PANOERASE: SPHERICAL GEOMETRY-AWARE PANORAMIC VIDEO OBJECT REMOVAL

Anonymous authors
Paper under double-blind review

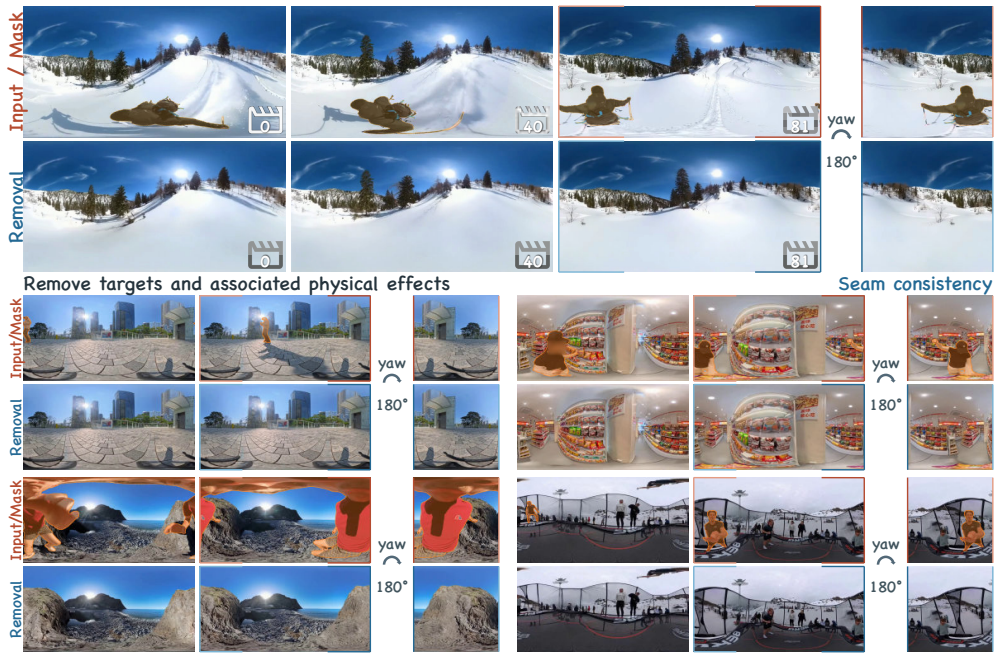

Figure 1: **Showcase of PanoErase.** Given a panoramic video and a target mask, PanoErase enables the removal of the target and its associated effects. The rightmost panel uses a 180° yaw shift to display the original ERP seam at the center.

ABSTRACT

Panoramic video object removal aims to remove targets and associated effects while reconstructing occluded content across the viewing sphere. However, existing perspective-based methods struggle with panoramic videos, where equirectangular projection (ERP) introduces seam discontinuities and polar distortions that challenge both spherical reconstruction and target-effect association, while paired panoramic removal data remain scarce. To address these challenges, we construct PANOVOR, a paired panoramic video removal dataset containing 1.53M frame pairs spanning 26.6 hours, with corresponding target masks. Building on PanoVOR, we propose PANOERASE, a unified model that incorporates spherical geometry into panoramic object and effect removal. PanoErase first learns panorama-aware completion from large-scale panoramic videos using *SC-RoPE*, which encodes longitude periodicity and longitude-independent polar limits. It is then refined with paired removal supervision through *SphericalControl*, which combines target-specific local evidence with panorama-wide context for effect-aware removal while preserving unrelated content. On PanoVOR-Eval, PanoErase improves target-region perceptual quality, effect removal, seam continuity, and polar reconstruction over the compared methods. On PanoVOR-Wild, it achieves the highest quality scores from both evaluators on sampled frames. The dataset, code and models will be released publicly. The anonymous project website is available at `https://panoerase-anonymous.pages.dev/`.

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.

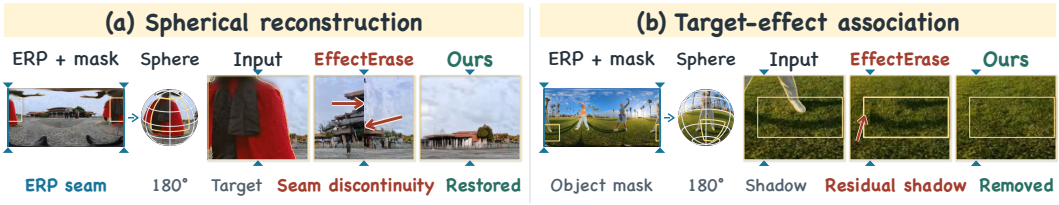

Figure 2: **Challenges in panoramic video object removal.** (a) Spherical reconstruction: EffectErase exhibits a seam discontinuity. (b) Target-effect association: the target's cast shadow remains after removal. PanoErase improves seam continuity and removes the associated shadow.

1 INTRODUCTION

Video object removal enables creators to eliminate unwanted content while preserving coherent and visually plausible scenes, benefiting applications such as content creation, post-production, and privacy protection. However, existing methods predominantly focus on perspective videos, leaving panoramic video object removal largely underexplored despite the growing adoption of panoramic cameras for immersive content.

An intuitive solution is to transfer perspective-based background reconstruction and target-effect removal to panoramic videos. However, the unique properties of equirectangular projection (ERP) pose challenges to both components. On one hand, spherical neighbors can be separated by the ERP seam or distorted near the poles, challenging contextual modeling for background reconstruction. On the other hand, full-sphere capture exposes target-associated effects across wider regions and varying distortions, complicating their association with the target. As shown in Figure 2, directly applying the perspective removal method EffectErase (Fu et al., 2026) to panoramas breaks the reconstructed scene at the ERP seam and leaves the cast shadow of the removed target.

To address these challenges, we begin by constructing PANOVOR, a large-scale paired dataset specifically designed for panoramic video object removal. Capturing aligned panoramic pairs in the real world is challenging because camera motion and unrelated scene dynamics are difficult to reproduce across the full viewing sphere. We therefore construct PanoVOR through paired rendering, re-simulating each camera trajectory in target-present and target-absent conditions while explicitly covering seam-crossing and polar configurations. After that, PanoVOR contains *1.53M* frame pairs spanning *26.6 hours*, with aligned source-clean videos and target masks covering camera motion, object motion, dynamic backgrounds, and challenging seam-crossing and polar configurations. Each scene is rendered with and without the target, allowing target-associated effects such as shadows, reflections, lighting interactions, deformation, and fluid effects to be naturally reflected in the paired supervision. To the best of our knowledge, PanoVOR is the only paired dataset dedicated to panoramic video object removal, providing systematic supervision and evaluation for full-sphere object-and-effect removal. For evaluation, we further derive PANOVOR-EVAL with clean references and PANOVOR-WILD with real in-the-wild videos.

Building upon the open-source Wan2.1-VACE model (Jiang et al., 2025), we propose PANOERASE, a unified framework that progressively addresses *spherical reconstruction* and *target–effect association* by two training stages. First, *Spherical Coupled RoPE (SC-RoPE)* models spherical positional relationships by preserving seam periodicity and polar consistency for panorama-aware reconstruction. Specifically, it couples rotary channels through spherical geometry, maintaining longitude continuity across the ERP seam while reducing longitude dependence toward the poles. Then, *SphericalControl* associates target-specific evidence with panorama-wide semantic and geometric context, enabling effect removal beyond the object mask. It combines mask-centered tangent views with full-ERP appearance and geometry, and employs scene-to-target attention to relate each panoramic location to the selected target. Together, the two stages progressively learn spherical reconstruction and target-conditioned removal within a unified generator.

On PANOVOR-EVAL, PANOERASE reduces target-region Tangent-LPIPS by 19% and Effect MAE by 13% relative to the strongest respective baselines, lowers the seam log-ratio from 1.9082 to 0.4801, and raises polar WS-PSNR by 1.06 dB. On PANOVOR-WILD, it achieves the highest Qwen and Gemini quality scores on matched sampled frames and is the only method whose seam response falls below that of the input. Our contributions are summarized as follows:

- We construct PANOVOR, a paired panoramic video removal dataset containing 1.53M frame pairs spanning 26.6 hours, with aligned source-clean pairs and target masks covering challenging seam-crossing and polar configurations.
- We identify two core challenges in panoramic removal including spherical reconstruction and target–effect association. We propose PANOERASE, a unified framework that addresses them through *SC-RoPE* for seam-periodic and polar-consistent positional modeling, and *Spherical-Control* for panorama-wide target–effect association beyond the object mask.
- We establish PanoVOR-Eval and PanoVOR-Wild for controlled and real-world evaluation, and conduct extensive experiments on reconstruction quality, effect removal, unrelated-content preservation, and seam continuity. PanoErase achieves the best target-region perceptual quality, effect removal, and seam continuity among the compared video object-removal methods.

