TOWARDS COHERENT SCENE RECONSTRUCTION IN VIDEO OBJECT REMOVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Video object removal requires coherent scene reconstruction after eliminating unwanted objects and their associated effects. Cross-frame observations can reveal occluded content, yet mixed input representations, misleading cues, and uneven reconstruction support limit their effective use. We propose **TIE**, a framework integrating *Type-Isolated Attention* and *Evidence-Guided Supervision* for coherent scene reconstruction. Type-Isolated Attention organizes the original video, mask, and noisy target into distinct token streams and directs their interaction, maintaining target access to conditions while preventing noisy-target feedback into condition representations. Beyond accessibility, Evidence-Guided Supervision uses paired edited targets and frozen cross-frame matching to assess whether observations provide valid support for the scene after removal. The resulting evidence calibrates the allocation of complementary point-wise and region-wise auxiliary supervision, emphasizing precise matching where support is strong and local appearance and variation constraints where it is weaker. This allocation supplements dense reconstruction to account for differences in support across locations. Extensive experiments demonstrate that TIE achieves state-of-the-art overall performance, enabling more complete removal of objects and their associated effects, more faithful reconstruction of occluded content, and improved temporal coherence while preserving the surrounding scene.

1 INTRODUCTION

Video object removal (VOR) aims to erase unwanted objects and all their associated effects while reconstructing occluded content and preserving the surrounding scene over time (Lu et al., 2021; Suhail et al., 2023; Wu et al., 2022). Recent methods adapt pretrained video generators with region guidance and specialized supervision to improve removal and background completion (Miao et al., 2025; Fu et al., 2026; Liu et al., 2026; Hu et al., 2026). These methods improve affected-region localization and removal consistency, but effectively exploiting cross-frame observations for coherent scene reconstruction remains challenging. Cross-frame observations can reveal occluded structure (Lee et al., 2019; Gao et al., 2020; Oh et al., 2019; Zou et al., 2021), yet may also contain content or effects that should disappear. This calls for selective access to scene information and supervision calibrated to its reconstruction support.

This selective access depends first on how the generator represents its inputs. Models built on pretrained video generators (Blattmann et al., 2023b;a; Chen et al., 2024; 2023) such as Wan (Team Wan et al., 2025) combine the original video, input mask, and noisy target to specify scene content, removal locations, and the evolving reconstruction. Channel concatenation is a common design among current generative removal methods, including ROSE and EffectErase. By joining spatially aligned latents before token projection, it supplies scene content and removal locations jointly to guide reconstruction. However, their fusion within shared tokens leaves attention to infer distinct input functions, which may allow irrelevant cross-frame information to interfere with background recovery. In contrast, sequence concatenation represents each input with a separate set of tokens, allowing attention to distinguish and separately weight scene observations, the mask, and the noisy target. Figure 1 supports this distinction quantitatively through improved reconstruction fidelity and spatial and temporal removal coherence. Qualitatively, token concatenation yields clearer background recovery while reducing responses to earlier masked content and retaining localized boundary-context

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.

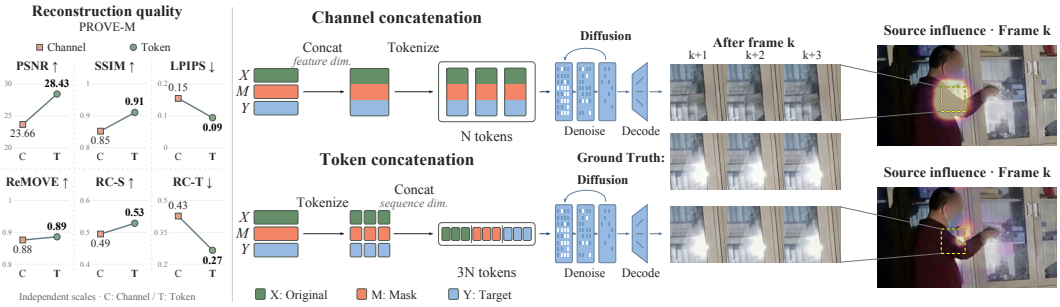

Figure 1: **Channel versus token concatenation.** We compare reconstruction quality and cross-frame input sensitivity under the same settings. Maps average the gradient sensitivity of aligned reconstructed regions across three subsequent frames to an earlier input latent slice.

responses. This pattern suggests more selective use of contextual cues for reconstruction with less interference from misleading cross-frame information.

However, more selective responses do not establish that the retained context is valid after removal. Unmasked regions may still contain object-associated effects, and some occluded locations lack reliable cross-frame support. Separate token sequences also leave the interaction direction unrestricted, allowing the noisy target to alter original-video and mask representations. Coherent scene reconstruction thus requires controlling information flow between input types and adapting supervision to the validity and strength of reconstruction support.

We propose **TIE**, a video object removal framework that integrates *Type-Isolated Attention* and *Evidence-Guided Supervision* to achieve coherent scene reconstruction despite misleading observations and uneven cross-frame support. First, we organize information access through Type-Isolated Attention. Rather than fusing heterogeneous inputs within shared tokens through channel concatenation, we adopt separate token sequences with learned type embeddings. In every self-attention block, target tokens access all streams, while original-video and mask tokens interact only within the condition streams. This preserves access to scene observations and removal guidance while preventing noisy-target feedback into condition representations during the forward pass. Second, we move from information access to reliable reconstruction support through Evidence-Guided Supervision. Compared with a single image, video provides additional cross-frame observations that may reveal occluded content, but these observations may retain object-associated effects or offer unequal support across locations. We therefore use paired edited targets to filter candidate observations and frozen cross-frame matching to estimate counterfactual-valid support. This evidence calibrates a geometric prior based on mask-boundary proximity to allocate complementary point-wise and region-wise auxiliary supervision. Stronger support favors precise matching, while weaker support emphasizes local appearance and variation constraints, supplementing dense reconstruction.

Our contributions are as follows:

- **Type-Isolated Attention.** We combine explicit token types with directed attention, preserving scene access while keeping condition representations independent of noisy target.
- **Evidence-Guided Supervision.** We calibrate point-wise and region-wise auxiliary supervision using counterfactual-valid cross-frame evidence of reconstruction support.
- **State-of-the-Art Performance.** Quantitative and qualitative evaluations establish TIE's overall advantage in video object removal, with ablations supporting its key designs.

