CreativeVR: Diffusion-Prior-Guided Approach for Structure and Motion Restoration in Generative and Real Videos

Anonymous CVPR submission

Paper ID 00009

Abstract

Generative video models are rapidly entering cinematic production pipelines, yet even state-of-the-art text-to-video (T2V) systems produce fine-scale structural artifacts, distorted faces and hands, warped backgrounds, and temporally inconsistent motion that prevent direct use in professional workflows. Re-generation is costly, non-deterministic, and risks deviating from creative intent, motivating a dedicated post-production refinement *stage. Classical video restoration and super-resolution (VR/VSR) methods, in contrast, are tuned for synthetic degradations such as blur and downsampling and tend to sharpen these artifacts rather than repair them, while diffusion-prior restorers are usually trained on photometric noise and offer little control over the trade-off between perceptual quality and fidelity. We introduce* CreativeVR, *a diffusion-prior-guided video restoration framework for AI-generated (AIGC) and real videos with severe structural and temporal artifacts. Built as a lightweight deep adapter on a frozen T2V DiT backbone, CreativeVR exposes a single* precision knob *that lets editors and artists smoothly trade off between faithful detail preservation and aggressive structure/motion-corrective synthesis, functioning as a creative control surface in the post-production pipeline. Our key technical contribution is a* temporally coherent synthetic degradation module *that composes morphing, directional motion blur, and grid-based warping to simulate realistic geometric failures during training, aligning the diffusion prior toward the hard failure modes of modern generators. We evaluate on the curated* AIGC54 *benchmark spanning outputs from five T2V models (Veo3, Sora, Pika, Firefly, Ray3), using FIQA, perceptual, and GPT-based multi-aspect scoring. CreativeVR achieves state-of-the-art AIGC refinement quality while remaining competitive on standard video restoration benchmarks, with practical throughput (*$\sim$*13 FPS @ 720p on a single 80 GB A100).*

1. Introduction

Modern video generation models have made significant advances in generating high-quality video content in terms of semantic composition and adherence to user prompts [2, 8, 10, 27]. With the increasing demand for high-quality content, these models are expected to produce production-level videos with high resolution and high frame rates, which requires highly detailed structures and realistic motion dynamics. They are strong overall but brittle at fine-scale structural details. Even cutting-edge models such as Veo3 [8], Ray3 [2], Kling [33], LTX-2 [21] and SORA2 [27] produce structural and temporal artifacts, including imperfect faces and hands, broken topology, warped backgrounds, cross-frame drift, missing object permanence, and fine-detail flicker (Fig. 1(b)). These artifacts hinder adoption in production, where re-generation or prompt iteration is costly, non-deterministic, and risks deviating from the original intent; this motivates positioning a dedicated *post-production refinement* stage that corrects structure and motion while preserving the semantics and identity of the source video.

Severe artifacts are not unique to AI-generated content (AIGC). Real-world footage, including legacy archives, smartphone videos under compression and low light, low-frame-rate captures, and scans of damaged material, often exhibits degraded detail and temporal instability Fig. 1(c). A practical solution should improve geometry and motion in both *generated* and *real* videos while remaining faithful to the source.

Classical video restoration and super-resolution (VR/VSR) have been highly effective for degradations such as blur, noise, and downsampling, commonly modeled by synthetic or stochastic processes (Fig. 1(a)) [7, 40]. However, these methods lack semantic priors; when presented with generative artifacts (e.g., a malformed hand or face), they tend to stabilize the artifact rather than repair it. Post-hoc fixes for T2V outputs, such as GAN or diffusion upsamplers and refiners like VideoGigaGAN [44] and VEnhancer [11], either alter composition

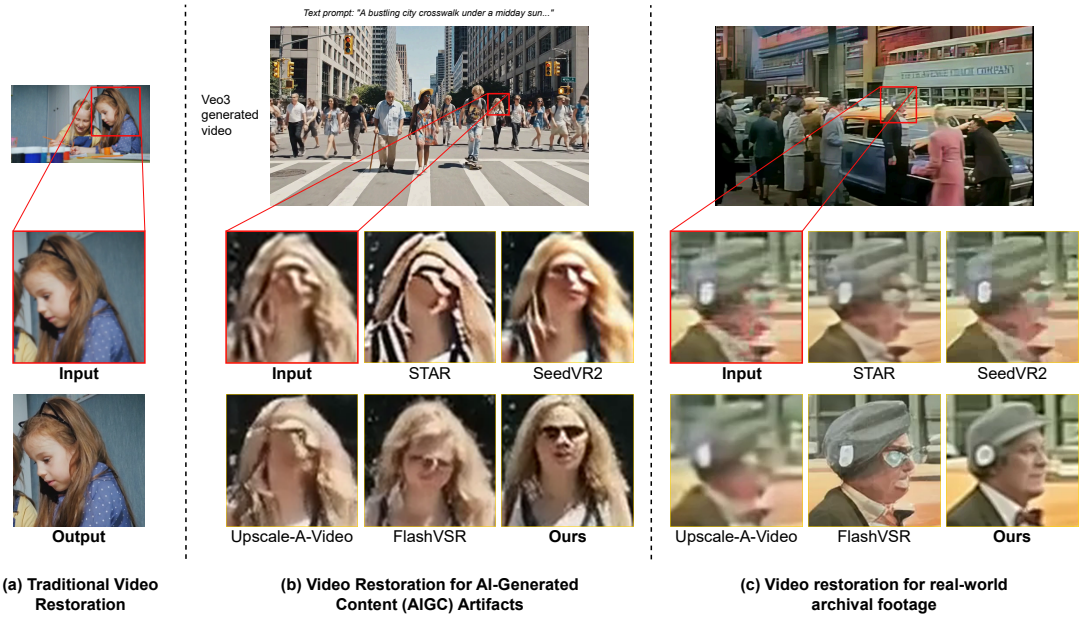

Figure 1. **Precise Restoration vs Structural-corrective Video Restoration.** (a) Traditional Video Restoration represents the *precision* regime: recover detail under synthetic degradations when geometry and motion are intact. (b) AIGC clips exemplify the *prior-guided spatiotemporal correction* regime, the goal is to fix geometric errors and restore temporal coherence while preserving semantics and identity. (c) Real-world archival footage lies between these extremes and benefits from both precision and prior-guided correction. Our method modulates this trade-off via a precision knob, behaving precise in case (a) and like a prior-guided corrector in cases (b) and (c).

