

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

Tejas Panambur^{1,*} Ishan Rajendrakumar Dave^{2,*} Chongjian Ge² Ersin Yumer² Xue Bai²

¹University of Massachusetts Amherst ²Adobe Inc.

tpanambur@umass.edu, {idave, cge, yumer, xubai}@adobe.com

<https://daveishan.github.io/creativevr-webpage/>

*Equal contribution †Work completed during an internship at Adobe Inc.

Overview

- Implementation details, including dataset descriptions and evaluation metrics, are provided in Section A.
- Additional quantitative results on further benchmarks are reported in Section B.
- Additional visualizations, qualitative comparisons, and diverse application examples are presented in Section C.

A. Datasets and Implementation Details

A.1. Datasets

For traditional video restoration, we *only evaluate* on three standard synthetic benchmarks REDS30 [4], SPMCS [10], and UDM10 [5] without any additional training or fine-tuning. Following prior work, we use their standard synthetic degradation pipelines and paired LR–HR sequences for evaluation, which cover diverse motion patterns and scene types and provide a controlled setting to measure both spatial fidelity and temporal consistency.

A.2. Metrics

For precision-oriented traditional video restoration with paired ground truth, we report four standard full-reference metrics: PSNR [3] and SSIM [8] to measure distortion fidelity, and LPIPS [12] and DISTS [1] to capture perceptual similarity. All scores are averaged over all frames and sequences in each dataset.

A.3. Synthetic Degradation pipeline

We define three temporal augmentation presets - *Light*, *Medium*, and *Strong*, that control the overall strength of temporal corruption. For each clip, we first apply a single spatial augmentation and then compose a small set of temporal operators in a fixed spatial-first ordering. On average, Light uses roughly two temporal operators per clip,

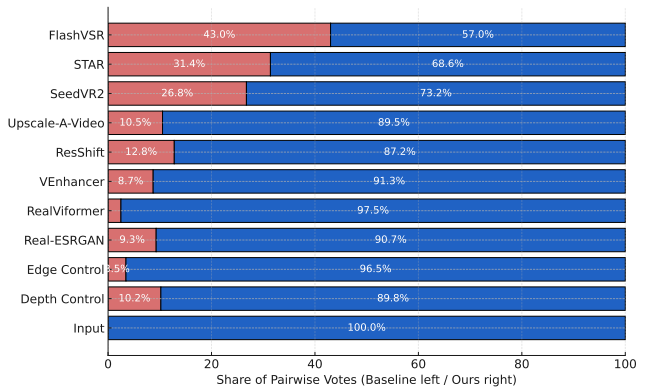


Figure 1. **Preference Study.** Bars show pairwise vote shares between *Ours* (right) and each baseline (left); higher right-side bars indicate stronger preference for our method.

Medium uses three, and Strong uses up to four, drawn from the pool described below.

Motion blur. To simulate camera and object motion, we apply a 2D motion blur with a random orientation in $[0^\circ, 360^\circ]$ and a kernel size sampled from a range on the order of 3–20 pixels. The blur strength is tied to the preset, with Light sampling from the lower end of this range and Strong emphasizing longer and more pronounced streaks.

Warped grid distortion. Geometric wobble and mild rolling-shutter artifacts are introduced via a grid-based spatial warp. We vary both the grid resolution (from coarse $\sim 4 \times 4$ to finer $\sim 12 \times 12$ control points) and the displacement amplitude (roughly 0.05 to 0.3 in normalized coordinates), with Strong using coarser grids and larger displacements than Light.

Temporal morphing. We synthesize nonlinear motion by interpolating between a small number of keyframes (two in our implementation) using a parametric morphing operator. A scalar strength parameter controls how far the interpolated motion deviates from the original trajectory; Light

uses mild perturbations, whereas Strong samples from the upper part of the range (up to around 0.5–0.6).

Stochastic frame dropping. Dropped or corrupted frames are simulated by randomly removing frames and reconstructing them via simple temporal interpolation. We control both the per-frame drop probability and the maximum number of dropped frames per clip, with Light using low drop rates (single-frame events) and Strong allowing multiple consecutive drops, leading to visibly irregular motion.

Temporal downsampling. Finally, we degrade temporal fidelity by subsampling the clip in time with a factor between roughly 1.5 and 3.5, followed by re-timing to the original length. Higher factors are reserved for the Strong preset and produce noticeable temporal aliasing and “jerky” motion, while Light uses only mild subsampling.