2 RELATED WORK

Panoramic generation and editing. Panoramic videos are commonly stored in equirectangular projection (ERP), whose left and right borders are adjacent on the sphere and whose rows stretch toward the poles. PanFusion (Zhang et al., 2024) connects panoramic and perspective diffusion branches through projection-aware attention, and 360DVD (Wang et al., 2024) adds a panoramic adapter to a video model. PanoWan (Xia et al., 2025) combines latitude-aware sampling, rotated denoising, and padded decoding, while PanoWorld-X (Yin et al., 2025) reprojects ERP features onto the sphere. SE360 (Zhong et al., 2025) and World-Shaper (Liang et al., 2026) extend such priors to panoramic image editing. Closer to our task, earlier work removes objects from 360° videos (MacQuarrie & Steed, 2015), and DAOVI (Seshimo & Isogawa, 2025) uses geodesic motion and depth-aware propagation for omnidirectional inpainting; neither addresses the effects a removed object leaves in the scene. These methods handle ERP geometry through sampling, reprojection, or padding; *SC-RoPE* instead encodes it in the rotary positional encoding, coupling rotary pairs so that each block is exactly periodic in longitude and becomes longitude-independent at the poles.

Video inpainting and object removal. E2FGVI (Li et al., 2022) reconstructs masked regions through cross-frame propagation, while DiffuEraser (Li et al., 2025), VideoPainter (Bian et al., 2025), and VACE (Jiang et al., 2025) exploit generative video priors. Object removal further requires removing what the target changes beyond its mask, such as shadows and reflections. Omnimatte methods decompose videos into object-associated layers (Lu et al., 2021; Lin et al., 2023; Samuel et al., 2025); ROSE (Miao et al., 2025) predicts affected regions, EffectErase (Fu et al., 2026) learns object–effect correlations from target features through cross-attention, and SVOR (Hu et al., 2026) improves robustness under imperfect conditions through a reconstruction-to-removal curriculum. These effect-aware methods rely on paired supervision: RORD (Sagong et al., 2022) and Video4Removal (Wei et al., 2025) capture pairs from real videos, ROSE renders them, and EffectErase's VOR combines both sources. All of these pairs are perspective, so none contains a target or effect split by the ERP seam or stretched near a pole. PanoVOR supplies paired supervision in exactly these configurations, and *SphericalControl* provides target evidence from both ERP and mask-centered tangent views, to associate distant effects with the selected target.

3 METHOD

PanoErase follows SVOR's reconstruction-to-removal protocol (Hu et al., 2026) (Fig. 3). Stage I adapts VACE (Jiang et al., 2025) to panoramic completion with *SC-RoPE* on unpaired videos. Stage II builds on this completion prior to learn target-conditioned removal with *SphericalControl* and paired source–clean supervision, including **PanoVOR** (Sec. 3.2). Both stages retain VACE's flow-matching objective $\mathcal{L}_{\text{vel}}$, predicting $\epsilon - z_0$ from $z_\sigma = (1-\sigma)z_0 + \sigma\epsilon$.

3.1 PRELIMINARIES AND PROBLEM FORMULATION

Equirectangular projection. A panoramic video records the full viewing sphere using longitude λ and latitude ϕ. ERP unwraps these spherical coordinates into a regular 2D grid. For a pixel (u, v) in an ERP frame of width W and height H, the corresponding spherical coordinates are

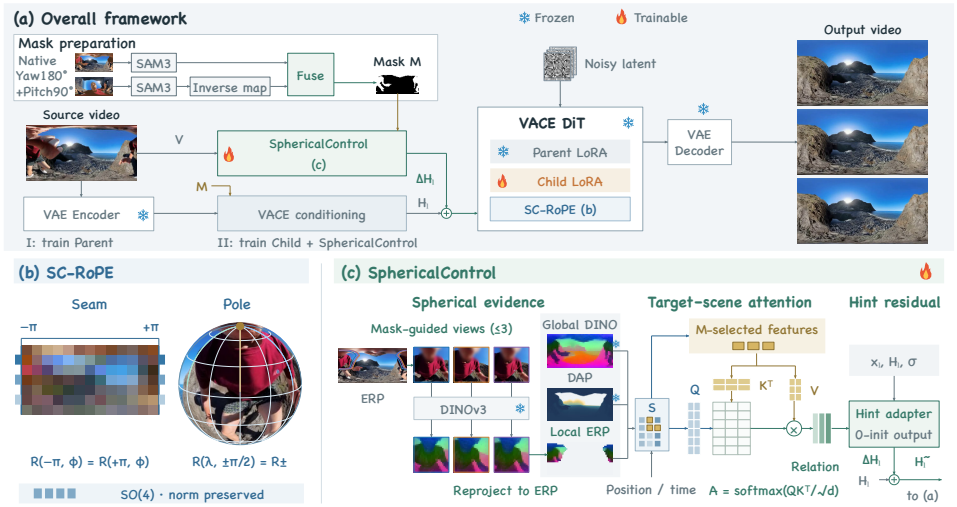

Figure 3: **PanoErase overview.** Stage I learns spherical reconstruction with SC-RoPE on unpaired panoramas; Stage II learns target-conditioned removal with SphericalControl on paired source–clean videos. All compared methods receive identical target masks (Sec. 4.1); the mask-extraction block only produces these masks and is detailed in Appendix A.2.

$$\lambda = 2\pi \frac{u + \frac{1}{2}}{W} - \pi, \qquad \phi = \pi \frac{v + \frac{1}{2}}{H} - \frac{\pi}{2}. \tag{1}$$

Two geometric properties are important for panoramic reconstruction and removal. Longitude is periodic across the left-right boundary, where neighboring spherical regions appear at opposite image borders. Meanwhile, longitudinal distances shrink toward the poles, where all longitudes converge.

Problem formulation. Given an ERP source video $V = \{I_t\}_{t=1}^{T}$ and an object-mask sequence $\mathcal{M}_{\text{obj}}$, PanoErase predicts $\widehat{V} = \mathcal{F}_\theta(V, \mathcal{M}_{\text{obj}})$ that removes the selected object and its associated effects, reconstructs the occluded background, and preserves unrelated content.

3.2 PanoVOR Dataset

Table 1: **Paired video removal datasets.** Perspective statistics follow EffectErase (Fu et al., 2026). PanoVOR reports named groups. Bold and underlining mark the largest and second-largest values.

Dataset	Persp.	360°	Effects	Dynamic Camera	Dynamic Object	Dynamic Background	Scenes / Groups	Image Pairs	Hours
RORD (Sagong et al., 2022)	✓	✗	✗	✗	✓	✗	24	516.7K	5.98
Video4Removal (Wei et al., 2025)	✓	✗	✓	✗	✓	✗	6	134.3K	1.55
ROSE (Miao et al., 2025)	✓	✗	✓	✓	✗	✗	25	1,501.0K	<u>27.79</u>
VOR (Fu et al., 2026)	✓	✗	✓	✓	✓	✓	<u>67</u>	**12,556.8K**	**145.33**
PanoVOR (ours)	✓	✓	✓	✓	✓	✓	105	<u>1,530.0K</u>	26.6

Paired supervision for panoramic removal requires the same full-sphere scene observed with and without the target under identical camera motion. Real capture makes this hard: two takes differ in sensor noise, lighting, and camera trajectory, and a 360° camera also records everything around it, including the operator and passers-by, so the takes rarely differ only in the target; perspective pairs, in turn, cannot be lifted to the sphere. Existing paired removal datasets are perspective-only and never place targets or effects at the ERP seam or the poles (Table 1). We therefore construct PanoVOR by rendering, obtaining **1.53M source–clean frame pairs spanning 26.6 hours** that cover camera motion, object motion, and dynamic backgrounds.

Paired rendering. We render each camera trajectory twice through AirSim360 (Ge et al., 2025) in Unreal Engine scenes with diverse targets, once with the target and once without it (Fig. 4).

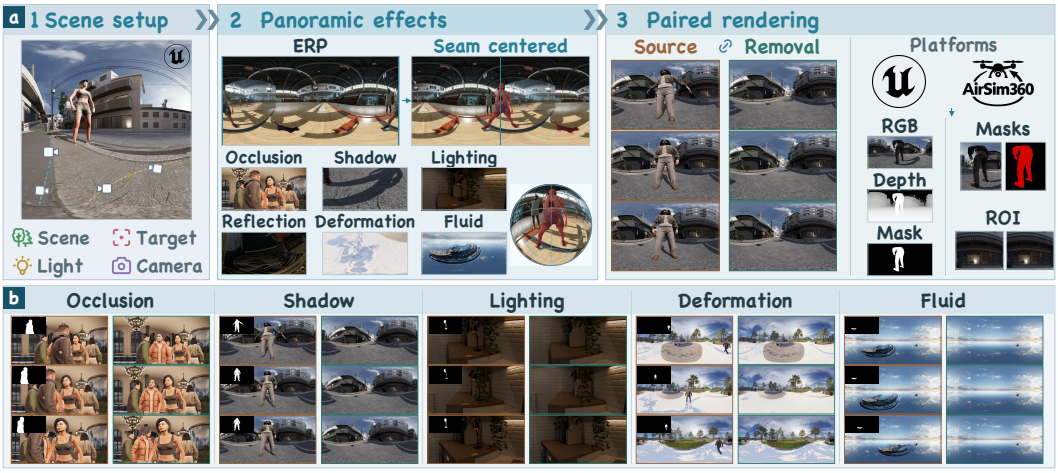

Figure 4: **Constructing PanoVOR.** (a) Each trajectory is rendered with and without the target, giving aligned source–clean ERP videos and masks. (b) Examples from PanoVOR.