and identity through uncontrolled hallucination or remain domain-specific and fail to generalize across both AI and real footage [11, 44]. Diffusion-prior restorers thus sit in an uncomfortable regime: they have strong generative priors but are typically trained on photometric degradations, with no mechanism to steer how much the model should trust the corrupted input versus the prior. As a result, they may sharpen textures without repairing geometry (Fig. 1(b)), or over-correct and drift from the source [36, 42, 54, 55].

In this work we propose *CreativeVR*, a diffusion-prior-guided video restoration framework that explicitly targets this trade-off. We build a deep adapter on top of a frozen text-to-video DiT backbone: the backbone provides a strong generative prior learned from large-scale T2V training, while the adapter is conditioned on the degraded input and trained for restoration. At training time, clean clips and their synthetically degraded counterparts are encoded into a shared latent space, and the adapter features are injected into alternating backbone blocks through a scalar precision control γ_ℓ. Small γ_ℓ values recover the familiar "precision" regime of Fig. 1(a), whereas larger γ_ℓ values let the adapter act as a structure- and motion-corrective prior for AIGC and real videos with severe artifacts, as in Fig. 1(b,c).

A key novelty is our temporally coherent synthetic degradation module, which we use as a *degradation-as-control* curriculum. Instead of training on simple noise or bicubic downsampling, we compose several coherent degradations such as temporal morphing, directional motion blur, grid-based warping, whose parameters evolve smoothly over time. This produces clips with realistic structural failures (e.g., warped faces, wobble, low-FPS blend, and rolling-shutter-like distortions) but without unnatural flicker, and explicitly aligns the diffusion prior toward the hard failure modes observed in modern video generators.

Finally, we evaluate CreativeVR both in the AIGC regime and on standard restoration benchmarks. We curate *AIGC54* test set from several state-of-the-art video generators and design an evaluation protocol that combines face-structure quality (FIQA), semantic and perceptual metrics, and GPT-based multi-aspect and arena-style preference scoring. Across AIGC clips, CreativeVR attains state-of-the-art results, including up to **+37%** relative FIQA improvement over the inputs and consistent gains across all metrics, and it also outperforms existing methods on real videos with severe spatio-temporal artifacts. Our method achieves practical throughput ($\sim$13 FPS @ 720p, $\sim$4 FPS @ 1080p on a single 80 GB A100) and generalizes in a zero-shot manner up to 1080p resolution.

Our main contributions are summarized below:

- We introduce **CreativeVR**, a diffusion-prior-guided deep adapter on a frozen T2V DiT backbone, equipped with a single precision knob that smoothly trades off between

precise restoration and structure-/motion-corrective refinement for both AIGC and real videos.

- We propose a **temporally coherent synthetic degradation** module that composes morphing, directional motion blur, grid-based warping, frame dropping, and spatio-temporal resampling to generate realistic structural artifacts without temporal flicker, providing a targeted training curriculum for diffusion priors.
- We curate the *AIGC54* benchmark with a metric suite combining FIQA, GPT-based multi-aspect, and VBench evaluation, and demonstrate state-of-the-art AIGC refinement while remaining competitive on standard VSR benchmarks, with throughput ($\sim$13 FPS @ 720p).

2. Related Work

2.1. Traditional (Precise) Video Restoration

A large body of image restoration work [18,20,22,24,30, 38, 48, 49] has shown that powerful generative priors significantly improve reconstruction under complex degradations. However, video restoration additionally requires *temporal consistency* in both structure and appearance, making frame-independent refiners inadequate for long-range distortions, flicker, and geometric inconsistencies.

Classical VR/VSR methods from deformable alignment (EDVR [40], TDAN [34]) through bidirectional propagation (BasicVSR++ [7]) and recurrent memory (RLSP [9], TMNet [43]) to GAN-based upscaling (VideoGigaGAN [45]) are trained under simplified degradations and remain fidelity-oriented, failing to generalize to the geometry-level artifacts of modern T2V models. CreativeVR addresses this gap with a frozen diffusion prior and degradation-conditioned adapters for structure-faithful restoration.

2.2. Diffusion Models for Video Restoration

Diffusion-based priors offer a more expressive alternative to classical VSR. Image-domain refiners such as StableSR [39] have been extended to video: STAR [42] adds temporal alignment layers, while VEnhancer [11] injects cross-frame attention. More recently, SeedVR [37], SeedVR2 [36], InfVSR [53], and FlashVideo [51] directly fine-tune large video diffusion transformers, achieving strong results but at the cost of substantial compute and potential degradation of the pretrained motion and composition priors. CreativeVR instead freezes the backbone and learns lightweight adapters, achieving efficient restoration without altering the original prior.

2.3. Efficient Diffusion Adaptation

Parameter-efficient techniques such as Adapters [12], LoRA [13], and ControlNet-style conditioning [50] enable lightweight specialization of large models. Deep adapter approaches (VACE [16], ResTuning [17]) provide modular conditioning with fewer trainable parameters. CreativeVR adopts a *degradation-as-control* design: the degraded video itself serves as an internal control signal via lightweight adapters trained under a synthetic degradation curriculum, turning adapter-based conditioning into a self-supervised restoration mechanism that preserves the frozen backbone's motion and composition priors.