Selection strategy. Given a preset, we randomly sample a subset of the above operators and apply them sequentially, ensuring that Light, Medium, and Strong correspond to increasing numbers and strengths of temporal corruptions while keeping the underlying action recognizable for training.

B. Additional Quantitative Results

B.1. Pairwise Preference Evaluation (Arena)

We further employ a pairwise arena-style evaluation using a GPT judge. For each comparison, the evaluator is shown two restored clips from the same input sequence and asked to indicate a preference based on temporal consistency and frame quality. The aggregated votes across all clips provide a direct perceptual ranking. As shown in Fig. 1, our method receives the majority of votes against both classical and diffusion-based baselines, reflecting strong perceptual preference in head-to-head comparisons. We also conducted a human check on randomly sampled pairs, using 10 pairs per head-to-head comparison ($\approx 20\%$ of all comparisons), where four independent users confirmed that GPT’s pairwise preferences were consistent with majority human choices, although the arena contributed limited additional insights beyond the multi-aspect scores. We also ran a small human check on 20% of the arena pairs, confirming that GPT’s preferences aligned with majority human choices, although the arena added limited additional insights beyond the multi-aspect scores.

B.2. Multi Aspect Scoring Face Quality

Beyond FaceIQA, which already shows that our method significantly improves facial quality on AIGC54, we further run an object-centric evaluation in which Qwen2.5-VL-72B acts as a geometry-aware judge on cropped face regions. For each crop, Qwen scores five aspects following our prompt rubric: *Geometry & Silhouette* (GE), *Edge Definition & Continuity* (ED), *Detail Realism* (DT), *Arti-*

fact Level (AR; reverse scale), and *Tonal Robustness* (TN). The model outputs integer scores in $[0, 10]$, and we average them across frames and dimensions to obtain an aggregate object-centric quality score. The full prompt and scoring rubric are provided in the rubric shown in Figure 2.

As summarized in Table 1, our method achieves the highest aggregate score (7.84), providing an independent confirmation of the strong facial quality observed with FaceIQA. Compared to prior video restoration baselines, we obtain the best scores on *Geometry & Silhouette*, *Edge Definition & Continuity*, *Artifact Level*, and *Tonal Robustness*, indicating more stable facial shapes, cleaner boundaries, and fewer structural artifacts while preserving photometric fidelity. Although FlashVSR attains slightly higher *Detail Realism*, our method remains competitive and does so without introducing spurious high-frequency artifacts, as reflected by our superior geometry and artifact scores.

B.3. Robustness to input degradations.

To assess whether the refiner can also operate as a precise video restoration model, we apply four levels of synthetic input corruption to the REDS30 and YouHQ40 benchmarks: purely spatial $4\times$ downscaling, joint spatial $4\times +$ temporal $2\times$ downsampling, and two spatio-temporal degradation settings Light and Strong defined in Section A.3. We then reconstruct the clean videos and evaluate using precision-oriented metrics (e.g., PSNR, SSIM, LPIPS, DISTs, NIQE, MUSIQ, CLIP-IQA, and DOVER), with results reported in Tables 2 and 3. Across both datasets, our Medium setting already matches or surpasses the best prior methods on most full-reference metrics for the spatial and mild spatio-temporal corruptions, while the Strong setting remains competitive and often leads under the most severe ST-Strong case. This consistent behavior across four corruption types indicates that the same refiner can robustly undo both synthetic spatial degradations and harder spatio-temporal artifacts without sacrificing perceptual quality.

