential tilt of the clean teacher distribution. Let c denote the condition and $p_{\mathrm{T},0}(x_0 \mid c)$ the clean teacher distribution. Given a scalar reward $r(x_0, c)$, we define the reward-tilted distribution as

$$p_0^\star(x_0 \mid c) \propto p_{\mathrm{T},0}(x_0 \mid c) \exp\big(\beta r(x_0, c)\big), \tag{1}$$

where $\beta > 0$ controls the strength of reward tilting. This reweighting increases the relative probability of higher-reward samples while remaining within the support of the teacher distribution. As $\beta \to 0$, $p_0^\star$ reduces to $p_{\mathrm{T},0}$, recovering the standard teacher-matching objective. Equivalently, Eq. 1 arises as the solution to a KL-regularized expected-reward objective; the derivation is provided in Appendix E.1.

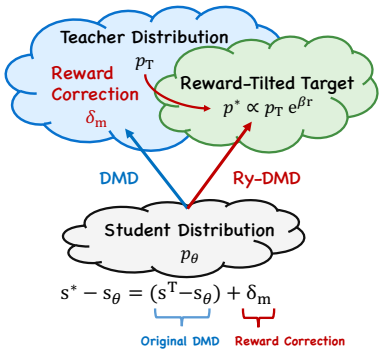


Figure 4: **Conceptual illustration of Ry-DMD.**

3.3.2 Local Reward-Tilted Teacher Target

The global target in Eq. 1 is not directly available during DMD training. Instead, at each noisy student state, we sample multiple teacher continuations and use their rewards to estimate how reward tilting would shift the local teacher target. During DMD training, the one-step generator predicts $x_0^\theta = G_\theta(z, c)$, which is re-noised to

$$x_\sigma = (1 - \sigma) x_0^\theta + \sigma \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

Starting from x_σ, a stochastic K-step teacher rollout produces $x_0 = \mathcal{T}_{\mathrm{T}}(x_\sigma, c; \eta)$, where η denotes the stochastic noise used during the rollout. These rollouts provide a Monte Carlo approximation of the conditional teacher endpoint distribution $q_{\mathrm{T}}^K(x_0 \mid x_\sigma, c)$. Extending the reward-tilting principle in Eq. 1 to this conditional distribution, we define a reward-reweighted rollout distribution as

$$q_{\mathrm{T},\beta}^K(x_0 \mid x_\sigma, c) \propto q_{\mathrm{T}}^K(x_0 \mid x_\sigma, c) \exp\big(\beta r(x_0, c)\big). \tag{3}$$

We estimate the resulting target shift using J independent teacher rollouts $\{x_0^{(j)}\}_{j=1}^J$ with rewards $r_j = r(x_0^{(j)}, c)$. Let π_j denote the normalized reward weight of the j-th rollout, with $\pi_j \propto \exp(\beta a_j)$ and $a_j = (r_j - \bar{r})/(s_r + \varepsilon)$, where rewards are standardized within each rollout group for robustness to reward scale. We then compute the original and reward-weighted teacher endpoint means as

$$\widehat{m}_{\mathrm{T}} = J^{-1} \textstyle\sum_{j=1}^J x_0^{(j)}, \qquad \widehat{m}_\beta = \textstyle\sum_{j=1}^J \pi_j x_0^{(j)}. \tag{4}$$

and use their difference

$$\delta m = \widehat{m}_\beta - \widehat{m}_{\mathrm{T}} \tag{5}$$

as the reward-induced shift of the local teacher target. Crucially, δm is estimated entirely from teacher samples and scalar reward evaluations, without differentiating through the reward function. The normalized form of the local tilt, the Monte Carlo estimator, and additional interpretation are provided in Appendix E.2.

3.3.3 Reward-Yield DMD Update

The reward-weighted rollouts tell us how the teacher endpoint target should move to favor higher-reward outputs. We incorporate this shift directly into the teacher term of DMD, while leaving the remaining distribution-matching procedure unchanged. Specifically, we convert the estimated endpoint displacement into a correction of the teacher term:

$$s_{\mathrm{T}}^{\mathrm{RY}} = s_{\mathrm{T}} + \lambda_{\mathrm{RY}} \frac{\delta m}{\max(\sigma, \varepsilon_\sigma)}, \tag{6}$$

where λ_{RY} controls the reward-correction strength and ε_σ stabilizes the update at small noise levels. The conversion from the endpoint displacement to the teacher correction is provided in Appendix E.3.