Recent timestep distillation methods [14,23,31,47] compress diffusion sampling to a few steps. CreativeVR is fully compatible: our adapters plug directly into a distilled backbone without retraining.

3. Method

We address the problem of diffusion–prior-guided video restoration. Given a clean target video $x \in \mathbb{R}^{T\times H\times W\times 3}$, we synthesize a degraded input video $\tilde{x}$ and learn a diffusion model that restores $\tilde{x}$ back to the high–quality space of x. At test time, the model receives only a low–quality video and produces a restored video $\hat{x}$ at the target resolution.

3.1. Architecture and Optimization

Our framework builds on a text-to-video (T2V) diffusion transformer (DiT) backbone, similar to VACE [16], augmented with a deep adapter branch (Fig. 2). The T2V DiT backbone is pretrained on a large-scale video generation task and kept frozen during training. The adapter is a lighter DiT with the same block design but only half as many blocks as the backbone.

Let E denote the video VAE encoder and P the patch embedder. For each clean training video x, we first construct a degraded counterpart $\tilde{x} = \mathcal{D}(x;\eta)$, where $\mathcal{D}$ is our synthetic degradation module with parameters η (Sec. 3.2). Both x and $\tilde{x}$ are mapped into the latent space using the same frozen VAE, yielding $z = E(x)$ and $\tilde{z} = E(\tilde{x})$. The degraded latent $\tilde{z}$ is further projected to conditioning tokens $c = P(\tilde{z})$. We follow a standard diffusion formulation over the clean latent video z. A noise scheduler samples a timestep t and produces a noisy latent

$$z_t = \alpha_t z + \sigma_t \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

which is passed to the frozen DiT denoiser. Let h_ℓ denote the hidden features at the ℓ-th DiT block of the backbone, and let $a_\ell = A_\ell(c)$ be the output of the corresponding adapter block, which processes the degraded-video tokens c. At a subset of layers $\ell \in \mathcal{L}_{\text{adapt}}$ we fuse backbone and adapter features via a residual modulation:

$$\tilde{h}_\ell = h_\ell + \gamma_\ell\, a_\ell, \tag{2}$$

where $\gamma_\ell \geq 0$ is a scalar *precision knob* that controls the adapter's influence on the backbone. Small values of γ_ℓ preserve the frozen T2V prior (strong generative bias, less

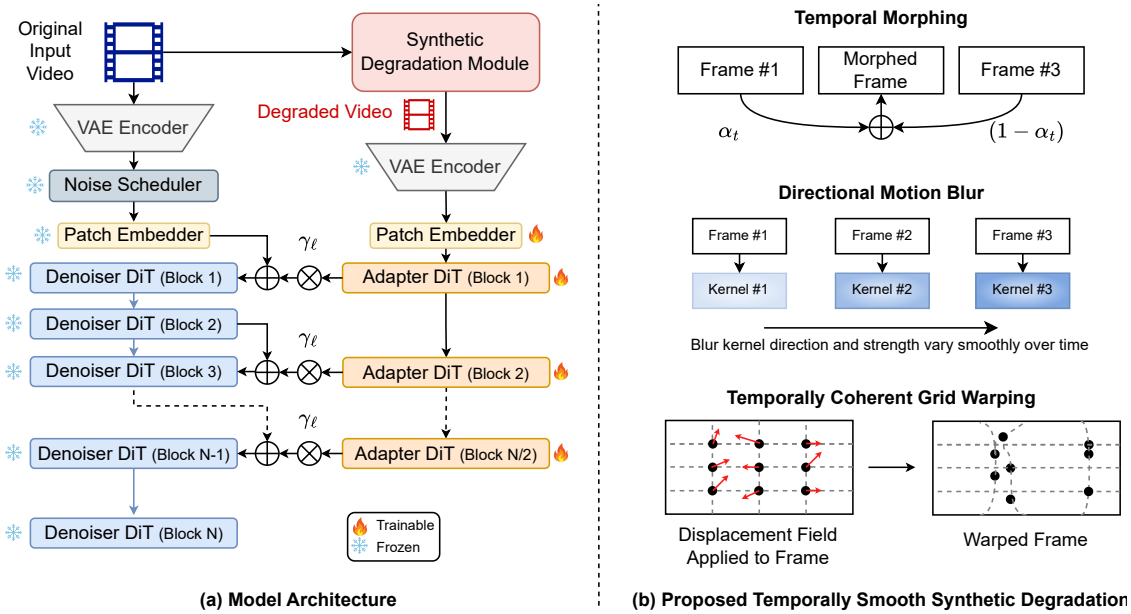

Figure 2. **CreativeVR overview: diffusion-prior-guided video restoration.** (a) During training, a clean input clip is passed through a temporally smooth synthetic degradation module to produce a degraded counterpart; both clips are encoded by a frozen VAE and processed by a frozen text-to-video DiT backbone augmented with a lightweight adapter DiT. The adapter is conditioned on degraded latents and injected into alternating backbone blocks via a precision knob γ_ℓ, allowing the model to trade off prior strength and faithfulness to the input. (b) The degradation module composes temporally coherent morphing, directional motion blur, and grid-based warping to mimic realistic structural and motion artifacts, guiding the adapter to learn structure- and motion-corrective restoration behavior.

dependence on degradation), while larger values increase fidelity to the degraded input. The complete denoiser $\varepsilon_{\theta,\phi}$ consists of the frozen backbone parameters θ and trainable adapter parameters ϕ. We optimize only ϕ with the standard noise-prediction diffusion loss

$$\mathcal{L}_{\text{diff}}(\phi) = \mathbb{E}_{x,t,\varepsilon}\left[\left\|\varepsilon - \varepsilon_{\theta,\phi}(z_t, t, c)\right\|_2^2\right], \tag{3}$$

encouraging the coupled backbone–adapter system to denoise towards the clean video latent z while being guided by the degraded-video tokens c.