2 Related Work

2.1 Video Inpainting and Object Removal

Video inpainting provides a basis for object removal by recovering masked content (Zeng et al., 2020; Liu et al., 2021; Li et al., 2022; Chang et al., 2019; Ren et al., 2022). ProPainter (Zhou et al., 2023) propagates information across frames, while VACE (Jiang et al., 2025) supports generative completion within a unified video-editing framework. Removing an object also requires eliminating

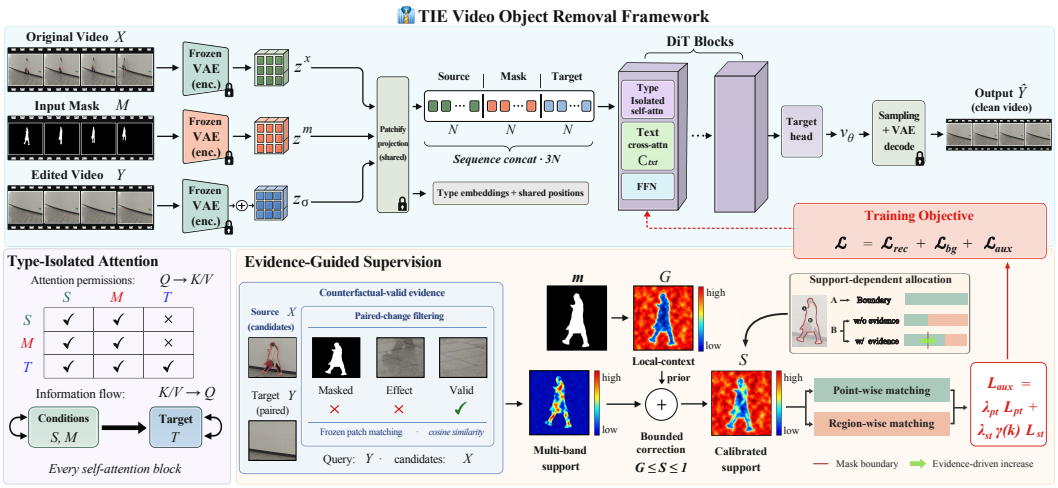

Figure 2: **Overview of TIE for coherent scene reconstruction.** Type-Isolated Attention organizes information flow among original-video, mask, and target tokens. Evidence-Guided Supervision calibrates complementary point-wise and region-wise matching. The support-dependent allocation inset illustrates how geometry and evidence determine their weights.

its effects beyond the mask, motivating methods that model affected regions (Lin et al., 2023; Lee et al., 2025). ROSE (Miao et al., 2025) uses synthetic pairs and affected-region prediction, and EffectErase (Fu et al., 2026) learns removal and insertion jointly with effect-consistency supervision. UnderEraser (Liu et al., 2026) models object–effect relations through distillation, while SVOR (Hu et al., 2026) improves localization through temporal mask aggregation and denoising-aware segmentation. These advances improve object and effect removal, but reconstructing the revealed scene remains challenging when cross-frame observations contain misleading content or provide insufficient support. TIE addresses this reconstruction challenge through Type-Isolated Attention and Evidence-Guided Supervision, organizing access to observations and adapting supervision to their reconstruction support.

2.2 Supervision for Reconstruction in VOR

Adapting pretrained generators (Ho et al., 2022; Yang et al., 2025; Guo et al., 2024; Yu et al., 2023) such as Wan to VOR requires visual guidance and reconstruction objectives that support background recovery. VACE introduces a context adapter, while VideoPainter (Bian et al., 2025) separately encodes background information. For object removal, ROSE combines channel-based conditioning with affected-region prediction. EffectErase couples task-aware cross-attention with effect-consistency supervision, and SVOR weights object-associated effects during training. These designs guide affected-region modeling through visual context and supervision. Coherent scene reconstruction further requires distinguishing useful observations from misleading content and accounting for unequal cross-frame support (Xu et al., 2019; Zhang et al., 2022a;b; Kang et al., 2022; Zhang et al., 2019). Our TIE addresses these needs through Type-Isolated Attention for selective information access and Evidence-Guided Supervision for allocating complementary point-wise and region-wise auxiliary losses according to counterfactual-valid reconstruction support.

3 Method

TIE is a video object removal framework for coherent scene reconstruction, as illustrated in Figure 2. Its two modules address selective access to observations and the reliability of reconstruction support. Type-Isolated Attention preserves distinct input types and directs their interaction. Evidence-Guided Supervision estimates counterfactual-valid cross-frame support and uses it to allocate complementary auxiliary losses within a shared reconstruction objective.

3.1 TASK FORMULATION

Given original video X and binary input mask M, we learn from paired edited target Y, where the object and its effects, potentially beyond M, are removed. A frozen VAE (Kingma & Welling, 2014; Rombach et al., 2022) encodes X, Y, and the RGB-replicated mask into z^x, z^y, and z^m. A separately resampled binary mask m weights losses at latent sites $r \in \Omega$. With Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and level $\sigma \in [0, 1]$, linear flow matching (Lipman et al., 2023; Liu et al., 2023) defines

$$\begin{aligned} z_\sigma &= (1-\sigma)z^y + \sigma\epsilon, \qquad u = \epsilon - z^y, \\ \hat{z}^y &= z_\sigma - \sigma v_\theta(z_\sigma, \sigma; z^x, z^m, c_{\text{txt}}). \end{aligned} \tag{1}$$

The predictor v_θ learns velocity u, and $\hat{z}^y$ is its clean endpoint estimate for reconstruction and auxiliary losses. Text condition c_{txt} encodes the fixed removal instruction in Appendix C.

3.2 TYPE-ISOLATED ATTENTION

Our Type-Isolated Attention module combines separately addressable input types with directed information flow. As motivated, original-video observations, removal masks, and the evolving target serve different functions. Explicit token types distinguish these functions, and directed attention prevents the noisy target from modifying condition representations.