The clean pass re-simulates the scene, so shadows, reflections, lighting interactions, deformation, and fluid effects are recomputed without the target and the occluded background is revealed; these effect types follow EffectErase (Fu et al., 2026), with fluid effects added. Beyond the effect types, target placement is designed for panoramas: targets and their associated effects are placed across the left–right ERP seam and in polar regions, configurations absent from perspective data.

Statistics and evaluation sets. The corpus comprises 5,432 sequences in 105 named groups, rendered at 2048×1024 and 16 fps. PanoVOR-Eval holds out 50 paired cases with clean references, and PanoVOR-Wild adds real in-the-wild panoramic videos to evaluate transfer beyond rendered scenes (Sec. 4.1). More details are provided in Appendix A.

3.3 Stage I: Spherical Reconstruction with SC-RoPE

Starting from the released SVOR removal adapter, we train the parent LoRA to reconstruct unpaired panoramas from masked inputs. Foreground-shaped masks follow SVOR's placement protocol, extended with spherical wrapping across ordinary, seam-crossing, and polar configurations. The velocity objective emphasizes masked completion while supervising visible context.

SC-RoPE. A periodic $\mathrm{SO}(2)$ block uses $e^{\mathrm{i}m\lambda}$ with integer m, ensuring periodicity while retaining longitude dependence at the poles. Scaling its phase by $\cos\phi$ removes polar dependence but generally breaks periodicity. For a continuous single-pair field whose same-latitude relative phase depends only on longitude differences, a longitude-independent polar limit forces $m = 0$.

SC-RoPE couples two rotary pairs so that latitude attenuates the longitude-dependent coefficient while preserving the integer longitude phase and the joint feature norm (Fig. 5). For an integer longitude mode m and a fixed coupling parameter $\alpha > 0$, define

$$D(\phi) = \sqrt{\cos^2\phi + \alpha^2 \sin^2\phi},$$
$$a_m = e^{\mathrm{i}m\lambda}\frac{\cos\phi}{D(\phi)}, \qquad b_m = \alpha\frac{\sin\phi}{D(\phi)}. \tag{2}$$

The coupled rotation acts on complex features z_1, z_2 as

$$z_1' = a_m z_1 - b_m \overline{z_2},$$
$$\overline{z_2'} = b_m z_1 + \overline{a_m}\,\overline{z_2}. \tag{3}$$

Because b_m is real and $|a_m|^2 + b_m^2 = 1$, the complex map on $(z_1, \overline{z_2})$ belongs to $\mathrm{SU}(2)$, and its corresponding real operator belongs to $\mathrm{SO}(4)$. Integer m ensures longitude periodicity. At the equator, $b_m = 0$ and $a_m = e^{im\lambda}$, recovering independent periodic rotations. As $\phi \to \pm\pi/2$, $a_m \to 0$ and $b_m \to \pm 1$, yielding a longitude-independent limit at each pole. The field is continuous on the sphere and relaxes exact relative-longitude dependence.

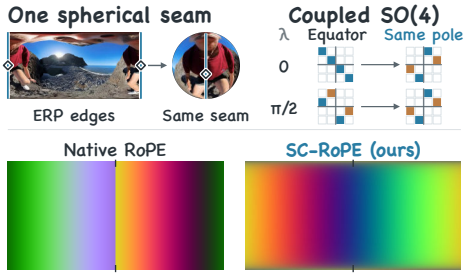

Figure 5: **SC-RoPE.** Integer longitude harmonics ensure seam continuity, while coupled rotations yield longitude-independent polar limits. Bottom: seam-centered visualizations of native RoPE and SC-RoPE.

Spherical attention. *SC-RoPE* replaces the horizontal query and key rotations with six coupled $\mathrm{SO}(4)$ blocks and three Cartesian-phase triplets of the unit direction, retaining the native temporal and vertical encodings and leaving value features unchanged. At each query, the completion-mask occupancy controls the blend between *SC-RoPE* and native attention outputs. Both paths attend to the full key set, providing spherical context to masked queries while retaining the native attention output for queries outside the mask.

3.4 Stage II: Target–Effect Association with SphericalControl

Stage I reconstructs the original scene, including visible shadows and reflections. Stage II learns the scene after removing the selected object and its associated effects. We freeze the backbone and parent adapter, then train a child LoRA and *SphericalControl* on paired data from PanoVOR, ROSE (Miao et al., 2025), and EffectErase (Fu et al., 2026). Paired source–clean supervision specifies the desired removal outcome, while *SphericalControl* provides an explicit interaction path between the selected target and panorama-wide scene features.

SphericalControl. ERP can split target evidence across the seam and stretch its appearance near the poles. *SphericalControl* combines local target evidence from mask-centered gnomonic views, which reduce deformation around the view center, with longitude-wrapped ERP features that retain context beyond these local views. A frozen DINOv3 encoder (Siméoni et al., 2025) supplies appearance and semantic features. Panoramic depth predicted by DAP (Lin et al., 2026) and its spatial gradients provide geometric context.

Panoramic appearance and geometry form H_{scene}. The object mask selects descriptors from ERP and tangent views into a target bank, which each scene location queries:

$$\begin{aligned} A_{\text{target}} &= \mathrm{Attn}(Q_{\text{scene}}, K_{\text{target}}, V_{\text{target}}), \\ H_{\text{rel}} &= \mathrm{LN}(H_{\text{scene}} + A_{\text{target}})\,. \end{aligned} \tag{4}$$

Queries come from H_{scene}; keys and values come from the target bank. The residual retains scene context while adding target evidence. We align H_{rel} to the generator's spatiotemporal grid and inject it through zero-initialized residual adapters into the VACE conditioning branch. Feature extractors remain frozen; *SphericalControl* and the child LoRA train jointly.

Training objective. The clean video supplies the full-frame velocity target for $\mathcal{L}_{\text{vel}}$. Paired differences (Sec. 3.2) estimate effect support E outside the dilated object mask and preservation support K. The regional loss $\mathcal{L}_E$ applies clean-target velocity supervision within E, while the Hard-K loss $\mathcal{L}_K$ prioritizes high-error patches within K to penalize localized damage.

A chart loss $\mathcal{L}_{\text{chart}}$ additionally compares predicted and clean-target velocity latents in mask-centered tangent views over trusted reveal and effect regions. The Stage II objective is

$$\mathcal{L} = \mathcal{L}_{\text{vel}} + \lambda_E \mathcal{L}_E + \lambda_K \mathcal{L}_K \quad + \lambda_{\text{chart}} \mathcal{L}_{\text{chart}}. \tag{5}$$

Inference. Given the source clip and target mask, *SphericalControl* conditions a single full-ERP denoising process with the trained parent and child adapters to output the edited clip.

4 EXPERIMENTS

Table 2: **Quantitative comparison on PanoVOR-Eval.** Reconstruction, temporal consistency, and panoramic geometry. Best results are bold, second best underlined; – denotes not evaluated.

	Reconstruction				Temporal and geometry		
Method	PSNR↑	LPIPS↓	T-LPIPS↓	ERG-H↑	Warp↓	Seam↓	Polar WS-PSNR↑
SVOR (Hu et al., 2026)	<u>27.9503</u>	.0867	.2284	.7933	9.6903	1.9785	<u>24.3745</u>
EffectErase (Fu et al., 2026)	26.3853	<u>.0803</u>	<u>.1987</u>	**.8194**	<u>9.0806</u>	<u>1.9082</u>	23.9433
UnderEraser (Liu et al., 2026)	27.1562	.0989	.3620	-.6511	19.7230	1.9203	13.6462
VACE (Jiang et al., 2025)	27.9351	.0958	.3443	.3907	18.1964	1.9914	18.2646
PanoGen	26.2431	.1214	.3599	.1500	16.7295	–	18.9541
PanoErase	**28.7060**	**.0744**	**.1614**	<u>.8124</u>	**7.1222**	**0.4801**	**25.4375**

Table 3: **Effect removal, preservation, and real-video quality.** Regional errors on PanoVOR-Eval and diagnostics on PanoVOR-Wild.