3.2. Synthetic Training Augmentations

We instantiate the synthetic degradation operator $\mathcal{D}(x;\eta)$ as a composition of temporally coherent augmentations that mimic natural capture artifacts such as motion blur, geometric wobble, low frame rate, and exposure-induced blending. Given a clean RGB clip $x = \{X_t\}_{t=1}^T$, the module produces a degraded clip $\tilde{x} = \mathcal{D}(x;\eta) = \{Y_t\}_{t=1}^T$, with augmentation parameters that evolve smoothly over time to avoid high-frequency flicker.

Structure-preserving spatiotemporal downsampling. We jointly downsample the clip in time and space by factors s_{temp} and s_{spat}, then reconstruct it back to the original resolution and frame rate by simple interpolation, mimicking low frame-rate and low-resolution capture while preserving global scene layout.

Temporal morphing. We interpolate between adjacent frames to mimic low-shutter or low-FPS blending,

$$Y_t = \alpha_t X_t + (1-\alpha_t) X_{t+1}, \tag{4}$$

where $\alpha_t \in [\alpha_{\min}, \alpha_{\max}]$ varies smoothly across time. This softly blends motion phases across frames, resembling exposure integration in real cameras.

Stochastic frame dropping. We sample a binary drop mask $m_t \sim \text{Bernoulli}(p_{\text{drop}})$ under a maximum run-length constraint and remove frames with $m_t = 0$; missing frames are then reconstructed by linear interpolation between the nearest retained neighbors. This simulates realistic temporal discontinuities observed in mobile or streaming videos while keeping motion trajectories plausible.

Directional motion blur. We model exposure-integrated motion by convolving frame with an oriented line kernel,

$$Y_t = K_t(\theta_t, \ell_t) * X_t, \tag{5}$$

where θ_t and ℓ_t denote the blur orientation and kernel length, respectively. By allowing θ_t and ℓ_t to change smoothly over time, we imitate camera or object motion that produces temporally varying, yet coherent, motion blur.

Grid-based spatial warping. To simulate rolling-shutter wobble and low-frequency geometric distortions, we generate a displacement field $d_t = (d_t^x, d_t^y)$ by upsampling

smooth low-resolution noise and apply it via backward warping,

$$Y_t(u) = X_t\big(u + d_t(u)\big), \tag{6}$$

where u indexes pixel coordinates. The resulting elastic, wavy deformations resemble lens wobble, heat haze, or hand-held jitter while keeping local geometry plausible.

Smooth parameter trajectories. All temporal parameters $(\alpha_t, \theta_t, \ell_t, d_t)$ are drawn from low-frequency trajectories obtained using sinusoidal or Perlin-noise bases and optionally smoothed with a 1D box filter. This design avoids abrupt parameter jumps and prevents synthetic flicker, yielding degradations that evolve smoothly like real-world capture effects.

Our augmentations are particularly helpful because we distort *structures* in the input, whereas the prior method uses only non-structural statistical corruptions (e.g., speckle noise, JPEG compression). When the base generation model receives structurally distorted guidance, it is encouraged to learn to correct structure.

3.3. Sampling

At inference time we condition only on a degraded input clip $\tilde{x}$: we encode $\tilde{x}$ to latents $\tilde{z} = E(\tilde{x})$, obtain conditioning tokens $c = P(\tilde{z})$, and run the diffusion sampler on $\tilde{z}$ to produce a restored latent $\hat{z}$, which is decoded by the VAE decoder into the final video $\hat{x}$. For efficient sampling, we load CausVid-distilled LoRA weights [47] into the same frozen Wan2.1 backbone, reducing the number of sampling steps from 50 to 4. Since the architecture and latent space remain identical, our trained adapter blocks A_ℓ apply without any modification or retraining. Qualitatively, we observe no visible degradation in restoration quality from this swap across our test sets.

4. Experiments

4.1. Dataset and Implementation Details

Datasets. We train CreativeVR on the Mixkit split of OpenSora-Plan v1.1.0 [26], which provides ∼10K open-source video clips spanning diverse scenes. All videos are sampled as 49-frame sequences and resized to 480×832 before augmentation. To evaluate refinement quality on production-relevant AIGC content, we curate the *AIGC54* evaluation set, which consists of 54 five-second clips collected from five state-of-the-art T2V systems Veo3 [8], Pika2.2 [19], Firefly [1], Ray3 [2], and Wan2.1 [35] spanning diverse cinematic scenarios including traditional dance, crowded street scenes, courtroom debates, sports events, and newsroom.

Evaluation metrics. To assess AIGC artifacts on faces, we score face crops with six Face Image Quality Assessment (FIQA) models: eDifFIQA [5], DifFIQA [4], CLIB-FIQA [28], CR-FIQA [6], MR-FIQA [29], and FaceQAN [3]. We additionally report an aesthetic score from a CLIP-LAION initialized model and the objectness confidence of a YOLOv8-based face detector to track perceptual appeal and detection reliability.

Implementation Details. We utilize the WAN2.1-based 1.3B-parameter DiT model as the base architecture, with deep adapters inserted into alternating DiT blocks. We train on $T = 49$ frames resized to 480×832 (divisible by latent strides). Training is fast and converges in approximately 5k iterations on 8 NVIDIA H100 GPUs, with a batch size of 1 per GPU. We use the AdamW optimizer with a learning rate of 10^{-4}, gradient accumulation $\times 2$, and global gradient norm clipping at 2.0. The parameter γ_ℓ is initialized to 1.0 during training.

During sampling, we replace the teacher backbone weights with the student CausVid model and use 4 sampling steps. By default, the parameter γ_ℓ is set to 0.4 for prior-guided restoration and 1.0 for precise video restoration benchmarks. Our method supports zero-shot inference for input resolutions ranging from 176p up to 1080p, achieving approximately 13 FPS at 720p and 4 FPS at 1080p on a single 80 GB A100 GPU.