4 EXPERIMENTS

4.1 COMPARISON WITH STATE-OF-THE-ART VSR METHODS

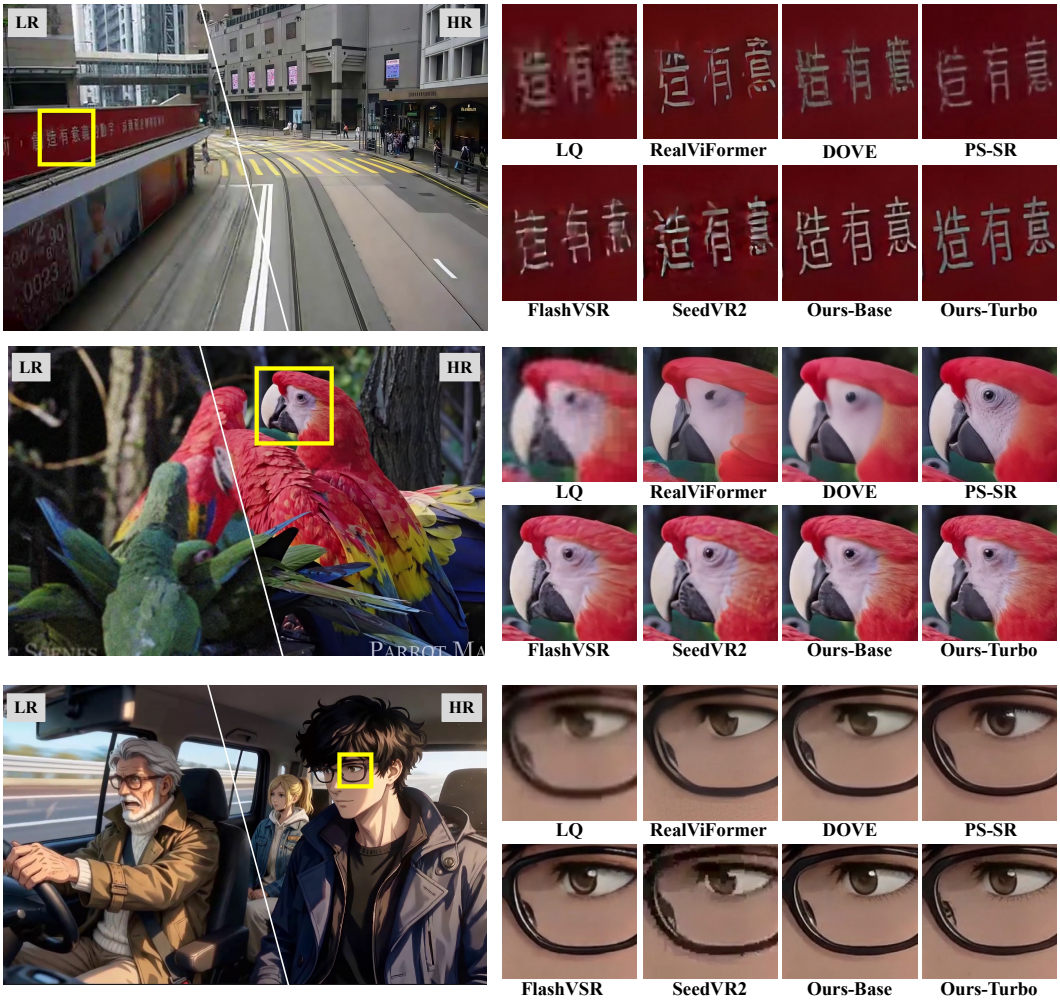


Figure 5: Qualitative comparison with representative VSR methods. Zoomed regions highlight the reconstruction of text, fine textures, and facial details.

We evaluate on synthetic benchmarks, including YouHQ40 (Zhou et al., 2024), SPMCS (Tao et al., 2017), and RealisVideo-4K (Zhao et al., 2025), as well as the real-world VideoLQ (Chan et al., 2022b) and AI-generated AIGC35 benchmarks. Table 2 compares HyR-VSR with representative VSR methods, including RealViFormer (Zhang & Yao, 2024), DOVE (Chen et al., 2026d), SeedVR2-3B (Wang et al., 2026), PS-SR (Wu et al., 2026), and FlashVSR (Zhuang et al., 2026). We report PSNR, SSIM, and LPIPS (Zhang et al., 2018) on datasets with ground truth, and MUSIQ (Ke et al., 2021), CLIPIQA (Wang et al., 2023), NIQE (Mittal et al., 2012), and DOVER (Wu et al., 2023a) on all datasets. HyR-VSR achieves strong overall restoration quality, with qualitative results in Fig. 5 showing better preservation of text, fine textures, and natural details with fewer spurious artifacts. We further conduct a blinded pairwise user study following the GSB protocol used in (Wang et al., 2026). Ours-Base serves as the anchor, and each method pair is evaluated on 20 videos randomly

Table 1: **Human preference study.** Preference margins relative to Ours-Base.

Method	Fidelity	Quality	Overall
RealViFormer	−70%	−72%	−73%
DOVE	−68%	−68%	−70%
DLoRAL	−38%	−28%	−35%
SeedVR2	−53%	−55%	−56%
PS-SR	−38%	−38%	−41%
FlashVSR	−15%	−10%	−15%
Ours-Turbo	−7%	−5%	−6%
Ours-Base	0%	0%	0%

Table 2: Quality comparison with state-of-the-art VSR methods under varying inference costs. Top-1/2/3 results are highlighted in red/orange/yellow.