Explicit token types and sequence concatenation. A shared frozen patch projection and flattening operator $\mathcal{P}$ produces N tokens of width d per latent. With learned type embeddings, $H^x = \mathcal{P}(z^x) + e_x$, $H^m = \mathcal{P}(z^m) + e_m$, and $H^y = \mathcal{P}(z_\sigma) + e_y$ are concatenated along the sequence dimension as $H^0 = [H^x; H^m; H^y] \in \mathbb{R}^{3N \times d}$. This preserves input identities for type-dependent attention. Corresponding tokens share rotary coordinates. Source and mask tokens use timestep zero, while target tokens use the timestep associated with σ.

Type-isolated information flow. In every self-attention block, source and mask tokens form conditions C and attend within C, while target tokens U attend to both C and U. Conditions thus remain independent of the noisy target during the forward pass while guiding reconstruction. For queries Q, keys K, and values V,

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + \mathcal{B}\right) V, \qquad \mathcal{B} = \begin{bmatrix} \mathbf{0}_{C,C} & -\boldsymbol{\infty}_{C,U} \\ \mathbf{0}_{U,C} & \mathbf{0}_{U,U} \end{bmatrix}. \tag{2}$$

Rows index queries and columns index keys, with head width d_h. In Figure 2, S, M denote the condition streams C, and T denotes target stream U. Backbone normalization and rotary encoding apply to Q, K. Text cross-attention and feed-forward layers are unchanged, and the output head predicts v_θ from target tokens.

3.3 EVIDENCE-GUIDED SUPERVISION

Unmasked observations may retain object-associated effects, and occluded locations may lack reliable correspondences. Evidence-Guided Supervision therefore estimates candidate validity and cross-frame support to allocate complementary point-wise and region-wise auxiliaries that supplement dense reconstruction. Figure 2 illustrates this allocation, while Figure 3 details the corresponding comparisons.

Counterfactual-valid evidence estimation. We first identify observations in the original video that should be preserved in the edited result. The evidence branch in Figure 2 filters out masked objects and object-associated effects before matching source candidates to the clean target. We compute paired RGB change between X and Y and pool it to latent sites, obtaining δ_r. With adaptive threshold τ_δ and stabilizer $\varepsilon > 0$, changed-region and preservation weights are

$$e_r = (1 - m_r)\,\text{clip}_{[0,1]}\left(\frac{\delta_r - \tau_\delta}{\max(\tau_\delta, \varepsilon)}\right), \qquad p_r = 1 - m_r - e_r. \tag{3}$$

The weights m_r, e_r, and p_r sum to one and allocate reconstruction supervision to object, changed, and preserved regions, respectively. Outside the mask, larger paired RGB changes shift weight from preservation to the changed region. The adaptive threshold τ_δ incorporates the median and median absolute deviation (Rousseeuw & Croux, 1993) of outside-mask differences to suppress minor paired discrepancies. High p_r thus indicates little change relative to this threshold. Appendix A provides its full definition.

Each clean-target token i queries local source neighborhoods using unit-normalized embeddings from frozen $\mathcal{P}$. Candidate j must lie outside the spatially dilated mask and satisfy $p_j \geq \tau_p$ on the token grid, where τ_p is the validity threshold. A temporal band groups neighborhoods at a given offset before and after the query time on the token grid. Each band combines cosine similarity, the best-to-second-best margin, and candidate validity, scoring zero with fewer than two usable candidates. We take the second-largest band score to require support from at least two bands and prevent an isolated match from dominating. After thresholding weak scores, interpolating to latent sites, and restricting to the mask, we obtain detached evidence $c_r \in [0, 1]$. High c_r indicates reliable support received from other frames. Evidence guides supervision without entering the predictor. Appendix A specifies scoring and thresholds.

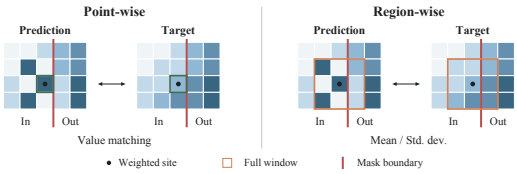

Figure 3: **Complementary reconstruction comparisons.** Point-wise matching compares values at the marked site. Region-wise matching compares local means and standard deviations in windows that may cross the red mask boundary. Dots mark weighted sites.

From evidence to calibrated support. We calibrate a geometric prior G_r with received evidence c_r to obtain support S_r, as illustrated in the support-dependent allocation inset of Figure 2. Mask-boundary proximity approximates local-context support, favoring point-wise fidelity near the boundary and region-wise matching inward. With spatial erosion depth $d_r = 1$ at the boundary and clipwise maximum $d_{\max}$,

$$G_r = (1 - m_r) + m_r \exp\left[-\frac{\max(d_r - 1, 0)}{\ell}\right], \qquad \ell = \max(\tau_g d_{\max}, 1). \tag{4}$$

The scale τ_g controls spatial decay. To retain local-context support when matching misses valid observations, we use G_r as a baseline and add a bounded evidence correction:

$$\alpha(\sigma) = \lambda_e \,\mathrm{sigmoid}\left(\frac{\sigma - \sigma_{\mathrm{mid}}}{T_\sigma}\right), \qquad S_r = G_r + \min\{\Delta_{\max}, \alpha(\sigma) c_r (1 - G_r)\}. \tag{5}$$

The factor $(1 - G_r)$ limits the correction, capped by $\Delta_{\max}$. For $\lambda_e \in [0, 1]$, $\Delta_{\max} \geq 0$, and $T_\sigma > 0$, $G_r \leq S_r \leq 1$, with $S_r = G_r$ when $c_r = 0$. The sigmoid emphasizes evidence at higher noise levels, when the target contains less usable structure. Outside the mask, $G_r = S_r = 1$, but the auxiliary weights below vanish.

In Figure 2, boundary site A has $G_r = 1$ and receives only point-wise auxiliary supervision. At interior site B, evidence increases point-wise weight and reduces region-wise weight relative to geometry alone, preserving their total.

Supervision allocation and overall objective. The green and orange segments in Figure 2 represent complementary point-wise and region-wise weights $w_r^{\mathrm{pt}} = m_r S_r$ and $w_r^{\mathrm{st}} = m_r(1 - S_r)$, satisfying $w_r^{\mathrm{pt}} + w_r^{\mathrm{st}} = m_r$.

