VIDEO GENERATION MODELS ARE ZERO-SHOT LOW-LEVEL VISION SOLVERS

Anonymous authors
Paper under double-blind review

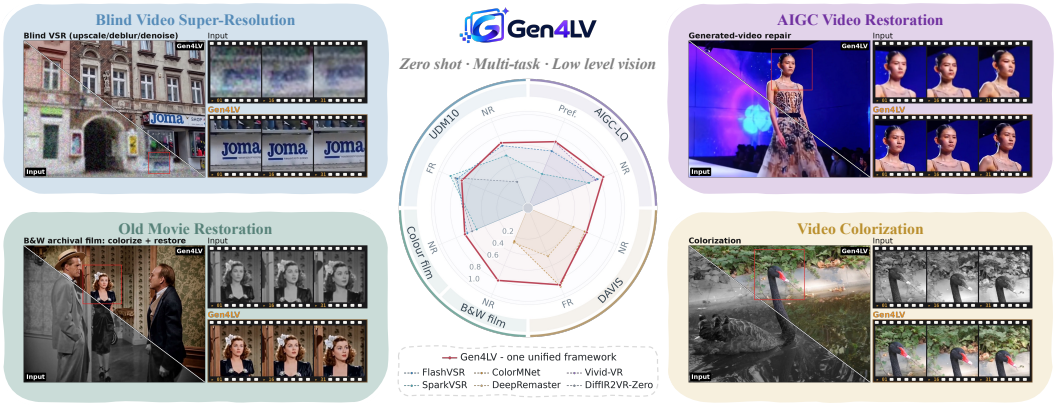

Figure 1: **One frozen video generator for diverse low-level vision tasks.** With the same frozen generator and no task-specific training, Gen4LV handles blind video super-resolution, the repair of generated video, archival film restoration and video colorization, among others. Each panel splits input and output along the diagonal and shows the first, middle and last frames of each. Centre: per-axis normalized scores on each benchmark, grouped into no-reference video quality (NR), full-reference fidelity (FR) and VLM preference (Pref.), for Gen4LV and the baselines of each task.

ABSTRACT

Video generation models learn rich priors over appearance and motion, enabling instruction-guided synthesis and editing. Here we investigate a central question: can these priors also support low-level vision without task-specific fine-tuning? We show that a single frozen video generator can serve as a zero-shot solver for a spectrum of low-level video tasks, including blind video super-resolution, colorization, archival film restoration, and restoration of AI-generated video. Our framework, Gen4LV, casts these tasks as reference-conditioned generation with complementary visual cues. That is, the degraded video guides motion, layout, and content, while a single frame restored by an image editing expert guides appearance. However, this conditioning scheme enables the generator to synthesize detailed video but does not, by itself, ensure fidelity to the input. For tasks requiring high fidelity, we further introduce frequency-gated data consistency to constrain the output using reliable observation information. In blind super-resolution, for instance, the gate restricts observation-based corrections to low frequencies, improving structural alignment while preserving generated texture and enabling inference-time control over the perception–fidelity trade-off. Across eight benchmarks spanning four tasks, Gen4LV delivers strong no-reference video quality, matching or surpassing specialized models in multiple restoration settings without task-specific fine-tuning. These results demonstrate that pretrained video generators can serve as generalists for low-level vision, pointing toward a future in which diverse visual tasks are addressed by specifying instructions, references, and constraints at inference time.

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

Video generation models learn to synthesize detailed appearance and coherent motion from large collections of visual experience (Yang et al., 2025; Wan et al., 2025; Ma et al., 2026; Zheng et al., 2024). These capabilities make pretrained video models a compelling source of priors for tasks beyond synthesis. Recent work has begun to reveal their utility for visual understanding, geometry, and reasoning (Wiedemer et al., 2025; King et al., 2026; Wan et al., 2025). However, low-level vision problems such as video restoration pose a different challenge: the output must recover useful detail while remaining tied to a particular observation. The same generative power that produces high-quality videos can also alter the content significantly that restoration is meant to preserve, thus producing hallucinated content (Bai et al., 2025b). This raises a simple question:

Does video generation pretraining also equip the model to solve diverse low-level vision tasks?

The answer is no when applied directly. Given a degraded video and an instruction to render it sharply, state-of-the-art generators fall far below task-specific restorers on every no-reference criterion (Table 1) and fail in two ways (Fig. 3): the output either retains the degradation, which the generator reproduces as content of its reference, or departs from the observation, repainting what the input does not show. Low-level vision is a stricter test than perception, since the output must agree with the observation pixel by pixel and frame by frame, and a generative prior's readiness to complete what it cannot see is here the primary failure mode. The failure lies in the interface rather than the prior, however: the same generator restores the video once its conditions are chosen differently (Sec. 4.5).

Existing video restoration takes one of two routes. First, fine-tuned models collapse prior and likelihood, $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x})$, into one network trained on synthetic degradation pairs (Wang et al., 2021; Chan et al., 2022), now initialized from video diffusion priors (Zhou et al., 2024; Xie et al., 2025; Chen et al., 2026; Wang et al., 2026a; Zhuang et al., 2026; Yu et al., 2026); this reentangles the prior with one degradation distribution, so their benchmarks are in distribution by construction. Second, training-free (e.g., plug-and-play)methods (Venkatakrishnan et al., 2013; Zhang et al., 2021; Chung et al., 2023; Wang et al., 2023b; Zhu et al., 2023) keep the prior frozen behind an explicit operator, but assume image priors and known operators: on video they flicker (Yeh et al., 2024; Kwon & Ye, 2025), and when the operator is blind they have no rule for what to trust. We take a third route: the generator stays frozen, the degradation is represented by an operator outside the network, and only the interface through which the observation enters is designed.

Based on this finding, we introduce Gen4LV, a novel training-free method that solves diverse low-level tasks using a single pretrained prior. Gen4LV elicits the latent ability with two components, each addressing one failure case (Fig. 2). Against the retained degradation, a *motion–appearance decoupled conditioning* strategy uses the degraded video as the reference for motion, layout and content only, and a single frame restored by an image expert to guide appearance. Against the departure from the observation, a *frequency-gated plug-and-play data consistency* step holds the output to the input only in the geometric band, where an approximate operator remains reliable; the gate extends plug-and-play consistency to blind degradations and avoids the failure of the textbook step, which overwrites synthesized mid-frequency texture with the upsampled observation (Sec. 3.3). Across eight benchmarks spanning four tasks, Gen4LV matches or exceeds specialized, in-distribution-trained models on no-reference video quality without seeing a restoration pair: the answer to our question is negative for the generator used as it is, and positive once its latent ability is elicited.

Our contributions are fourfold.

- **Diagnosing the limitations of generative video restoration.** We show that directly conditioning video generators on degraded videos leads to either degradation retention or deviations from the observed content. Our analysis identifies the conditioning interface, rather than the generative prior, as the source of these failures (Secs. 3.1 and 4.2).
- **Motion–appearance decoupled conditioning.** We develop a conditioning strategy that separates structural and appearance guidance through complementary visual references: the degraded video provides motion, layout, and content cues, while a frame restored by an image expert supplies appearance cues. This design reduces degradation transfer while maintaining spatial and temporal correspondence with the input (Sec. 3.2).