Dataset	Metric	RealViFormer	DOVE	SeedVR2-3B	PS-SR	FlashVSR	Ours-Base	Ours-Turbo
YouHQ40	PSNR ↑	*23.50*	20.91	23.34	22.35	22.53	**23.64**	23.51
	SSIM ↑	0.6051	0.5532	0.6206	0.5899	0.5787	**0.6250**	*0.6097*
	LPIPS ↓	0.3528	0.3514	0.2768	0.3122	0.2721	**0.2698**	*0.2728*
	MUSIQ ↑	60.36	60.82	57.17	**66.20**	65.91	*64.71*	63.91
	CLIPIQA ↑	0.3823	0.3964	0.3703	**0.5063**	0.4701	0.4257	*0.4322*
	NIQE ↓	3.490	4.581	4.213	3.72	**3.385**	*3.560*	3.737
	DOVER ↑	10.42	10.97	12.23	12.93	**13.47**	*13.33*	13.37
SPMCS	PSNR ↑	22.64	**23.09**	22.24	21.16	21.35	22.34	*22.43*
	SSIM ↑	0.5651	**0.6109**	0.5867	0.5452	0.5398	*0.5714*	0.5654
	LPIPS ↓	0.3368	0.2738	**0.2521**	0.3240	*0.2680*	0.2700	0.2647
	MUSIQ ↑	61.91	68.93	65.63	69.47	*69.59*	70.09	**70.13**
	CLIPIQA ↑	0.3416	0.4605	*0.4633*	0.4650	0.4618	**0.4753**	0.4546
	NIQE ↓	3.781	4.234	3.601	3.971	**3.389**	3.401	*3.411*
	DOVER ↑	7.428	8.352	8.462	8.907	*9.250*	9.676	**9.811**
Realis-4K	PSNR ↑	30.81	23.26	*30.88*	28.35	30.84	31.23	**32.06**
	SSIM ↑	0.8240	0.7841	0.8206	0.8517	*0.8596*	0.8639	**0.8724**
	LPIPS ↓	0.1631	0.213	0.1929	0.1970	0.1671	*0.1639*	**0.1625**
	DOVER ↑	11.35	12.60	12.93	12.20	*13.27*	**13.43**	13.30
VideoLQ	NIQE ↓	4.485	4.915	5.114	5.00	**4.309**	*4.619*	4.784
	MUSIQ ↑	49.40	45.79	38.70	45.34	**49.70**	*47.64*	46.93
	CLIPIQA ↑	0.3117	0.3009	0.233	*0.3066*	**0.3140**	0.2939	0.3013
	DOVER ↑	7.217	7.542	5.955	7.769	*7.811*	**8.045**	7.827
AIGC35	NIQE ↓	4.983	5.173	**4.816**	4.923	*4.888*	4.851	5.014
	MUSIQ ↑	48.08	48.78	47.40	*50.08*	**50.83**	50.30	49.53
	CLIPIQA ↑	0.3871	0.4130	0.3945	**0.4546**	*0.4058*	0.4012	0.4001
	DOVER ↑	10.91	12.21	11.86	11.76	12.63	**13.01**	*12.57*

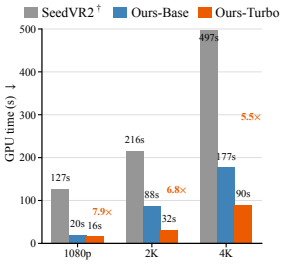
sampled from RealisVideo-4K, VideoLQ, and AIGC35 benchmarks. Participants compare the two outputs in terms of visual fidelity, perceptual quality, and overall preference. Table 1 reports the resulting preference margins relative to Ours-Base. Ours-Base is preferred over all compared methods, while Ours-Turbo remains close despite its higher inference efficiency.

Figure 6: Efficiency comparison on 89-frame videos. (a) GPU time and peak memory on NVIDIA H200. (b) End-to-end wall-clock inference time on NVIDIA RTX 4090, including CPU KV-cache offloading and asynchronous prefetching overhead.

(a) NVIDIA H200

Method	Backbone Params (B)	GPU Time (s) ↓			Peak Memory (GB) ↓		
		1080p	2K	4K	1080p	2K	4K
DOVE	5.0	89.38	165.3	374.7	26.54	26.22	26.56
PS-SR	1.3	228.4	1123.1	1751.4	47.79	50.63	50.63
SeedVR2	3.0	51.46	157.8†	410.5†	111.21	49.57†	51.76†
Ours-Base‡	1.3	7.91	17.28	46.34	13.11	21.29	45.00
Ours-Turbo	1.3	**6.63**	**12.31**	**30.59**	**8.72**	**13.23**	**26.33**

(b) NVIDIA RTX 4090



† SeedVR2 full-frame inference OOMs so we use tiled inference. ‡ Ours-Base shares the same architecture and inference procedure as FlashVSR on H200. FlashVSR's attention operators are unsupported on RTX 4090.

We achieve substantial efficiency gains for high-resolution VSR. We evaluate computational efficiency in Figure 6 using 89-frame videos at 1080p, 2K, and 4K resolutions on NVIDIA H200 and consumer RTX 4090 GPUs. We substantially reduce both inference latency and peak memory compared with the base model, enabling high-resolution restoration with much lower computational

overhead. In particular, with additional memory optimizations, we enable 4K VSR inference on a single RTX 4090; implementation details are provided in Appendix B.

4.2 Efficient High-Resolution VSR with Hybrid Attention

We evaluate hybrid attention both before and after adaptation. The training-free setting tests how much of the pretrained attention can be localized directly, while the adapted setting measures the resulting quality and efficiency after training the modified architecture.

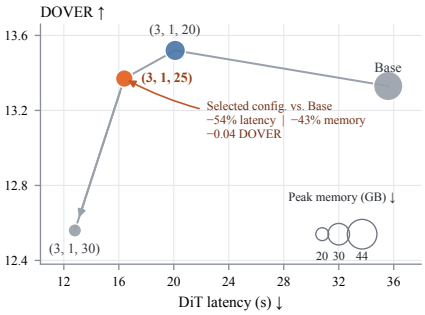


Figure 7: **Quality and efficiency of trained hybrid attention.** Increasing localization reduces DiT latency and peak memory while preserving quality up to $(3, 1, 25)$; further localization causes a clear quality drop.

We first consider the training-free setting. We use $(3, 1, 15)$ as a conservative reference and vary one factor at a time. Since the attention pattern is modified without retraining, we only count clear structural failures as artifacts, including block or grid-shaped corruption and severe temporal ghosting; minor changes in sharpness or texture are not considered failures. The pretrained model tolerates $(3, 1, 15)$, but reducing the spatial or temporal context, or increasing the number of localized blocks, quickly introduces artifacts. In particular, $(1, 1, 15)$, $(3, 0, 15)$, and configurations with $b \geq 20$ fail under direct replacement, motivating explicit adaptation of the modified attention structure.