Higher support favors point-wise fidelity. The left comparison in Figure 3 matches predicted and target endpoints at site r. With channel-averaged error $\ell_r^y = \mathrm{mean}_c \, |\hat{z}_{r,c}^y - z_{r,c}^y|$,

$$\mathcal{L}_{\mathrm{pt}} = \frac{\sum_r w_r^{\mathrm{pt}} \ell_r^y}{\max(\sum_r m_r, \varepsilon)}. \tag{6}$$

The fixed mask denominator preserves allocated supervision strength. The region-wise loss $\mathcal{L}_{\mathrm{st}}$ matches local means and standard deviations, as shown in Figure 3, plus spatial and temporal variation energies from weighted squared adjacent-site differences. Statistics use the full latent field: windows may cross the mask boundary, while w_r^{st} weights the discrepancy at center r. The normalized statistics discrepancy is scaled by $\sum_r w_r^{\mathrm{st}} / \max(\sum_r m_r, \varepsilon)$ and is zero for zero total weight. These constraints preserve local appearance and variation without identical values at every site. Appendix B gives the definitions and reductions.

We combine these auxiliaries with dense reconstruction $\mathcal{L}_{\mathrm{rec}}$ and background preservation $\mathcal{L}_{\mathrm{bg}}$. Reconstruction includes regional flow matching, endpoint fidelity, and target temporal-difference

Table 1: **Quantitative comparison on PROVE-M.**

Method	PSNR↑	SSIM↑	LPIPS↓	ReMOVE↑	CFD↓	RC-S↑	RC-T↓
ProPainter (Zhou et al., 2023)	22.1846	0.8768	0.1559	0.8676	**0.2774**	0.4427	0.5951
VACE (1.3B) (Jiang et al., 2025)	20.0826	0.8654	0.1545	0.8117	0.3283	0.4036	0.5217
ROSE (1.3B) (Miao et al., 2025)	26.1333	0.9003	0.1212	0.8803	0.3364	0.4924	0.6538
EffectErase (1.3B) (Fu et al., 2026)	27.0049	0.9098	0.1142	<u>0.8841</u>	0.3412	<u>0.527</u>	<u>0.2728</u>
UnderEraser (14B) (Liu et al., 2026)	<u>28.3325</u>	<u>0.9156</u>	0.0981	0.8824	0.2986	0.5188	0.3276
SVOR (1.3B) (Hu et al., 2026)	27.4289	**0.9239**	**0.0839**	0.8836	<u>0.2794</u>	0.5236	0.2987
TIE (Ours, 1.3B)	**28.9711**	0.9099	<u>0.0871</u>	**0.8849**	0.2971	**0.5294**	**0.249**

matching. Object and changed-region endpoints match the edited-target latent z^y, while preserved endpoints match the original-video latent z^x. Background preservation deliberately reinforces preserved-region velocity and source-endpoint errors with full-volume normalization, retaining background-area influence. The overall objective is

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{bg}} + \mathcal{L}_{\text{aux}}(k), \\ \mathcal{L}_{\text{aux}}(k) &= \lambda_{\text{pt}}\mathcal{L}_{\text{pt}} + \lambda_{\text{st}}\gamma(k)\mathcal{L}_{\text{st}}, \qquad \gamma(k) = \min(1, k/K_w). \end{aligned} \tag{7}$$

The coefficients balance the auxiliaries. Here $k \geq 1$ counts iterations after auxiliary activation, and $\gamma(k)$ warms up region-wise supervision over K_w iterations. We average this objective over videos, noise levels, and Gaussian noise. Appendix B details reconstruction and background losses.

3.4 Training procedure

During training, Type-Isolated Attention processes original-video, mask, and noisy-target tokens to predict velocity and estimate the clean endpoint. Paired changes and frozen cross-frame matching provide evidence for auxiliary supervision allocation. The objective in Eq. (7) combines reconstruction, background preservation, and evidence-guided auxiliaries for optimization.

4 Experiments

4.1 Experiment settings

Training dataset. We use 32K paired samples from publicly released datasets: 16K rendered samples from ROSE and 16K real-world samples randomly selected from EffectErase. Each sample contains an original video, a removal mask, and a paired target with the designated object and its associated effects removed. These targets supervise scene reconstruction and identify removal-related changes for counterfactual-valid evidence estimation.

Benchmarks and evaluation protocols. We evaluate on PROVE-M, the paired subset of PROVE-Bench (Li et al., 2026), whose 80 real-world video pairs provide aligned object-free references for assessing background reconstruction. Synchronized camera-motion augmentation maintains alignment among the input videos, masks, and ground-truth videos. We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), ReMOVE (Chandrasekar et al., 2024), CFD (Yu et al., 2025), and PROVE's RC-S and RC-T. RC-S compares local feature distributions between restored regions and surrounding background, while RC-T measures changes across jointly restored regions in adjacent frames. These metrics assess reconstruction fidelity and removal coherence.

We further evaluate on VORD-L, the extension of the dataset introduced by VOR-Bench (Huang et al., 2026), which provides 300 paired videos and graffiti masks covering diverse objects, scenes, and associated effects. We follow its Qwen3-VL-based vision-language model (VLM) (Bai et al., 2025) assessment protocol to score removal completeness, background consistency, and logical consistency, together with an overall score.

Implementation details. We build TIE on Wan2.1-Fun InP 1.3B, training only LoRA adapters (Hu et al., 2022) and type embeddings. Optimization uses AdamW on four NVIDIA A100 GPUs with 89-frame clips at 832×480 resolution. The model accumulates 13,500 optimization updates, and evaluation uses 30 denoising steps. Appendix C provides more details.