- **Frequency-gated data consistency.** We introduce a plug-and-play data consistency step that restricts corrections to the low-frequency structural band, improving alignment with the observation under blind degradations while preserving generated texture (Sec. 3.3).
- **A unified framework with one frozen generator.** We integrate these components into Gen4LV, a zero-shot restoration framework that uses a single frozen video generator for blind video super-resolution, colorization, archival film restoration, and AI-generated video restoration. Across these tasks, we demonstrate performance that matches or exceeds that of task-specific methods (Fig. 1; Sec. 4).

2 Related Work

Video restoration from generative priors. To surpass conventional regression, recent methods leverage diffusion models as restoration priors. Early efforts adapted image diffusion to the temporal domain (Rota et al., 2024; Yang et al., 2024a; Zhou et al., 2024), whereas contemporary approaches directly repurpose native video generative models (Xie et al., 2025; Wang et al., 2025; Chen et al., 2026; Bai et al., 2026), alongside optimizations for accelerated or streaming inference (Chen et al., 2026; Wang et al., 2026a; Zhuang et al., 2026; Yu et al., 2026). However, by relying predominantly on synthetic pairs derived from hand-crafted degradation pipelines (Wang et al., 2021; Chan et al., 2022), this supervised paradigm tightly couples the generative prior to preconceived degradation distributions. Consequently, performance gains on standard benchmarks, which frequently share these training degradations (Zhou et al., 2024), conflate the prior's intrinsic capacity with dataset adaptation, obscuring its independent contribution.

Zero-shot restoration with frozen diffusion priors. Zero-shot approaches bypass fine-tuning by coupling a frozen generative prior with explicit data consistency. Traditional methods integrate classical plug-and-play techniques (Venkatakrishnan et al., 2013; Zhang et al., 2021) into the reverse sampling loops of image diffusion models (Chung et al., 2023; Wang et al., 2023b; Zhu et al., 2023). To extend this paradigm to video, recent adaptations combine image priors with temporal propagation (Yeh et al., 2024; Kwon & Ye, 2025) or exploit latent video consistency models (Spagnoletti et al., 2026). While these efforts avoid degradation-specific training, they rigidly enforce global reconstruction constraints assuming a perfectly known degradation operator. Consequently, they fail in practical blind settings, as they cannot balance an imperfect observation against a strong generative prior, unable to distinguish which structural components to trust versus which textures to synthesize. To address this, Sec. 3.3 builds on DiffPIR with a frequency-gated consistency mechanism to explicitly control this balance.

Instruction-following video generation. Modern video generators provide unified interfaces driven by heterogeneous visual and textual prompts (Yang et al., 2025; Wan et al., 2025; Ma et al., 2026; Agarwal et al., 2026), increasingly serving as frozen priors for visual perception and geometric reasoning (Wiedemer et al., 2025; Wang et al., 2026b; King et al., 2026; Jiang et al., 2025; Mai et al., 2025). While this versatility naturally invites their application to low-level video restoration, direct instruction following fails. Unlike high-level semantic tasks, restoration demands strict pixel-level fidelity to an imperfect observation; lacking this explicit structural grounding, current models inevitably oscillate between hallucinating unobserved textures and retaining original degradations. In contrast, our framework elicits this latent restoration capability without fine-tuning, employing the decoupled conditioning and data consistency mechanisms introduced in Sec. 3.

3 Method

3.1 Preliminaries

Direct application of video generators. We examine whether pretrained video generators can perform restoration directly by evaluating Wan2.2, CogVideoX1.5, LingBot-Video, and MiniMax-H3 through their native video-conditioning interfaces. Given degraded videos and restoration prompts, these models exhibit two recurring failure modes: degradation retention and deviation from the observed content (Sec. 4.2).

Backbone selection and architecture. We adopt MiniMax-H3, a state-of-the-art open-weight video generation model, for its strong generative prior and flexible multi-modal conditioning in-

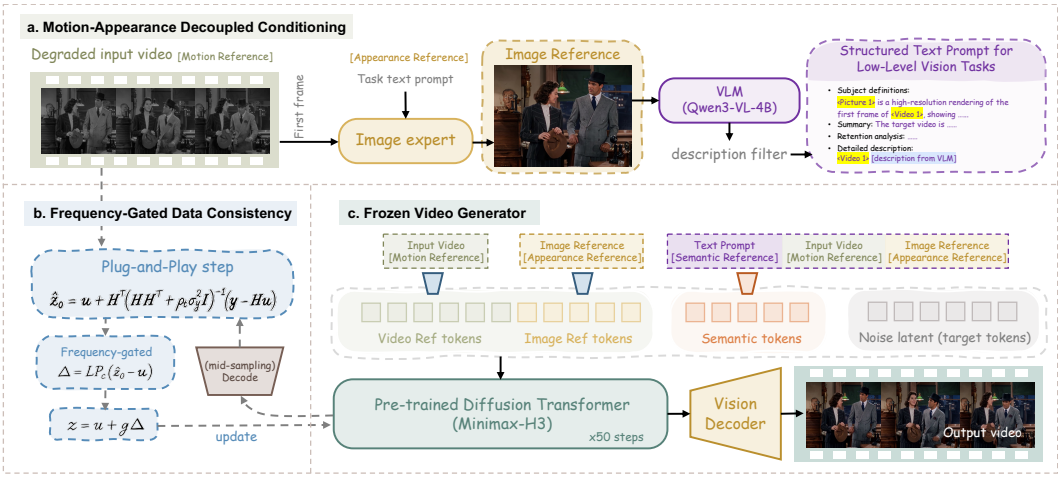

Figure 2: **Overview of Gen4LV.** (a) Motion–appearance decoupled conditioning: the degraded video guides motion and content, an expert-restored first frame guides appearance, and a filtered VLM description supplies the structured prompt. (b) Frequency-gated data consistency: at selected sampling steps, the current estimate is decoded and corrected, with only the low-pass component of the consistency update returned to the sampler. (c) Frozen video generator: visual references provide VAE tokens and, together with the prompt, semantic conditioning. The generator denoises the target latent over 50 steps and decodes the restored video.

terface. Its joint support for image, video, and text inputs enables complementary guidance for motion and appearance. The model is a flow-matching diffusion transformer in which visual references provide semantic conditioning through a vision-language encoder and clean reference tokens through a video autoencoder. We freeze all generator parameters and construct its conditions using an off-the-shelf image expert, queried once per clip, and a VLM captioner.

Setting and notation. Let $\boldsymbol{x}$ denote the target video, $\boldsymbol{y}$ the observation, and $\mathcal{A}$ the degradation operator, which is unknown in the blind setting. We specify a forward model $\boldsymbol{H}$ and estimate its mismatch level σ_y on a separate probe set (Appendix B). Under flow matching, $\boldsymbol{x}_t = (1-\sigma_t)\boldsymbol{x}_0 + \sigma_t\boldsymbol{\epsilon}$ with velocity $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{x}_0$, yielding the clean estimate $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_t - \sigma_t\boldsymbol{v}$. The output retains the input length, frame rate, and timestamps. We use *zero-shot* to indicate that inference operates on $\boldsymbol{y}$ and conditions derived from it without access to the true degradation operator or its noise level; *training-free* refers specifically to the generator.