We next adapt these hybrid-attention configurations. As shown in Fig. 7, adaptation substantially extends the usable range of hybrid attention. Increasing the number of local-attention blocks reduces latency and peak memory while preserving restoration quality up to $(3, 1, 25)$, whereas further localization to $(3, 1, 30)$ causes a clear quality drop. We therefore adopt $(3, 1, 25)$ as the default configuration. The more restrictive $(1, 1, 15)$ and $(3, 0, 15)$ configurations still exhibit visible artifacts after adaptation and are therefore excluded from the figure.

4.3 Reward-Guided VSR Post-Training

We evaluate Ry-DMD on OCR-guided VSR post-training using a synthetic text-focused dataset with annotated text ground truth; details of the data generation pipeline are provided in Appendix F. All methods start from the same pretrained one-step VSR model.

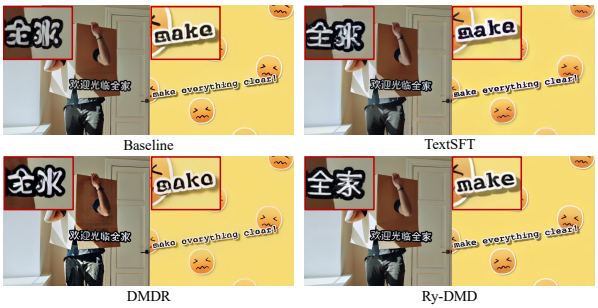


Figure 8: **Qualitative comparison of OCR-guided VSR post-training.** Ry-DMD restores more recognizable text, while DMDR often introduces artifacts.

Following the standard practice in Flow-GRPO (Liu et al., 2026), given target text y and the text $\hat{y}$ recognized by PaddleOCR (Cui et al., 2025), we compute the OCR reward as $r_{\text{OCR}} = 1 - \frac{\min(d_{\text{Lev}}(\hat{y}, y), |y|)}{|y|}$, where both strings are stripped of spaces before evaluation.

Table 3: **Quantitative results of OCR-guided post-training.**

Method	OCR ↑	PSNR ↑	LPIPS ↓	MUSIQ ↑	DOVER ↑
Baseline	0.624	23.94	0.246	75.35	9.79
TextSFT	0.656	22.02	0.295	**76.41**	10.58
DMDR	0.527	19.01	0.374	74.17	8.48
Ry-DMD	**0.776**	**24.86**	**0.240**	75.80	**10.90**

We compare four post-training strategies. Baseline denotes the original pretrained model without further adaptation. TextSFT continues standard DMD training on the text-focused dataset, testing whether task-specific data alone is sufficient. DMDR (Jiang et al., 2026) directly introduces an OCR optimization objective into DMD training; since the OCR reward is non-differentiable, we use a CTC-based differentiable surrogate and backpropagate through the OCR model. Instead, Ry-DMD directly uses the scalar OCR reward without requiring its gradient. As shown in Table 3, Ry-DMD achieves the strongest OCR improve-

ment while preserving restoration quality. The qualitative results in Fig. 8 further show clearer text restoration with fewer artifacts.

4.4 Generalization to Text-to-Image Generation

We further evaluate whether Ry-DMD generalizes beyond restoration by applying it to one-step SDXL text-to-image generation. We initialize the student from a one-step DMD2 model (Yin et al., 2024a) distilled from SDXL (Podell et al., 2024). Training uses OCR prompts from Flow-GRPO (Liu et al., 2026) together with a disjoint training split of DrawBench (Saharia et al., 2022). We optimize a combined reward consisting of OCR and PickScore. Ry-DMD relies only on scalar reward evaluations and does not backpropagate through either scorer.

Table 4: Reward-guided text-to-image generation on SDXL.

Model	NFE ↓	OCR ↑	PickScore ↑	HPSv2 ↑	ImageReward ↑	Aesthetic ↑
SDXL	50	0.125	22.2	27.5	0.87	**5.48**
DMD2	4	0.098	22.5	29.2	0.83	5.47
One-step DMD2	1	0.087	22.0	26.8	0.67	5.39
Ry-DMD (ours)	1	**0.241**	**22.7**	**29.6**	**1.02**	5.44

We report PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023b), ImageReward (Xu et al., 2023), and the LAION aesthetic predictor (Schuhmann et al., 2022). As shown in Table 4, Ry-DMD improves the OCR score from 0.087 to 0.241, surpassing the 50-step teacher, while maintaining or improving complementary quality metrics. This demonstrates that Ry-DMD also applies beyond VSR. Ry-DMD is robust to its hyperparameter choices, as further analyzed on SDXL in Appendix C. The same gradient-free formulation also extends naturally to few-step generation, with results on four-step SD3.5-M reported in Appendix D.

4.5 Rethinking DMD Scaling for One-Step VSR

To understand how to effectively scale up one-step VSR training, we first investigate whether allocating a larger optimization budget can consistently improve DMD performance. To reduce domain bias, we construct a fixed held-out evaluation set by sampling from multiple benchmarks, including AIGC, YouHQ, and VideoLQ, covering both synthetic and real-world content with diverse degradations. Intermediate checkpoints are evaluated using MUSIQ and DOVER.