C. Additional Visualizations

C.1. Additional qualitative comparisons

We provide further qualitative examples from the AIGC benchmark in Figures 3 and 4, showing restorations from all competing methods. Across diverse prompts and source generators, our approach visibly outperforms prior methods like FlashVSR, Real-Viformer, ResShift, SeedVR2, STAR, Upscale-A-Video, and VEnhancer: it recovers clearer object and face structures, suppresses characteristic artifacts, and produces more natural overall aesthetics, yielding videos that better preserve the intended content.


```

1 # Qwen2.5-VL prompt template for object-centric quality
2 JUDGE_PROMPT = """You are a meticulous computer-vision rater for a CVPR paper.
3 Judge only the cropped object/person region. Be strict about geometric
4 plausibility, boundary stability, and structural artifacts (wobble, double
5 edges, ghosting, halos, smeared surfaces, plastic textures, ringing). If an
6 'INPUT' image or caption is provided, **ignore it** completely -- judge each
7 candidate independently. Return ONLY valid JSON. No extra text. If a criterion
8 is not visible, set it to null and add a brief tag in "notes".
9
10 TASK
11 Score each candidate crop A...K using 0-10 integers (0 = very poor,
12 10 = excellent). Use labels exactly as given (A, B, C, ...).
13
14 Rubric (0-10, integers):
15 - GE (Geometry & Silhouette): Judge geometry in the object's semantic frame.
16   Check canonical parts/primitives for the category (e.g., limb/joint
17   alignment, straight planar borders, right-angle corners, circular rims/wheels,
18   text baselines, symmetry where applicable). Penalize distorted contours,
19   wobble/jitter, double edges, part warping, melted forms, bent straight lines,
20   ovalized circles, or misshapen letters/logos.
21 - ED (Edge Definition & Continuity): Assess edges along semantic boundaries
22   (main silhouette, structural lines, text/glyph strokes, circular/straight
23   borders). Edges must trace the true shape with continuous, coherent paths.
24   Do not reward edge contrast by itself. Penalize fraying, ghost/double edges,
25   bright/dark halos/ringing, staircase/zippering, or overshoot that shifts the
26   perceived boundary.
27 - DT (Detail Realism): Reward fine detail only when it follows the
28   form/material (fabric folds conforming to shape, hair/filament strands,
29   micro-structure, legible glyph strokes). Penalize plasticiness,
30   over-smoothing, and especially invented crunchy micro-patterns or grain that
31   do not align with the material or introduce false pores/texture.
32 - AR (Artifact Level; reverse scale): Structural artifacts such as
33   halos/ringing/overshoot, checkerboard/tiling, zippering, deconvolution
34   ripples, temporal smears, motion trails, compression-like textures,
35   patch seams, moire. (10 = no artifacts, 0 = severe artifacts)
36 - TN (Tonal Robustness): Consider only major tonal failures -- blown highlights,
37   crushed shadows, or hue shifts that harm form perception. Ignore
38   vibrance/saturation; do not reward colorfulness.
39
40 Scoring rules:
41 - Integers only (0-10). When uncertain, choose the lower score.
42 - Prioritize shape correctness over sharpness. If you observe
43   halo/overshoot/double contours or invented high-frequency texture, give low
44   ED and AR (0-2) and do not boost DT for "crispness".
45 - Focus on structure and boundary integrity; ignore style/brightness preferences
46   or artistic color.
47 - Heavily penalize geometry wobble, melted forms, halos/ghosting, and
48   broken/duplicated edges.
49 - If a criterion cannot be judged, set it to null and add a tag in "notes"
50   (e.g., "NOT_VISIBLE", "OCCLUDED").
51
52 OUTPUT JSON SCHEMA (strict):
53 {
54   "per_image": {
55     "A": {"GE": int|null, "ED": int|null, "DT": int|null,
56         "AR": int|null, "TN": int|null, "notes": ["TAG1", "TAG2"]},
57     "B": {...}
58   },
59   "overall_ranking": ["C", "A", "B", ...],
60   "confidence": "high" | "medium" | "low"
61 }
62
63 IMAGES:
64 {image_descriptions}
65 """

```

Figure 2. Qwen2.5-VL prompt template for multi-aspect face-quality (rubric) scoring.

Restoration Method	FlashVSR [15]	Real-ESRGAN [7]	Real-Viformer [13]	ResShift [11]	SeedVR2 [6]	STAR [9]	Upscale-A-Video [14]	VENhancer [2]	Ours
Geometry & Silhouette	7.64	5.98	4.50	2.46	4.28	5.20	3.82	6.48	8.04
Edge Definition & Continuity	8.10	6.20	5.08	2.84	4.50	5.42	4.08	6.64	8.24
Detail Realism	8.22	5.54	4.14	2.34	3.90	4.80	3.58	6.02	7.46
Artifact Level (reverse)	5.99	5.64	4.18	2.34	4.00	4.92	3.64	6.18	7.64
Tonal Robustness	6.62	5.90	4.26	2.40	4.10	5.00	3.68	6.26	7.80
Aggregate Score	7.31	5.85	4.43	2.48	4.16	5.07	3.76	6.32	7.84

Table 1. **Multi-Aspect Scoring for Face Quality judged by Qwen2.5-VL-72B**. Scores (0–10) are given for geometry & silhouette, edge definition & continuity, detail realism, artifact level (reverse), and tonal robustness. Our method achieves the best aggregate score and leads on four out of five dimensions, corroborating the strong facial quality observed with FaceIQa.