3.2 Motion–Appearance Decoupled Conditioning

As illustrated in Fig. 2(a), we decouple structural and appearance guidance through two complementary visual references. The degraded video provides motion, layout, and content cues, while a restored frame supplies appearance cues that are missing from the observation. Together, they guide the generator to restore visual quality while preserving correspondence with the input.

Appearance reference. We restore the first frame of $\boldsymbol{y}$ using an off-the-shelf image expert, Nano Banana Pro, and use the result as the appearance reference. This frame provides cues for texture and fine detail, which the video prior uses to synthesize temporally coherent appearance across frames. We set its short-edge length to twice that of the video reference to balance the token allocation between the two references. We empirically find this resolution choice important for effective appearance guidance.

Motion reference. We use the degraded video to guide motion, spatial layout, and scene content. We upsample it to the output canvas and set its reference resolution to the target short-edge length, avoiding additional resizing within the conditioning pipeline. This preserves the spatial alignment of the video reference with the target output and provides temporal guidance that the appearance reference alone cannot supply.

Semantic conditioning. Following the prompting setup of MiniMax-H3, we use Qwen3-VL-4B-Instruct to generate a textual description of the restored appearance reference. We then remove descriptions of camera and subject motion, leaving the video reference to provide temporal guidance

and avoiding potentially conflicting textual instructions. The filtered description is inserted into a fixed prompt template that identifies the appearance reference as the restored first frame of the video and explicitly instructs the generator to preserve the input framing, motion, and timing.

3.3 FREQUENCY-GATED DATA CONSISTENCY

Although decoupled conditioning improves appearance and content preservation, it does not explicitly enforce agreement with the observation. We therefore introduce a frequency-gated data consistency step that corrects structural deviations while limiting changes to generated texture(Fig. 2(b)).

Plug-and-play consistency. Plug-and-play methods alternate between prior updates and data consistency steps to enforce agreement with the observation (Venkatakrishnan et al., 2013; Zhang et al., 2021). At sampling step t, we decode the clean estimate as $\boldsymbol{u} = \mathrm{Dec}(\hat{\boldsymbol{x}}_0)$ and solve $\min_{\boldsymbol{z}} \|\boldsymbol{H}\boldsymbol{z} - \boldsymbol{y}\|^2 + \rho_t \sigma_y^2 \|\boldsymbol{z} - \boldsymbol{u}\|^2$, balancing observation consistency against proximity to the generated estimate. The closed-form solution is

$$\hat{\boldsymbol{z}}_0 = \boldsymbol{u} + \boldsymbol{H}^\top \big(\boldsymbol{H}\boldsymbol{H}^\top + \rho_t\, \sigma_y^2\, \boldsymbol{I}\big)^{-1} (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{u})\,, \qquad \rho_t = \lambda \frac{(1-\sigma_t)^2}{\sigma_t^2}, \tag{1}$$

where the schedule places greater emphasis on observation consistency at high noise levels and on the generated estimate at low noise levels (Zhang et al., 2021). Although the correction leaves the null space of $\boldsymbol{H}$ unchanged, its components in $\mathrm{range}(\boldsymbol{H}^\top)$ can still alter texture-bearing frequencies. Consequently, unrestricted correction can improve geometric alignment while suppressing synthesized detail (Table 5; Sec. 4.5).

Frequency-gated correction. We control the frequency content of the consistency update by filtering its residual:

$$\boldsymbol{\Delta} = \mathrm{LP}_c(\hat{\boldsymbol{z}}_0 - \boldsymbol{u})\,, \qquad \boldsymbol{z} = \boldsymbol{u} + g\,\boldsymbol{\Delta}, \tag{2}$$

where LP_c is a Gaussian low-pass filter at the target resolution, c controls the cutoff, and g determines the correction strength. The filter emphasizes coarse structural corrections while attenuating changes to texture-bearing frequencies, without directly smoothing $\boldsymbol{u}$ or $\boldsymbol{y}$. The cutoff and gain therefore separately control the frequency range and strength of consistency, allowing fidelity to be adjusted at inference time (Fig. 6). We incorporate the correction into sampling through the latent residual $\mathrm{Enc}(\boldsymbol{z}) - \mathrm{Enc}(\boldsymbol{u})$, as summarized in Algorithm 1.

Consistency under blind degradations. Frequency gating also controls how strongly we rely on different frequency bands of the observation. With a known forward operator, the cutoff provides a means of adjusting the perception–distortion trade-off (Blau & Michaeli, 2018). Under blind degradations, we restrict corrections to a narrower band, where coarse structure remains informative despite mismatch between $\boldsymbol{H}$ and the true degradation. Our ablations support this choice (Sec. 4.5), and the probe analysis examines sensitivity to the forward model (Appendix B). We specify $\boldsymbol{H}$, σ_y, and the correction bandwidth for each task, omitting the consistency step when no suitable forward model is available (Sec. 4.1).

Algorithm 1 Frequency-gated data consistency.

for $t = 0$ **to** $T-1$ **do**
  $\hat{\boldsymbol{x}}_0 \leftarrow \boldsymbol{x}_t - \sigma_t\, \mathrm{DiT}(\boldsymbol{x}_t)$
  if $t \in \mathcal{S}$ **then**
    $\boldsymbol{u} \leftarrow \mathrm{Dec}(\hat{\boldsymbol{x}}_0)$
    $\rho_t \leftarrow \lambda(1-\sigma_t)^2/\sigma_t^2$
    $\boldsymbol{\delta} \leftarrow \boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{H}^\top + \rho_t\sigma_y^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{u})$
    $\boldsymbol{\Delta} \leftarrow \mathrm{LP}_c(\boldsymbol{\delta})$
    $\boldsymbol{z} \leftarrow \boldsymbol{u} + g\,\boldsymbol{\Delta}$
    $\hat{\boldsymbol{x}}_0 \leftarrow \hat{\boldsymbol{x}}_0 + \mathrm{Enc}(\boldsymbol{z}) - \mathrm{Enc}(\boldsymbol{u})$
  end if
  $\boldsymbol{x}_{t+1} \leftarrow \mathrm{Euler}(\hat{\boldsymbol{x}}_0, \sigma_t, \sigma_{t+1})$
end for
return $\boldsymbol{x}_T$

4 EXPERIMENTS

4.1 OVERVIEW

Tasks and datasets. We evaluate four low-level tasks. *Blind video super-resolution* uses UDM10 (Yi et al., 2019), SPMCS (Tao et al., 2017), and YouHQ40 (Zhou et al., 2024). While these benchmarks share the synthetic degradations (Wang et al., 2021; Chan et al., 2022) underlying the baselines' training data, rendering them in-distribution, they remain strictly zero-shot for us. *Colorization* relies on DAVIS (Perazzi et al., 2016) and REDS4 (Nah et al., 2019). Finally, *archival*

Table 1: **Blind** $\times 4$ **video super-resolution.** *Task-specific* models train in-distribution; *zero-shot* methods do not. Grey rows denote unranked references of off-the-shelf generators using native defaults (UDM10 here; SPMCS and YouHQ40 in Table 9). **Bold**/underline: best/second.