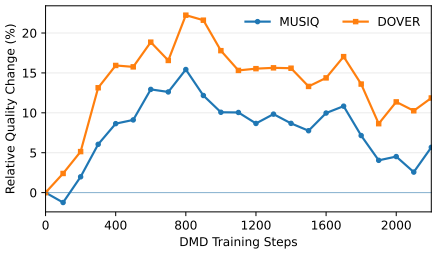


Figure 9: **Ablation on DMD steps.**

As shown in Fig. 9, both metrics improve rapidly during early DMD training. However, the gains do not persist: after an early peak, performance becomes non-monotonic and can decline with continued optimization. Simply increasing the number of DMD updates therefore does not reliably yield a better one-step VSR model.

We next examine two other components of the distillation setup: student initialization and teacher capacity. As shown in Table 5, both a stronger initialization and a larger teacher consistently improve the final student, with their combination performing best. These results indicate that effective DMD scaling depends on the overall distillation setup rather than optimization duration alone.

Table 5: **Ablation on student initialization and teacher capacity.**

Init.	Teacher	MUSIQ ↑	DOVER ↑
Base	1.3B	45.87	8.512
Strong	1.3B	48.20	8.738
Base	14B	51.76	9.144
Strong	14B	**52.78**	**9.417**

5 Conclusion

We introduce HyR-VSR, combining hybrid spatiotemporal attention with Ry-DMD for efficient and reward-guided one-step VSR. The framework improves the quality efficiency trade-off and enables spatially untiled 4K inference on a single 24GB RTX 4090. Ry-DMD enables black-box reward optimization without reward backpropagation, with benefits demonstrated in both VSR and text-to-image generation.

AI USE STATEMENT

In this work, we used generative AI tools to edit and polish the manuscript for grammar, wording, and readability. The tools were not used to generate research ideas, design the methodology or experiments, interpret experimental results, or formulate mathematical claims. All AI-assisted edits were reviewed and revised by the authors to ensure technical accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237. IEEE, 2018.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4945–4954. IEEE, 2021.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5962–5971. IEEE, 2022a.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5952–5961. IEEE, 2022b.

Junsong Chen, Jincheng Yu, Yitong Li, Shuchen Xue, Haozhe Liu, Jingyu Xin, Yuyang Zhao, Tian Ye, Zhangjie Wu, Zian Wang, et al. Sana-video 2.0: Hybrid linear attention with attention residuals for efficient video generation. *arXiv preprint arXiv:2607.21553*, 2026a.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Zhou Daquan, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. In *International Conference on Learning Representations*, volume 2026, pp. 38973–38997, 2026b.

Junyang Chen, Jiangxin Dong, Long Sun, Yixin Yang, and Jinshan Pan. Stcdit: Spatio-temporally consistent diffusion transformer for high-quality video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38281–38290, 2026c.

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. *Advances in Neural Information Processing Systems*, 38:85218–85237, 2026d.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9232–9241. IEEE, 2024.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

Linwei Dong, Ruoyu Guo, Ge Bai, Zehuan Yuan, Yawei Luo, and Changqing Zou. Guiding distribution matching distillation with gradient-based reinforcement learning. *arXiv preprint arXiv:2604.19009*, 2026.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in neural information processing systems*, 36:79858–79885, 2023.

Yushi Huang, Xiangxin Zhou, Ruoyu Wang, Chi Zhang, Jun Zhang, and Tianyu Pang. Reinforcing few-step generators via reward-tilted distribution matching. *arXiv preprint arXiv:2605.26108*, 2026.

Zhiwei Jia, Yuesong Nan, Huixi Zhao, and Gengdai Liu. Reward fine-tuning two-step diffusion models via learning differentiable latent-space surrogate reward. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12912–12922. IEEE, 2025.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, Changsheng Lu, Zhen Li, Mengmeng Wang, et al. Distribution matching distillation meets reinforcement learning. In *European Conference on Computer Vision*, pp. 281–299. Springer, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Xiaohui Li, Yihao Liu, Shuo Cao, Ziyan Chen, Shaobin Zhuang, Xiangyu Chen, Yinan He, Yi Wang, and Yu Qiao. Diffvsr: Revealing an effective recipe for taming robust video super-resolution against complex degradations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15319–15328. IEEE, 2025.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: (o(nlogn)) sparse attention with energy decay for long video generation. *Advances in Neural Information Processing Systems*, 38:16822–16852, 2026.

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. *Advances in neural information processing systems*, 35:378–393, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Yong Liu, Jinshan Pan, Yinchuan Li, Qingji Dong, Chao Zhu, Yu Guo, and Fei Wang. Ultravsr: Achieving ultra-realistic video super-resolution with efficient one-step diffusion space. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7785–7794, 2025.

Zhengyao Lv, Menghan Xia, Xintao Wang, and Kwan-Yee K Wong. Duo-vsr: Dual-stream distillation for one-step video super-resolution. *arXiv preprint arXiv:2603.22271*, 2026.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Claudio Rota, Marco Buzzelli, and Joost van de Weijer. Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models. In *European Conference on Computer Vision*, pp. 36–53. Springer, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Xuan Shen, Chenxia Han, Yufa Zhou, Yanyue Xie, Yifan Gong, Quanyi Wang, Yiwei Wang, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. Draftattention: Fast video diffusion via low-resolution attention guidance. *arXiv preprint arXiv:2505.14708*, 2025.

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.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

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. *Advances in Neural Information Processing Systems*, 38:172821–172841, 2026.

Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4482–4490. IEEE, 2017.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

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. In *International Conference on Learning Representations*, volume 2026, pp. 40987–41009, 2026.

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1954–1963. IEEE, 2019.