	PanoVOR-Eval			PanoVOR-Wild				
Method	Effect↓	Stable↓	Hard-K↓	QScore-Q↑	QScore-G↑	TRR↓	Warp↓	Seam-Δ ↓
SVOR (Hu et al., 2026)	<u>23.1698</u>	<u>5.4298</u>	21.6937	<u>6.8710</u>	<u>2.7903</u>	.0618	4.6420	1.1362
EffectErase (Fu et al., 2026)	24.0433	6.8989	26.8868	6.8145	2.7339	<u>.0602</u>	**4.5905**	<u>1.0313</u>
UnderEraser (Liu et al., 2026)	28.0212	**5.2669**	**21.0076**	6.6774	2.2742	.2214	7.5872	1.3247
PanoGen	29.3251	6.4173	25.3956	6.6613	2.2500	**.0566**	6.9093	–
PanoErase	**20.1271**	5.5465	<u>21.4265</u>	**7.0484**	**3.1371**	.0656	<u>4.6259</u>	**-0.6356**

Effect, Stable, and Hard-K are MAEs on the 0–255 scale. QScore-Q and QScore-G are rated by Qwen2.5-VL-7B-Instruct and Gemini 2.5 Pro with one 10-point prompt; TRR needs a detectable input target. Seam-Δ is relative to the input. Bold and underlining mark the best and second-best results; – denotes not evaluated.

4.1 IMPLEMENTATION DETAILS

We initialize PanoErase from Wan2.1-VACE-1.3B (Jiang et al., 2025) with the released SVOR removal LoRA (Hu et al., 2026) (rank 128, $\alpha = 64$). Stage I trains on public and in-house panoramic data at a learning rate of 10^{-5}, with *SC-RoPE*, circular seam padding, and pole padding. Stage II freezes the backbone and parent adapter, then trains *SphericalControl* and a rank-128 child LoRA at a learning rate of 10^{-4} on PanoVOR, ROSE (Miao et al., 2025), and EffectErase (Fu et al., 2026). Both stages use 8 H20 GPUs, bf16 precision, seed 42, and 81-frame clips at 832×416 and 16 fps.

PanoVOR-Eval is scene-disjoint from training, with no shared source trajectories or their derived samples. We evaluate all methods on cases with common metric support. PanoVOR-Wild provides real-video diagnostics without clean references. We compare with the removal methods SVOR (Hu et al., 2026), EffectErase (Fu et al., 2026), and UnderEraser (Liu et al., 2026), the VACE backbone (Jiang et al., 2025), and PanoGen, a panoramic inpainting baseline fine-tuned from Wan2.2-Fun-5B-InP on panoramic videos with the latitude-aware sampling of PanoWan (Xia et al., 2025), without paired removal supervision. All methods receive the same source clips and masks and generate 81 frames at 832×416 with their released conditioning settings. Masks are used as supplied, without dilation or consolidation. All methods use seed 42. We score the generated MP4 clips.

PSNR and LPIPS assess full-frame reconstruction, while Tangent-LPIPS and ERG-H assess recovery within the target region. Effect MAE measures error in target-associated effect regions outside the object mask, and Stable and Hard-K MAE measure preservation error over the keep support and its worst decile. Flow-warp AUC measures temporal consistency. The seam log-ratio and polar WS-PSNR assess reconstruction at the ERP cut and the poles. Following EffectErase (Fu et al., 2026), we compute QScore with Qwen2.5-VL-7B-Instruct on a 10-point scale, rating removal completeness and visual artifacts in matched tangent-view frame pairs. On PanoVOR-Wild we additionally report the conditional target-residual rate (TRR), flow-warp AUC, and seam log-ratio.

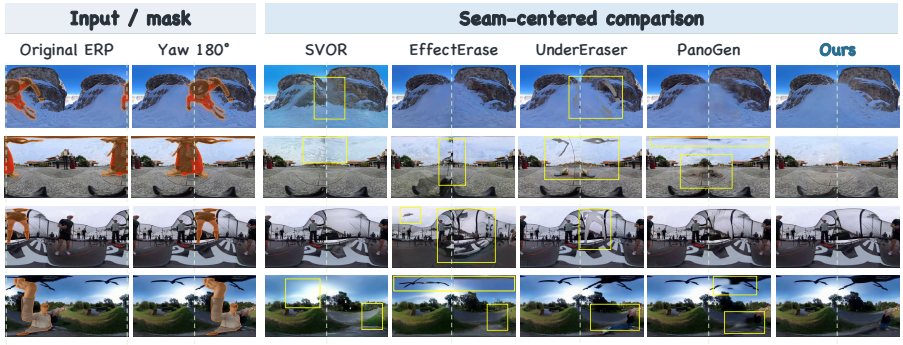

Figure 6: **Qualitative comparison on PanoVOR-Wild.** Seam-centered views (dashed line); boxes mark residual targets, remaining effects, and seam discontinuities.

4.2 Comparison with Baselines

Reconstruction and temporal consistency. On PanoVOR-Eval (Table 2), PanoErase achieves the highest full-frame PSNR and the lowest LPIPS, target-region Tangent-LPIPS, and flow-warp AUC. Its Tangent-LPIPS of 0.1614, compared with 0.1987 for EffectErase, indicates better perceptual quality within the removed region; EffectErase obtains the highest ERG-H.

Consistency under panoramic geometry. PanoErase achieves the lowest seam log-ratio and the highest polar WS-PSNR (Table 2). Its seam score is 0.4801, compared with 1.9082 for EffectErase, while polar WS-PSNR improves by 1.06 dB over SVOR. These results show gains at both the ERP boundary and the poles. On PanoVOR-Wild, the negative Seam-Δ of -0.6356 indicates a lower seam response than the input.

Effect removal and scene preservation. PanoErase reduces Effect MAE by 13.1% relative to SVOR, with comparable Stable error and lower Hard-K error (Table 3). UnderEraser obtains the lowest preservation errors, while PanoErase achieves the lowest Effect MAE and the second-lowest Hard-K MAE. On PanoVOR-Wild, PanoErase receives the highest QScores from both evaluators on matched frame pairs and the second-lowest flow-warp AUC.

Qualitative comparison. Fig. 6 shows real panoramas under a 180° yaw that places the ERP seam at the image center, so that residual targets, remaining effects, and broken structures appear in one view. PanoErase removes the target together with its effects and keeps the surrounding structure continuous across the seam. Fig. 7 shows cases from four scenes. Appendix C provides evaluation details, the complete comparison, and results from eight scenes spanning outdoor, indoor, underwater, aerial, and night conditions.

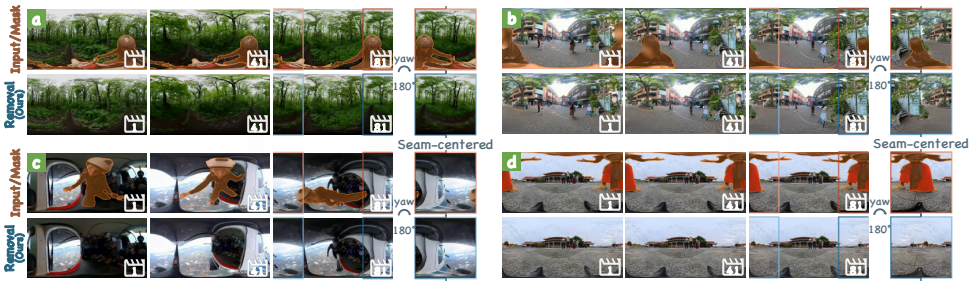

Figure 7: **Removal results on PanoVOR-Wild.** Input with the target mask (top) and our result (bottom) at frames 1, 41, and 81; the rightmost column shows frame 81 with the ERP seam centered.

4.3 Ablation Study

Spherical position encoding. We train Stage-I models with periodic $\mathrm{SO}(2)$ and *SC-RoPE* using equal training budgets (Table 4). Both encodings are periodic in longitude; the coupled encoding additionally has longitude-independent polar limits. On Polar50, *SC-RoPE* raises polar PSNR by $1.56\,\mathrm{dB}$ and reduces polar MAE by 6.81, supporting its design for polar reconstruction. It also improves hole WS-PSNR, hole Tangent-LPIPS, and seam-band PSNR on PanoVOR-Eval. Fig. 8(a) compares native RoPE with *SC-RoPE* and illustrates improved seam continuity.

Table 4: **SC-RoPE ablation.** Results on PanoVOR-Eval and Polar50 (50 paired polar cases, $|\phi| \geq 75°$).

Metric	Periodic $\mathrm{SO}(2)$	SC-RoPE
PanoVOR-Eval		
Hole WS-PSNR↑	19.6451	**19.8575**
Hole T-LPIPS↓	0.2107	**0.2059**
Seam PSNR↑	19.2045	**19.3732**
Polar50		
Polar PSNR↑	16.9501	**18.5062**
Polar MAE↓	35.7850	**28.9741**
Hole WS-PSNR↑	18.4977	**19.5969**

Table 5: **SphericalControl ablations.** Tangent cues: local DINO features and coverage; panorama cues: panoramic DINO features. Bold and underlining mark the best and second-best results.