Corruption Type	Metric	FlashVSR [15]	Real-ESRGAN [7]	Real-Viformer [13]	ResShift [11]	SeedVR2 [6]	STAR [9]	Upscale-A-Video [14]	VENhancer [2]	Ours (Medium)	Ours (Strong)
Spatial Downsampling	PSNR ↑	25.87	25.95	26.03	25.73	22.59	24.33	24.03	15.46	27.12	26.02
	SSIM ↑	0.73	0.72	0.72	0.73	0.65	0.70	0.64	0.44	0.79	0.76
	LPIPS ↓	0.38	0.40	0.43	0.18	0.36	0.21	0.28	0.38	0.13	0.16
	DISTS ↓	0.15	0.18	0.17	0.08	0.13	0.08	0.12	0.10	0.06	0.06
	NIQE ↓	2.91	6.63	6.51	3.09	5.94	3.39	2.49	4.20	3.00	3.17
	MUSIQ ↑	66.40	36.48	67.39	27.11	31.62	64.45	66.79	48.33	62.81	62.98
	CLIP-IQA ↑	0.53	0.37	0.54	0.33	0.35	0.48	0.50	0.40	0.46	0.44
	DOVER ↑	31.97	21.85	33.67	14.99	17.80	35.83	24.41	24.37	33.28	32.39
Spatio-Temporal Downsampling	PSNR ↑	15.48	25.13	25.86	24.57	25.43	24.07	23.41	22.40	25.47	25.51
	SSIM ↑	0.44	0.69	0.71	0.69	0.72	0.69	0.62	0.63	0.74	0.74
	LPIPS ↓	0.39	0.41	0.43	0.21	0.38	0.23	0.30	0.39	0.16	0.18
	DISTS ↓	0.10	0.19	0.17	0.09	0.14	0.09	0.13	0.15	0.07	0.07
	NIQE ↓	2.95	6.58	6.52	3.09	5.83	3.43	2.53	4.34	2.97	3.21
	MUSIQ ↑	66.34	34.02	26.60	66.13	33.01	60.65	66.57	43.92	62.56	61.71
	CLIP-IQA ↑	0.53	0.37	0.54	0.34	0.36	0.46	0.50	0.39	0.46	0.44
	DOVER ↑	32.36	22.09	36.98	14.94	18.39	33.64	24.10	23.19	37.02	31.56
Spatio-Temporal Light	PSNR ↑	20.73	20.84	20.65	20.24	20.59	20.56	19.91	19.92	20.73	21.18
	SSIM ↑	0.58	0.59	0.58	0.56	0.59	0.59	0.54	0.56	0.58	0.59
	LPIPS ↓	0.64	0.59	0.37	0.46	0.48	0.54	0.47	0.66	0.43	0.51
	DISTS ↓	0.34	0.32	0.20	0.25	0.26	0.29	0.25	0.36	0.13	0.24
	NIQE ↓	8.23	8.13	4.03	4.44	5.96	6.85	3.88	8.61	3.20	6.62
	MUSIQ ↑	15.64	22.04	46.37	44.34	23.97	21.88	51.76	21.47	63.89	21.34
	CLIP-IQA ↑	0.24	0.31	0.40	0.49	0.28	0.29	0.43	0.30	0.51	0.30
	DOVER ↑	5.86	11.91	22.05	20.30	12.96	13.61	17.54	9.75	27.99	10.53
Spatio-Temporal Strong	PSNR ↑	20.59	15.63	20.48	20.12	20.41	20.42	19.73	19.75	20.69	20.95
	SSIM ↑	0.58	0.43	0.58	0.56	0.58	0.58	0.53	0.56	0.58	0.59
	LPIPS ↓	0.64	0.59	0.38	0.51	0.49	0.55	0.48	0.66	0.47	0.43
	DISTS ↓	0.34	0.32	0.20	0.26	0.26	0.29	0.25	0.36	0.24	0.13
	NIQE ↓	8.28	8.12	3.89	4.47	5.92	6.84	3.74	8.59	6.56	3.15
	MUSIQ ↑	15.75	21.74	47.60	43.47	23.94	21.83	53.17	21.50	21.95	63.19
	CLIP-IQA ↑	0.24	0.31	0.41	0.49	0.28	0.30	0.43	0.29	0.30	0.51
	DOVER ↑	5.48	11.11	22.50	19.51	12.83	13.54	17.76	9.25	10.48	28.39

Table 2. **Robustness under synthetic corruptions on REDS30 [4]**. Higher is better for PSNR, SSIM, MUSIQ, CLIP-IQA, and DOVER; lower is better for LPIPS, DISTS, and NIQE.

