HARNESSING VIDEO DIFFUSION MODEL AS PRIOR AND SUPERVISION FOR VIDEO RESTORATION

Anonymous authors
Paper under double-blind review

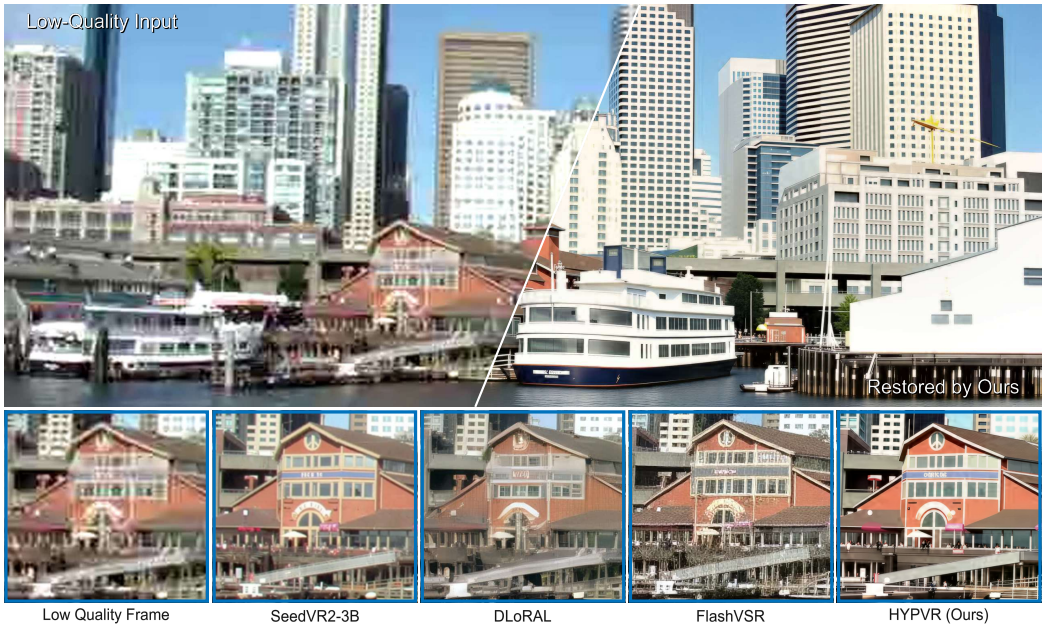


Figure 1: Visual comparison on a video with large camera motion. HYPVR combines a pretrained video diffusion backbone with a video distribution matching loss and sampled-frame image adversarial supervision to restore fine-grained structures and textures from the low-quality input.

ABSTRACT

Adversarial transfer has proven an effective way of exploiting pretrained diffusion priors for image restoration: the restoration model is initialized from a diffusion prior and optimized adversarially against a discriminator that is itself a strong pretrained image model, which yields single step restoration at high perceptual quality. Extending this recipe to video is not straightforward. A pretrained video diffusion model can initialize the restoration network, but pretrained image discriminators act frame by frame, and a comparably capable pretrained video discriminator is not readily available to judge spatial appearance and temporal dynamics jointly. We, therefore, use distribution matching with the score of a pretrained video generation model to supervise temporal dynamics without requiring a video discriminator. This also avoids VSR-specific diffusion pretraining, reducing training costs and complexity. For spatial supervision, we apply image adversarial training to one randomly sampled frame per clip, whose gradient largely supports the full-frame image adversarial update. Extensive experiments demonstrate that the proposed HYPVRachieves state-of-the-art spatial quality and temporal consistency with a single forward pass, making effective use of the prior for efficient video restoration.

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.

1 INTRODUCTION

Pretrained diffusion models have become the primary source of generative priors for image restoration. Recent work shows that initializing a restoration network from a pretrained diffusion generative model and transferring the generative prior to the restoration task through adversarial training yields high quality single step image restoration model (Lin et al., 2025). We aim to leverage generative priors for video restoration with similar effectiveness, seeking to obtain a model capable of restoring rich visual details and temporal coherence through a single forward pass.

Supplying effective spatiotemporal supervision for generative video restoration is a challenging problem. Early works use optical flow to establish correspondences between frames and propagate information about existing content through alignment and fusion (Wang et al., 2019; Tian et al., 2020; Chan et al., 2021b; 2022a). Yet, generative restoration must also recover details absent from the input, and correspondences between existing content alone are insufficient to assess whether these newly added details evolve plausibly across frames. Another line of work characterizes the video distribution implicitly with a learned model such as a spatiotemporal discriminator (Wang et al., 2025a), which has to be finely discriminative in space and sensitive in time simultaneously. How to reconcile temporal consistency with the high visual quality that generative restoration demands, at an affordable training cost, is the central question we address.

We propose video distribution matching (Video DM) as an alternative form of video-prior supervision. Video DM provides temporal supervision from a native video generative prior, constraining how generated details evolve across frames. In our method, this supervision comes from a general purpose pretrained video model, and by selecting a noise timestep range suited to the restoration task, it is optimized jointly with a spatially aware objective. This differs from the common use of DM for distillation, which transfers a multi-step teacher's generative ability to a few-step model (Yin et al., 2024b;a; Zhuang et al., 2025). We show that VSR-specific diffusion pretraining is not required for high-quality single-step video restoration: a restorer initialized from a pretrained video model can learn directly from the distributional supervision of a general-purpose video teacher, reducing training cost and complexity while better balancing generation quality and temporal consistency.

Even with spatiotemporal supervision from distribution matching, further improving per-frame visual quality remains challenging. As noted by Lin et al. (2025), video priors remain limited in visual quality compared with image priors, and video-level supervision alone may struggle to capture fine textures and local realism. Per-frame image adversarial supervision can help address these limitations, but the costs of high-resolution decoding, discrimination, and backpropagation scale with clip length. We find that the image-adversarial gradient from a single sampled frame largely aligns with the update obtained by supervising all frames, suggesting redundancy in per-frame spatial supervision. We therefore propose applying the image adversarial loss to one randomly sampled frame per clip, varying its position across iterations to approximate the all-frame spatial objective at a low pixel-domain cost. This makes high-quality image supervision practical for video training and, together with full-clip Video DM, improves both spatial fidelity and temporal consistency.

This gives our method, HYPVR. A pretrained video model both provides the initialization prior for the restoration network and, through distribution matching, supplies spatiotemporal supervision; the image adversarial objective on sampled frames and the fidelity objective together improve spatially aware quality and maintain correspondence with the input content. All of this supervision acts on a single video restoration model, and inference requires only one generator forward pass. Extensive experiments demonstrate that, given the same video generative prior, HYPVRachieves state-of-the-art spatial quality and temporal consistency while maintaining high inference efficiency, making effective use of the prior for video restoration.

2 RELATED WORK

Video Restoration Networks Early video restoration methods model spatiotemporal relations through 3D convolutions, feature alignment, recurrent propagation, and optical-flow-guided bidirectional propagation (Caballero et al., 2017; Jo et al., 2018; Wang et al., 2019; Tian et al., 2020; Fuoli et al., 2019; Haris et al., 2019; Chan et al., 2021b; 2022a). Vision transformers subsequently strengthen long-range spatiotemporal modeling (Dosovitskiy et al., 2020; Liang et al., 2024), with patch-level alignment or masked intra- and inter-frame attention reducing redundant computation

(Shi et al., 2022; Zhou et al., 2024b; Zhang & Yao, 2024). These methods use deterministic objectives and idealized degradations, often oversmoothing results under complex real-world degradation.

Generative Priors for Restoration Generative models can supplement severely degraded inputs with perceptually realistic detail. GANs reduce the discrepancy between generated and real distributions through adversarial training (Goodfellow et al., 2020), and have been widely used for image restoration (Ledig et al., 2017; Wang et al., 2018; 2021; Pan et al., 2021; Chan et al., 2021a; Zhu et al., 2022) and extended to video with temporal constraints (Lucas et al., 2019; Chu et al., 2020; Brooks et al., 2022; Xu et al., 2025). Diffusion models have become still stronger image restoration priors (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022; Dhariwal & Nichol, 2021; Saharia et al., 2022; Wang et al., 2024; Lin et al., 2024; Yu et al., 2024; Yang et al., 2024a; Chen et al., 2025; Dong et al., 2025). HYPIR initializes a restoration network from a pretrained diffusion model and adapts it adversarially for high-quality single-step image restoration (Lin et al., 2025). Applying image diffusion models independently to video, however, causes temporal inconsistency and motivates priors that jointly model spatial content and temporal variation.

Recent generative video restoration methods take different routes. STAR adapts a T2V diffusion model with local information enhancement and a dynamic frequency loss (Xie et al., 2025); DLoRAL adapts the spatial and temporal capabilities of a text-to-image foundation model with dual LoRAs (Sun et al., 2025; Hu et al., 2022); FlashVSR uses staged distillation for one-step streaming VSR (Zhuang et al., 2025); and SeedVR improves the attention efficiency of video diffusion (Wang et al., 2025b), while SeedVR2 further uses adversarial post-training for single-step restoration (Wang et al., 2025a). These methods use generative priors but obtain spatial and temporal capabilities differently.

Where Spatial and Temporal Priors Come From Spatial realism is consistently drawn from large-scale generative models, whether used as inference-time samplers (Wang et al., 2024; Lin et al., 2024; Yu et al., 2024), restoration initialization (Lin et al., 2025), or video backbones (Wang et al., 2025b; Xie et al., 2025; Zhuang et al., 2025). Temporal priors, by contrast, enter restoration systems through diverse mechanisms, each with distinct limitations. Explicit alignment and propagation effectively transfer structures and textures observable in the low-quality input (Wang et al., 2019; Tian et al., 2020; Chan et al., 2021b; 2022a; Yang et al., 2024b), but cannot fully determine how missing high-frequency detail should be generated and evolve. Temporal attention, motion modules, and parameter-efficient adaptation (Zhou et al., 2024a; Chen et al., 2024; Guo et al., 2023; Sun et al., 2025), temporally correlated noise (Chang et al., 2025; Burgert et al., 2025), and temporal latent representations (Kouzelis et al., 2025) improve cross-frame coherence, yet plausible motion and low-frequency structure do not necessarily stabilize newly generated textures. Pretrained video backbones directly supply learned temporal capabilities (Wang et al., 2025b; Xie et al., 2025; Zhuang et al., 2025), but standard restoration objectives need not preserve their alignment with the original video distribution during strong perceptual optimization.