Variant	Effect MAE↓	Hole T-LPIPS↓	PSNR↑
w/o SphericalControl	24.8250	0.1776	28.6906
w/o tangent cues	21.8018	0.1745	<u>28.7150</u>
w/o panorama cues	<u>21.0421</u>	<u>0.1674</u>	**28.8083**
w/ *SphericalControl*	**20.1271**	**0.1614**	28.7060



Figure 8: **Qualitative ablations.** (a) Native RoPE versus SC-RoPE; crops are centered on the ERP seam. (b) Matched-schedule training without and with SphericalControl.

Target-conditioned scene control. All variants in Table 5 are retrained. *SphericalControl* reduces Effect MAE by 18.9% and improves hole Tangent-LPIPS. Removing either tangent or panorama cues worsens both metrics, with greater degradation when tangent cues are removed. Full-frame PSNR remains close to the model without control, while the individual cue ablations attain slightly higher values. Fig. 8(b) shows the corresponding reduction in shadow residue.

5 Conclusion and Future Work

We introduced PanoVOR, a paired full-sphere dataset for panoramic object and effect removal, and PanoErase, which combines panoramic reconstruction through *SC-RoPE* with target-conditioned removal through *SphericalControl*. PanoErase achieves the best target-region fidelity, effect removal, and seam continuity on PanoVOR-Eval and the highest quality scores on real PanoVOR-Wild videos. Future work will explore reconstructing the physical effects of previously occluded objects and removing spatially extensive effects using larger, higher-quality paired datasets.

AI Use Statement

Large language models assisted manuscript organization and revision, LaTeX formatting, engineering support, and review of claims and references. Vision-language models also contributed to the automated dataset audit and QScore evaluation; these outputs are reported as automated assessments. Authors remain responsible for verifying the methods, numerical records, and claims.

REFERENCES

Anand Bhattad, Konpat Preechakul, and Alexei A. Efros. Visual jenga: Discovering object dependencies via counterfactual inpainting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 20

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. VideoPainter: Any-length video inpainting and editing with plug-and-play context control, 2025. URL https://arxiv.org/abs/2503.05639. 3

Yuqing Chen, Lin Liu, Haisu Wu, Xiaopeng Zhang, Yaowei Wang, Yujiu Yang, and Qi Tian. Dualeraser: Joint video object and effect removal via balanced text-mask guidance and decoupled locator-preserver, 2026. URL https://arxiv.org/abs/2605.30045. 20

Haoran Feng, Ruiyang Zhang, Longyi Zhang, Dizhe Zhang, and Lu Qi. Enhancing in-context panoramic generation via geometric-aware pretraining, 2026. URL https://arxiv.org/abs/2607.08765. 20

Yang Fu, Yike Zheng, Ziyun Dai, and Henghui Ding. EffectErase: Joint video object removal and insertion for high-quality effect erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2005–2014, June 2026. 2, 3, 4, 5, 6, 7, 17, 19, 20

Xian Ge, Yuling Pan, Yuhang Zhang, Xiang Li, Weijun Zhang, Dizhe Zhang, Zhaoliang Wan, Xin Lin, Xiangkai Zhang, Juntao Liang, Jason Li, Wenjie Jiang, Bo Du, Ming-Hsuan Yang, and Lu Qi. AirSim360: A panoramic simulation platform within drone view, 2025. 4, 14

Jiagao Hu, Yuxuan Chen, Fuhao Li, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. From ideal to real: Stable video object removal under imperfect conditions, 2026. URL https://arxiv.org/abs/2603.09283. 3, 7, 16, 19

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, October 2025. 2, 3, 7, 16, 17, 19

Saksham Singh Kushwaha, Sayan Nag, Yapeng Tian, and Kuldeep Kulkarni. Object-wiper: Training-free object and associated effect removal in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 20

Yao-Chih Lee, Erika Lu, Sarah Rumbley, Michal Geyer, Jia-Bin Huang, Tali Dekel, and Forrester Cole. Generative omnimatte: Learning to decompose video into layers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12522–12532, June 2025. 20

Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting, 2025. URL https://arxiv.org/abs/2501.10018. 3, 17, 20

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17562–17571, June 2022. 3, 20

Dong Liang, Yuhao Liu, Jinyuan Jia, Youjun Zhao, and Rynson W. H. Lau. World-shaper: A unified framework for 360° panoramic editing, 2026. URL https://arxiv.org/abs/2602.00265. 3, 20

Geng Lin, Chen Gao, Jia-Bin Huang, Changil Kim, Yipeng Wang, Matthias Zwicker, and Ayush Saraf. Omnimatterf: Robust omnimatte with 3d background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23471–23480, 2023. 3, 20

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 6

Dingming Liu, Wenjing Wang, Chen Li, Jing Lyu, and Haotian Dong. From understanding to erasing: Towards complete and stable video object removal, 2026. URL https://arxiv.org/abs/2604.01693. 7, 19, 20

Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T. Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4507–4515, 2021. 3, 20

Andrew MacQuarrie and Anthony Steed. Object removal in panoramic media. In *Proceedings of the 12th European Conference on Visual Media Production*, CVMP 2015, pp. 1–10. ACM, November 2015. doi: 10.1145/2824840.2824844. URL https://doi.org/10.1145/2824840.2824844. 3, 20

Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Hantang Liu, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. ROSE: Remove objects with side effects in videos. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 149140–149162. Curran Associates, Inc., 2025. doi: 10.52202/085713-4988. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/db2808c312de68420ef20f055b726549-Paper-Conference.pdf. 3, 4, 6, 7, 17, 20

Min-Cheol Sagong, Yoon-Jae Yeo, Seung-Won Jung, and Sung-Jea Ko. RORD: A real-world object removal dataset. In *33rd British Machine Vision Conference (BMVC)*. BMVA Press, 2022. 3, 4

Dvir Samuel, Matan Levy, Nir Darshan, Gal Chechik, and Rami Ben-Ari. Omnimattezero: Fast training-free omnimatte with pre-trained video diffusion models, 2025. Accepted to SIGGRAPH ASIA 2025. 3, 20

Ryosuke Seshimo and Mariko Isogawa. DAOVI: Distortion-aware omnidirectional video inpainting. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_334/paper.pdf. 3, 20

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, et al. Dinov3, 2025. 6

Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6913–6923, June 2024. 3, 20

Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, and Zhanyu Ma. OmniEraser: Remove objects and their effects in images with paired video-frame data, 2025. URL https://arxiv.org/abs/2501.07397v1. 3, 4

Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *Computer Vision – ECCV 2024*, volume 15135 of *Lecture Notes in Computer Science*, pp. 112–129. Springer, 2024. 20

Feier Wu, Wanke Xia, Xu He, Zilang Zhou, Si Chen, Dongxia Liu, Liyang Chen, Qimeng Wu, Zhengbo Zhang, Wenming Yang, and Zhiyong Wu. Effectlearner: World-aware object-effect reasoning for real-world video object removal, 2026a. URL https://arxiv.org/abs/2608.05565. 20

Ziyi Wu, Daniel Watson, Andrea Tagliasacchi, David J. Fleet, Marcus A. Brubaker, and Saurabh Saxena. 360anything: Geometry-free lifting of images and videos to 360°. In *European Conference on Computer Vision (ECCV)*, 2026b. URL https://arxiv.org/abs/2601.16192. 20

Yifei Xia, Shuchen Weng, Siqi Yang, Jingqi Liu, Chengxuan Zhu, Minggui Teng, Zijian Jia, Han Jiang, and Boxin Shi. PanoWan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 148027–148046. Curran Associates, Inc., 2025. doi: 10.52202/085713-4949. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d9a12369739e3d26e63e7fe66201c16f-Paper-Conference.pdf. 3, 7, 20

Yinzhe Xu, Huajian Huang, Yingshu Chen, and Sai-Kit Yeung. 360VOTS: Visual object tracking and segmentation in omnidirectional videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):9785–9797, 2025. 16

Shilin Yan, Xiaohao Xu, Renrui Zhang, Lingyi Hong, Wenchao Chen, Wenqiang Zhang, and Wei Zhang. PanoVOS: Bridging non-panoramic and panoramic views with transformer for video segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 346–365. Springer, 2024. 16

Liu Yang, Huiyu Duan, Yucheng Zhu, Xiaohong Liu, Lu Liu, Zitong Xu, Guangji Ma, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Omni^2: Unifying omnidirectional image generation and editing in an omni model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 10103–10112. ACM, October 2025. doi: 10.1145/3746027.3755405. URL https://doi.org/10.1145/3746027.3755405. 20

Yuyang Yin, Haoxiang Guo, Fangfu Liu, Mengyu Wang, Hanwen Liang, Eric Li, Yikai Wang, Xiaojie Jin, Yao Zhao, and Yunchao Wei. Panoworld-x: Generating explorable panoramic worlds via sphere-aware video diffusion, 2025. 3

Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting, 2025. 17

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6347–6357, June 2024. 3, 20

Jixin Zhao, Zhouxia Wang, Peiqing Yang, and Shangchen Zhou. Precise object and effect removal with adaptive target-aware attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 17, 20

Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, and Wei-Shi Zheng. Panorama generation from nfov image done right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21610–21619, June 2025. 20

Haoyi Zhong, Fang-Lue Zhang, Andrew Chalmers, and Taehyun Rhee. Se360: Semantic edit in 360° panoramas via hierarchical data construction, 2025. URL https://arxiv.org/abs/2512.19943. 3, 20

Shangchen Zhou, Chongyi Li, Kelvin C.K. Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 17

Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. MiniMax-Remover: Taming bad noise helps video object removal. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 75518–75547. Curran Associates, Inc., 2025. doi: 10.52202/085713-2536. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6d255ddeb1f4a5ebb8e448b3998bb8e5-Paper-Conference.pdf. 17, 20

APPENDIX CONTENTS

A Dataset Details **14**

A.1 Rendering and Pairing Audit 14

A.2 Spherical Mask Extraction 15

B Method Details **15**

B.1 Coordinates and SC-RoPE 15

B.2 Stage-I Training 16

B.3 Stage-II Conditioning and Objectives 16

B.4 Inference 17

C Experimental Details and Additional Results **17**

C.1 Additional Quantitative Results 17

C.2 Additional Qualitative Results 18

C.3 Evaluation Protocol 19

C.4 Metric Definitions 19

D Extended Related Work **20**

A Dataset Details



Figure 9: **Paired video examples from PanoVOR.** Each row shows three frames from one case: source frames with target masks on the left and corresponding clean ground truth on the right. Examples include wakes, seam-crossing targets, selective removal, shadows near the poles, and crowded scenes.

A.1 Rendering and Pairing Audit

PanoVOR pairs two AirSim360 renders (Ge et al., 2025) of each camera trajectory, with and without the selected target (Fig. 9). Re-simulating the clean scene reveals the occluded background and updates target-dependent shadows, reflections, lighting, deformation, and fluid motion. The passes use synchronized ERP frames at 2048×1024 and 16 fps. Object masks are exported by the renderer; effect and keep supports are estimated from source–clean differences (Sec. C.4). The paired changes cover the six families illustrated in Fig. 4: occlusion, shadow, lighting, reflection, deformation, and fluid effects.

To screen for unintended changes between passes, we inspect a seeded, stratified sample of 250 pairs at five frames per pair. Two independent adjudication passes check unrelated-instance deletion, pose drift, and target-associated scene changes. Sources that fail the screen are removed as a whole from training and evaluation. Within the sampled pairs, 86% have source–clean differences covering at least 256 native pixels outside the target mask. Differences reach at least one ERP border band in 49% of sequences and both bands in the same frame in 35%. These statistics quantify how often removal supervision extends beyond the object and reaches the ERP boundary.

A.2 Spherical Mask Extraction

A target may be split by the ERP seam or stretched near a pole, making it difficult to segment from one ERP view. We query a promptable segmenter on the original frame and on a rotated view obtained by composing a 180° yaw with a 90° pitch. The rotation moves the original seam and polar regions into the interior of the view. We map both predictions back to the original ERP coordinates, retain the prompted target, and merge its masks by union. The removal comparisons use their supplied canonical masks.

B Method Details

B.1 Coordinates and SC-RoPE

Equation 1 uses pixel centers, with $u = 0$ at the left boundary and $v = 0$ at the southernmost row. The unit viewing direction is

$$\mathbf{s}(u, v) = (\cos\phi\cos\lambda,\ \cos\phi\sin\lambda,\ \sin\phi)^\top. \tag{6}$$

Longitude wraps at the ERP boundary, and spherical-area integration weights each pixel by $\cos\phi$.

Single-pair stationary longitude encoding. Consider a continuous rotary field on one complex pair whose relative phase at a common latitude depends only on $\lambda_q - \lambda_k$. At each latitude, this condition gives an angle of the form $\theta = m(\phi)\lambda + c(\phi)$. Longitude periodicity requires $m(\phi)$ to be an integer. Continuity in latitude then makes m constant, and a longitude-independent limit at either pole forces $m = 0$. Thus, under this stationarity condition, a single rotary pair cannot retain a nonzero longitude mode and satisfy both boundary properties. ***SC-RoPE*** couples two pairs and relaxes exact relative-longitude dependence, while preserving the rotation norm.

Real operator and boundary behavior. Writing the coefficients in Eq. 2 as $a_m = x + \mathrm{i}y$ and $b_m = b \in \mathbb{R}$, Eq. 3 acts on $(\mathrm{Re}\, z_1, \mathrm{Im}\, z_1, \mathrm{Re}\, z_2, \mathrm{Im}\, z_2)^\top$ through

$$\mathbf{R}_m(\mathbf{s}) = \begin{bmatrix} x & -y & -b & 0 \\ y & x & 0 & b \\ b & 0 & x & -y \\ 0 & -b & y & x \end{bmatrix}. \tag{7}$$

Since $x^2+y^2+b^2 = 1$, $\mathbf{R}_m^\top\mathbf{R}_m = \mathbf{I}$ and $\det\mathbf{R}_m = 1$. Integer m gives $\mathbf{R}_m(\lambda+2\pi, \phi) = \mathbf{R}_m(\lambda, \phi)$. As $\phi \to \pm\pi/2$, $(x, y, b) \to (0, 0, \pm 1)$, so the polar limits are independent of longitude.

Rotary allocation and attention. Each head has 21 horizontal complex pairs. Six coupled blocks use pairs $(0, 7), (1, 8), (2, 9), (3, 10), (4, 5), (6, 11)$, modes $(8, 5, 3, 2, 1, 1)$, and $\alpha = 0.2$. The remaining pairs form Cartesian triplets $(12, 13, 14)$, $(15, 16, 17)$, and $(18, 19, 20)$, with component-wise phases $e^{\mathrm{i}gs_c}$ and gains $g = 1/4, 1/16, 1/64$, respectively. Time and height rotary pairs remain unchanged.

The query output blends native and spherical attention according to the condition-mask occupancy $g_{\mathrm{hole}} \in [0, 1]$:

$$\mathbf{Y} = \mathbf{Y}_{\mathrm{native}} + g_{\mathrm{hole}}\,\beta(\mathbf{Y}_{\mathrm{SC}} - \mathbf{Y}_{\mathrm{native}}). \tag{8}$$

Both paths attend to all keys. Here g_{hole} is the patch-average mask occupancy, and $\beta = u^2(3 - 2u)$ for normalized warm-up progress $u \in [0, 1]$, with $\beta = 1$ at inference. Queries outside the condition mask retain the native output.

B.2 Stage-I Training

Stage I continues the released SVOR perspective removal LoRA (Hu et al., 2026) on the 1.3B VACE backbone (Jiang et al., 2025), using rank 128, LoRA scale $\alpha = 64$, bf16 precision, and a learning rate of 10^{-5}. Circular seam padding and pole padding accompany *SC-RoPE*.

The completion data contain 24,843 clips from 19,714 source videos, drawn from PanoVOS (Yan et al., 2024), the 360VOS set of 360VOTS (Xu et al., 2025), and captured footage, plus an auxiliary branch of 105,201 panoramic images. Video samples have 81 frames at 832×416 and 16 fps. Foreground mask tracks transferred from other clips define moving holes; yaw placement avoids the recipient's foreground, and the unmodified recipient video supplies the reconstruction target. Regular clips and clean-background clips with large holes are sampled equally.

ERP-periodic masks cover ordinary, seam-crossing, and polar placements. Seam- and pole-touching clips are oversampled by factors of two and four. Mask dilation and corruption provide coarse-mask examples. The velocity objective weights the hole interior, its boundary, and visible content by 5, 3, and 1, respectively, with temporal pixel, temporal statistic, and flow-consistency weights of 0.1, 0.05, and 0.05. Reconstruction of the original video supplies both completion and preservation supervision.

B.3 Stage-II Conditioning and Objectives

Conditioning architecture. *SphericalControl* extracts frozen DINOv3 features from the source ERP frames and up to three mask-centered 100° tangent views, alongside predicted panoramic depth and its spatial gradients. Four dispersed frames each contribute one global and up to eight mask-selected local target descriptors. Scene-to-target attention produces the relation field in Eq. 4. Longitude sampling is periodic, and a validity gate suppresses padded locations. The field is pooled to the latent hint grid, using temporal maximum pooling where source frames share a latent slot, and injected through zero-initialized residual adapters at VACE hints 0, 7, and 14. Stage II trains these adapters and the child LoRA, with the backbone, parent LoRA, and feature extractors frozen. All conditioning is computed from the source video and target mask.