		no-reference, video			no-reference, per-frame			full-reference		
Dataset / type	Method	DOVER↑	FastVQA↑	Q-Align↑	MUSIQ↑	NIQE↓	CLIPIQA+↑	PSNR↑	SSIM↑	LPIPS↓
UDM10	Wan2.2-5B	0.2093	0.2439	2.20	28.01	8.53	0.2555	18.24	0.5939	0.6093
video generation	LingBot-Video (refiner)	0.1258	0.1357	1.86	20.74	8.95	0.2357	21.40	0.6712	0.5604
	LingBot-Video	0.1219	0.1847	1.82	22.58	9.26	0.2367	20.62	0.6509	0.5860
	CogVideoX1.5-5B	0.1358	0.1414	1.93	22.43	9.20	0.2316	17.91	0.5810	0.6467
	MiniMax-H3	0.1941	0.3192	2.18	28.36	7.04	0.2843	17.70	0.5899	0.5465
task-specific	DOVE	0.5863	0.8142	3.96	61.17	5.22	0.5093	**26.44**	0.7720	0.2674
	SeedVR2-3B	0.4101	0.5945	3.28	46.72	5.88	0.3895	25.98	0.7624	0.2649
	SeedVR2-7B	0.4383	0.6537	3.51	50.23	5.48	0.4221	26.01	0.7683	**0.2418**
	FlashVSR	0.5831	0.8213	4.17	**65.67**	3.75	0.5643	23.59	0.6975	0.2906
	FlashVSR-Tiny	0.5690	0.8141	4.12	62.96	3.98	0.5396	23.86	0.7079	0.2798
	SparkVSR	0.5567	0.7793	3.83	56.97	5.96	0.4692	26.40	**0.7727**	0.2802
zero-shot	DiffIR2VR-Zero	0.3148	0.6115	2.91	47.21	5.55	0.5328	24.36	0.6706	0.4688
	Gen4LV (ours)	**0.6612**	**0.8472**	**4.27**	65.49	**3.30**	**0.5826**	21.68	0.6002	0.3636
SPMCS	DOVE	0.4922	0.7268	4.20	69.18	4.77	0.6289	21.15	0.5523	0.2962
task-specific	SeedVR2-3B	0.4484	0.6890	4.15	64.90	3.90	0.5598	21.01	0.5576	0.2865
	SeedVR2-7B	0.4368	0.6858	4.07	64.11	4.13	0.5573	**21.32**	**0.5691**	**0.2780**
	FlashVSR	0.4684	0.7434	3.97	69.59	**3.15**	0.6208	20.38	0.5035	0.3073
	FlashVSR-Tiny	0.4872	0.7376	3.95	68.27	3.40	0.6016	20.48	0.5093	0.2977
	SparkVSR	0.4648	0.7009	3.98	66.19	5.41	0.5555	21.15	0.5565	0.3200
zero-shot	DiffIR2VR-Zero	0.3597	0.6657	3.50	63.94	4.73	0.6639	20.86	0.4925	0.4253
	Gen4LV (ours)	**0.5246**	**0.7457**	**4.30**	**71.22**	3.40	**0.7043**	18.91	0.4345	0.3313
YouHQ40	DOVE	0.6687	0.8421	4.19	63.40	4.93	0.5608	**23.94**	**0.6483**	0.3042
task-specific	SeedVR2-3B	0.6779	0.8459	4.36	63.35	4.04	0.5507	22.97	0.6310	0.2746
	SeedVR2-7B	0.6858	0.8466	4.48	65.39	3.84	0.5526	23.13	0.6381	**0.2570**
	FlashVSR	0.6892	**0.8616**	**4.51**	**70.45**	**3.05**	0.6329	21.73	0.5565	0.3004
	FlashVSR-Tiny	0.6824	0.8549	4.47	67.87	3.31	0.6079	21.99	0.5660	0.2953
	SparkVSR	0.6429	0.8093	4.03	58.81	5.75	0.4967	23.82	0.6441	0.3388
zero-shot	DiffIR2VR-Zero	0.5500	0.7777	3.78	62.42	4.74	**0.6412**	22.62	0.5554	0.4091
	Gen4LV (ours)	**0.7060**	0.8555	4.48	68.74	3.42	0.6379	20.73	0.5273	0.3197



Figure 3: **Blind** $\times 4$ **super-resolution on YouHQ40.** Top: an input/ground-truth overview and cropped results at frames 1, 16, and 31 for the input, task-specific baselines, Gen4LV (orange), and ground truth. Bottom: full-frame comparisons between directly applied video generators and Gen4LV at frame 16, with the blue regions enlarged. Gen4LV is shown without the consistency step.

film restoration and *AI-generated video restoration* evaluate real-world and generative artifacts using MovieLQ (Yu et al., 2026) alongside two newly curated subsets: MovieLQ-BW and AIGC-LQ (Lin et al., 2025). Appendix C details more dataset curation and evaluation protocols.

Metrics. Blind super-resolution uses full-reference fidelity (PSNR, SSIM, LPIPS (Zhang et al., 2018)) and no-reference quality, anchoring our perceptual claims on video-level DOVER (Wu et al., 2023), FastVQA (Wu et al., 2022), and Q-Align (Wu et al., 2024), supplemented by frame-level MUSIQ (Ke et al., 2021), NIQE (Mittal et al., 2012), and CLIPIQA+ (Wang et al., 2023a). Colorization adds DISTS (Ding et al., 2020), chrominance (PSNR-ab, ΔE_{00} (Sharma et al., 2005)), and temporal consistency (CDC (Liu et al., 2024)). Archival films rely on no-reference scores, plus input deviation and color metrics for grayscale clips. For AI-generated videos, we use HPSv3 (Ma et al., 2025) and Qwen3-VL-32B (Bai et al., 2025a) to assess realism, alongside CLIP (Radford et al., 2021) and RAFT (Teed & Deng, 2020) for source agreement. For ablations we also report *dispersion*: each output frame is registered to the upsampled input frame with an affine ECC alignment (Evangelidis & Psarakis, 2008), and dispersion is the mean absolute deviation of the per-frame

Table 2: **Colorization, DAVIS.** The two protocols are not comparable. REDS4 is in Appendix A.

		full-reference			colorization-specific			no-reference, video		
	Method	PSNR↑	LPIPS↓	DISTS↓	PSNR-ab↑	ΔE_{00} ↓	CDC↓	DOVER↑	FastVQA↑	Q-Align↑
automatic	DDColor	24.01	0.1867	0.0861	27.73	9.77	0.0080	**0.5740**	0.7768	3.90
	ColorSurge-Tiny	**28.82**	**0.1190**	**0.0721**	**32.65**	**5.18**	0.0039	0.5238	0.7711	3.87
	DDNM	23.64	0.2065	0.1349	27.05	11.83	0.0065	0.4170	0.7181	3.50
	Gen4LV (ours)	26.02	0.1489	0.0726	29.62	8.01	**0.0037**	0.5681	**0.7833**	**4.05**
exemplar	ColorMNet	**32.86**	0.0700	0.0438	**37.32**	**3.20**	0.0041	**0.5457**	**0.7825**	3.97
	DeepRemaster 768	24.43	0.1881	0.1247	30.19	7.98	0.0099	0.4764	0.7528	3.58
	DeepRemaster 320	25.57	0.2582	0.1498	33.28	5.65	0.0094	0.4052	0.5888	3.59
	Gen4LV (ours)	32.39	**0.0683**	**0.0375**	35.86	3.53	**0.0037**	0.5179	0.7733	**4.04**

translation from its clip average, in output pixels; it measures frame-to-frame geometric instability relative to the observation.