Temporal supervision still largely relies on pixel differences, optical-flow warping, and flow-based consistency losses (Lai et al., 2018; Teed & Deng, 2020). Such deterministic objectives cannot cleanly separate flicker from legitimate high-frequency variation caused by motion, and may reduce error by smoothing texture. Task-specific video discriminators provide distributional supervision, but must learn spatial and temporal discrimination from restoration data, making them costly and sensitive to optimization. We instead use distribution matching (Yin et al., 2024b;a) to introduce the joint spatiotemporal distribution of a pretrained video model directly as restoration supervision. Related to score-distillation objectives (Poole et al., 2023; Wang et al., 2023b), it requires neither a video discriminator nor a teacher pretrained for the full VSR mapping; instead, a general-purpose video score constrains generated detail across space and time.

3 METHOD

3.1 OVERALL FRAMEWORK

Given a low quality (LQ) video $\mathbf{y} = \{y_i\}_{i=1}^{T}$, HYPVR aims to recover a high quality (HQ) video $\hat{\mathbf{x}} = \{\hat{x}_i\}_{i=1}^{T}$ with natural spatial detail and plausible temporal evolution. We initialize the restoration generator G_θ from a pretrained video diffusion model. With E and D denoting the video VAE encoder and decoder, we write the restoration process as $\hat{\mathbf{z}} = G_\theta(E(\mathbf{y}))$ and $\hat{\mathbf{x}} = D(\hat{\mathbf{z}})$. This

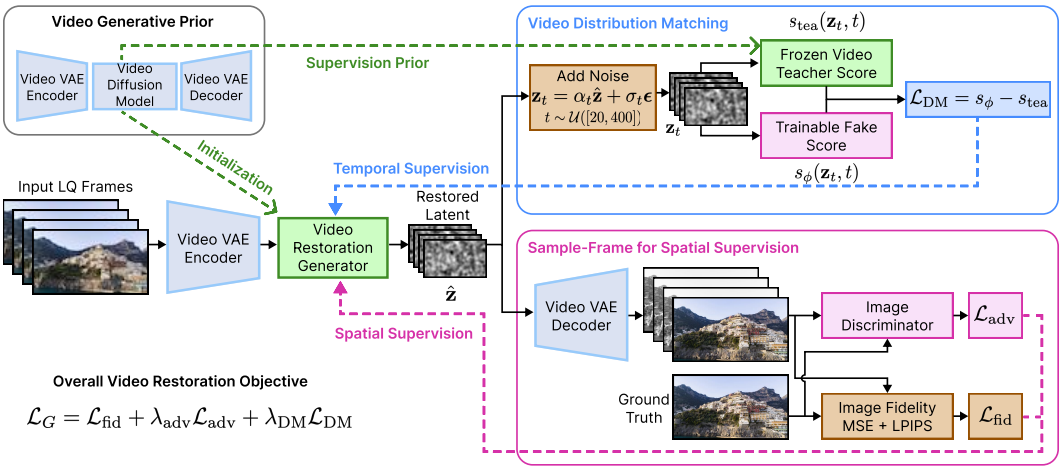


Figure 2: Overview of HYPVR. A pretrained video model initializes the single-step restoration generator and provides a frozen teacher score for Video DM on the full latent clip; a trainable fake score tracks the generated distribution. Image adversarial and fidelity losses act on one sampled decoded frame, combining efficient spatial supervision with full-clip spatiotemporal supervision.

notation omits input resolution adaptation, latent noise augmentation, and other model conditions. The generator processes all frames jointly, restoring the full latent sequence in one forward pass.

Figure 2 summarizes two complementary forms of distributional supervision. Video DM uses a general purpose pretrained video teacher to constrain the joint spatiotemporal relations among generated details over the complete output latent sequence. Image adversarial supervision acts on one randomly sampled decoded frame per clip, exploiting redundancy in spatial gradients across frames to improve per frame perceptual quality efficiently. A fidelity objective on paired high quality video maintains correspondence between the restored result and the input content.

We train the single step mapping directly, without additional VSR specific diffusion pretraining. The general purpose video teacher remains frozen and supplies distribution level update directions through its score, so pretrained video knowledge informs both initialization and supervision.

3.2 Video Distribution Matching as Temporal Supervision

Temporal supervision must constrain the appearance and evolution of missing detail, both underspecified by the LQ input. We use a general purpose video model's joint spatiotemporal distribution for content dependent supervision. Let $p_{\theta,t}$ denote the marginal distribution of generator outputs over training inputs after noise perturbation, and $p_{\text{tea},t}$ the video distribution characterized by the teacher at the same noise level. The idealized distribution matching objective is

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{t\sim q(t)}\left[w(t)\, D_{\text{KL}}(p_{\theta,t} \,\|\, p_{\text{tea},t})\right], \tag{1}$$

where $q(t)$ is the sampling distribution over supervised timesteps and $w(t)$ is a time weight. The shared text condition is omitted for brevity. The general purpose video teacher supplies a natural video distribution, while paired data teaches the restoration network the input to output mapping. The teacher need not learn the full VSR task or demonstrate that mapping through multi step restoration. Matching the complete clip constrains inter frame dependencies while jointly adjusting local structure and detail; we optimize this spatially aware temporal supervision with the image adversarial objective.

Concretely, we perturb the generated latent sequence with noise as $\mathbf{z}_t = \alpha_t\hat{\mathbf{z}} + \sigma_t\boldsymbol{\epsilon}$, where $\alpha_t = 1-\sigma_t$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The frozen teacher and fake network share text but not LQ conditioning; we convert their flow velocities to clean-latent estimates and obtain scores $s_{\text{tea}}(\mathbf{z}_t, t)$ and $s_\phi(\mathbf{z}_t, t)$ as in DMD (Yin et al., 2024b). To track generated latents, the fake velocity v_ϕ minimizes $\mathcal{L}_{\text{fake}} = \mathbb{E}_{\mathbf{y},t,\boldsymbol{\epsilon}}\|v_\phi(\mathbf{z}_t, t) - (\boldsymbol{\epsilon} - \hat{\mathbf{z}})\|_2^2$, with $\hat{\mathbf{z}}$ detached in both $\mathbf{z}_t$ and the target. The Video DM gradient with respect to the generator is

$$\nabla_\theta \mathcal{L}_{\text{DM}} \approx \mathbb{E}_{\mathbf{y},t,\boldsymbol{\epsilon}}\left[w(t)\alpha_t J_\theta^\top \left(s_\phi(\mathbf{z}_t, t) - s_{\text{tea}}(\mathbf{z}_t, t)\right)\right], \qquad J_\theta = \frac{\partial \hat{\mathbf{z}}}{\partial \theta}. \tag{2}$$

Generator updates detach both score predictions, propagating their difference to θ through the generated latents; the teacher remains frozen.

3.3 Choosing the Noise Range for Restoration Supervision

The noise level determines how much of the restored content remains visible to the teacher: high noise removes existing structure and emphasizes its generative prior, whereas low noise preserves the output and yields content-specific adjustment directions. In HYPVR, video initialization and fidelity supervision already anchor main content and motion, while image adversarial training improves per-frame quality. We therefore choose the Video DM noise range to complement these objectives and constrain plausible spatiotemporal relations among generated details.

Following the model's diffusion timestep indexing, we sample Video DM timesteps uniformly from $\{20, \ldots, 400\}$ instead of the wider $\{20, \ldots, 980\}$. This affects only training-time teacher supervision; inference remains single step. At $t_{\max} = 400$, no-reference quality improves over the wider range while fidelity and warping error remain comparable; lowering it to 200 degrades both fidelity and temporal consistency. Thus, limiting high-noise supervision benefits spatially aware optimization, but using only very low noise cannot stabilize the cross-frame relations among generated details. We select $[20, 400]$ to balance video distributional supervision and image perceptual optimization.

Low-noise Video DM lets a general purpose video teacher supervise restoration: VSR-specific pretraining yields no meaningful overall gain. Balancing quality and cost, we use WAN2.1-T2V-1.3B as the fixed teacher with CFG scale 5 (Table 7 in Appendix C).

3.4 Sampled Frame Spatial Supervision from Gradient Redundancy

Video DM constrains spatiotemporal relations, while fine texture and local realism also require strong image-level perceptual supervision. Applying image adversarial training to every frame incurs high-resolution decoding, feature extraction, and backpropagation. We exploit redundancy across frames to retain this supervision at lower cost.

Frames within a clip often share subjects, scenes, and textures, so their image-adversarial supervision tends to point in similar directions. This motivates estimating the full-frame spatial objective from sampled frames.

For a fixed video, network state, and other randomness, let ℓ_i be the image adversarial loss on frame i; the full-frame mean has parameter gradient $g_{\text{all}} = \frac{1}{T}\sum_{i=1}^{T} \nabla_\theta \ell_i$. Sampling $i \sim \mathcal{U}\{1, \ldots, T\}$ gives $\hat{g} = \nabla_\theta \ell_i$. If sampled decoding uses the same differentiable computation as the full-frame objective, $\mathbb{E}_i[\hat{g}] = g_{\text{all}}$, with variance $\mathbb{E}_i[\|\hat{g} - g_{\text{all}}\|_2^2] = \frac{1}{T}\sum_{i=1}^{T} \|\nabla_\theta \ell_i - g_{\text{all}}\|_2^2$. Thus, uniform sampling is unbiased, while similarity among per-frame parameter gradients determines its variance. We quantify latent-space directional agreement with positive gradient contribution ratio (PGCR): the share of absolute coordinate-wise inner products aligned with the full-frame image-adversarial gradient (formula in Appendix B.1). Figure 3 shows that one-frame gradients remain predominantly aligned throughout training.