Full-frame and regional supervision. Let $\varepsilon_i = \|v_\theta(i) - v^\star(i)\|^2$ be channel-averaged velocity error at latent cell i, and let $w_i \in [0, 1]$ indicate valid source content. The clean video supplies $v^\star$, giving

$$\mathcal{L}_{\text{vel}} = \frac{\sum_i w_i \varepsilon_i}{\sum_i w_i}. \tag{9}$$

For regional supervision, E is the paired-difference support outside the dilated object mask, expanded spherically by four latent cells. The keep support K is the trusted remainder with E removed. A clip indicator $\tau \in \{0, 1\}$ selects identity-aligned pairs with trusted difference supports. The regional terms are

$$\mathcal{L}_E = \frac{\sum_i \tau E_i \varepsilon_i}{\sum_i \tau E_i}, \qquad \mathcal{L}_K = \frac{\sum_{i \in \mathcal{P}^\star} \tau K_i \varepsilon_i}{\sum_{i \in \mathcal{P}^\star} \tau K_i}. \tag{10}$$

The set $\mathcal{P}^\star$ contains the highest-error 20% of 4×4 latent patches with positive keep support, ranked by detached K-weighted mean error. It concentrates preservation supervision on local damage. Loss numerators and denominators are pooled across the training batch; empty supports contribute zero. The evaluation Hard-K metric instead uses the highest-error 10% of keep-region pixels (Sec. C.4).

Tangent-chart supervision. Let P_j bilinearly sample the ERP velocity latents into tangent view j. Define $\delta_{j,i} = P_j(v_\theta)_i - P_j(v^\star)_i$ and $q_{j,i} = P_j(\tau \,\text{clip}(R + E, 0, 1))_i$, where R is the revealed-region support and the target branch is detached. The loss compares values and horizontal/vertical first differences:

$$\mathcal{L}_{\text{chart}} = \frac{\sum_{j,i} A_i q_{j,i} \|\delta_{j,i}\|^2}{C \sum_{j,i} A_i q_{j,i}} + \frac{\sum_j \sum_{(i,k) \in \mathcal{N}} A_i q_{j,i} q_{j,k} \|\delta_{j,i} - \delta_{j,k}\|^2}{C \sum_j \sum_{(i,k) \in \mathcal{N}} A_i q_{j,i} q_{j,k}}. \tag{11}$$

Here C is the channel count and $\mathcal{N}$ contains horizontal and vertical neighbor pairs. For tangent-plane coordinates (a_i, b_i), $A_i \propto (1 + a_i^2 + b_i^2)^{-3/2}$ is the mean-normalized solid-angle weight; each

difference uses the weight at its starting cell. Sums include frames and batch samples. We use up to three 48×48 charts with a 100° field of view, sharing the conditioning-view axes and periodic ERP sampling. Equation 5 uses $\lambda_E = \lambda_K = 0.5$ and $\lambda_{\text{chart}} = 0.3$.

Training schedule. Stage I uses 1,000 updates, followed by 800 Stage-II updates for the main model. Its chart supervision starts at Stage-II update 451 on ERP pairs with valid reveal/effect support.

B.4 Inference

The model receives the source ERP video, canonical object masks, and a fixed generic removal prompt. It predicts the complete edited video in one denoising process. Reported results use the supplied masks directly and the generated full frames, without mask consolidation or source-frame compositing.

C Experimental Details and Additional Results

C.1 Additional Quantitative Results

Table 6: **ROSE-Benchmark.** Published results from Fu et al. (2026) and a separate PanoErase diagnostic. Bold and underlining rank only the published-protocol block.

Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓
Published protocol				
ObjectClear (Zhao et al., 2026)	29.535	<u>0.920</u>	<u>0.076</u>	742.829
OmniPaint (Yu et al., 2025)	27.569	0.910	0.085	809.645
ProPainter (Zhou et al., 2023)	27.200	0.915	0.095	171.020
DiffuEraser (Li et al., 2025)	26.502	0.898	0.128	167.483
VACE (Jiang et al., 2025)	20.805	0.694	0.174	254.117
MiniMax-Remover (Zi et al., 2025)	26.770	0.905	0.099	137.840
ROSE (Miao et al., 2025)	<u>31.122</u>	0.917	0.077	<u>72.177</u>
EffectErase (Fu et al., 2026)	**32.161**	**0.948**	**0.039**	**55.578**
PanoErase diagnostic protocol				
PanoErase	28.449	0.922	0.080	—

Table 7: **EffectErase VOR-Eval.** Published results from Fu et al. (2026) and a separate PanoErase diagnostic. Bold and underlining rank only the published-protocol block.

Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓
Published protocol				
ObjectClear (Zhao et al., 2026)	22.583	0.787	<u>0.190</u>	1391.858
OmniPaint (Yu et al., 2025)	21.511	0.781	0.201	1439.867
ProPainter (Zhou et al., 2023)	21.975	0.800	0.225	589.012
DiffuEraser (Li et al., 2025)	21.946	<u>0.802</u>	0.214	559.497
VACE (Jiang et al., 2025)	17.677	0.591	0.294	806.476
MiniMax-Remover (Zi et al., 2025)	21.963	<u>0.802</u>	0.217	539.427
ROSE (Miao et al., 2025)	<u>22.966</u>	0.792	0.203	<u>383.084</u>
EffectErase (Fu et al., 2026)	**23.750**	**0.806**	**0.170**	**342.871**
PanoErase diagnostic protocol				
PanoErase	21.952	0.774	0.212	—

Perspective-video diagnostics. Tables 6 and 7 summarize published perspective results and a separate PanoErase evaluation. The published block reproduces Table 2 of Fu et al. (2026). Our diagnostic evaluates the final PanoErase model on 60 ROSE and 43 VOR-Eval cases with 20 denoising steps, placing 704×396 content on an 832×416 canvas and scoring at 480×270. Rankings apply only within the published-protocol block; FVD is not computed for PanoErase.

Full-frame fidelity. Complementing Table 2, SSIM is 0.8628 for SVOR, 0.8389 for EffectErase, 0.8583 for UnderEraser, 0.8716 for VACE, and 0.8780 for PanoErase.

Table 8: **Training-stage comparison on PanoVOR-Eval.** Full-frame fidelity, target-region perceptual quality, and temporal consistency.

Model	PSNR↑	LPIPS↓	Hole T-LPIPS↓	Warp AUC↓
Stage I (SC-RoPE)	28.8736	0.0580	0.2059	12.7772
Stage II (PanoErase)	28.7060	0.0744	0.1614	7.1222

From reconstruction to removal. Table 8 compares the Stage-I model with the final Stage-II model on PanoVOR-Eval. Stage II lowers hole Tangent-LPIPS from 0.2059 to 0.1614 and flow-warp AUC from 12.7772 to 7.1222, while full-frame PSNR decreases by 0.17 dB and LPIPS increases. The comparison shows improved target-region perceptual quality and temporal consistency alongside a trade-off in full-frame fidelity.

C.2 Additional Qualitative Results

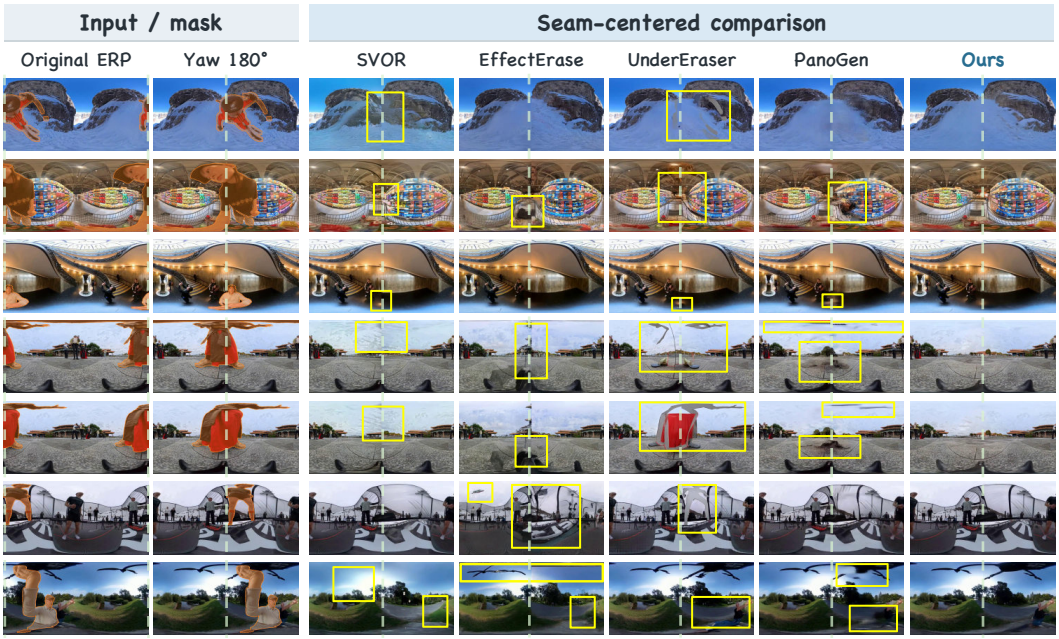

Figure 10: **Complete qualitative comparison on PanoVOR-Wild.** Seven cases under a 180° yaw that centers the ERP seam (dashed line). Boxes highlight residual targets, remaining effects, and discontinuities in the surrounding scene.