Aiqiu Wu, Zhaofan Qiu, Ting Yao, and Tao Mei. Ps-sr: Pseudo-single-step video super-resolution via speculative diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38218–38227, 2026.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20087–20097. IEEE, 2023a.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025b.

Haoxuan Xu, Yi Liu, Tianfu Li, Ruolin Shen, Boyuan Jiang, Jinlong Peng, Donghao Luo, Xiaobin Hu, Shuicheng Yan, and Haoang Li. Irpo: Boosting image restoration via post-training grpo. *arXiv preprint arXiv:2512.00814*, 2025a.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. Xattention: Block sparse attention with antidiagonal scoring. *arXiv preprint arXiv:2503.16428*, 2025b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2139–2149. IEEE, 2025c.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *Advances in Neural Information Processing Systems*, 38:96965–96991, 2026.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, and Lei Zhang. Gdpo-sr: Group direct preference optimization for one-step generative image super-resolution. *arXiv preprint arXiv:2603.16769*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *European conference on computer vision*, pp. 412–428. Springer, 2024.

Shijie Zhao, Xuanyu Zhang, Bin Chen, Weiqi Li, Qunliang Xing, Kexin Zhang, Yan Wang, Junlin Li, Li Zhang, Jian Zhang, et al. Oars: Process-aware online alignment for generative real-world image super-resolution. *arXiv preprint arXiv:2603.12811*, 2026.

Weisong Zhao, Jingkai Zhou, Xiangyu Zhu, Weihua Chen, Xiao-Yu Zhang, Zhen Lei, and Fan Wang. Realisvsr: Detail-enhanced diffusion for real-world 4k video super-resolution. *arXiv preprint arXiv:2507.19138*, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2535–2545. IEEE, 2024.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Diffusion reinforcement learning via centered reward distillation. *arXiv preprint arXiv:2603.14128*, 2026.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43482–43493, 2026.

A Training Details

In Stage 1, we train the 14B full-attention teacher for 4 days using LoRA with rank 384, AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, a learning rate of 2×10^{-5}, and a batch size of 16. In Stage 2, we initialize the hybrid-attention student from pretrained FlashVSR stage-2 checkpoint and perform full-parameter training for 2 days using AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, a learning rate of 1×10^{-5}, a batch size of 32, and EMA. In Stage 3, we perform full-parameter one-step distillation for 600 steps using the 14B teacher, AdamW with $(\beta_1, \beta_2) = (0, 0.999)$, a learning rate of 1×10^{-5}, a batch size of 16, and EMA. Training uses a filtered coreset together with the text-SFT augmentation. For Ry-DMD, we set $\beta = 0.1$, $\lambda_{\mathrm{RY}} = 0.2$, $J = 4$, and $K = 3$.

B RTX 4090 Implementation

The attention operators used in our H200 implementation are not directly supported on RTX 4090 GPUs. We therefore implement a lightweight Triton adaptation based on FlashAttention-2 to support the same hybrid-attention computation on Ada GPUs.

For 4K inference, the main memory bottleneck comes from the persistent KV cache of the retained global-attention layers. We selectively offload these KV caches to pinned CPU memory and asynchronously transfer them on a dedicated CUDA stream. Before the corresponding layer is executed in the next streaming chunk, the cache is prefetched back to GPU. This substantially reduces persistent GPU memory usage while overlapping most data transfers with computation. Importantly, this adaptation does not introduce spatial tiling: each frame is processed at its full spatial resolution, while only the KV-cache residency is changed. This enables spatially untiled 4K inference on a single 24GB RTX 4090.

C Ry-DMD Hyperparameter Analysis

We study the sensitivity of Ry-DMD to four key hyperparameters on SDXL: the reward temperature β, the correction weight λ_{RY}, the number of teacher rollouts J, and the rollout depth K. Unless otherwise specified, we use $\beta = 0.1$, $\lambda_{\mathrm{RY}} = 0.2$, $J = 4$, and $K = 3$. We vary one parameter at a time while keeping the others fixed. As shown in Table 6, Ry-DMD is stable under moderate variations of its hyperparameters, while overly strong reward weighting or correction degrades performance.

Table 6: **Hyperparameter analysis of Ry-DMD on SDXL.** We vary one hyperparameter at a time from the default configuration $(\beta, \lambda_{\mathrm{RY}}, J, K) = (0.1, 0.2, 4, 3)$.

Setting	β	λ_{RY}	J	K	OCR ↑	ImageReward ↑
Default	0.1	0.2	4	3	0.241	1.02
β	0.05	–	–	–	0.212	0.98
	0.2	–	–	–	0.236	1.01
	0.4	–	–	–	0.203	0.92
λ_{RY}	–	0.05	–	–	0.201	0.98
	–	0.1	–	–	0.237	1.01
	–	0.4	–	–	0.198	0.89
J	–	–	2	–	0.227	1.00
	–	–	8	–	0.248	1.04
K	–	–	–	2	0.239	1.01
	–	–	–	4	0.239	1.03

D Extending Ry-DMD to Few-Step Generation

Ry-DMD is not restricted to one-step generators. We further apply it to a 4-step SD3.5-M model for OCR-guided post-training. As shown in Table 7, Ry-DMD improves OCR accuracy from 0.501

Table 7: Ry-DMD on 4-step SD3.5-M. Ry-DMD substantially improves OCR accuracy while preserving the original 4-step sampling budget and improving general image-quality and preference metrics.