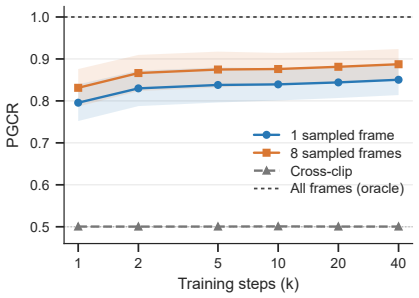


Figure 3: Positive Gradient Contribution Ratio (PGCR) of sampled- versus full-frame latent gradients. Values above 0.5 indicate dominant aligned contributions; shading shows variability.

In the implementation, we sample one output frame uniformly per video at each iteration and compute pixel domain supervision on that frame alone. Following existing sampled decoding practice, the output of that frame is written as $\hat{x}_i = D_i(\hat{\mathbf{z}})$, where D_i comprises the decoding operations, together with the necessary temporal context, that are required to obtain the ith output frame. We retain the differentiable decoding path only for the pixel domain losses on that frame.

Let A_ψ denote the image discriminator. The spatial adversarial objective is

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{y},i}\left[\ell_{\text{adv}}\left(A_\psi(\hat{x}_i)\right)\right]. \tag{3}$$

The discriminator is built on pretrained 2D visual features, and ℓ_{adv} follows the generator adversarial objective of HYPIR. The discriminator is updated with high quality real frames and generated frames. Resampling frames each iteration spreads spatial supervision throughout the video over training.

Fidelity supervision is applied to the sampled frame as well. Given a high quality reference video $\mathbf{x}$ paired with the input, we use

$$\mathcal{L}_{\text{fid}} = \mathbb{E}_{(\mathbf{y},\mathbf{x}),i}\left[\lambda_{\text{mse}}\|\hat{x}_i - x_i\|_2^2 + \lambda_{\text{lpips}}\,\text{LPIPS}(\hat{x}_i, x_i)\right]. \tag{4}$$

With $\lambda_{\text{mse}} = 1$ and $\lambda_{\text{lpips}} = 2$, this objective preserves reference content and improves spatial quality alongside image adversarial supervision.

What frame sampling reduces is the cost of pixel domain supervision and of the associated decoding. The generator still processes the complete video latent sequence jointly, and Video DM still acts on the complete clip. Thus, high quality image supervision incurs little training cost without losing video level spatiotemporal modeling or distributional constraints.

3.5 Joint Training and Inference

The joint training objective for the generator is

$$\mathcal{L}_G = \mathcal{L}_{\text{fid}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{DM}}\mathcal{L}_{\text{DM}}, \tag{5}$$

where λ_{adv} and λ_{DM} weight image adversarial supervision and Video DM, respectively; the latter uses the approximate gradient in Eq. 2. We set $\lambda_{\text{adv}} = 0.5$ and $\lambda_{\text{DM}} = 1.0$ throughout training.

During training the generator first produces the complete video latent sequence. We compute the fidelity and image adversarial losses on one randomly sampled frame per clip, and the Video DM gradient over the complete clip within the selected noise timestep range. The generator receives the joint update from these objectives; the image discriminator is updated with real and generated frames; the fake score network is updated with stop gradient generated latent sequences; and the general purpose video teacher is kept fixed.

At inference, the image discriminator, the fake score network and the video teacher take no part in the computation, and only the restoration generator and the video VAE are retained. The input video is encoded, the generator produces the restored latent sequence in a single forward pass, and the decoder outputs the high quality video.

4 Experiments

4.1 Experimental Setup

Implementation Details. We train on a further-cleaned subset of an initial collection of 62K high-quality YouTube clips, processing videos at 1024×1024 and training exclusively on random 512×512 crops. Low-quality inputs strictly follow the original RealBasicVSR degradation (Chan et al., 2022b). We initialize from pretrained WAN2.1-1.3B (Wan et al., 2025) without additional VSR diffusion pretraining, and apply HYPIR's latent noise augmentation (Lin et al., 2025) to the encoded LQ input $E(\mathbf{y})$ before the generator, with $t_{\text{aug}} \sim \mathcal{U}[200, 800]$. Training uses 29-frame clips for 20K steps, then 61-frame clips for 10K, with rank-64 LoRA, RMSProp at 10^{-5}, total batch size 16, and EMA decay 0.999; it takes about 43 hours on one machine with eight NVIDIA H200 GPUs.

Datasets and Degradations. We evaluate on synthetic POND25, SPMC30 (Tao et al., 2017), UDM10 (Yi et al., 2019), and YouHQ40 (Zhou et al., 2024a), and real-world VideoLQ (Chan et al., 2022b). POND25 comprises 25 professionally captured Pond5[1] clips, including scenes with multiple moving subjects and large camera motion. We downsample its 2560×1440 sources to 1280×720 before degradation to challenge fine-detail recovery and temporal stability. Synthetic inputs use the original RealBasicVSR degradation (Chan et al., 2022b); VideoLQ receives no synthetic degradation.

Evaluation Metrics. PSNR (Wikipedia contributors, 2024), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) measure full-reference fidelity and perceptual distance; CLIP-IQA (Wang et al., 2023a), MUSIQ (Ke et al., 2021), and DOVER (Wu et al., 2023) assess no-reference quality; and warping error (E-Warp) measures temporal consistency. VideoLQ lacks ground-truth frames, so full-reference metrics and E-Warp, which requires ground-truth-derived flow, are unavailable.

[1] https://www.pond5.com/

Table 1: Quantitative comparison on synthetic and real-world benchmarks. Best and second-best results are red and blue; UAV denotes Upscale-A-Video. HYPVR balances perceptual quality and temporal consistency on synthetic sets and leads DOVER on all five benchmarks.

Dataset	Metric	RealViformer	MGLD-VSR	UAV	STAR	SeedVR2-3B	DLoRAL	FlashVSR	HYPVR (Ours)
POND25	LPIPS↓	0.3864	0.3722	0.4608	0.5329	0.2661	0.3673	0.3049	0.3027
	MUSIQ↑	63.3772	57.1638	63.7041	30.6886	58.7561	67.2696	63.4249	68.2838
	DOVER↑	0.5056	0.4330	0.4526	0.2682	0.5487	0.4579	0.5613	0.6329
	E-Warp↓	5.3201	5.4009	8.4695	2.7547	3.8804	7.3952	5.3837	4.5978
SPMC30	LPIPS↓	0.3785	0.3948	0.4761	0.5903	0.2774	0.3734	0.3352	0.3180
	MUSIQ↑	63.1467	57.4914	68.3485	33.8998	66.4695	66.8331	70.8832	70.7314
	DOVER↑	0.3159	0.3178	0.4565	0.1853	0.4553	0.3371	0.4781	0.5109
	E-Warp↓	4.8088	4.1941	6.9460	3.0933	3.2161	8.0673	4.8652	3.6861
UDM10	LPIPS↓	0.3673	0.3568	0.4596	0.4775	0.2462	0.3603	0.2821	0.3230
	MUSIQ↑	56.9349	52.8591	62.0505	30.5382	56.9412	64.6576	65.6088	67.6176
	DOVER↑	0.4603	0.4041	0.5177	0.2196	0.4953	0.5081	0.6010	0.6615
	E-Warp↓	4.1632	5.0314	6.5879	3.4635	3.4372	6.3255	4.4325	3.9865
YouHQ40	LPIPS↓	0.4137	0.4034	0.4432	0.5072	0.2976	0.3919	0.3076	0.3213
	MUSIQ↑	63.4022	52.2742	65.3396	30.5615	56.7780	66.5687	66.1301	66.8486
	DOVER↑	0.6269	0.5760	0.6038	0.4233	0.6719	0.5942	0.6920	0.7271
	E-Warp↓	5.1663	4.9954	8.5808	2.9235	4.1604	7.8108	5.5620	4.7001
VideoLQ	MUSIQ↑	52.2132	45.6323	48.7480	37.1804	38.6966	59.4151	51.8961	54.8766
	DOVER↑	0.4997	0.5285	0.4584	0.5078	0.4750	0.5279	0.5651	0.6415

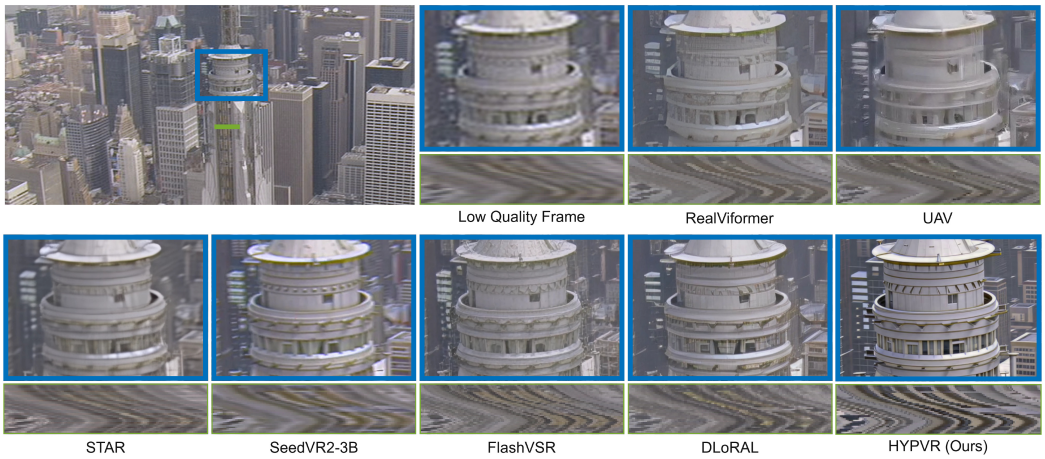


Figure 4: Qualitative and temporal profile comparison on VideoLQ. HYPVR restores rich, semantically correct details while maintaining smooth temporal evolution.