Baselines. Blind super-resolution evaluates task-specific models, zero-shot methods, and general-purpose video generators on UDM10 (Chen et al., 2026; Wang et al., 2026a; Zhuang et al., 2026; Yu et al., 2026; Yeh et al., 2024; Wan et al., 2025; Ma et al., 2026; Yang et al., 2025). Colorization compares automatic, exemplar-based, and zero-shot approaches (Kang et al., 2023; Zhao et al., 2025; Yang et al., 2024b; Iizuka & Simo-Serra, 2019; Wang et al., 2023b). Archival clips inherit the super-resolution suite for color footage and the video colorizers for grayscale. AI-generated restoration adds generative-enhancement baselines (He et al., 2024; Bai et al., 2026).

Settings. All tasks use a single frozen generator for 50 steps. General baselines use their native defaults, while MiniMax-H3 follows Sec. 4.5; additional baseline settings are provided in Appendix A. Frequency-gated consistency is applied only for known operators: bicubic $\times 4$ for blind super-resolution (c=0.06, g=0.5, N=10 updates, with σ_y estimated as described in Appendix B) and the exact grayscale operator, full band, for colorization. The update configuration follows Sec. 4.5. Consistency is disabled for operator-free archival and AI-generated tasks. Their references remain unchanged, except the motion reference for AI-generated videos, which is Gaussian low-pass filtered (f_c=0.031) and noised (σ=80) to suppress its generative artifacts.

4.2 Blind video super-resolution

Table 1 details the quantitative results. Off-the-shelf video generators fail severely, yielding at most 0.21 DOVER and 2.20 Q-Align on UDM10 (versus 0.66 and 4.27 for Gen4LV) while exhibiting severe hallucination and degradation retention. Across all benchmarks, Gen4LV dominates no-reference quality without requiring training pairs, maintaining full-reference fidelity competitive with supervised restorers. It consistently outperforms the zero-shot baseline, confirming the generative prior's efficacy when properly conditioned. Qualitatively, Gen4LV recovers legible, temporally stable text where task-specific restorers produce smeared or fragmented strokes, as shown in Fig. 3.

4.3 Colorization and archival film

By changing or omitting the consistency operator, we apply the same frozen generator to colorization and film restoration. On DAVIS (Table 2; Fig. 7b, Appendix E), Gen4LV leads LPIPS, DISTS, CDC, and Q-Align on DAVIS under the exemplar protocol, outperforming the DAVIS-trained ColorMNet on these metrics. Under the automatic protocol, it leads Q-Align, FastVQA, and temporal color consistency. Additional REDS4 results appear in Appendix A.

For joint restoration and colorization of archival black-and-white film, we omit data consistency and evaluate without ground truth (Table 3; Fig. 4, top). Gen4LV leads all three video-level quality metrics and temporal color consistency, achieving 4.52 Q-Align and 0.135 CDC versus the best baseline scores of 3.41 and 0.232, respectively, with chroma comparable to ColorMNet. On color film at $\times 4$, it ranks first on Q-Align and NIQE and second on DOVER and MUSIQ (Table 6, Appendix A).

4.4 Artifacts of generated video

Generated video is corrupted at its source rather than observed through an operator, so $\boldsymbol{y}$ is trusted only for layout and motion; we low-pass and noise the motion reference so that appearance comes from the appearance reference while low-frequency layout and motion survive. Against the super-resolution set, VEnhancer and Vivid-VR, the last trained for generated video (Table 4), Gen4LV

Table 3: **Black-and-white archival film.** Ten clips, no ground truth; the last two columns are ranked between the two arms that restore.

	no-reference, video			colorization			departure from input	
Method	DOVER↑	FastVQA↑	Q-Align↑	Chroma↑	CDC↓	Drift-ab↓	Luma corr.↑	DISTS-g↓
DeepRemaster-768	0.4014	0.4697	3.20	15.02	3.216	**0.952**	**0.9977**	**0.0278**
ColorMNet (exemplar)	0.4606	0.5704	3.15	**30.59**	0.286	1.617	0.9996	0.0017
ColorSurge-Tiny	0.2740	0.4038	3.41	22.80	0.232	1.703	0.9997	0.0018
Gen4LV (ours)	**0.6417**	**0.8425**	**4.52**	28.84	**0.135**	1.784	0.9203	0.1397



Figure 4: **Black-and-white archival film (top) and generated video (bottom)**, no ground truth for either; ColorMNet receives our exemplar.

Table 4: **Artifacts of generated video.** Ten clips, no ground truth; the judge shares its encoder with our generator.

	VLM judge			no-reference, video			agreement with source	
Method	P(real)↑	structure↑	HPSv3↑	Q-Align↑	DOVER↑	FastVQA↑	CLIP↑	EPE %↓
VEnhancer	0.337	3.84	7.61	3.74	0.589	0.808	0.906	0.055
SeedVR2-7B	0.368	3.81	7.72	4.28	0.649	0.816	**0.956**	0.039
Vivid-VR	0.595	4.20	9.13	4.35	**0.733**	**0.887**	0.863	0.056
FlashVSR	0.598	4.14	8.78	4.42	0.725	0.881	0.868	0.055
SparkVSR	0.176	3.46	7.82	4.12	0.678	0.857	0.899	**0.033**
Gen4LV (ours)	**0.805**	**4.72**	**9.17**	**4.62**	0.695	0.869	0.913	0.102

is first on all three preference columns, the VLM judge's realness and structural plausibility and HPSv3, winning 9 or 10 of 10 paired comparisons on realness against every baseline, and takes Q-Align among the video-level criteria while keeping CLIP agreement with the source second only to SeedVR2. The bottom row of Fig. 4 shows the visual comparison: the baselines largely preserve the softness and distortion of the source face, whereas Gen4LV renders a plausible, sharp face and costume, stable across frames.

4.5 Ablations

We build the method up on UDM10 (Table 5), conditioning first and then the consistency step, and sweep the gate on REDS4, where the operator is exact, and UDM10, where it is nominal (Fig. 5).

Conditioning. Naive conditioning yields low perceptual quality and poor alignment: 28.4 MUSIQ and 14.2 px dispersion (Table 5, top). The appearance reference alone improves MUSIQ to 67.5 but increases dispersion to 28.3 px. Decoupled conditioning retains most of this perceptual gain while reducing dispersion to 1.2 px, demonstrating the complementary roles of the two references.

Consistency and the gate. Ungated consistency improves PSNR by 4.0 dB and reduces dispersion to 0.28 px, but lowers MUSIQ by 18.4 (Table 5, bottom), consistent with the texture suppression discussed in Sec. 3.3. Frequency gating preserves MUSIQ and DISTS while improving PSNR by 1.9 dB and reducing dispersion by 38% relative to conditioning alone. The deployed $N{=}10$ schedule retains a 1.6 dB PSNR gain.