Table 2: **Vision-language model evaluation results on VORD-L.**

Method	Removal compl.↑	BG consist.↑	Logical consist.↑	Overall↑
ProPainter	3.32	3.14	3.67	3.38
VACE	3.28	3.31	3.45	3.35
ROSE	4.23	4.15	4.58	4.32
EffectErase	4.25	4.25	<u>4.65</u>	4.38
UnderEraser	4.23	4.2	4.52	4.31
SVOR	<u>4.31</u>	**4.3**	<u>4.65</u>	<u>4.42</u>
TIE (Ours)	**4.37**	<u>4.28</u>	**4.69**	**4.44**

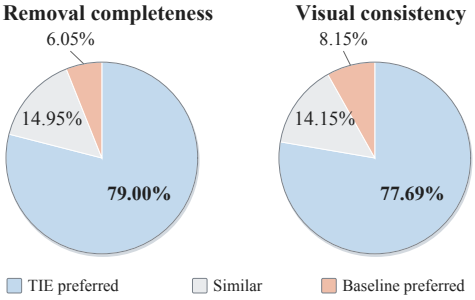

Figure 4: **Human study results.**

4.2 Quantitative comparison

Evaluation on PROVE-M. TIE achieves state-of-the-art overall performance among the evaluated methods in Table 1. Compared with recent methods such as SVOR, TIE delivers more accurate background recovery in PSNR and stronger spatial and temporal removal coherence in RC-S and RC-T. Its leading ReMOVE score further supports effective removal, demonstrating consistent advantages across reconstruction and editing quality.

VLM-based generalization evaluation. We evaluate on VORD-L to examine whether TIE's benefits extend to diverse removal scenarios under perceptually oriented VLM assessment. TIE achieves the best overall score in Table 2, leading in removal completeness and logical consistency while maintaining strong background consistency. These findings support its generalization to varied scenes, with edits that better satisfy perceptual and semantic expectations.

Analysis of quantitative results. The complementary evaluations show how reconstruction quality translates into a coherent removal result. RC-S and RC-T capture contextual compatibility and temporal consistency beyond frame-level similarity. Their improvement alongside reference-based fidelity suggests that TIE reconstructs content that both matches the reference and integrates with the surrounding scene over time. Favorable VLM judgments further connect these reconstruction benefits with complete removal and plausible scene logic. This combined evidence supports TIE's use of selective information access and support-adaptive supervision to recover a coherent scene from potentially misleading cross-frame observations.

4.3 Qualitative comparison

Visual quality. Figure 5 compares our method with representative inpainting and strong generative removal baselines. Their outputs retain foreground remnants, wakes, or mirror reflections after removal. Our method more completely removes these residual effects while preserving surrounding appearance. Full video comparisons are provided in the supplementary material.

Temporal coherence of scene reconstruction. Coherent scene reconstruction requires restored content to remain stable during playback. Ghosting, such as that in SVOR's last-row result in Figure 5, is difficult to convey through isolated frames. Figure 6 complements these views by aggregating local DINO feature discrepancies between adjacent outputs over 18 frame pairs. Brighter regions indicate larger temporal discrepancies, helping locate inconsistent reconstruction. SVOR and UnderEraser exhibit pronounced hotspots in the first and third rows, whereas TIE shows weaker discrepancies in the corresponding regions and across much of the second row. These results support TIE's effectiveness in maintaining temporal coherence after removal, complementing the visual evidence of reduced ghosting. Appendix D.4 details the computation.

Human study. We conduct an anonymous A/B comparison with representative baselines, assessing removal completeness and visual consistency with balanced method positions. Figure 4 shows that TIE is preferred in 79.00% and 77.69% of comparisons on these criteria, respectively, indicating a clear human preference for its more complete removal and coherent scene reconstruction. Detailed evaluation settings are provided in the supplementary material.



Figure 5: **Qualitative comparison.** Selected video frames with input masks marked in green. Full video comparisons are provided in the supplementary material.

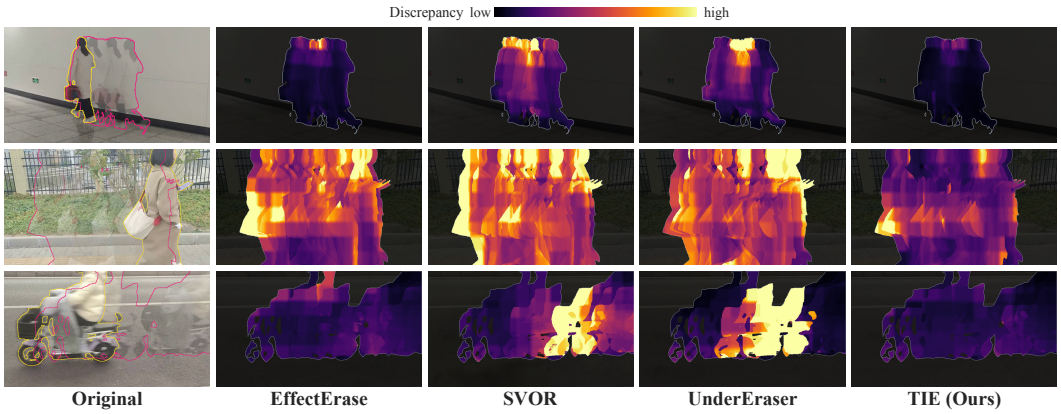

Figure 6: **Temporal discrepancies after object removal.** Each row shows one video. The left column shows an original frame with translucent earlier object positions, yellow current-mask contours, and magenta temporal-support contours. Heatmaps on the right aggregate local feature discrepancies between adjacent output frames for each method. Brighter colors indicate larger discrepancies.

4.4 Ablation Studies

Contribution of framework components. Table 3 shows the complementary benefits of input organization, attention control, and evidence-guided supervision. Token concatenation improves recovery over channel concatenation by making inputs separately accessible. Adding type-isolated attention improves perceptual quality and temporal coherence, supporting directed information flow that prevents noisy-target feedback. Evidence-guided supervision further strengthens reconstruction and coherence, demonstrating the benefit of calibrating supervision to reconstruction support.

Table 3: **Framework component ablations on PROVE-M.** Both concatenation baselines omit Type-Isolated Attention and Evidence-Guided Supervision (EGS). TIE w/o EGS adds only Type-Isolated Attention.

Method	PSNR↑	SSIM↑	LPIPS↓	ReMOVE↑	CFD↓	RC-S↑	RC-T↓
Channel concat.	23.6559	0.8521	0.1549	0.8757	0.3082	0.4945	0.4263
Token concat.	28.4314	**0.9107**	0.0943	**0.8857**	0.3022	0.5282	0.2653
TIE w/o EGS	28.477	0.9036	0.0927	0.8848	**0.2939**	0.529	0.2578
TIE (Full)	**28.9711**	0.9099	**0.0871**	0.8849	0.2971	**0.5294**	**0.249**

Table 4: **Evidence and supervision ablations on PROVE-M.** Region-wise indicates whether $\mathcal{L}_{\text{st}}$ is enabled. Valid and Raw denote counterfactual-valid estimation and local token matching, respectively. None disables cross-frame evidence correction.