Inference and Baselines. Inference uses overlapping spatial tiles of 512×512 pixels with 128-pixel overlap and 61-frame temporal windows with 29-frame overlap, temporal reflection padding, and $t_{\text{aug}} = 200$. We compare with RealViFormer (Zhang & Yao, 2024), Upscale-A-Video (Zhou et al., 2024a), MGLD-VSR (Yang et al., 2024b), STAR (Xie et al., 2025), SeedVR2-3B (Wang et al., 2025a), DLoRAL (Sun et al., 2025), and FlashVSR (Zhuang et al., 2025), using each method's official inference settings except for wavelet-based color correction (Wang et al., 2024), which we apply to all methods with identical hyperparameters.

4.2 Main Comparison

For space, Tables 1 and 2 focus on LPIPS, MUSIQ, DOVER, and E-Warp as representative perceptual-fidelity, image-quality, video-quality, and temporal-consistency metrics; full results, including PSNR, SSIM, and CLIP-IQA, appear in Appendices A and B.

Qualitative Comparison. Figure 4 compares restorations on VideoLQ; more synthetic and real-world examples appear in Appendix G and Figs. 10–15. Most methods recover the building structure, but FlashVSR and DLoRAL show discontinuous high-frequency textures in the temporal profiles, while STAR and SeedVR2 produce blurrier crops with less detail. HYPVR combines coherent

Table 2: Ablations on POND25. (a) Image Adv. complements Video DM; (b) Video DM balances perceptual quality and temporal consistency; (c) the image discriminator outperforms its 3D counterpart; (d) one sampled frame suffices. Best and second-best results per panel appear in red and blue.

(a) Training objectives

Image Adv.	Video DM	LPIPS↓	MUSIQ↑	DOVER↑	E-Warp↓
		0.2782	61.1013	0.5303	3.3365
	✓	0.2821	61.3995	0.5349	3.1440
✓		0.2922	63.9428	0.5804	5.4008
✓	✓	0.2960	64.6228	0.5751	3.9835

(b) Temporal supervision

Temp. loss	LPIPS↓	MUSIQ↑	DOVER↑	E-Warp↓
None	0.2922	63.9428	0.5804	5.4008
Pixel diff.	0.3177	59.9419	0.5644	3.7174
Flow warp	0.4675	50.5535	0.4889	1.4037
RAFT flow	0.2920	63.1272	0.5665	4.1041
Video DM	0.2960	64.6228	0.5751	3.9835

(c) Adversarial supervision

Discriminator	LPIPS↓	MUSIQ↑	DOVER↑	E-Warp↓
None	0.2821	61.3995	0.5349	3.1440
3D video	0.3289	64.2093	0.5638	4.7760
Image Adv.	0.2960	64.6228	0.5751	3.9835

(d) Sampled-frame efficiency

Frames	Time/step (s)	MUSIQ↑	E-Warp↓
1	3.62	64.6228	3.9835
8	5.83	64.1589	3.9083
29	9.38	64.0297	3.8687

structure and rich detail with smooth temporal evolution, demonstrating the complementary visual and temporal effects of image adversarial supervision and Video DM.

Quantitative Comparison. HYPVR's LPIPS ranks among the top three on every synthetic benchmark. HYPVR ranks first or second in MUSIQ and CLIP-IQA (Table 3) and first in DOVER on all five benchmarks. Its DOVER reaches 0.6329 on POND25 and 0.6415 on VideoLQ, versus second-best scores of 0.5613 and 0.5651. STAR has the lowest E-Warp on three of four synthetic benchmarks, but its much weaker perceptual scores suggest over-smoothed outputs; richer detail generally makes pixel-level temporal alignment harder. Compared with the two strongest detail-generating baselines, DLoRAL and FlashVSR, HYPVR matches perceptual quality while lowering E-Warp on every synthetic benchmark—from 8.0673 and 4.8652 to 3.6861 on SPMC30, for example. This balance is consistent with Video DM constraining temporal variation at little cost to visual quality (Sec. 4.3).

4.3 Ablation Study

Unless stated otherwise, all ablations train directly on 29-frame clips for 20K iterations and use POND25 for evaluation. We disable latent noise augmentation to isolate the compared objectives and supervision strategies from this auxiliary engineering component.

Training Objectives. Table 2(a) isolates image adversarial supervision and Video DM; all variants retain the fidelity objective. Fidelity alone produces smooth outputs with grid artifacts; image adversarial supervision improves detail but harms inter-frame consistency, whereas Video DM alone improves consistency with limited perceptual gain. Together, they retain Image Adv.'s perceptual quality while reducing its E-Warp. The crops and temporal profiles in Fig. 6(a) illustrate this: Image Adv. adds rich texture, while Video DM stabilizes it across frames.

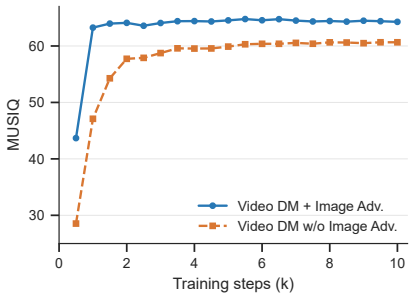
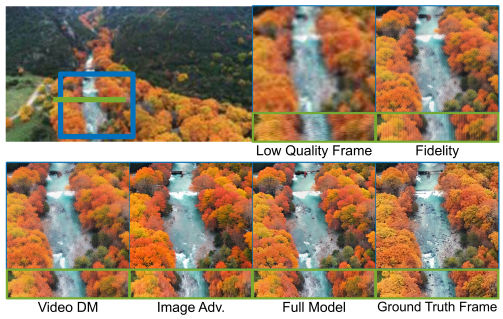
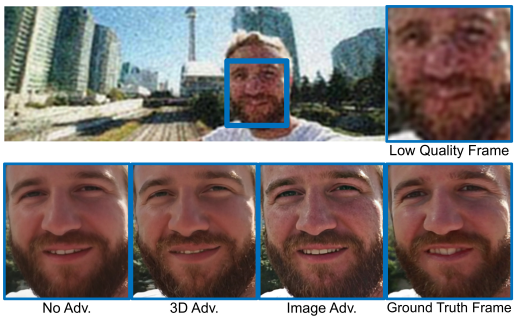


Figure 5: Image adversarial supervision accelerates the visual-quality convergence of Video DM training.

Adversarial Supervision. We compare the HYPIR-derived 2D image discriminator, a video-latent 3D discriminator, and no adversarial supervision. Without adversarial supervision, no-reference perceptual quality falls markedly (Table 2(c)). The 3D discriminator recovers some quality, but the image discriminator achieves the best MUSIQ and DOVER, with better LPIPS and E-Warp than the 3D variant. We hypothesize that the gap reflects slower, less stable 3D adversarial convergence, which may prevent a comparably strong solution within the 20K-step ablation schedule. Figure 6(b) shows the corresponding advantage in local structure and texture; Fig. 5 shows that image adversarial supervision also accelerates the visual-quality convergence of Video DM training, rather than merely shifting the final operating point.



(a) Complementarity of Objectives



(b) Adversarial Objective Comparison



(c) Temporal Objective Comparison

Figure 6: Qualitative ablations on POND25. (a) Image adversarial supervision adds detail; Video DM stabilizes it. (b) The image discriminator yields sharper structure than its 3D counterpart. (c) Video DM suppresses temporal fluctuation without smoothing away detail.

Efficient Sampled-Frame Image Supervision. Decoding and discriminating every frame is costly; Table 2(d) compares supervision of one, eight, or all 29 frames. One frame takes 3.62 seconds per step versus 9.38 seconds for 29, while supervising more frames yields only small, mixed metric changes. We therefore sample one frame per clip for image supervision.

Temporal Supervision. We compare Video DM with inter-frame pixel difference, flow-warping consistency, and RAFT-based flow regression (Teed & Deng, 2020). In the actual restoration setting, we estimate each objective's prediction-gradient ℓ_2 norm and reweight the losses to give comparable gradient contributions. Compared with the Image Adv.-only baseline in Table 2(b), Video DM lowers E-Warp while retaining comparable visual quality. Among the added temporal losses, it offers the strongest perceptual–temporal balance. Flow-warping consistency achieves the lowest E-Warp (1.4037, below the ground-truth reference of 2.6142), but over-smooths outputs and removes fine detail, showing that E-Warp alone does not establish better restoration. Figure 6(c) shows deterministic losses suppress texture variation for consistency, while Video DM suppresses flicker but retains sharp detail. Thus, Video DM provides temporal supervision without over-smoothing.

Video DM Design. Appendix C (Table 7) examines its timestep range, teacher, and guidance scale. Restricting $t_{\max}$ to 400 improves perceptual quality over 980 without materially changing fidelity or E-Warp, whereas 200 degrades both. VSR-specific teacher pretraining and scaling to 14B bring little consistent gain; stronger CFG improves perceptual quality.

5 Conclusion

HYPVR pairs a pretrained video generator with Video DM and sampled-frame image adversarial supervision for single-step restoration. A general-purpose video teacher's score stabilizes detail across frames without a video discriminator or VSR-specific diffusion pretraining. HYPVR leads DOVER on four synthetic benchmarks and real-world VideoLQ with competitive perceptual quality and temporal consistency. Ablations show Video DM improves stability without the over-smoothing of deterministic temporal losses, while one-frame supervision yields comparable quality at far lower per-step training time than full-frame supervision.

AI USE STATEMENT

We used generative AI tools as writing aids: for language polishing of author-written text and for assistance with LaTeX formatting. We also used them as coding assistants for parts of the experimental and evaluation scripts. We did not use generative AI to generate research ideas, to design the method, to produce or alter any experimental result, or to write the scientific claims of this paper. All AI-assisted text and code were reviewed by the authors, and all reported numbers were produced by the pipelines described above and verified against the raw logs. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model initialization, Video DM objective and gradient, sampled-frame supervision, noise range, and joint training objective. Section 4.1 reports the training schedule, optimization and hardware settings, degradation pipeline, inference tiling, baselines, and evaluation metrics. Appendix E details the training-data selection and POND25 construction; Appendices A and B provide complete quantitative results; Appendix C reports Video DM design choices; and Appendices D and F describe the runtime and user-study protocols. We will release the training and testing code.

REFERENCES

Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *NeurIPS*, 35: 31769–31781, 2022.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, pp. 13–23, 2025.

Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, pp. 4778–4787, 2017.

Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pp. 14245–14254, 2021a.

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 *CVPR*, pp. 4947–4956, 2021b.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, pp. 5972–5981, 2022a.

Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022b.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: A temporally-correlated noise prior for diffusion models. *arXiv preprint arXiv:2504.03072*, 2025.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, pp. 28188–28197, 2025.

Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In *CVPR*, pp. 9232–9241, 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 TOG*, 39(4):75–1, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

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 *CVPR*, pp. 23174–23184, 2025.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCVW*, pp. 3476–3485, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, pp. 3897–3906, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

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

Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, pp. 3224–3232, 2018.

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*, pp. 5148–5157, 2021.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. In *ICML*, pp. 31648–31666, 2025.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, pp. 170–185, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pp. 4681–4690, 2017.

Jingyun Liang, Jiezhang 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*, 2024.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, pp. 430–448, 2024.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics (TOG)*, 44(6):1–21, 2025.

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.

Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE TPAMI*, 44(11):7474–7489, 2021.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 45(4):4713–4726, 2022.

Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. In *NeurIPS*, volume 35, pp. 36081–36093, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

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.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pp. 402–419, 2020.

Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, pp. 3360–3369, 2020.

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.

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, 2023a.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 132(12):5929–5949, 2024.

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*, 2025a.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *CVPR*, pp. 2161–2172, 2025b.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018.

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, pp. 1954–1963, 2019.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pp. 1905–1914, 2021.

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

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.

Wikipedia contributors. Peak signal-to-noise ratio — Wikipedia, the free encyclopedia, 2024. URL https://en.wikipedia.org/w/index.php?title=Peak_signal-to-noise_ratio&oldid=1210897995. [Online; accessed 4-March-2024].

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*, pp. 20144–20154, 2023.

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.

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 *CVPR*, pp. 2139–2149, 2025.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, pp. 74–91, 2024a.

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *ECCV*, pp. 224–242, 2024b.

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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3106–3115. IEEE, 2019.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 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 *CVPR*, pp. 6613–6623, 2024b.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, pp. 25669–25680, 2024.

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*, pp. 586–595, 2018.

Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *ECCV*, pp. 412–428, 2024.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *CVPR*, pp. 2535–2545, 2024a.

Xingyu Zhou, Leheng Zhang, Xiaorui Zhao, Keze Wang, Leida Li, and Shuhang Gu. Video super-resolution transformer with masked inter&intra-frame attention. In *CVPR*, pp. 25399–25408, 2024b.

Jiapeng Zhu, Deli Zhao, Bo Zhang, and Bolei Zhou. Disentangled inference for gans with latently invertible autoencoder. *IJCV*, 130(5):1259–1276, 2022.

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.

A Complete Quantitative Comparison

Table 3 reports all seven metrics for each synthetic benchmark and the available no-reference metrics for VideoLQ, adding PSNR, SSIM, and CLIP-IQA to the compact main-paper results.

Table 3: Complete quantitative comparison on synthetic and real-world benchmarks. Best and second-best results are marked in red and blue, respectively. UAV denotes Upscale-A-Video; VideoLQ lacks ground truth for full-reference metrics and E-Warp.

Dataset	Metric	RealViformer	MGLD-VSR	UAV	STAR	SeedVR2-3B	DLoRAL	FlashVSR	HYPVR (Ours)
POND25	PSNR↑	23.0262	22.9366	21.1377	22.7054	23.6961	22.3196	22.2854	21.7376
	SSIM↑	0.5873	0.5911	0.4549	0.6162	0.6394	0.5742	0.5803	0.5745
	LPIPS↓	0.3864	0.3722	0.4608	0.5329	0.2661	0.3673	0.3049	0.3027
	MUSIQ↑	63.3772	57.1638	63.7041	30.6886	58.7561	67.2696	63.4249	68.2838
	CLIPIQA↑	0.3862	0.3175	0.4736	0.2157	0.3895	0.5640	0.4370	0.5423
	DOVER↑	0.5056	0.4330	0.4526	0.2682	0.5487	0.4579	0.5613	0.6329
	E-Warp↓	5.3201	5.4009	8.4695	2.7547	3.8804	7.3952	5.3837	4.5978
SPMC30	PSNR↑	22.0154	21.8249	20.1115	20.7290	21.9925	21.2016	20.9222	20.8250
	SSIM↑	0.5303	0.5306	0.3849	0.4832	0.5469	0.4890	0.4800	0.4832
	LPIPS↓	0.3785	0.3948	0.4761	0.5903	0.2774	0.3734	0.3352	0.3180
	MUSIQ↑	63.1467	57.4914	68.3485	33.8998	66.4695	66.8331	70.8832	70.7314
	CLIPIQA↑	0.3839	0.3263	0.6071	0.2484	0.5250	0.5885	0.5956	0.6167
	DOVER↑	0.3159	0.3178	0.4565	0.1853	0.4553	0.3371	0.4781	0.5109
	E-Warp↓	4.8088	4.1941	6.9460	3.0933	3.2161	8.0673	4.8652	3.6861
UDM10	PSNR↑	24.4148	24.1031	21.8547	24.0201	25.0531	23.3454	23.1419	22.4244
	SSIM↑	0.6795	0.6725	0.5111	0.6769	0.7124	0.6464	0.6384	0.6258
	LPIPS↓	0.3673	0.3568	0.4596	0.4775	0.2462	0.3603	0.2821	0.3230
	MUSIQ↑	56.9349	52.8591	62.0505	30.5382	56.9412	64.6576	65.6088	67.6176
	CLIPIQA↑	0.3960	0.3342	0.5159	0.1985	0.4409	0.6260	0.5087	0.6007
	DOVER↑	0.4603	0.4041	0.5177	0.2196	0.4953	0.5081	0.6010	0.6615
	E-Warp↓	4.1632	5.0314	6.5879	3.4635	3.4372	6.3255	4.4325	3.9865
YouHQ40	PSNR↑	22.9901	23.0433	21.1247	22.5528	23.3818	22.0229	22.2515	21.3217
	SSIM↑	0.5748	0.5872	0.4426	0.5831	0.6209	0.5440	0.5567	0.5314
	LPIPS↓	0.4137	0.4034	0.4432	0.5072	0.2976	0.3919	0.3076	0.3213
	MUSIQ↑	63.4022	52.2742	65.3396	30.5615	56.7780	66.5687	66.1301	66.8486
	CLIPIQA↑	0.4509	0.3490	0.5655	0.2602	0.4244	0.6420	0.5165	0.5848
	DOVER↑	0.6269	0.5760	0.6038	0.4233	0.6719	0.5942	0.6920	0.7271
	E-Warp↓	5.1663	4.9954	8.5808	2.9235	4.1604	7.8108	5.5620	4.7001
VideoLQ	MUSIQ↑	52.2132	45.6323	48.7480	37.1804	38.6966	59.4151	51.8961	54.8766
	CLIPIQA↑	0.3460	0.2807	0.3206	0.2635	0.2357	0.5794	0.3777	0.4310
	DOVER↑	0.4997	0.5285	0.4584	0.5078	0.4750	0.5279	0.5651	0.6415

B Complete Ablation Results

Tables 4–6 report the complete metrics for the objective, discriminator, sampled-frame, and temporal-supervision ablations on POND25.

Table 4: Complete ablations of the training objectives and adversarial supervision on POND25. Best and second-best results within each panel are marked in red and blue, respectively. All variants in (a) use the fidelity objective; those in (b) retain fidelity and Video DM.

(a) Complementarity of Training Objectives

Image Adv.	Video DM	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIP-IQA↑	DOVER↑	E-Warp↓
		24.0448	0.6446	0.2782	61.1013	0.3780	0.5303	3.3365
	✓	23.7516	0.6391	0.2821	61.3995	0.3855	0.5349	3.1440
✓		22.9266	0.6065	0.2922	63.9428	0.4661	0.5804	5.4008
✓	✓	22.9722	0.6088	0.2960	64.6228	0.4763	0.5751	3.9835

(b) Comparison of Adversarial Objectives

Adversarial supervision	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIP-IQA↑	DOVER↑	E-Warp↓
None	23.7516	0.6391	0.2821	61.3995	0.3855	0.5349	3.1440
3D video discriminator	22.0477	0.5700	0.3289	64.2093	0.4519	0.5638	4.7760
HYPIR image discriminator	22.9722	0.6088	0.2960	64.6228	0.4763	0.5751	3.9835

Table 5: Complete efficiency comparison of sampled-frame image supervision on POND25. Best and second-best metric values are marked in red and blue.

Supervised frames	Time/step (s)	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIP-IQA↑	DOVER↑	E-Warp↓
1	3.62	22.9722	0.6088	0.2960	64.6228	0.4763	0.5751	3.9835
8	5.83	22.9937	0.6093	0.2907	64.1589	0.4717	0.5808	3.9083
29	9.38	23.0374	0.6123	0.2923	64.0297	0.4707	0.5748	3.8687

Table 6: Complete comparison of temporal supervision strategies on POND25. All variants retain fidelity and image adversarial supervision; "None" removes only temporal supervision. Best and second-best results are marked in red and blue, respectively.

Temporal supervision	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIP-IQA↑	DOVER↑	E-Warp↓
None	22.9266	0.6065	0.2922	63.9428	0.4661	0.5804	5.4008
Inter-frame pixel difference ($\lambda = 30$)	23.4445	0.6233	0.3177	59.9419	0.4316	0.5644	3.7174
Flow warping consistency ($\lambda = 80$)	18.6089	0.5628	0.4675	50.5535	0.3586	0.4889	1.4037
RAFT flow regression ($\lambda = 10$)	22.9436	0.6096	0.2920	63.1272	0.4417	0.5665	4.1041
Video DM (ours)	22.9722	0.6088	0.2960	64.6228	0.4763	0.5751	3.9835

B.1 Positive Gradient Contribution Ratio

Let g average latent gradients over T frames and $\hat{g}_k$ over a k-frame sample S_k:

$$g = \frac{1}{T}\sum_{i=1}^{T} \nabla_{\hat{\mathbf{z}}} \ell_i, \qquad \hat{g}_k = \frac{1}{k}\sum_{i \in S_k} \nabla_{\hat{\mathbf{z}}} \ell_i. \tag{6}$$

PGCR measures the fraction of the absolute coordinate-wise inner-product contribution that agrees in sign with the full-frame gradient:

$$\mathrm{PGCR}(\hat{g}_k, g) = \frac{\sum_j \max(\hat{g}_{k,j} g_j, 0)}{\sum_j |\hat{g}_{k,j} g_j| + \epsilon}. \tag{7}$$

Here j indexes latent coordinates and ϵ is a small numerical constant. A value above 0.5 means that aligned contributions dominate, while a value of one indicates that all nonzero coordinate-wise contributions agree in sign.