Method	NFE ↓	OCR ↑	HPSv2 ↑	ImageReward ↑	Aesthetic ↑
SD3.5-M (4-step)	4	0.501	30.3	1.02	5.52
+ Ry-DMD	4	**0.824**	**31.2**	**1.23**	**5.54**

to 0.824 while preserving the original 4-step sampling budget. Importantly, the improvement is not obtained by sacrificing general generation quality: HPSv2, ImageReward, and Aesthetic scores all improve after post-training.



Figure 10: **Qualitative comparison of Ry-DMD on 4-step SD3.5-M.** Ry-DMD substantially improves text correctness and legibility while preserving the intended scene content. Beyond text rendering, the resulting images often exhibit cleaner structures, more coherent materials and lighting, and improved overall visual quality.

Figure 10 provides qualitative examples. Ry-DMD substantially improves the correctness and legibility of rendered text, while preserving the semantic content specified by the prompt. We also observe cleaner object structures, more coherent materials and lighting, and more visually plausible scene composition in these examples. Together with the improvements in general-purpose reward metrics, these results suggest that optimizing the OCR objective with Ry-DMD does not induce evident reward hacking, and can improve text rendering without degrading overall image quality.

E Additional Derivations for Ry-DMD

This section provides supporting derivations for Ry-DMD introduced in Section 3.3. We first show that the global reward-tilted target arises from a KL-regularized expected-reward objective. We then derive the local reward-tilted endpoint mean and its finite-rollout estimator, followed by the connection between the resulting endpoint shift and the DMD teacher correction.

E.1 Derivation of the Reward-Tilted Target

For a fixed condition c, consider the KL-regularized expected-reward objective

$$\mathcal{J}(q) = \mathbb{E}_{x_0 \sim q(\cdot|c)}\left[r(x_0, c)\right] - \frac{1}{\beta} D_{\mathrm{KL}}\left(q(\cdot \mid c) \,\|\, p_{\mathrm{T},0}(\cdot \mid c)\right), \tag{7}$$

subject to $\int q(x_0 \mid c)\,\mathrm{d}x_0 = 1$. Introducing a Lagrange multiplier for the normalization constraint and setting the functional derivative with respect to q to zero gives

$$\log \frac{q^\star(x_0 \mid c)}{p_{\mathrm{T},0}(x_0 \mid c)} = \beta r(x_0, c) + \text{const.} \tag{8}$$

Therefore,

$$p_0^\star(x_0 \mid c) = \frac{p_{\mathrm{T},0}(x_0 \mid c)\exp(\beta r(x_0, c))}{Z(c)}, \qquad Z(c) = \mathbb{E}_{p_{\mathrm{T},0}(\cdot|c)}\left[\exp(\beta r(x_0, c))\right], \tag{9}$$

which recovers the reward-tilted target in Eq. 1. The exponential tilt changes the relative density of teacher-supported samples according to their rewards while preserving the support of the teacher distribution. In the limit $\beta \to 0$, $p_0^\star(\cdot \mid c) \to p_{\mathrm{T},0}(\cdot \mid c)$.

E.2 Local Reward-Tilted Mean Estimation

During DMD training, reward tilting is applied locally to the conditional endpoint distribution induced by stochastic teacher rollouts from the current noisy state. For fixed (x_σ, c), let $q_{\mathrm{T}}^K(x_0 \mid x_\sigma, c)$ denote the endpoint distribution induced by a stochastic K-step teacher sampler. The normalized form of Eq. 3 is

$$q_{\mathrm{T},\beta}^K(x_0 \mid x_\sigma, c) = \frac{q_{\mathrm{T}}^K(x_0 \mid x_\sigma, c)\exp(\beta r(x_0, c))}{Z_\beta(x_\sigma, c)}, \tag{10}$$

where

$$Z_\beta(x_\sigma, c) = \mathbb{E}_{x_0 \sim q_{\mathrm{T}}^K(\cdot|x_\sigma,c)}\left[\exp(\beta r(x_0, c))\right]. \tag{11}$$

The mean endpoint under the reward-tilted distribution is therefore

$$m_\beta(x_\sigma, c) = \frac{\mathbb{E}_{x_0 \sim q_{\mathrm{T}}^K(\cdot|x_\sigma,c)}\left[x_0 \exp(\beta r(x_0, c))\right]}{\mathbb{E}_{x_0 \sim q_{\mathrm{T}}^K(\cdot|x_\sigma,c)}\left[\exp(\beta r(x_0, c))\right]}, \tag{12}$$

while the original conditional teacher mean is

$$m_{\mathrm{T}}(x_\sigma, c) = \mathbb{E}_{x_0 \sim q_{\mathrm{T}}^K(\cdot|x_\sigma,c)}[x_0]. \tag{13}$$

We denote the population-level displacement induced by reward tilting as

$$\Delta m = m_\beta(x_\sigma, c) - m_{\mathrm{T}}(x_\sigma, c). \tag{14}$$

In practice, we draw J independent teacher rollouts

$$x_0^{(j)} \overset{\text{i.i.d.}}{\sim} q_{\mathrm{T}}^K(\cdot \mid x_\sigma, c), \qquad j = 1, \ldots, J, \tag{15}$$

with rewards $r_j = r(x_0^{(j)}, c)$. Without reward standardization, the corresponding self-normalized Monte Carlo estimates are

$$\widehat{m}_\beta = \sum_{j=1}^{J} \pi_j x_0^{(j)}, \qquad \pi_j = \frac{\exp(\beta r_j)}{\sum_{k=1}^{J}\exp(\beta r_k)}, \tag{16}$$

and

$$\widehat{m}_{\mathrm{T}} = \frac{1}{J}\sum_{j=1}^{J} x_0^{(j)}. \tag{17}$$

The finite-rollout estimate of the reward-induced endpoint shift is thus

$$\delta m = \widehat{m}_\beta - \widehat{m}_{\mathrm{T}}, \tag{18}$$

corresponding to Eq. 5. The expression above gives the unstandardized estimator of Δm. In practice, we compute δm using the group-standardized weights below, as in Eq. 5.