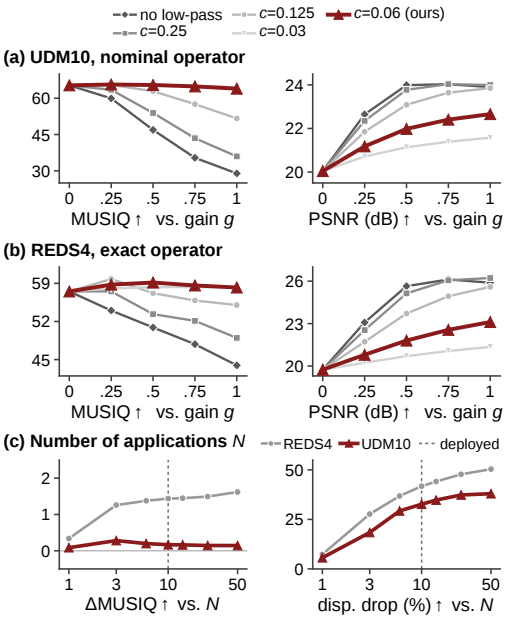

Figure 5: **Ablations of frequency-gated data consistency.** (a,b) Effects of gain and cutoff on UDM10 (nominal operator) and REDS4 (exact operator). (c) Effect of the number of corrections N at $c = 0.06$ and $g = 0.5$, with matched total correction weight.

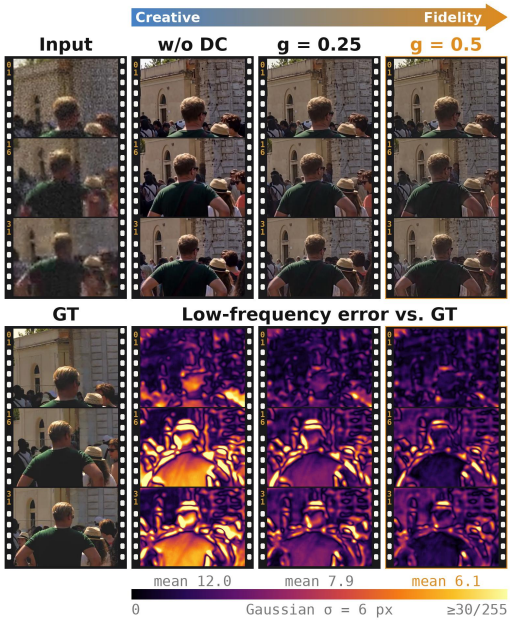

Figure 6: **Controlling the creativity–fidelity trade-off.** Outputs at increasing gain (top) and their low-frequency errors (bottom), shown for frames 1, 16, and 31 of a UDM10 clip. The cutoff is fixed at $c = 0.06$; errors are measured after Gaussian filtering with $\sigma = 6$ pixels.

Parameters of the gate. Frequency gating stabilizes perceptual quality across gains (Fig. 5a,b): increasing g from 0.25 to 1.0 reduces MUSIQ by 31 without gating but only 1.7 at c=0.06 on UDM10. The optimal cutoff narrows from 0.125 with the exact operator to 0.06 with the nominal one, consistent with Sec. 3.3; we use c=0.06 and g=0.5. With matched total correction weight, timing changes MUSIQ by at most 0.15 (Table 8), while N=10 retains 86–87% of the geometric improvement of all-step correction at one-fifth of its correction cost (Fig. 5c). Gating also reduces sensitivity to σ_y (Appendix B).

Table 5: **Conditioning and consistency ablations on UDM10.** Baseline: MiniMax-H3 with the degraded video and a restoration prompt. Dispersion (Sec. 4.1) in px. ‡: all 50 steps, g=0.5. Last row: deployed N=10 schedule used in Table 1.

	MUSIQ↑	LPIPS↓	DISTS↓	PSNR↑	disp.↓
MiniMax-H3 (baseline)	28.36	0.547	0.245	17.70	14.19
appearance reference only	67.47	0.568	0.205	14.48	28.28
+ decoupled conditioning	65.32	0.374	0.158	20.04	1.17
+ DC, full band‡	46.92	0.435	0.231	23.99	0.28
+ DC, frequency-gated‡	65.46	0.363	0.159	21.98	0.73
+ frequency-gated DC (ours)	65.49	0.364	0.158	21.68	0.79

A creativity–fidelity knob. At a fixed cutoff, the gain controls fidelity without retraining (Fig. 6). Increasing g from 0 to 0.5 reduces low-frequency error from 12.0 to 6.1 while preserving generated texture (Appendix D).

5 Conclusion

We asked whether instruction-following video generation already subsumes low-level vision. Used directly, it does not: state-of-the-art generators instructed to restore a degraded video either retain the degradation or depart from the observation. The ability is nonetheless latent. Conditioned on the degraded video for motion and on one restored frame for appearance, and held to the observation by a frequency-gated plug-and-play step, one frozen generator with no task-specific training handles four tasks, matching or exceeding in-distribution-trained specialists on no-reference video quality across eight benchmarks while narrowing, though not closing, the fidelity gap; improvements to the generator carry over to every task without retraining.

AI USE STATEMENT

In this work, we used generative AI models for video restoration, appearance reference generation, visual captioning, and automated evaluation, as described in the methodology and experiments. Additionally, we used LLM assistance for English-language editing, LaTeX formatting, and feedback on methodological descriptions, mathematical explanations, and interpretations of experimental results. We reviewed AI-assisted manuscript revisions for technical accuracy and consistency with the implemented method and reported results. The use of generative models and VLM-based evaluation is documented in the corresponding sections. We take responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Haoran Bai, Xiaoxu Chen, Canqian Yang, Zongyao He, Sibin Deng, and Ying Chen. Vivid-VR: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YV5Zgv8pdg.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=MNSW6U5zUA.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 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, 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, 2022.

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, 2026.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Georgios D. Evangelidis and Emmanouil Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008.

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

Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 01–14. IEEE, 2025.

Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. Ddcolor: Towards photo-realistic image colorization via dual decoders. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 328–338. IEEE, 2023.

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.

Michael King, Aravindh Mahendran, Matthew Koichi Grimes, Fedor Kitashov, Adham Elarabawy, Pedro Velez, Maks Ovsjanikov, and Viorica Pătrăucean. Gen4u: Unifying video generation and understanding via diffusion. *arXiv preprint arXiv:2607.06856*, 2026.

Taesung Kwon and Jong Chul Ye. Vision-xl: High definition video inverse problem solver using latent image diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10465–10474. IEEE, 2025.

Jiahao Lin, Weixuan Peng, Bojia Zi, Yifeng Gao, Xianbiao Qi, Xingjun Ma, and Yu-Gang Jiang. Brokenvideos: A benchmark dataset for fine-grained artifact localization in ai-generated videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13421–13427, 2025.

Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *Computational Visual Media*, 10(2):375–395, 2024.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, et al. Scaling mixture-of-experts video pretraining for embodied intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

Jinjie Mai, Wenxuan Zhu, Haozhe Liu, Bing Li, Cheng Zheng, Jürgen Schmidhuber, and Bernard Ghanem. Can video diffusion model reconstruct 4d geometry? *arXiv preprint arXiv:2503.21082*, 2025.

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

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1996–2005. IEEE, 2019.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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.

Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(1):21–30, 2005.

Alessio Spagnoletti, Andrés Almansa, and Marcelo Pereyra. Lvtino: Latent video consistency inverse solver for high definition video restoration. In *International Conference on Learning Representations*, volume 2026, pp. 100074–100103, 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.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, pp. 402–419, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58535-8. doi: 10.1007/978-3-030-58536-5_24. URL https://doi.org/10.1007/978-3-030-58536-5_24.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pp. 945–948. IEEE, 2013.