C VIDEO DM DESIGN ANALYSIS

We further analyze the noise range, teacher configuration, and classifier-free guidance (CFG) scale used by Video DM. All variants follow the ablation protocol in Sec. 4.3 and change only the design choice indicated in Table 7.

Noise Range. Image adversarial supervision is responsible for generating high-frequency texture, while the fidelity objective anchors low-frequency content to the LQ input. Video DM is therefore primarily needed to regularize the spatiotemporal consistency of the generated detail rather than to reconstruct the complete low-frequency structure. This division of labor suggests that supervision over the entire diffusion trajectory may be unnecessary: restricting Video DM to lower-noise timesteps may better preserve the current restoration while focusing its update on local detail and its evolution across frames. The noise-range rows of Table 7 are consistent with this interpretation. Reducing $t_{\max}$ from 980 to 400 improves MUSIQ, CLIP-IQA, and DOVER while retaining comparable fidelity and E-Warp; reducing it further to 200 noticeably degrades fidelity and temporal stability. We consequently select $t_{\max} = 400$ as the best perceptual–temporal balance.

Video Teacher. The low-noise operating range may also explain the limited benefit of VSR-specific diffusion pretraining in the teacher search of Table 7. One possible reason is that such pretraining strengthens LQ-conditioned content recovery along the denoising trajectory, whereas our fidelity objective already supplies this constraint and Video DM does not use most of that trajectory. The general-purpose WAN2.1-T2V-1.3B teacher (Wan et al., 2025) is therefore sufficient. Scaling the teacher to WAN2.1-T2V-14B likewise produces no consistent improvement. We hypothesize that the larger model's additional capacity is most useful for capabilities such as text–image alignment, global spatial layout, camera motion, and motion understanding, which are not strongly exercised when Video DM operates mainly as a low-noise consistency constraint on generated high-frequency detail.

Classifier-Free Guidance. CFG remains effective despite this restricted role. Increasing its scale from 1 to 5 steadily improves LPIPS, MUSIQ, and CLIP-IQA, with scale 5 providing the strongest perceptual results and scale 3 the best DOVER, as shown by the CFG rows of Table 7. Video DM is still a joint spatiotemporal distributional constraint, so guidance strengthens the plausibility and coherence of generated detail across both space and time. At the same time, it does not impose deterministic frame-to-frame correspondence, allowing it to improve temporal consistency with less disruption to natural temporal evolution than explicit temporal objectives.

Table 7: Joint ablation of Video DM design choices on POND25. Best and second-best results within each search dimension are marked in red and blue, respectively. For the noise-range search, timesteps are sampled from $\mathcal{U}([20, t_{\max}])$. Our final configuration uses $t_{\max} = 400$, the WAN2.1-T2V-1.3B teacher, and a CFG scale of 5.

Search dimension	Setting	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIP-IQA↑	DOVER↑	E-Warp↓
Noise range	$t_{\max} = 980$	22.9996	0.6102	0.2937	63.6589	0.4594	0.5614	4.1296
	$t_{\max} = 800$	22.9971	0.6101	0.2943	63.8627	0.4648	0.5647	4.1119
	$t_{\max} = 600$	23.0240	0.6101	0.2961	63.7375	0.4626	0.5676	3.9011
	$t_{\max} = 400$	22.9722	0.6088	0.2960	64.6228	0.4763	0.5751	3.9835
	$t_{\max} = 200$	21.8857	0.5795	0.3202	64.1414	0.4859	0.5685	4.9478
Video teacher	WAN2.1-T2V-1.3B	22.9722	0.6088	0.2960	64.6228	0.4763	0.5751	3.9835
	VSR-pretrained teacher	22.9300	0.6077	0.2995	63.9224	0.4703	0.5714	3.9573
	WAN2.1-T2V-14B	22.8561	0.6048	0.2985	63.4966	0.4793	0.5730	4.2434
CFG scale	1	22.9899	0.6104	0.2980	63.6808	0.4643	0.5718	3.8802
	3	22.9873	0.6099	0.2962	64.1196	0.4668	0.5786	3.8940
	5	22.9722	0.6088	0.2960	64.6228	0.4763	0.5751	3.9835

D Inference Efficiency

We compare inference efficiency on a 61-frame input at 512×512 resolution. All runtimes are measured on a single NVIDIA H200 GPU under an identical evaluation environment. Table 8 reports the number of restoration steps, wall-clock runtime, computation, and parameter count.

Table 8: Inference efficiency for restoring 61 frames at 512×512 resolution, measured on a single NVIDIA H200 GPU. HYPVR combines single-step restoration with the temporal compression of a T2V foundation model, running comparably to FlashVSR and faster than other generative methods.

Method	Steps	Runtime (s)	Compute (TFLOPs)	Parameters
RealViformer	1	1.03	4.9	5.3M
Upscale-A-Video	60	85.25	9,482.7	1.4B
MGLD-VSR	50	34.61	2,851.2	1.4B
STAR	30	153.47	16,272.9	2.5B
SeedVR2-3B	1	12.12	606.0	3.6B
DLoRAL	1	23.55	334.4	1.3B
FlashVSR	1	1.68	190.1	1.8B
HYPVR (Ours)	1	1.65	309.1	1.4B

HYPVR first avoids the iterative denoising cost of conventional diffusion-based VSR by performing restoration in a single forward step. Its T2V foundation model also compresses the temporal dimension by a factor of four in the video latent space. This is substantially more efficient than T2I-based foundations such as the one used by DLoRAL, whose image representation must process the video frame by frame. Compared with SeedVR2-3B, HYPVR uses a smaller model and substantially less computation, leading to a correspondingly lower runtime.

HYPVR and the full version of FlashVSR exhibit similar wall-clock speed at this input size, despite their different nominal FLOP counts. A plausible explanation is hardware utilization: FlashVSR's sparse attention can incur masking, indexing, and irregular memory-access overhead while exposing less parallelism than dense kernels. At 512×512, reduced arithmetic may not fully offset sparse-execution overhead, while HYPVR's denser computation can better utilize the GPU. This could explain their similar runtimes despite FlashVSR's lower FLOP count.

E Dataset Details

E.1 Video Training Dataset

We collect approximately 2,000 hours of high-resolution YouTube videos spanning portraits, architecture, animals, plants, and natural scenes. The source videos range from 4K to 8K resolution and have frame rates above 24 fps; we retain only content with sufficiently high bitrates and no visible compression artifacts. Before scene segmentation, videos containing black borders or an effective frame height below 2,048 pixels are removed. We detect scene boundaries with PySceneDetect and discard scenes shorter than 64 frames at 24 fps, obtaining an initial pool of approximately 840K clips.

Each clip is then assessed along two axes: frame quality and scene dynamics. For frame quality, we sample frames at 3 fps and average their MUSIQ and LAION-Aesthetic scores. For scene dynamics, we use the magnitude of optical flow estimated by RAFT (Teed & Deng, 2020) as a measure of temporal motion. Joint ranking by these criteria produces a candidate set of 62K high-quality clips, after which we perform an additional cleaning pass to remove residual quality issues and obtain the final training collection. All retained clips are downsampled by a factor of two to improve frame fidelity and are decoded at 24 fps for training.

E.2 POND25 Validation Dataset

POND25 contains 25 professionally captured clips selected from Pond5, covering humans, animals, urban environments, and natural landscapes. Each source clip is approximately four seconds long,

recorded at 24 fps and 2560×1440, and contains no scene transition. We evaluate at 1280×720 so that the synthetic degradation removes more fine-scale information and better exposes a model's ability to recover and temporally stabilize detail. As illustrated in Figs. 7 and 8, the benchmark includes aerial views, multi-subject scenes, rich depth structure, and rapid camera motion. Unlike benchmarks dominated by handheld footage, POND25 consists of professionally filmed sequences with smooth camera motion and without pronounced shake or jitter.

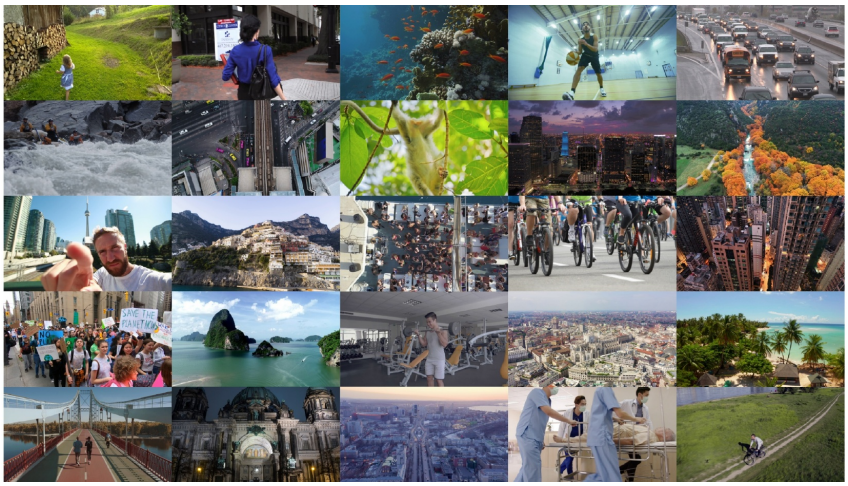


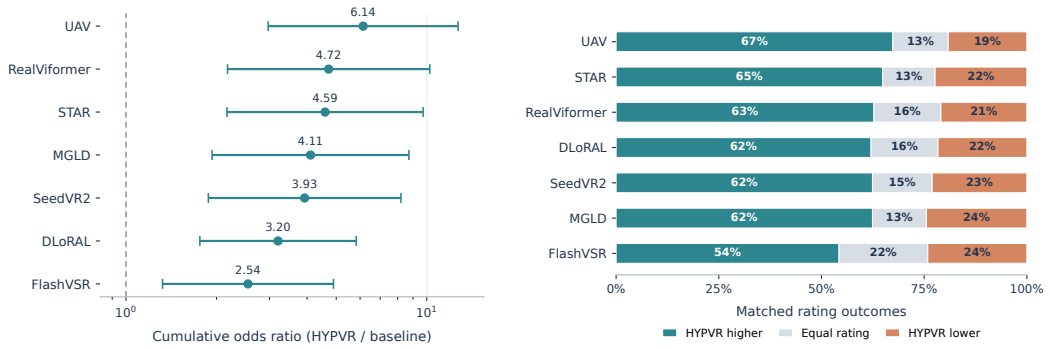
Figure 7: Representative content from POND25. The benchmark covers diverse subjects and scene layouts, including people, animals, urban structures, and natural environments.