Reward standardization. In the practical algorithm, rewards are standardized within each rollout group as

$$a_j = \frac{r_j - \bar{r}}{s_r + \varepsilon}, \qquad \pi_j = \frac{\exp(\beta a_j)}{\sum_{k=1}^{J} \exp(\beta a_k)}. \tag{19}$$

Centering by $\bar{r}$ does not affect the normalized exponential weights, since the common multiplicative factor cancels. Dividing by $s_r + \varepsilon$ is equivalent to using a rollout-dependent effective temperature

$$\beta_{\text{eff}} = \frac{\beta}{s_r + \varepsilon}, \tag{20}$$

which reduces sensitivity to the numerical scale of different reward functions while preserving their within-group ordering. The standardized weights therefore implement a group-adaptive tilt, while the fixed-β population formulas above describe the unstandardized case.

Small-β interpretation. For additional intuition, a first-order expansion of Eq. 12 gives

$$\Delta m = \beta \operatorname{Cov}_{q_{\mathrm{T}}^{K}(\cdot|x_\sigma,c)} (x_0, r(x_0, c)) + \mathcal{O}(\beta^2). \tag{21}$$

Thus, in the small-β regime, reward tilting moves the conditional teacher endpoint mean along variations in x_0 that covary positively with the reward.

E.3 From Endpoint Shift to the DMD Teacher Correction

We finally derive how the endpoint displacement estimated above translates into the reward correction in Eq. 6. In this section, s_{T} denotes the parameterized teacher term used in our DMD update. For $\sigma > 0$, it is related to its associated clean-endpoint target$\mu_{\mathrm{T}}(x_\sigma, c)$ by

$$s_{\mathrm{T}}(x_\sigma, c) = \frac{\mu_{\mathrm{T}}(x_\sigma, c) - x_\sigma}{\sigma}. \tag{22}$$

Ry-DMD modifies the clean teacher target using the estimated reward-induced endpoint displacement:

$$\mu_{\mathrm{T}}^{\mathrm{RY}}(x_\sigma, c) = \mu_{\mathrm{T}}(x_\sigma, c) + \lambda_{\mathrm{RY}} \delta m, \tag{23}$$

where λ_{RY} controls the correction strength. Substituting this target into Eq. 22 yields

$$\begin{aligned} s_{\mathrm{T}}^{\mathrm{RY}}(x_\sigma, c) &= \frac{\mu_{\mathrm{T}}(x_\sigma, c) + \lambda_{\mathrm{RY}} \delta m - x_\sigma}{\sigma} \\ &= s_{\mathrm{T}}(x_\sigma, c) + \lambda_{\mathrm{RY}} \frac{\delta m}{\sigma}. \end{aligned} \tag{24}$$

Thus, under the above parameterization, the factor $1/\sigma$ follows directly from converting a clean-endpoint displacement into a displacement of the teacher term. For numerical stability at small noise levels, we use

$$d_{\mathrm{RY}} = \frac{\delta m}{\max(\sigma, \varepsilon_\sigma)}, \tag{25}$$

which gives the practical update

$$s_{\mathrm{T}}^{\mathrm{RY}} = s_{\mathrm{T}} + \lambda_{\mathrm{RY}} d_{\mathrm{RY}}, \tag{26}$$

equivalent to Eq. 6. Thus, Ry-DMD leaves the remaining DMD optimization procedure unchanged and requires only sampled teacher endpoints and scalar reward evaluations, without differentiating through the reward function. Here, μ_{T} remains the original teacher target, and δm is used as a rollout-estimated residual correction; the construction does not require $\mu_{\mathrm{T}} = m_{\mathrm{T}}$.

F Synthetic Text-OCR VSR Dataset

To study OCR-guided post-training, we construct a synthetic text-focused VSR dataset from the same video corpus used in our VSR pipeline. Synthetic text is rendered onto decoded high-quality frames to form the target video, and the corresponding low-quality input is generated with the standard VSR degradation pipeline. Since the text is procedurally generated, its content and spatial annotations are known exactly.

Text rendering. Each sample contains one or two text instances sampled from a multilingual corpus covering English, Chinese, numeric strings, and several additional languages. We randomize font, size, color, outline, opacity, and shadow. Text remains temporally coherent across the video and follows simple motion patterns such as static placement, translation, scale/rotation changes, or partial motion outside the frame. Per-frame locations and visibility are retained as annotations.

Filtering and OCR supervision. We discard invalid renderings with unsupported characters, insufficient size, or negligible visible area. For OCR reward computation, only sufficiently visible annotated text regions are evaluated, and a small number of valid frames are sampled from each video for efficiency. All compared post-training methods use the same source videos, rendered targets, degradation process, and filtering rules.

Each sample additionally stores lightweight text metadata used for OCR supervision. A schematic example is shown below:

```
{
  "source_path": "/videos/example.mp4",
  "gt_video": "[F, 3, H, W]",
  "lq_video": "[F, 3, H, W]",
  "text_metadata": {
    "instances": [{
      "text": "EXAMPLE 42",
      "language": "en",
      "motion_type": "translate",
      "frames": [
        {"frame_index": 0,
         "bbox": [x0, y0, x1, y1],
         "visible_ratio": 1.0}
      ]
    }]
  }
}
```