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, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2161–2172. IEEE, 2025.

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, 2026a.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, et al. Video generation models are general-purpose vision learners. In *European Conference on Computer Vision*, pp. 608–627. Springer, 2026b.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=mRieQgMtNTQ.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pp. 538–554. Springer, 2022.

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, 2023.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=PHjkVjR78A.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17108–17118. IEEE, 2025.

Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2(3):4, 2024.

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European conference on computer vision*, pp. 224–242. Springer, 2024a.

Yixin Yang, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Colormnet: A memory-based deep spatial-temporal feature propagation network for video colorization. In *European conference on computer vision*, pp. 336–352. Springer, 2024b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Chang-Han Yeh, Hau-Shiang Shiu, Chin-Yang Lin, Zhixiang Wang, Chi-Wei Hsiao, Ting-Hsuan Chen, and Yu-Lun Liu. Diffir2vr-zero: Zero-shot video restoration with diffusion-based image restoration models. *arXiv preprint arXiv:2407.01519*, 2024.

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.

Jiongze Yu, Xiangbo Gao, Pooja Verlani, Akshay Gadde, Yilin Wang, Balu Adsumilli, and Zhengzhong Tu. Sparkvsr: Interactive video super-resolution via sparse keyframe propagation. In *European Conference on Computer Vision*, pp. 397–415. Springer, 2026.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

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.

Hongbo Zhao, Jiaxing Li, Peiyi Zhang, Peng Xiao, Jianxin Lin, and Yijun Wang. Colorsurge: bringing vibrancy and efficiency to automatic video colorization via dual-branch fusion. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2535–2545. IEEE, 2024.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1219–1229. IEEE, 2023.

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 Additional quantitative results

Table 6: **Colour archival film, $\times 4$.** Ten clips of old film, 240×176 in, 960×704 out, no ground truth. The panel is entirely no-reference and blind to temporal behaviour; the reading is in Sec. 4.3.

	no-reference, video			no-reference, per-frame		
Method	Q-Align↑	DOVER↑	FastVQA↑	MUSIQ↑	NIQE↓	CLIPIQA+↑
DOVE	3.9652	0.4676	0.7713	60.72	4.8982	0.5371
SeedVR2-7B	3.1743	0.2608	0.4282	47.51	6.0817	0.3960
FlashVSR	4.1130	**0.5263**	**0.7954**	**65.84**	3.8073	0.5983
SparkVSR	3.7850	0.4384	0.7430	57.00	5.2154	0.4950
DiffIR2VR-Zero	3.9443	0.4237	0.7374	62.20	5.0623	**0.6220**
Gen4LV (ours)	**4.1983**	0.4835	0.7617	62.99	**3.7999**	0.5974

Table 7: **Colorization, REDS4**, outside ColorMNet's training distribution. Same columns and protocols as Table 2; the ordering holds on the perceptual columns.

		full-reference			colorization-specific			no-reference, video		
	Method	PSNR↑	LPIPS↓	DISTS↓	PSNR-ab↑	ΔE_{00} ↓	CDC↓	DOVER↑	FastVQA↑	Q-Align↑
automatic	DDColor	25.27	0.1265	**0.0556**	28.95	7.28	0.0078	**0.5644**	**0.7690**	**4.46**
	ColorSurge-Tiny	26.55	**0.1177**	0.0682	**30.23**	**6.45**	0.0059	0.4838	0.7686	4.22
	DDNM	24.32	0.1877	0.1133	27.52	11.83	0.0058	0.3867	0.7157	3.94
	Gen4LV (ours)	**26.68**	0.1265	0.0612	30.23	6.92	**0.0055**	0.5184	0.7636	4.18
exemplar	ColorMNet	**33.32**	**0.0541**	**0.0364**	**37.46**	**2.85**	0.0058	**0.4755**	**0.7745**	**4.53**
	DeepRemaster 768	24.89	0.1866	0.1207	30.52	6.84	0.0092	0.3904	0.7179	3.87
	DeepRemaster 320	25.52	0.3113	0.1552	33.20	5.45	0.0087	0.3689	0.6402	3.94
	Gen4LV (ours)	31.06	0.0933	0.0417	34.43	4.77	**0.0055**	0.4154	0.7485	4.08

Table 8: **Where the correction is applied**, change against the configuration without the step. Cutoff c=0.06; the gain of each window is compensated so that all five inject the same dose, that of steps 0–43 at g=0.5. Dispersion change in %. REDS4 on two clips, UDM10 on nine.

	REDS4 (n=2)		UDM10 (n=9)	
Steps receiving the correction	ΔMUSIQ	Δdisp.	ΔMUSIQ	Δdisp.
0–43	+0.73	−23	+0.17	−7
44–47	+0.35	−12	+0.15	−6
48	+0.30	−10	+0.16	−5
49	+0.42	−9	+0.29	−4
ten steps spread over the schedule	+0.46	−11	+0.30	−4

Where the correction is applied. Under the flow shift of 12 the weight κ_t with which a correction lands varies by a factor of 589 across the 50 steps, and step 49 alone carries 41% of the total, so an early and a late window at the same gain differ mainly in dose. With the dose equalized, the five windows of Table 8 differ by 0.43 MUSIQ on REDS4 and 0.15 on UDM10, both within the MUSIQ noise floor of 0.73. The earliest window repairs the most geometry on REDS4, consistent with leaving the prior time to redraw texture around the corrected skeleton, but on two clips this is a hypothesis.

Frequency and strength on REDS4. With the exact operator (Fig. 5b) the ungated step again collapses at every gain, from 54.0 to 43.9 MUSIQ, while the gated settings hold or raise MUSIQ

over the conditioning alone (57.5) at low gain. The best setting moves to the wider band, c=0.125 at g=0.25 (+2.3 MUSIQ, +2.0 dB PSNR), against c=0.06 with the nominal operator of UDM10; the deployed c=0.06, g=0.5 still gains 1.6 MUSIQ and 2.1 dB here.

Table 9: **Video generators used directly, SPMCS and YouHQ40**; the UDM10 block is in Table 1. Same protocol and scoring pipeline as Table 1, each clip on its own canvas; Gen4LV is repeated from Table 1 for reference and nothing is ranked.

		no-reference, video			no-reference, per-frame			full-reference		
Dataset / type	Method	DOVER↑	FastVQA↑	Q-Align↑	MUSIQ↑	NIQE↓	CLIPIQA+↑	PSNR↑	SSIM↑	LPIPS↓
SPMCS	Wan2.2-5B	0.1840	0.2095	2.15	27.91	8.57	0.2943	16.63	0.4310	0.7028
video generation	LingBot-Video (refiner)	0.1645	0.1569	1.90	23.72	9.27	0.3141	19.54	0.4994	0.6712
	LingBot-Video	0.1572	0.1692	1.87	23.42	9.51	0.3240	19.11	0.4865	0.6906
	CogVideoX1.5-5B	0.1564	0.1824	2.02	26.38	9.15	0.2843	16.52	0.4419	0.7239
	MiniMax-H3	0.2206	0.3913	2.53	41.60	6.04	0.4035	19.79	0.5109	0.4639
zero-shot	**Gen4LV (ours)**	0.5246	0.7457	4.30	71.22	3.40	0.7043	18.91	0.4345	0.3313
YouHQ40	Wan2.2-5B	0.4018	0.3173	2.71	27.32	9.20	0.2973	17.75	0.5049	0.6441
video generation	LingBot-Video (refiner)	0.3292	0.2126	2.49	24.09	9.53	0.2703	20.41	0.5596	0.6283
	LingBot-Video	0.3223	0.1998	2.44	23.42	9.80	0.2651	19.82	0.5459	0.6441
	CogVideoX1.5-5B	0.3815	0.2504	2.74	26.70	9.52	0.2802	17.19	0.4975	0.6878
	MiniMax-H3	0.3146	0.3557	2.59	30.81	7.06	0.3131	20.24	0.5483	0.4819
zero-shot	**Gen4LV (ours)**	0.7060	0.8555	4.48	68.74	3.42	0.6379	20.73	0.5273	0.3197