Figure 8: Frames sampled at successive time points from three POND25 clips, illustrating subject movement and camera motion in the benchmark.

F User Study

We conduct a video-only user study with 97 participants recruited through Prolific and ten scenes selected from VideoLQ. The study was administered through two aligned questionnaires, with every participant ultimately evaluating all ten videos. For each scene, participants assess restored 2K video clips of 4–6 seconds produced by the compared methods. The method identities are hidden, and participants rate the overall visual quality of each video on a scale from 1 to 10. Participants have no formal training in computer vision. This yields 970 ratings per method (97 participants $\times$ 10 videos).



(a) Model-estimated ordinal rating advantage.

(b) Matched rating outcomes.

Figure 9: User-study comparison of HYPVR with competing methods. (a) Points denote cumulative odds ratios from separate cumulative-logit models fitted to the original 1–10 ratings, and horizontal lines denote pointwise 95% confidence intervals. The models include participant and video fixed effects, with standard errors clustered by participant; an odds ratio greater than one indicates a shift toward higher ratings for HYPVR. (b) Each bar reports the percentage of 970 participant–video pairs in which HYPVR receives a higher, equal, or lower rating than the corresponding baseline. Methods are ordered by decreasing odds ratio in (a) and decreasing net advantage in (b). The confidence intervals are pointwise and are not adjusted for the seven comparisons.

We first compare HYPVR with each baseline using a separate cumulative-logit ordinal regression model fitted to the original 1–10 ratings. Each model includes participant and video fixed effects to account for systematic differences in rating tendency and video difficulty, respectively, while standard errors are clustered by participant to account for repeated ratings. We report $\exp(\beta)$ as the cumulative odds ratio, with pointwise 95% confidence intervals computed as $\exp(\beta \pm 1.96\,\mathrm{SE})$. For any rating threshold k, an odds ratio above one indicates that HYPVR has higher odds of receiving a score above k rather than at or below k, under the proportional-odds assumption.

We additionally perform a matched comparison within every participant–video pair. For each baseline, the HYPVR score is classified as higher than, equal to, or lower than the corresponding baseline score, producing 970 matched pairs per comparison. Methods are ordered by net advantage, defined as the proportion of higher ratings minus the proportion of lower ratings.

As shown in Fig. 9a, HYPVR exhibits a consistent shift toward higher ordinal ratings than every competing method. The cumulative odds ratios range from 2.54 against FlashVSR to 6.14 against UAV, and all pointwise 95% confidence intervals remain above one. These ratios quantify the common change in odds across rating thresholds; they do not represent score differences or direct win-probability ratios. The matched analysis in Fig. 9b provides a complementary sample-level view: HYPVR receives a higher rating in 54%–67% of participant–video pairs, an equal rating in 13%–22%, and a lower rating in 19%–24%. The largest matched advantage is observed against UAV, while FlashVSR provides the closest comparison; even against FlashVSR, higher ratings for HYPVR remain substantially more frequent than lower ratings.

G More Visual Comparisons

Figures 10 and 11 provide additional comparisons on synthetic benchmarks, while Figs. 12–15 report real-world results on VideoLQ.



Figure 10: Additional qualitative comparisons on synthetic benchmarks.

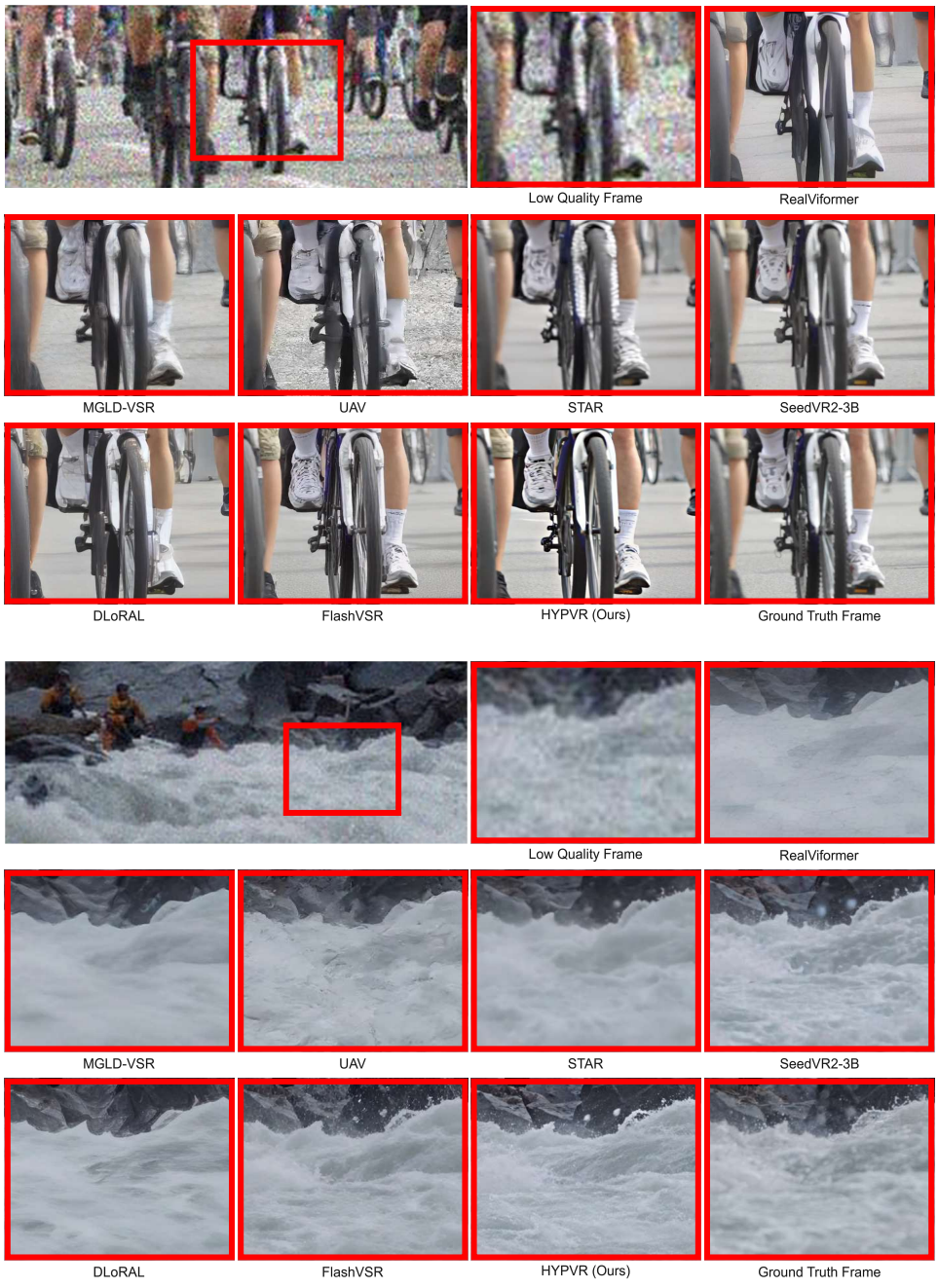


Figure 11: Additional qualitative comparisons on synthetic benchmarks.