Variant	Region-wise	Evidence	PSNR↑	SSIM↑	LPIPS↓	ReMOVE↑	CFD↓	RC-S↑	RC-T↓
a	–	Valid	28.7706	0.90823	0.0895	**0.8849**	0.2992	0.5285	0.253
b	✓	Raw	28.7677	0.90821	0.0893	0.8847	0.2959	0.5288	0.25
c	✓	None	28.7628	0.90816	0.0894	0.8847	**0.2953**	0.5284	0.2531
TIE	✓	Valid	**28.9711**	**0.9099**	**0.0871**	**0.8849**	0.2971	**0.5294**	**0.249**

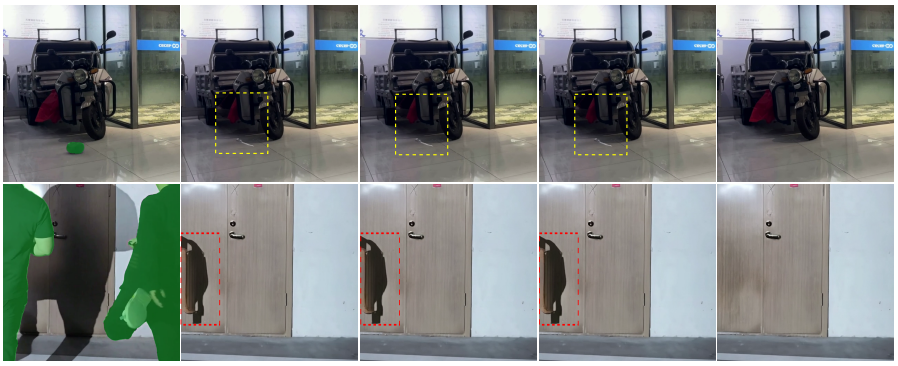

Figure 7: **Qualitative supervision ablations.** Columns show the original video, variants a–c from Table 4, and TIE. Green overlays mark input masks, and dashed boxes highlight differences.

Evidence and supervision allocation. We evaluate region-wise supervision and evidence estimation in Table 4: variant a removes region-wise matching, b uses raw local matching, and c retains only geometric support. TIE achieves the best reconstruction fidelity, perceptual quality, and temporal coherence, supporting the effectiveness of counterfactual-valid evidence beyond raw matching or geometry alone and the complementary role of region-wise supervision. Figure 7 further illustrates these benefits: columns 2–4 alter unaffected content in the first row and retain dark remnants in the second, whereas TIE in column 5 better preserves the scene and completes removal.

5 Conclusion

We presented TIE, a video object removal framework that integrates Type-Isolated Attention and Evidence-Guided Supervision for coherent scene reconstruction. Experiments on PROVE-M and VORD-L demonstrate state-of-the-art overall performance among the evaluated methods, with component ablations supporting the key designs. The results support selective information access and support-adaptive supervision for coherent scene reconstruction while removing objects and their associated effects. Building on this framework, future work will explore coordinated removal of multiple objects with overlapping effects, extending evidence-guided reconstruction to more complex object interactions.

6 AI USAGE DISCLOSURE

In this work, we used generative AI tools to assist with identifying and reviewing relevant literature and checking for potential overlap between our independently developed method and prior work. Additionally, we used these tools to make minor adjustments to individual components of scientific figures and to polish the wording of a small portion of the manuscript for clarity and readability. We did not use generative AI for any of the tasks designated as requiring disclosure under the conference policy. We reviewed all AI-assisted outputs: literature-related suggestions were checked against the original sources, and figure adjustments and wording changes were examined to ensure that they accurately reflected our work and preserved the intended meaning. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025. doi: 10.1145/3721238.3730673. URL https://doi.org/10.1145/3721238.3730673.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a. URL https://arxiv.org/abs/2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Blattmann_Align_Your_Latents_High-Resolution_Video_Synthesis_With_Latent_Diffusion_Models_CVPR_2023_paper.html.

Aditya Chandrasekar, Goirik Chakrabarty, Jai Bardhan, Ramya Hebbalaguppe, and Prathosh AP. ReMOVE: A reference-free metric for object erasure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7901–7910, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/EvGenFM/papers/Chandrasekar_ReMOVE_A_Reference-free_Metric_for_Object_Erasure_CVPRW_2024_paper.pdf.

Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3D gated convolution and temporal PatchGAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9066–9075, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Chang_Free-Form_Video_Inpainting_With_3D_Gated_Convolution_and_Temporal_PatchGAN_ICCV_2019_paper.html.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. URL https://arxiv.org/abs/2310.19512.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_VideoCrafter2_Overcoming_Data_Limitations_for_High-Quality_Video_Diffusion_Models_CVPR_2024_paper.html.

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*, 2026. URL https://henghuiding.com/EffectErase/.

Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision*, pp. 713–729. Springer, 2020. URL https://www.chengao.vision/FGVC/.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/94894cf9bab20c00f6eaaff3b7f2be70-Abstract-Conference.html.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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. *arXiv preprint arXiv:2603.09283*, 2026. URL https://arxiv.org/abs/2603.09283.

Haonan Huang, Tianrui Qiu, Xianghao Zang, Yinan Du, Zhixiang He, Chi Zhang, Hao Sun, Zhongjiang He, Tianwei Cao, Xuchong Zhang, Hongbin Sun, Kongming Liang, and Zhanyu Ma. VOR-Bench: A human perception-driven benchmark for video object removal. *arXiv preprint arXiv:2609.16878*, 2026. URL https://arxiv.org/abs/2609.16878.

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*, pp. 17191–17202, 2025. URL https://ali-vilab.github.io/VACE-Page/.