C.2. Ablation with Precision Knob during Sampling

At test time we expose a per-layer control scale γ_ℓ that rescales all adapter gains, modulating how strongly the

Augmentation Type	Metric	FlashVSR [15]	Real-ESRGAN [7]	Real-Viformer [13]	ResShift [11]	SeedVR2 [6]	STAR [9]	Upscale-A-Video [14]	VENhancer [2]	Ours (Medium)	Ours (Strong)
Spatial Downsampling	PSNR \uparrow	25.6	<u>26.4</u>	25.62	25.98	24.2	16.06	24.45	22.18	27.2	26.31
	SSIM \uparrow	0.722	0.715	<u>0.727</u>	0.713	0.702	0.417	0.633	0.63	0.767	0.726
	LPIPS \downarrow	<u>0.193</u>	0.367	0.31	0.203	0.378	0.274	0.285	0.402	0.182	0.354
	DISTS \downarrow	0.09	0.189	0.137	<u>0.1</u>	0.131	0.115	0.133	0.169	0.109	0.154
	NIQE \downarrow	<u>3.29</u>	6.99	6.18	3.6	3.84	4.91	3.04	5.58	3.76	6.89
	MUSIQ \uparrow	<u>68.73</u>	43.28	42.75	70.42	64.7	56	67.92	41.62	66.54	35.15
	CLIP-IQA \uparrow	<u>0.568</u>	0.396	0.412	0.609	0.513	0.425	0.506	0.406	0.512	0.366
	DOVER \uparrow	51.12	38.64	36.12	<u>53.39</u>	53.03	50.24	43.22	40.92	53.92	32.17
Spatio-Temporal Downsampling	PSNR \uparrow	16.19	25.23	25.51	24.67	24.77	23.59	23.57	21.85	25.84	<u>24.64</u>
	SSIM \uparrow	0.418	0.689	0.717	0.709	0.7	0.684	0.612	0.619	<u>0.715</u>	0.674
	LPIPS \downarrow	0.387	0.385	0.215	0.304	<u>0.209</u>	0.293	0.307	0.43	0.359	0.236
	DISTS \downarrow	0.136	0.194	0.121	0.133	0.095	0.126	0.143	0.191	0.156	0.114
	NIQE \downarrow	3.34	6.98	3.75	5.94	3.91	4.96	3.04	5.77	6.89	<u>3.68</u>
	MUSIQ \uparrow	<u>69.01</u>	41.06	64.51	45.77	63.28	52.74	67.54	38.15	34.97	69.07
	CLIP-IQA \uparrow	0.572	0.394	0.507	0.421	0.507	0.413	0.51	0.401	0.369	0.605
	DOVER \uparrow	52.05	38.53	<u>53.65</u>	37.26	53.1	47.68	42.99	40.13	32.44	53.26
Spatio-Temporal Light	PSNR \uparrow	21.56	21.9	21.65	21.21	21.47	21.17	21.27	16.49	20.43	<u>21.9</u>
	SSIM \uparrow	0.595	0.602	0.602	0.573	0.597	0.6	0.572	0.426	0.576	0.608
	LPIPS \downarrow	0.547	0.53	0.388	0.445	0.437	0.543	0.474	0.602	<u>0.407</u>	0.432
	DISTS \downarrow	0.298	0.302	0.217	0.26	0.242	0.305	0.27	0.345	0.149	<u>0.214</u>
	NIQE \downarrow	8.36	8.4	4.76	4.95	6.31	7.69	5.41	8.74	3.62	<u>6.47</u>
	MUSIQ \uparrow	17.32	24.42	45.66	<u>47.71</u>	27.53	24.87	42.74	23.47	67.23	30.24
	CLIP-IQA \uparrow	0.278	0.349	0.458	0.569	0.32	0.31	0.404	0.336	<u>0.55</u>	0.37
	DOVER \uparrow	22.27	30.22	43.64	43.63	36.11	33.97	31.95	29.67	49.22	<u>30.02</u>
Spatio-Temporal Strong	PSNR \uparrow	21.69	22.04	21.75	21.37	16.46	21.31	21.29	20.55	21.73	<u>21.92</u>
	SSIM \uparrow	0.606	<u>0.608</u>	0.607	0.581	0.431	0.606	0.571	0.581	0.601	0.611
	LPIPS \downarrow	0.426	0.521	0.371	0.431	0.54	0.527	0.446	0.587	<u>0.401</u>	0.421
	DISTS \downarrow	0.238	0.298	<u>0.208</u>	0.25	0.293	0.294	0.248	0.33	0.147	0.209
	NIQE \downarrow	6.22	8.35	<u>4.61</u>	4.85	8.32	7.48	4.81	8.48	3.55	6.27
	MUSIQ \uparrow	28.88	24.87	48.3	<u>48.39</u>	23.94	26	47.25	23.65	66.84	33.77
	CLIP-IQA \uparrow	0.331	0.35	0.469	0.569	0.282	0.318	0.431	0.33	<u>0.546</u>	0.386
	DOVER \uparrow	32.94	29.03	40.45	<u>40.94</u>	32.25	20.59	27.11	28.47	47.35	27.93