Video generators used directly. Each generator takes the bicubic-upsampled input at its default strength, 0.7 for Wan2.2 and LingBot-Video, 0.8 for CogVideoX1.5 and t=0.85 for the LingBot-Video refiner, with a description of the clip, written by the same VLM from the degraded input, and a request for a sharp, detailed rendering as the prompt; MiniMax-H3 receives the degraded video as its only reference, at the canvas short edge. Table 9 extends the grey rows of Table 1 to SPMCS and YouHQ40 under this protocol. The picture is the one on UDM10: every video-level score stays far below both the task-specific models and Gen4LV, a DOVER of at most 0.22 on SPMCS and 0.40 on YouHQ40 against our 0.52 and 0.71, and so does MUSIQ. PSNR and SSIM can exceed ours, on SPMCS by up to 0.9 dB, because an output that barely departs from the blurred input inherits its fidelity; LPIPS, which penalises the blur, is 0.46–0.72 against our 0.32–0.33. Three YouHQ40 clips are 1024×1024, beyond the positional range of CogVideoX1.5; for these its rotary table is extended, which the model has not seen in training.

Alternatives that fail. Injecting the observation as a noised initial state, as in SDEdit, injects usable and corrupted bands together, so no noise level keeps the conditioning and discards the blur at once; appending a sentence naming the artifacts leaves the output bit-identical. Neither failure is a tuning matter.

B Probing the operator mismatch

The mismatch level σ_y in Eq. 1 is measured, not tuned. For each track a probe set of ground-truth videos and their degraded inputs is formed from the training split of the corresponding benchmark, disjoint from every test sequence, and the nominal operator $\boldsymbol{H}$ is applied to the ground truth; σ_y is the root-mean-square residual $\|\boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}\|$ over the probe set. The fit is reported as a PSNR. On REDS4, whose degradation is the exact $\times 4$ operator, $\boldsymbol{H}$ fits to 56.6 dB and σ_y sits at the quantization floor. On UDM10, whose degradation is the blind pipeline of Sec. 4.1, seven candidate resamplers (bicubic, bilinear, Lanczos and area averaging with and without antialiasing) all fit to 25–26 dB and are indistinguishable, which is why the deployed $\boldsymbol{H}$ is bicubic and the trusted band is narrow. Once the correction is frequency-gated, substituting the REDS4 value of σ_y for the UDM10 one changes MUSIQ by less than the noise floor.

C Implementation details

The prior. MiniMax-H3 generates video and audio jointly; we use its Ref2VA mode and discard the audio stream. Semantically, a Qwen3-VL encoder truncated to its first 50 of 64 layers is run once over a single sequence in which each reference contributes a marker and a run of placeholder tokens filled from its pixels, followed by the prompt, and its last hidden state conditions the 50-block denoiser. Spatially, the same reference is encoded by the video autoencoder ($16\times$ spatial, $4\times$ temporal, 24 latent channels) into clean reference tokens held fixed over the 50 denoising steps.

Numerics of the consistency step. The correction returns to latent space as a residual, $\boldsymbol{z}_{\mathrm{lat}} = \hat{\boldsymbol{x}}_0 + \mathrm{Enc}(\boldsymbol{z}) - \mathrm{Enc}(\boldsymbol{u})$, so the autoencoder round trip cancels to first order and $g = 0$ is bit-exact. The

Euler update is written incrementally, as the plain step plus $(\boldsymbol{z}_{\mathrm{lat}} - \hat{\boldsymbol{x}}_0)\kappa_t$ with $\kappa_t = (\sigma_t - \sigma_{t+1})/\sigma_t$, since the algebraically identical rewrite accumulates 38 dB of noise over 50 steps in half precision.

Archival film. MovieLQ clips are cut to single-shot 32-frame windows; MovieLQ-BW consists of ten 32-frame clips we cut from six public-domain black-and-white films.

AIGC-LQ. The ten clips, selected from BrokenVideos as representative generation-side degradations, span Vidu (Bao et al., 2024), CogVideoX (Yang et al., 2025), Open-Sora (Zheng et al., 2024), MiniMax and EasyAnimate (Xu et al., 2024), and are normalized to a common input of 32 frames at 384×256 and restored at $\times 4$ to 1536×1024; the native frames serve only as a reference for agreement.

D A CREATIVITY–FIDELITY KNOB

Table 10: **The gain as a creativity–fidelity knob**, UDM10, ten clips, c=0.06, correction at every step; g=0 is the conditioning alone; dispersion in px.

g	0	0.25	0.5	0.75	1.0
PSNR↑	20.04	21.18	21.98	22.41	22.66
LPIPS↓	0.374	0.364	0.363	0.367	0.370
disp.↓	1.17	0.98	0.73	0.56	0.45
MUSIQ↑	65.32	65.66	65.46	64.87	63.94

At g=0 the output is the conditioning alone: the prior interprets the input freely, with the richest texture and a layout and tone that may drift from the observation; raising g pulls the low frequencies toward the observation and leaves the texture to the prior. On the clip of Fig. 6 the low-frequency error falls from 12.0 to 7.9 and 6.1 (in units of $1/255$, averaged over the three crops) and PSNR rises from 17.9 to 19.6 and 20.9 dB at an unchanged MUSIQ (61.2, 62.9, 62.7); the figure applies the correction at every step, and the deployed ten-step schedule behaves alike (Fig. 5c). Over all ten UDM10 clips (Table 10), PSNR and dispersion improve monotonically with g while MUSIQ declines slowly, by 1.4 at g=1. The gate is what makes the knob usable: without it, raising g from 0.25 to 1 costs 31 MUSIQ (Fig. 5a). In colorization, where the operator is exact, the cutoff plays the same role: widening it from c=0.015 to the full band raises PSNR by 2.5 dB on DAVIS in both protocols at a cost of 2.6–5.1 MUSIQ; we deploy the full band there, the fidelity end.

E ADDITIONAL QUALITATIVE RESULTS



(a) Blind $\times 4$ super-resolution, YouHQ40



(b) Automatic colorization, DAVIS



(c) Colour archival film, $\times 4$, MovieLQ



(d) Black-and-white archival film, MovieLQ-BW



(e) Generated video, AIGC-LQ

Figure 7: **Additional qualitative results, one clip per task**, complementing Figs. 3 and 4. Each row shows a full frame with the crop marked, the ground truth where it exists (a, b) and the input otherwise (c–e), then frames 1, 16 and 31 of the input, the baselines, Gen4LV (orange) and, where it exists, the ground truth. In (a) Gen4LV is shown without the consistency step.