Jaeyeon Kang, Seoung Wug Oh, and Seon Joo Kim. Error compensation framework for flow-guided video inpainting. In *European Conference on Computer Vision*, pp. 375–390. Springer, 2022. doi: 10.1007/978-3-031-19784-0_22. URL https://link.springer.com/chapter/10.1007/978-3-031-19784-0_22.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,

pp. 4413–4421, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Lee_Copy-and-Paste_Networks_for_Deep_Video_Inpainting_ICCV_2019_paper.html.

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*, pp. 12522–12532, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_Generative_Omnimatte_Learning_to_Decompose_Video_into_Layers_CVPR_2025_paper.html.

Fuhao Li, Shaofeng You, Jiagao Hu, Yu Liu, Yuxuan Chen, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. PROVE: A perceptual RemOVal cohErence benchmark for visual media. In *Proceedings of the ACM International Conference on Multimedia*, 2026. URL https://arxiv.org/abs/2605.14534. Accepted for publication.

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*, pp. 17562–17571, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_Towards_an_End-to-End_Framework_for_Flow-Guided_Video_Inpainting_CVPR_2022_paper.html.

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*, pp. 23471–23480, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Lin_OmnimatteRF_Robust_Omnimatte_with_3D_Background_Modeling_ICCV_2023_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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

Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. FuseFormer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14040–14049, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_FuseFormer_Fusing_Fine-Grained_Information_in_Transformers_for_Video_Inpainting_ICCV_2021_paper.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

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*, pp. 4507–4515, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lu_Omnimatte_Associating_Objects_and_Their_Effects_in_Video_CVPR_2021_paper.html.

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. *arXiv preprint arXiv:2508.18633*, 2025. URL https://arxiv.org/abs/2508.18633.

Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *Proceedings of the IEEE/CVF International Conference on Computer*

Vision, pp. 4403–4412, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Oh_Onion-Peel_Networks_for_Deep_Video_Completion_ICCV_2019_paper.html.

Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. DLFormer: Discrete latent transformer for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3511–3520, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Ren_DLFormer_Discrete_Latent_Transformer_for_Video_Inpainting_CVPR_2022_paper.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408. URL https://doi.org/10.1080/01621459.1993.10476408.

Mohammed Suhail, Erika Lu, Zhengqi Li, Noah Snavely, Leonid Sigal, and Forrester Cole. Omnimatte3D: Associating objects and their effects in unconstrained monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 630–639, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Suhail_Omnimatte3D_Associating_Objects_and_Their_Effects_in_Unconstrained_Monocular_Video_CVPR_2023_paper.html.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://www.cns.nyu.edu/~lcv/ssim/.

Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D^2NeRF: Self-supervised decoupling of dynamic and static objects from a monocular video. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32653–32666, 2022. URL https://papers.nips.cc/paper/2022/hash/d2cc447db9e56c13b993c11b45956281-Abstract-Conference.html.

Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Xu_Deep_Flow-Guided_Video_Inpainting_CVPR_2019_paper.html.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18456–18466, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yu_Video_Probabilistic_Diffusion_Models_in_Projected_Latent_Space_CVPR_2023_paper.html.

Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. OmniPaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17324–17334, 2025. URL https://arxiv.org/abs/2503.08677.

Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision*, pp. 528–543. Springer, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/2590_ECCV_2020_paper.php.

Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2720–2729, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Zhang_An_Internal_Learning_Approach_to_Video_Inpainting_ICCV_2019_paper.html.

Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*, pp. 74–90. Springer, 2022a. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/1456_ECCV_2022_paper.php.

Kaidong Zhang, Jingjing Fu, and Dong Liu. Inertia-guided flow completion and style fusion for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5982–5991, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhang_Inertia-Guided_Flow_Completion_and_Style_Fusion_for_Video_Inpainting_CVPR_2022_paper.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://richzhang.github.io/PerceptualSimilarity/.

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*, pp. 10477–10486, 2023. URL https://arxiv.org/abs/2309.03897.

Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16448–16457, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zou_Progressive_Temporal_Feature_Alignment_Network_for_Video_Inpainting_CVPR_2021_paper.html.

A EVIDENCE CONSTRUCTION DETAILS

We compute candidate validity from paired RGB changes before VAE encoding. For a latent grid of size $F_z \times H_z \times W_z$, we uniformly sample F_z frames from the preprocessed source and target clips and bilinearly resize them to $4H_z \times 4W_z$. Denote these RGB tensors, scaled to $[0, 1]$, by $\widetilde{X}$ and $\widetilde{Y}$. We average the channel-wise squared difference, take its square root, and pool the result onto the latent grid:

$$\delta = \mathrm{AvgPool}_{F_z,H_z,W_z}\left[\sqrt{\frac{1}{3}\sum_{c=1}^{3}(\widetilde{X}_c - \widetilde{Y}_c)^2}\right]. \tag{8}$$

Here c indexes the three color channels and $\mathrm{AvgPool}_{F_z,H_z,W_z}$ denotes adaptive average pooling to the indicated grid. To account for pair discrepancies, let $\bar{\delta}_m = \mathcal{R}(\delta; m)$ be the mean masked change, $\delta_{\mathrm{med}} = \mathrm{median}_{m=0}(\delta)$ the outside-mask median, and $\delta_{\mathrm{MAD}} = \mathrm{median}_{m=0}\,|\delta - \delta_{\mathrm{med}}|$ its median absolute deviation. The threshold is

$$\tau_\delta = \max\{\kappa\bar{\delta}_m,\ \delta_{\mathrm{med}} + 1.4826\lambda_{\mathrm{MAD}}\delta_{\mathrm{MAD}},\ \delta_{\min}\}. \tag{9}$$

Here κ scales the masked change, λ_{MAD} controls the outside-mask noise threshold, and $\delta_{\min}$ is a lower bound. For an empty mask, the global 95th percentile replaces $\bar{\delta}_m$. For a full mask, we use the global median and set $\delta_{\mathrm{MAD}} = 0$. Together with Eq. (3), this gives $p_r = 1 - m_r - e_r$.