4.2. Quantitative Comparison

We consider two main set of evaluation based on prior guided corrective which requires severe structural-motion artifact correction and precision-based traditional video restoration.

4.2.1 Prior-guided corrective Video Restoration.

Structural Integrity Evaluation. We quantitatively assess AIGC artifacts on the *AIGC54* dataset and compare against state-of-the-art video restoration methods: FlashVSR [55], Real-ESRGAN [41], Real-Viformer [52], ResShift [48], SeedVR2 [36], STAR [42], Upscale-A-Video [54], and VEnhancer [11]. Because generated videos lack paired references, we evaluate with no-reference, face-centric quality signals that are sensitive to structural plausibility. Concretely, we uniformly sample 16 frames per video, detect faces on the *input* using a YOLOv8-based face detector, expand boxes by 10% on each side, and extract the identical boxes from each restored output (no re-detection) to avoid selection bias. We report per-crop *relative gain* over the input, i.e., $\Delta = \text{score}_{\text{out}} - \text{score}_{\text{in}}$, averaged over crops and frames (see Table 1(a)).

To target structural integrity, we adopt six FIQA metrics: diffusion-prior robustness (DifFIQA/eDifFIQA [4,5]), recognition-embedding separability and confidence calibration (CR-FIQA/CLIB-FIQA [6, 28]), and multi-reference plus adversarial-noise sensitivity (MR-FIQA/FaceQAN [3, 29]); together, these provide reliable signals of geometric/edge correctness, texture realism, and identity preservation under AIGC artifacts. Notably, sharpening-oriented baselines sometimes fail to improve face quality (or even

Quality Metric \ Restoration Method	FlashVSR [55]	Real-ESRGAN [41]	Real-Viformer [52]	ResShift [48]	SeedVR2 [36]	STAR [42]	Upscale-A-Video [54]	VEnhancer [11]	Ours
(a) Structural Integrity *— relative improvement (%) over input videos*									
eDifFIQA	19.60	-0.60	9.40	4.50	21.60	28.90	-6.80	28.90	**35.60**
DifFIQA	-0.40	0.40	0.10	-0.10	0.40	**1.10**	-0.30	0.50	0.70
CLIB-FIQA	9.20	-1.70	2.70	1.10	8.90	7.00	-0.50	10.90	**14.80**
CR-FIQA	3.80	6.30	0.70	3.90	6.90	12.70	0.30	12.90	**17.20**
MR-FIQA	16.10	-6.20	-1.30	-1.70	20.40	22.50	-8.10	18.60	**37.40**
FaceQAN	2.50	-2.00	-8.70	-7.00	13.40	**34.40**	-11.30	10.60	30.60
Aesthetic Score	8.32	-2.29	2.15	0.72	4.16	4.16	-1.72	10.22	**11.19**
Objectness	3.45	0.00	0.04	0.00	1.61	1.16	0.00	2.06	**4.24**
(b) Multi-Aspect Video Quality *— GPT-based scoring (0–10 scale)*									
Visual Quality	7.86	6.91	6.73	7.05	7.36	8.14	6.73	7.41	**8.05**
Temporal Consistency	8.05	6.82	7.00	7.32	7.32	8.27	7.95	7.91	**8.55**
Face Quality	**6.64**	5.14	4.95	5.05	5.50	5.77	4.41	5.27	6.23
Motion Realism	7.95	6.68	7.18	7.09	7.14	8.05	7.68	7.41	**8.27**
Lighting & Atmosphere	**8.55**	8.05	8.05	8.36	8.23	8.50	7.82	8.36	8.04
Detail Preservation	7.73	6.55	6.23	6.73	7.00	7.73	6.45	6.95	**7.82**
Overall Mean	7.80	6.69	6.69	6.93	7.10	7.74	6.84	7.22	**7.82**
(c) VBench Perceptual Quality *— higher is better*									
Aesthetic Quality	0.6313	0.6338	0.6334	0.6284	**0.6519**	0.6157	0.6085	0.6314	0.6340
Background Consistency	0.9486	0.9481	0.9477	0.9503	0.9416	0.9506	0.9493	**0.9565**	0.9493
Imaging Quality	0.7229	0.7201	0.7075	0.6927	0.6963	**0.7347**	0.7078	0.6999	0.7311
Motion Smoothness	0.9841	0.9804	0.9853	0.9843	0.9850	0.9827	0.9856	**0.9907**	0.9879
Dynamic Degree	0.9118	0.8824	0.8761	0.8655	0.8529	0.8800	0.8529	0.8824	**0.9444**
Subject Consistency	0.9381	0.9385	0.9384	0.9384	0.9343	0.9363	0.9375	**0.9413**	0.9387
Average	0.8561	0.8506	0.8481	0.8433	0.8437	0.8500	0.8403	0.8504	**0.8642**

Table 1. **Comprehensive AIGC-artifact evaluation on AIGC54. (a)** Structural integrity via six FIQA metrics, aesthetic score, and face-detection confidence (relative gain % over input). **(b)** GPT-based multi-aspect scoring (0–10). **(c)** VBench perceptual quality. **Bold** = best, underline = second best.