Table 3. **Robustness under learned augmentation strengths on YouHQ40 [14]**. Higher is better for PSNR, SSIM, MUSIQ, CLIP-IQA, and DOVER; lower is better for LPIPS, DISTS, and NIQE.

degraded video steers the frozen prior. Larger γ_ℓ values bias the model toward precise, high-fidelity restoration that closely follows the input, while smaller γ_ℓ increase the influence of the prior and enable more aggressive corrective synthesis. Figure 5 visualizes this trade-off across different settings of γ_ℓ for different input samples.

C.3. Diverse Applications of CreativeVR

CG-to-real translation. At lower values of the inference scale γ_ℓ , our CreativeVR refiner can also act as a CG-to-real translator, converting stylized computer-generated footage into more realistic renderings, as illustrated in Figure 6. The identity of the CG character and the overall scene layout are preserved, while textures, lighting, and shading become more photorealistic. Higher-resolution visualizations are provided in the attached video examples.

Slow-motion generation. Beyond correcting artifacts, CreativeVR can also be used to synthesize slow-motion, high-frame-rate versions of existing videos. We first up-sample the input sequence in time using simple linear interpolation, which introduces morphing and blur artifacts in the intermediate frames (top row of Figure 7). Passing this interpolated clip through our refiner yields temporally smooth, detail-preserving frames (bottom row), effectively turning the original footage into a visually coherent slow-motion video. This suggests that CreativeVR can serve as a general frame-rate enhancement module for arbitrary input videos. Videos are attached in the supplementary.

Our method also enhances temporal stability by correcting individual frames using the overall temporal context; corresponding examples are provided in the attached videos under *Applications*.