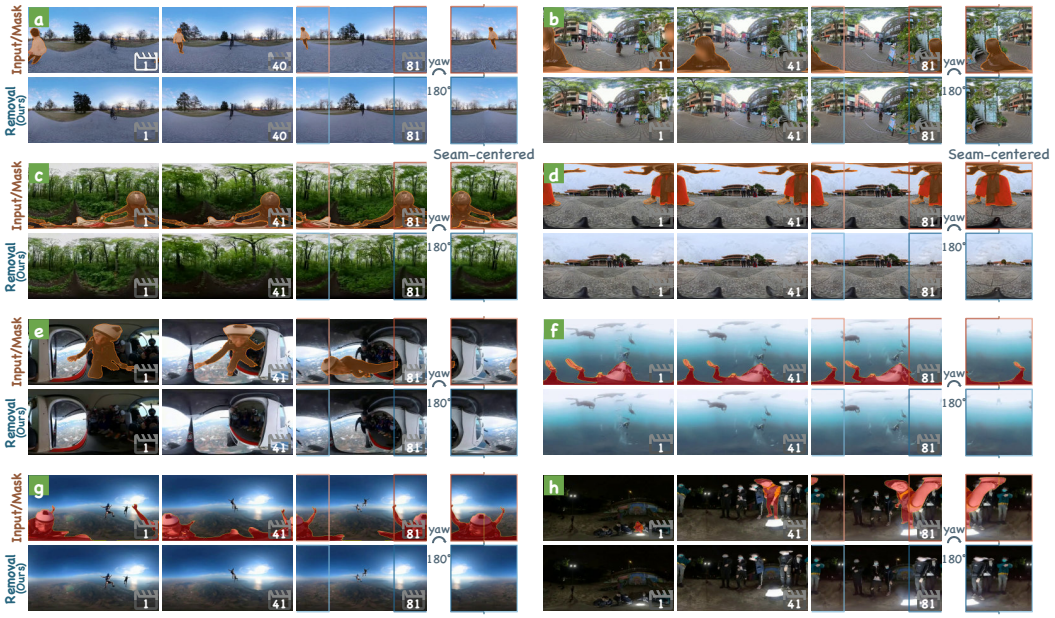

Figure 11: **Additional removal results on PanoVOR-Wild.** **(a)** a daylight park, **(b)** a pedestrian street, **(c)** a forest trail, **(d)** a temple plaza, **(e)** an aircraft cabin, **(f)** an underwater scene, **(g)** skydiving, and **(h)** a night scene. Top rows show the input with the target mask and bottom rows our result at the marked frames; the rightmost column repeats frame 81 under a 180° yaw that centers the ERP seam. Identifiable faces in panel (b) are blurred for privacy.

Fig. 10 gives all seven seam-centered comparison cases, of which Fig. 6 shows four, and Fig. 11 gives eight scenes, of which Fig. 7 shows four.

C.3 Evaluation Protocol

PanoVOR-Eval contains 50 paired cases and is scene-disjoint from training, with no shared source trajectories or their derived samples. The main comparison uses 49 cases with common metric support; one mask has no valid projected support. PanoVOR-Wild evaluates removal on 124 real panoramic clips excluded from both training stages, without clean references. All methods receive the same videos and canonical masks and produce 81 frames at 832×416 and 16 fps. SVOR (Hu et al., 2026), EffectErase (Fu et al., 2026), UnderEraser (Liu et al., 2026), and VACE (Jiang et al., 2025) use their released conditioning procedures; PanoGen uses its inpainting conditioning with the same masks. Reconstruction, regional MAE, flow-warp, and seam metrics score MP4 outputs under shared decoding and scoring procedures.

Scores are computed per video and averaged with equal weight over eligible cases. Eligibility follows the metric support rules below.

C.4 Metric Definitions

Reconstruction and recovery. Full-frame PSNR, SSIM, and LPIPS measure fidelity to the clean reference. Hole WS-PSNR weights ERP pixels by the cosine of latitude. Hole Tangent-LPIPS uses 90° tangent views and solid-angle-weighted aggregation over the projected object mask. For input X, clean reference G, prediction Y, and canonical mask H, ERG-H measures improvement over the unedited input:

$$\text{ERG-H} = 1 - \frac{\text{WS-MSE}(Y, G; H)}{\text{WS-MSE}(X, G; H)}. \tag{12}$$

The input scores 0 and the clean reference 1; cases with an invalid denominator are excluded. The main polar WS-PSNR uses the object mask intersected with the polar caps, $|\phi| \geq 60°$.

Temporal consistency. On PanoVOR-Eval, optical flow F is estimated from the clean reference and shared across methods. With W denoting warping, the residual error at frame interval Δt is

$$L(\Delta t) = \underset{\text{valid}\cap H}{\text{mean}} \left| [Y_t - W(Y_{t+\Delta t}, F)] - [G_t - W(G_{t+\Delta t}, F)] \right|. \tag{13}$$

Flow-warp AUC integrates L over $\Delta t \in \{1, 4, 8, 16\}$ on a normalized $\log_2 \Delta t$ axis. Valid support intersects the current and warped future masks, in-frame pixels, and forward–backward flow agreement. A case is eligible only when every offset has at least 20% coverage. PanoVOR-Wild has no clean reference: its Warp AUC measures prediction consistency using flow estimated from the predicted frames. The two datasets therefore use distinct temporal diagnostics.

Effect removal and preservation. Let H_d be the canonical mask dilated by four pixels with horizontal ERP wrapping, and define $D = \mathbb{1}[\|X - G\|_2 > 0.09]$ on RGB values in $[0, 1]$. The evaluation supports are $E = D(1 - H_d)$ and $K = (1 - H_d)(1 - E)$. Effect and Stable MAE average channel-wise absolute error over E and K, respectively, on the 0–255 scale. Hard-K MAE averages errors at or above the 90th percentile within each video's keep region. These pixel-space metrics use equal pixel weights and are distinct from the latent-patch training losses in Eq. 10.

Seam continuity. Seam evaluation uses rows where the canonical mask touches a band within eight pixels of either ERP border, expanded vertically by three pixels. A six-column Scharr operator measures discontinuity DS over these rows. We report

$$\text{Seam} = \log \frac{\text{DS}_{\text{pred}} + 10^{-3}}{\text{DS}_{\text{ref}} + 10^{-3}}. \tag{14}$$

The reference is the clean video on Eval and the source video on Wild. Cases with $\text{DS}_{\text{ref}} < 10^{-2}$ are excluded. Zero denotes reference-level discontinuity; negative values denote a smaller seam response.

Reference-free removal quality. Following EffectErase (Fu et al., 2026), QScore uses Qwen2.5-VL-7B-Instruct; Gemini 2.5 Pro supplies a second evaluator. Each rates removal completeness and visual artifacts on a 10-point scale from matched tangent-view input/result JPEGs at the first nonempty-mask frame. Conditional target-residual rate (TRR) measures residual detections at a 0.5 threshold on the input-detectable subset, comprising 68 videos and 4,050 frames.

D Extended Related Work

Object removal and effect association. Video inpainting uses temporal propagation or generative priors to reconstruct masked content (Li et al., 2022; 2025; Zi et al., 2025). Target-associated changes outside the mask require additional association: Omnimatte and its extensions use video layers (Lu et al., 2021; Lin et al., 2023; Lee et al., 2025; Samuel et al., 2025), ROSE predicts target-affected regions (Miao et al., 2025), and EffectErase uses supervised region guidance (Fu et al., 2026). Object-WIPER, DualEraser, EffectLearner, and UnderEraser further study effect localization and selective editing (Kushwaha et al., 2026; Chen et al., 2026; Wu et al., 2026a; Liu et al., 2026). Counterfactual image supervision supports ObjectClear and ObjectDrop (Zhao et al., 2026; Winter et al., 2024), while Visual Jenga examines scene dependence through inpainting (Bhattad et al., 2025).

Panoramic synthesis and editing. PanFusion and PanoDecouple adapt diffusion models to panoramic generation and field-of-view expansion (Zhang et al., 2024; Zheng et al., 2025). SE360, Omni^2, World-Shaper, and Canvas360 study controlled panoramic image generation and editing (Zhong et al., 2025; Yang et al., 2025; Liang et al., 2026; Feng et al., 2026). For video, 360DVD, PanoWan, and 360Anything develop panoramic generation or full-sphere expansion (Wang et al., 2024; Xia et al., 2025; Wu et al., 2026b). Panoramic removal also has direct precedents in the work of MacQuarrie and Steed (MacQuarrie & Steed, 2015) and distortion-aware omnidirectional video inpainting in DAOVI (Seshimo & Isogawa, 2025).