Dataset	Metrics	FlashVSR [55]	Real-ESRGAN [41]	Real-Viformer [52]	ResShift [48]	SeedVR2 [36]	STAR [42]	Upscale-A-Video [54]	VEnhancer [11]	Ours (Medium)	Ours (Strong)
UDM10 [32]	PSNR ↑	25.212	29.359	29.561	28.944	28.634	28.335	28.124	25.308	**29.680**	27.350
	SSIM ↑	0.532	0.855	0.864	0.839	0.843	0.838	0.821	0.784	**0.871**	0.836
	LPIPS ↓	0.473	0.251	0.189	0.211	0.229	0.189	0.242	0.288	**0.135**	0.234
	DISTS ↓	0.141	0.151	0.111	0.107	0.112	0.095	0.129	0.141	**0.069**	0.111
SPMCS [46]	PSNR ↑	**26.405**	20.089	20.092	19.382	19.147	18.130	19.440	19.272	26.286	26.112
	SSIM ↑	0.389	0.540	0.493	0.492	0.484	0.446	0.494	0.507	**0.854**	0.849
	LPIPS ↓	0.353	0.365	0.229	0.234	0.196	0.304	0.289	0.345	**0.133**	0.144
	DISTS ↓	0.151	0.209	0.135	0.132	0.104	0.153	0.157	0.167	**0.066**	0.069
REDS30 [25]	PSNR ↑	25.566	25.527	26.146	24.474	**26.380**	20.918	24.898	22.879	26.022	26.036
	SSIM ↑	0.397	0.727	0.760	0.677	**0.783**	0.594	0.684	0.643	0.779	0.776
	LPIPS ↓	0.389	0.359	0.159	0.231	0.159	0.284	0.240	0.356	0.159	**0.139**
	DISTS ↓	0.115	0.161	0.080	0.105	0.077	0.132	0.113	0.138	0.074	**0.060**

Table 2. **Reference-based benchmarks across datasets.** Method columns appear in alphabetical order. ↑ higher is better, ↓ lower is better.

regress) because they enhance high-frequency detail without correcting warped geometry; in contrast, diffusion-prior-guided approaches tend to yield consistent positive Δ on FIQA. Our method achieves the best gains over FIQA up

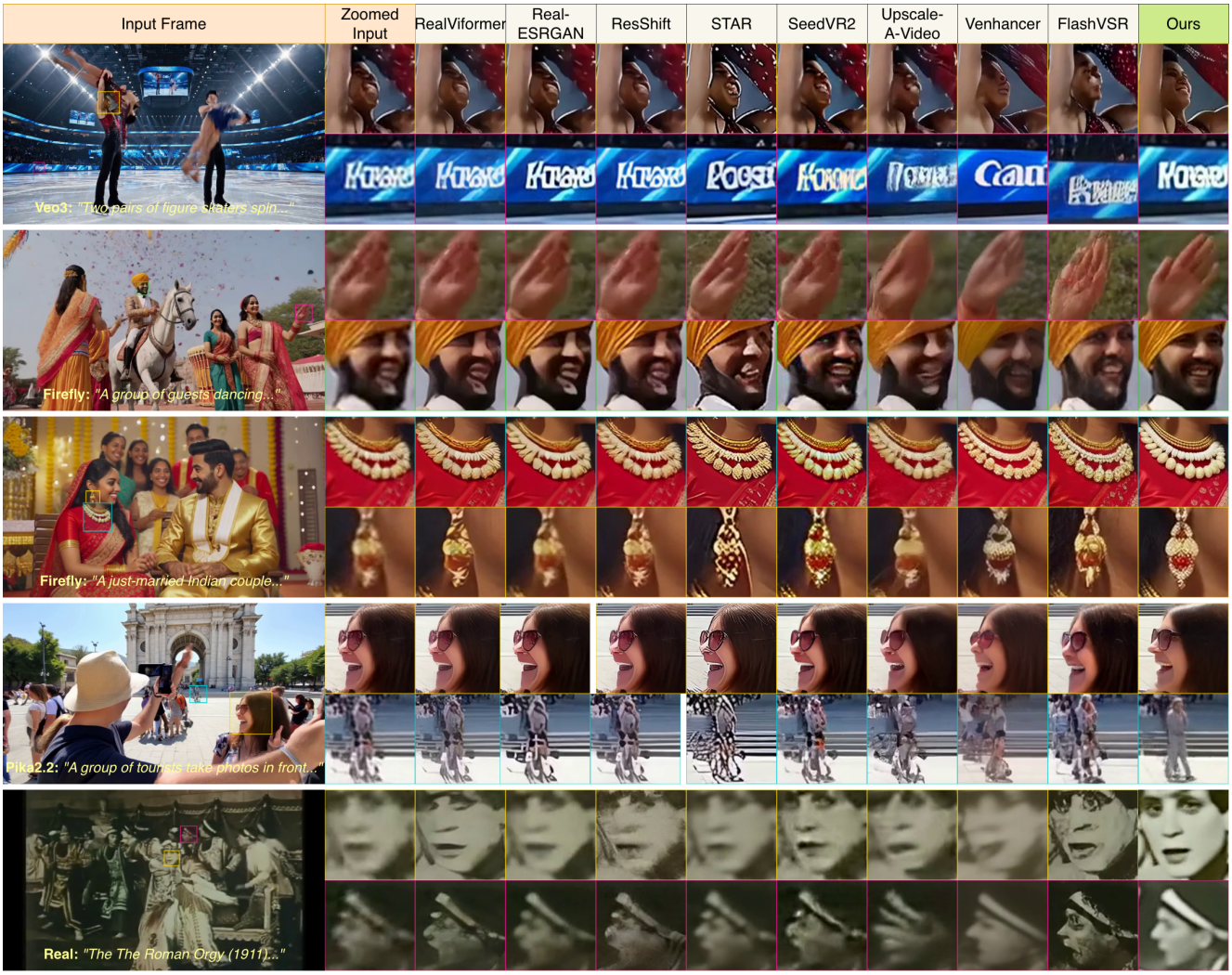

Figure 3. Qualitative Results. We comprehensively compare our method against a wide range of video restoration competitors.

to **+37%**, indicating strong structural correction.

Beyond FIQA, we also report an aesthetic score from a CLIP-LAION-initialized model which captures overall perceptual appeal of the crops and the objectness confidence of the YOLOv8 face detector as a proxy for semantic reliability. Our method attains the highest aesthetic gain and the largest increase in detection confidence, corroborating that structural corrections translate to perceptually cleaner and more reliably recognized faces.