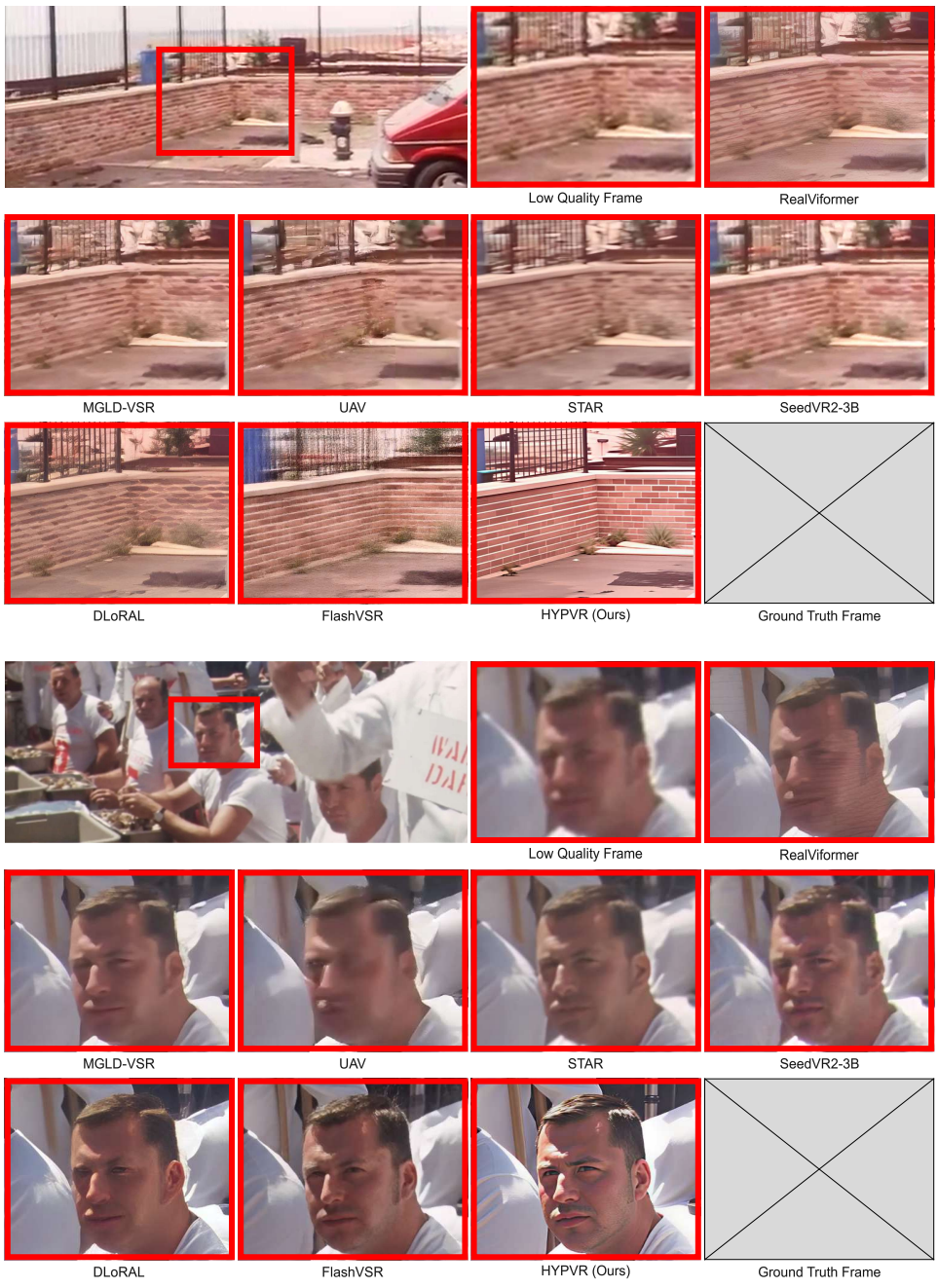


Figure 12: Additional qualitative comparisons on the real-world VideoLQ benchmark.

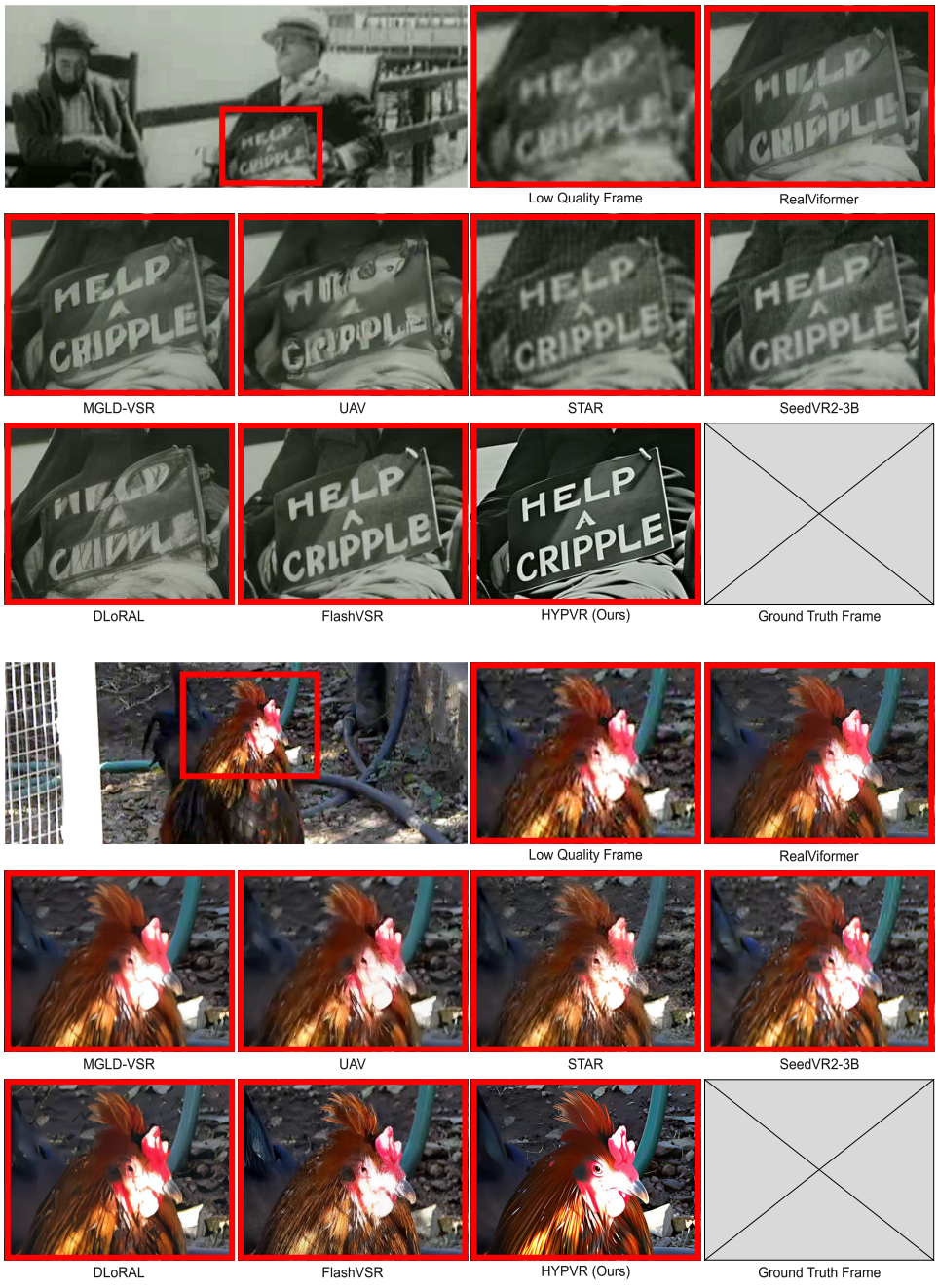


Figure 13: Additional qualitative comparisons on the real-world VideoLQ benchmark.



Figure 14: Additional qualitative comparisons on the real-world VideoLQ benchmark.

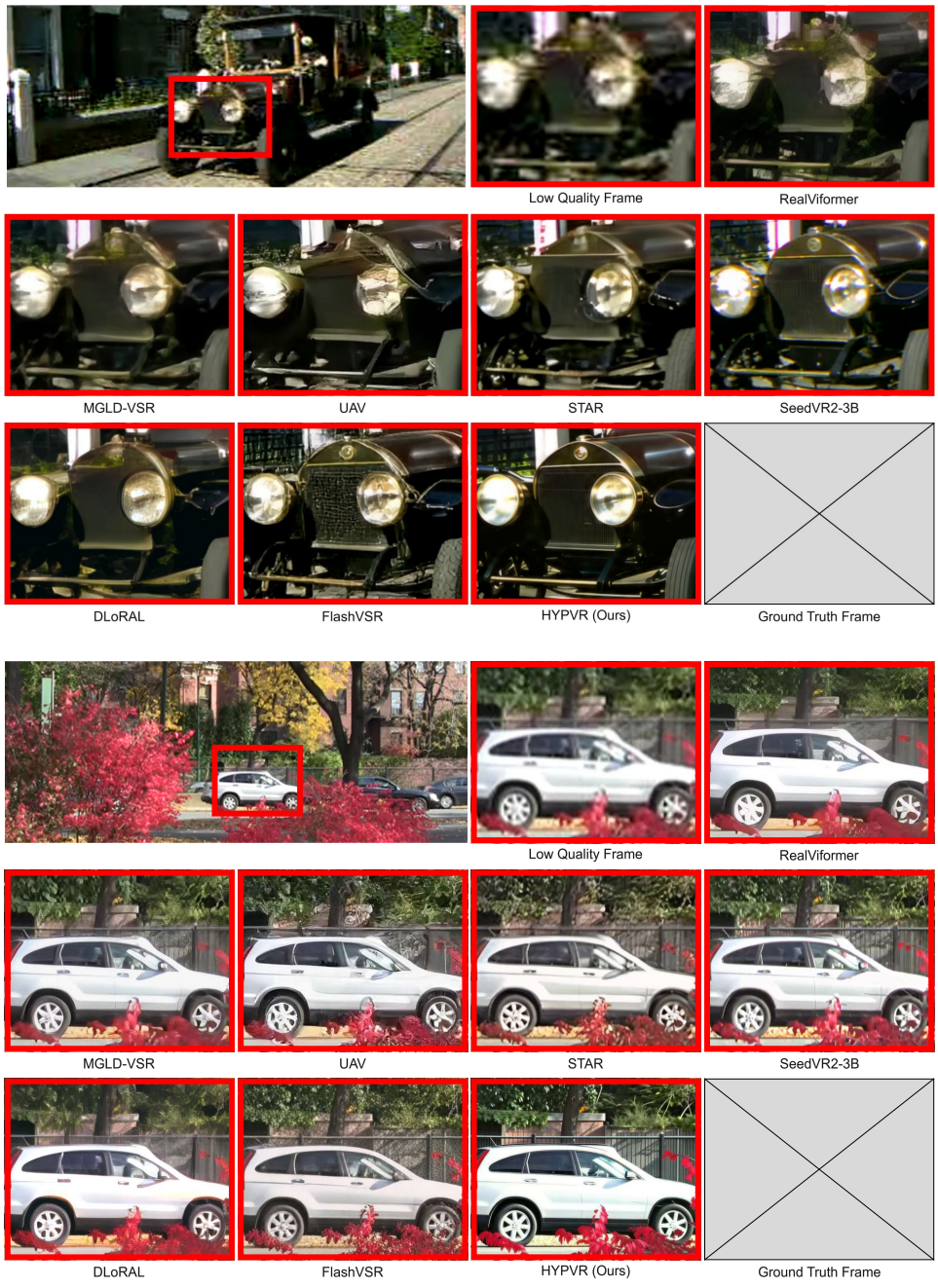


Figure 15: Additional qualitative comparisons on the real-world VideoLQ benchmark.