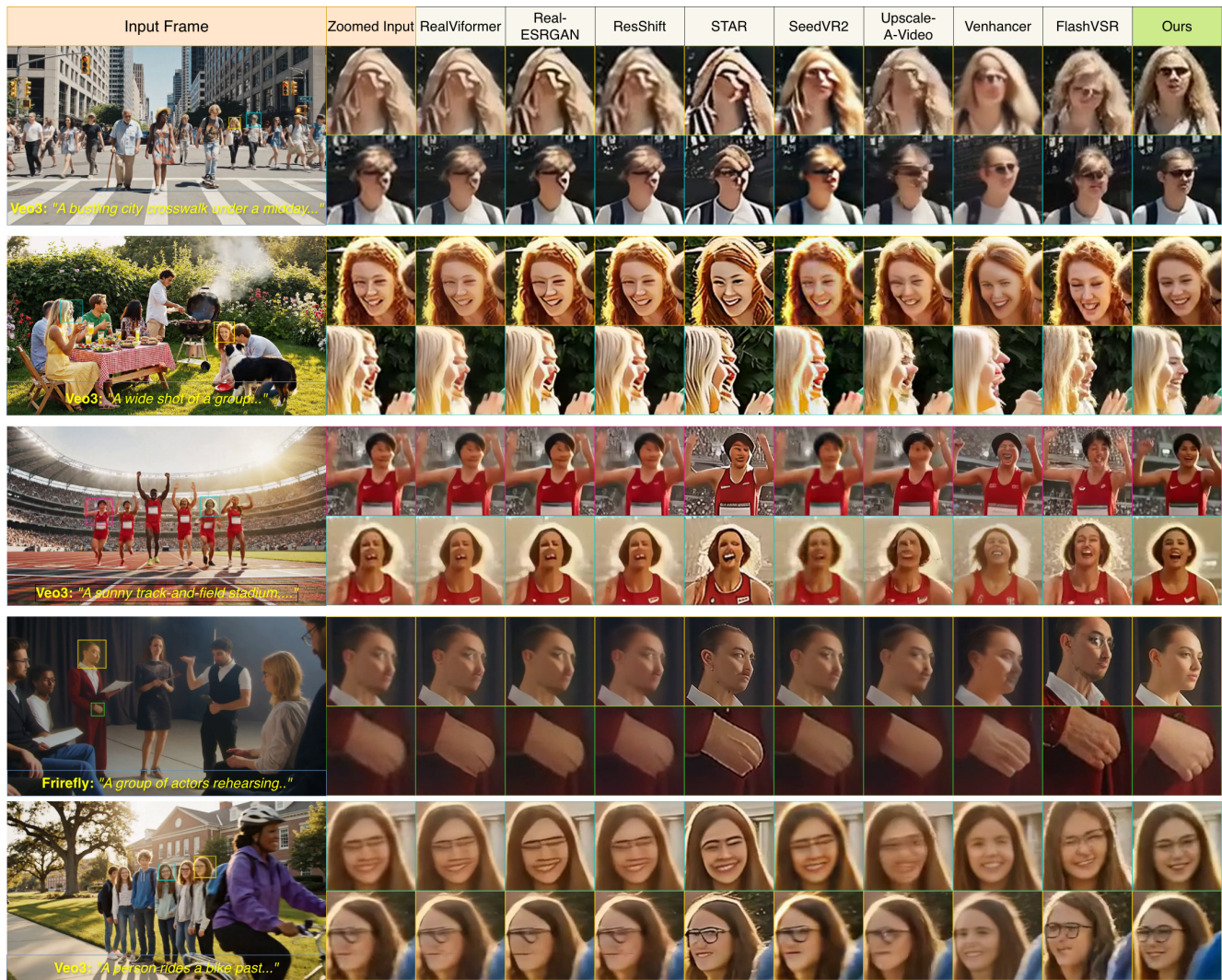


Figure 3. Qualitative Results. We comprehensively compare our method against a wide range of video restoration competitors.