Multi-Aspect Video Quality Assessment To complement structural metrics, we employ a GPT-based evaluator that scores each clip along six perceptual dimensions: visual quality, temporal consistency, face quality, motion realism, lighting and atmosphere, and detail preservation. Scores follow a 0–10 scale and are averaged to obtain an overall quality measure. This protocol captures perceptual factors that extend beyond distortion-oriented metrics. Table 1(b) summarizes the results across all baselines. Our method attains the highest mean score and shows consistent improvements across temporal coherence, visual quality, and detail preservation, indicating the effectiveness of diffusion priors in correcting structure while maintaining scene semantics. To verify reliability, four independent users evaluated 10 random test videos ($\approx$ 20% of the AIGC dataset), and GPT's rankings showed strong qualitative agreement with majority human preference

Perceptual video quality (VBench). To assess quality dimensions beyond face-centric metrics, we evaluate on VBench [15], which decomposes video quality into human-aligned dimensions. Table 1(c) reports six dimensions on AIGC54 set. Our method achieves the highest average score (0.8642), with notable gains in dynamic degree (0.9444) and imaging quality (0.7311), while remaining competitive on temporal consistency and subject preservation.

4.2.2 Traditional Video Restoration.

On standard video restoration benchmarks, CreativeVR attains PSNR/SSIM comparable to strong SR/VSR baselines and competitive LPIPS/DISTS scores (Table. 2). This shows that a model optimized for challenging AIGC and real-world artifacts also transfers well to classical degradations without task-specific tuning.

4.3. Qualitative Comparison

Prior methods fail to remove AIGC or real-world artifacts with blurred or fused facial profiles, smeared hands and jewelry, and distorted signage as shown in Fig. 3 . CreativeVR restores clean, geometrically plausible faces, fingers, and fine details while preserving pose and scene.

4.4. Ablation Study

Synthetic degradation strength during training. We ablate three augmentation levels: *Light*, *Medium*, and *Strong*. All variants share the same backbone and training recipe; only the corruption schedule (blur, warping, morphing, frame dropping, temporal downsampling) is varied. As shown in Table. 3, *Strong* yields the largest structure gains across all FIQA metrics on the AIGC54 benchmark, while *Light* offers only mild sharpening and *Medium* provides moderate improvements. GPT-based pairwise preferences (Table. 4) follow the same trend. Qualitatively (Fig. 4), *Strong* best removes geometric distortions without altering pose or identity.

Besides, Table. 2 indicates that on distortion-oriented SR/VR metrics (*e.g.*, PSNR, SSIM on SPMCS and REDS30), *Medium* slightly outperforms *Strong*, likely because heavy synthetic degradations induce prior-driven corrections that deviate from pixel-wise ground truth.

Inference control scale. At test time we expose a single control scale γ_ℓ that rescales all adapter gains, modulating how strongly the degraded video guides the frozen prior. As shown in Fig. 5, higher γ_ℓ values yield more precise, high-fidelity restoration close to the input, while lower γ_ℓ allow more prior-driven, highly creative re-synthesis.

This controllable behavior also supports different production workflows: lower γ_ℓ values favor more prior-driven re-synthesis, which can be useful for applications such as CG-to-real translation and slow-motion refinement, whereas higher γ_ℓ values preserve inputs more faithfully and are better suited for super-resolution tasks.

5. Conclusion

We introduced CreativeVR, a unified diffusion-prior framework that refines both AIGC and real-world videos while preserving structure, motion, and details. By combining plug-and-play adapter modules with a temporally coherent corruption curriculum, CreativeVR corrects severe structural and temporal artifacts that traditional VR/VSR or post-hoc refiners fail to address. A controllable precision knob enables a smooth trade-off between fidelity-oriented restoration and prior-guided structural correction, making the method broadly applicable across diverse degradation regimes. Experiments on AIGC-artifact benchmarks, real degraded footage, and standard VR datasets show that CreativeVR achieves state-of-the-art restoration quality with practical throughput and strong zero-shot scalability. As a lightweight enhancement layer for frozen T2V backbones, CreativeVR provides a practical path toward production-ready video refinement in modern creative pipelines.

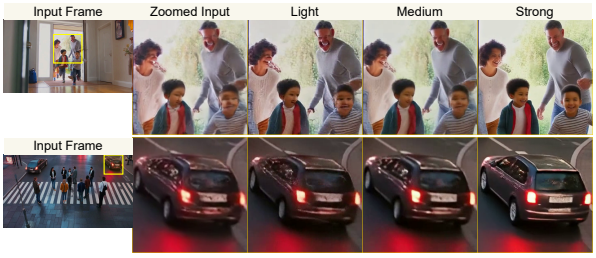

Figure 4. **Effect of degradation strength.** Stronger augmentations yield cleaner faces and sharper details.

Figure 5. **Inference Precision knob.** High precision preserves input details; low precision enables stronger corrective synthesis.

Quality Metric	Degradation Strength		
	Light	Medium	Strong
eDifFIQA	5.63	3.16	**34.34**
DifFIQA	0.07	0.39	**0.46**
CLIB-FIQA	2.15	4.04	**12.60**
CR-FIQA	6.01	7.03	**15.95**
FaceQAN	5.02	9.45	**26.14**

Table 3. **FIQA gains vs. training strength.** Values are relative improvements (%) over the input videos.