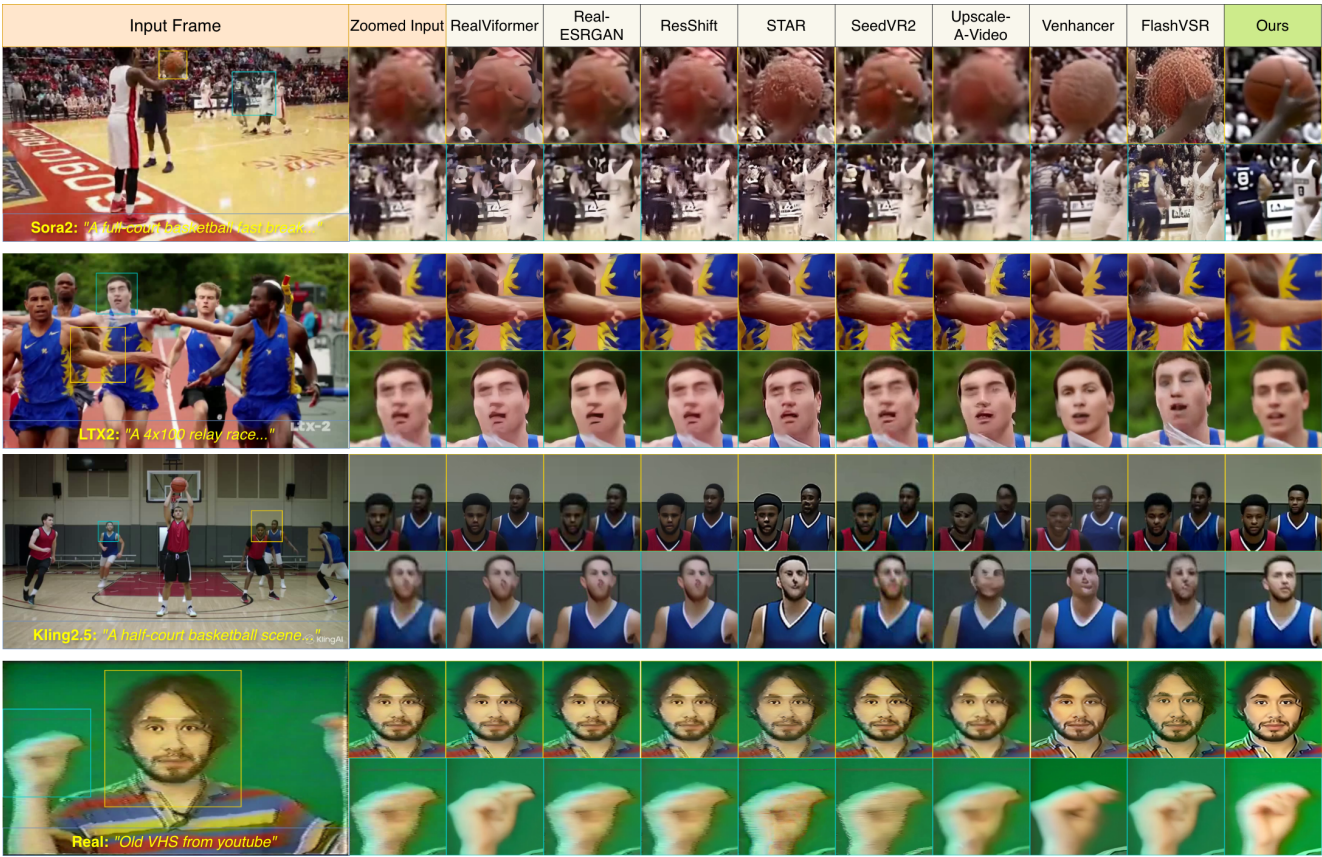


Figure 4. Qualitative Results. We comprehensively compare our method against a wide range of video restoration competitors.



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

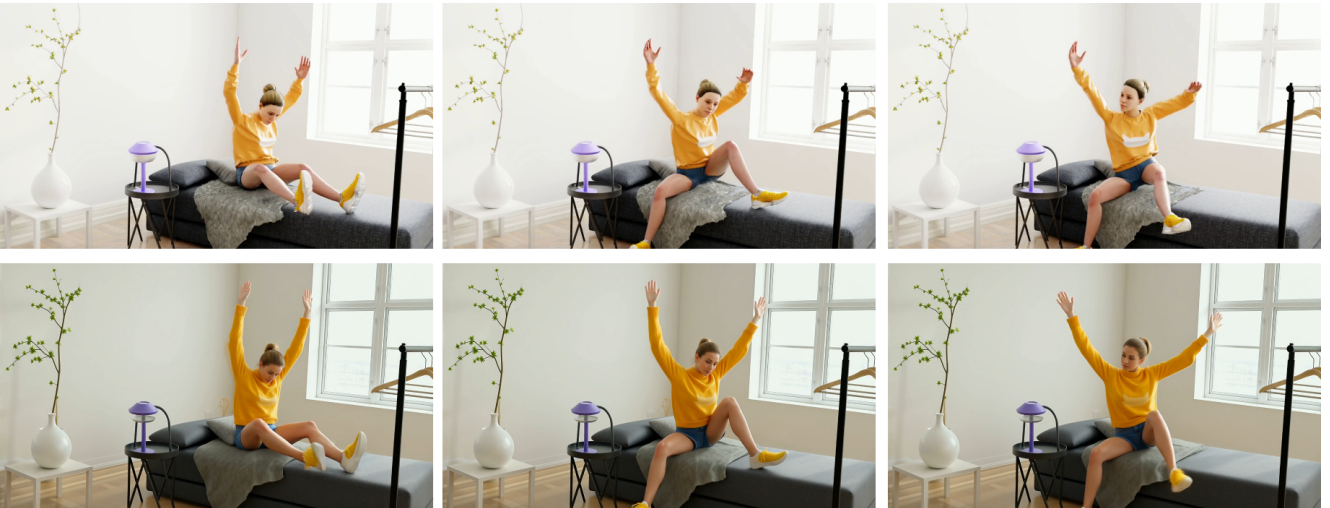


Figure 6. **CG-to-real translation with CreativeVR.** **Top:** input CG video frames. **Bottom:** outputs from our refiner at a low γ_e , which enhance realism while preserving the original character and scene composition.



Figure 7. **Slow-motion generation with CreativeVR.** **Top:** linearly interpolated frames used to upsample the input video in time, which exhibit noticeable morphing and blur artifacts. **Bottom:** refined output from CreativeVR, showing sharper details and more temporally consistent motion, resulting in a high-frame-rate slow-motion sequence.

References

- [1] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2022. [1](#)
- [2] 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). [4](#), [5](#)
- [3] Artur Hore and Dimitrios Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, pages 2366–2369, 2010. [1](#)
- [4] 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. [1](#), [4](#)
- [5] 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. [1](#)
- [6] 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. [4](#), [5](#)
- [7] 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. [4](#), [5](#)
- [8] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TPAMI*, 26(4):600–612, 2004. [1](#)
- [9] 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. [4](#), [5](#)
- [10] 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. [1](#)
- [11] 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. [4](#), [5](#)
- [12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. [1](#)
- [13] 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. [4](#), [5](#)
- [14] 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. [4](#), [5](#)
- [15] 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. [4](#), [5](#)