Scoring Aspect	Degradation Strength		
	Light	Medium	Strong
Visual Quality	7.69	7.89	**8.20**
Temporal Consistency	8.52	8.46	**8.77**
Face Quality	5.66	5.65	**5.99**
Motion Realism	8.18	8.05	**8.49**
Lighting & Atmosphere	8.60	8.56	**8.84**
Detail Preservation	7.61	7.74	**8.14**
Overall Mean	7.71	7.73	**8.07**

Table 4. Preference scores (1–10) vs. training strength

References

[1] Adobe. Adobe firefly. https://www.adobe.com/sensei/generative-ai/firefly.html, 2025. Accessed: 2025-11-13. 5

[2] Luma AI. Ray3: Intelligent video model with hdr and visual reasoning. https://lumalabs.ai/ray, 2025. Accessed: 2025-11-13. 1, 5

[3] Žiga Babnik, Peter Peer, and Vitomir Štruc. Faceqan: Face image quality assessment through adversarial noise exploration. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 748–754. IEEE, 2022. 5

[4] Žiga Babnik, Peter Peer, and Vitomir Štruc. Diffiqa: Face image quality assessment using denoising diffusion probabilistic models. In *2023 IEEE international joint conference on biometrics (IJCB)*, pages 1–10. IEEE, 2023. 5

[5] Žiga Babnik, Peter Peer, and Vitomir Štruc. ediffiqa: towards efficient face image quality assessment based on denoising diffusion probabilistic models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(4):458–474, 2024. 5

[6] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. Cr-fiqa: face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5836–5845, 2023. 5

[7] Kelvin C.K. Chan, Shiyu Zhou, Xintao Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proc. CVPR*, pages XXX–XXX, 2022. 1, 3

[8] Google DeepMind. Veo 3. https://deepmind.google/technologies/veo/, 2024. Accessed: 2025-01. 1, 5

[9] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 3

[10] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 1

[11] J. He, T. Xue, D. Liu, X. Lin, P. Gao, D. Lin, Y. Qiao, W. Ouyang, and Z. Liu. Venhancer: Generative space-time enhancement for video generation. OpenReview (ICLR 2025 submission), 2024. OpenReview / ICLR submission (see provided OpenReview link). 1, 2, 3, 5, 6

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Ben Morrone, Quentin De Laroussilhe, Antonio Gesmundo, Moe Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. *ICML Workshop / arXiv*, 2019. 3

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint, 2021. 3

[14] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 3

[15] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7

[16] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 3

[17] Zeyinzi Jiang, Chaojie Mao, Ziyuan Huang, Ao Ma, Yiliang Lv, Yujun Shen, Deli Zhao, and Jingren Zhou. Res-tuning: A flexible and efficient tuning paradigm via unbinding tuner from backbone. *Advances in Neural Information Processing Systems*, 36:42689–42716, 2023. 3

[18] Lingshun Kong, Dongqing Zou, Fu Lee Wang, Jimmy Ren, Xiaohe Wu, Jiangxin Dong, Jinshan Pan, et al. Deblurdiff: Real-word image deblurring with generative diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[19] Pika Labs. Pika 2.2. https://pika.art, 2025. Accessed: 2025-11-13. 5

[20] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement: a comprehensive survey. *International Journal of Computer Vision*, pages 1–31, 2025. 3

[21] Lightricks. Ltx-2: The next-generation multimodal ai video foundation model. https://ltx.video/, 2025. Accessed: 2025-11-14. 1

[22] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 3

[23] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 3

[24] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023. 3

[25] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6

[26] Open-Sora Plan Team. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 5

[27] OpenAI. Sora 2. https://openai.com/sora, 2025. Accessed: 2025-11-13. 1

[28] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. Clib-fiqa: Face image quality assessment with confidence

calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1694–1704, 2024. 5

[29] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. Mr-fiqa: Face image quality assessment with multi-reference representations from synthetic data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12915–12925, 2025. 5

[30] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 3

[31] Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. *Advances in Neural Information Processing Systems*, 37:36046–36070, 2024. 3

[32] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 4472–4480, 2017. 6

[33] Kuaishou Technology. Kling: A text-to-video generation model. https://klingai.com/, 2025. Accessed: 2025-11-14. 1

[34] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3360–3369, 2020. 3

[35] 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. 5

[36] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv preprint arXiv:2506.05301*, 2025. 2, 3, 5, 6

[37] J. Wang, Z. Lin, M. Wei, Y. Zhao, C. C. Loy, L. Jiang, and C. Yang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. arXiv preprint, 2025. 3

[38] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 3

[39] J. Wang, Z. Yue, S. Zhou, K. C. K. Chan, and C. C. Loy. Exploiting diffusion prior for real-world image super-resolution (stablesr). IJCV (preprint) / arXiv, 2024. 3

[40] Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutions. In *Proceedings of the CVPR Workshops (NTIRE) 2019*, 2019. 1, 3

[41] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 5, 6

[42] R. Xie, Y. Liu, P. Zhou, C. Zhao, J. Zhou, K. Zhang, Z. Zhang, J. Yang, Z. Yang, and Y. Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. arXiv preprint, 2025. 2, 3, 5, 6

[43] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6388–6397, 2021. 3

[44] Y. Xu, T. Park, R. Zhang, Y. Zhou, E. Shechtman, F. Liu, J.-B. Huang, and D. Liu. Videogigagan: Towards detail-rich video super-resolution. arXiv preprint, 2024. 1, 2

[45] Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2139–2149, 2025. 3

[46] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3106–3115, 2019. 6

[47] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22963–22974, 2025. 3, 5

[48] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36:13294–13307, 2023. 3, 5, 6

[49] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4791–4800, 2021. 3

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3

[51] Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yida Zhang, Yi Jiang, Zehuan Yuan, Binyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. *arXiv preprint arXiv:2502.05179*, 2025. 3

[52] Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *European Conference on Computer Vision*, pages 412–428. Springer, 2024. 5, 6

[53] Ziqing Zhang, Kai Liu, Zheng Chen, Xi Li, Yucong Chen, Bingnan Duan, Linghe Kong, and Yulun Zhang. Infvsr: Breaking length limits of generic video super-resolution. *arXiv preprint arXiv:2510.00948*, 2025. 3

[54] P. Zhou et al. Upscale-a-video: Text-guided latent diffusion for video upscaling. arXiv preprint, 2024. Text-guided upscaling / prompt-driven texture synthesis; replace with preferred paper details. 2, 5, 6

[55] Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025. 2, 5, 6