We extract unit-normalized features $f_i^x = \mathrm{norm}(\mathcal{P}(z^x)_i)$ and $f_i^y = \mathrm{norm}(\mathcal{P}(z^y)_i)$. For target token $i = (t, h, w)$, temporal band b includes source times $t - b$ and $t + b$ and spatial offsets up to r_b:

$$\begin{aligned}\mathcal{N}_b(i) = \{j = (t \pm b, h + \Delta h, w + \Delta w) : |\Delta h|, |\Delta w| \le r_b,\ j \text{ eligible}\},\\ (b, r_b) \in \{(1, 1), (2, 1), (4, 2), (8, 2)\}.\end{aligned} \tag{10}$$

Here j indexes a source token, $(\Delta h, \Delta w)$ is its spatial displacement from i, and r_b is the band-specific search radius. All offsets use the transformer token grid produced by $\mathcal{P}$, not the VAE latent grid or raw video frames. To define eligibility, we nearest-neighbor resample the mask to the token grid, dilate it spatially by one token, and trilinearly resample p. Candidates must lie outside this mask and satisfy $p_j \ge \tau_p = 0.85$. Out-of-bounds candidates are excluded without wrapping.

Let $s_{ib}^{(1)}$ and $s_{ib}^{(2)}$ be the two largest cosine similarities $(f_i^y)^\top f_j^x$ in band b, and let $p_{ib}^\star$ be the preservation weight of the best match. We compute

$$a_{ib} = \chi_{ib}\, p_{ib}^\star\, \mathrm{sigmoid}\left(\frac{s_{ib}^{(1)} - \tau_s}{T_s}\right)\mathrm{sigmoid}\left(\frac{s_{ib}^{(1)} - s_{ib}^{(2)} - \tau_m}{T_m}\right). \tag{11}$$

The similarity and margin thresholds are τ_s and τ_m, with temperatures T_s and T_m. The indicator χ_{ib} requires two usable matches. The implementation initializes unavailable similarities to -1 and tests $s_{ib}^{(2)} > -0.999$.

Writing $a_{i,(2)}$ for the second-largest score across bands, we aggregate the evidence as

$$c_i^{\mathrm{tok}} = a_{i,(2)}\,\mathbf{1}[a_{i,(2)} \ge \tau_c], \qquad c_r = m_r\ \mathrm{Interp}(c^{\mathrm{tok}})_r. \tag{12}$$

Scores below $\tau_c = 0.20$ are discarded. The second-largest score requires support from at least two bands, which need not lie on opposite sides of the query time. Trilinear interpolation maps evidence to latent sites, and multiplication by m_r restricts it to the mask. The frozen feature extraction and evidence calculation use no gradient.

B LOSS DEFINITIONS AND REDUCTIONS

The following reductions are defined per video, matching the local batch size of one. We use $\varepsilon = 10^{-6}$. Let Ω be the set of latent sites indexed by r and C_z the number of latent channels. For a nonnegative region weight ν, its mass is $M_\nu = \sum_{r\in\Omega} \nu_r$. The region average of scalar site error f_r is

$$\mathcal{R}(f;\nu) = \mathbf{1}[M_\nu > \varepsilon]\, \frac{\sum_{r\in\Omega} \nu_r f_r}{\max(M_\nu, \varepsilon)}. \tag{13}$$

Errors are first averaged over channels. A region with mass no greater than ε contributes zero. To match temporal changes, define $\Delta_t z_t = z_{t+1} - z_t$ and use the union weight $\nu_t^{\cup} = \max(\nu_{t+1}, \nu_t)$ for each adjacent pair:

$$\mathcal{T}(\nu) = \mathcal{R}\left(\frac{1}{C_z}\left\|\Delta_t \hat{z}^y - \Delta_t z^y\right\|_1 ; \nu^{\cup}\right). \tag{14}$$

The union includes sites belonging to the weighted region in either adjacent frame. For channel-valued f, define

$$\langle f \rangle_w = \frac{\sum_{r,c} w_r f_{r,c}}{C_z \max(\sum_r w_r, \varepsilon)}. \tag{15}$$

For region-wise matching, $\mu(z)$ and $s(z)$ are channel-wise means and standard deviations in $1\times7\times7$ windows, with padded entries excluded. We compute $s(z) = \sqrt{\max(\mathrm{AvgPool}(z^2) - \mu(z)^2, 0)}$. Moments use the full latent field before center-wise weighting by w, so windows can cross the mask boundary as illustrated in Figure 3.

Let $\mathcal{A}$ contain all axes in $\{t, h, w\}$ with length greater than one. For each $a \in \mathcal{A}$, define adjacent sites $\Omega_a = \{r : r + \mathbf{e}_a \in \Omega\}$, where $\mathbf{e}_a$ is a unit step, and pair weights $w_r^{(a)} = \min(w_r, w_{r+\mathbf{e}_a})$. The difference energy is

$$E_a(z;w) = \frac{\sum_{r\in\Omega_a,c} w_r^{(a)} (z_{r+\mathbf{e}_a,c} - z_{r,c})^2}{C_z \max(\sum_{r\in\Omega_a} w_r^{(a)}, \varepsilon)}. \tag{16}$$

Pair weights require both endpoints to contribute. With $w = w^{\mathrm{st}}$, the region-wise objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{st}} = \rho_w \Bigg[&\beta_\mu \left\langle |\mu(\hat{z}^y) - \mu(z^y)| \right\rangle_w + \beta_s \left\langle |s(\hat{z}^y) - s(z^y)| \right\rangle_w \\ &+ \frac{\beta_E}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \min\left\{ K, \left| \log \frac{E_a(\hat{z}^y; w) + \varepsilon}{E_a(z^y; w) + \varepsilon} \right| \right\} \Bigg], \\ \rho_w = &\frac{\sum_r w_r}{\max(\sum_r m_r, \varepsilon)}.\end{aligned} \tag{17}$$

The coefficients weight mean, standard-deviation, and variation-energy matching. The cap K bounds each log-energy discrepancy, while ρ_w preserves the allocated mask fraction. Empty allocations contribute zero. Axes of length one are excluded, and the energy term is zero if no axis remains.

For the reconstruction objective, use channel-averaged errors $\ell_r^v = C_z^{-1}\|v_{\theta,r} - u_r\|_2^2$, $\ell_r^y = C_z^{-1}\|\hat{z}_r^y - z_r^y\|_1$, and $\ell_r^x = C_z^{-1}\|\hat{z}_r^y - z_r^x\|_1$. The objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{rec}} = \omega(\sigma) &\sum_{\nu\in\{m,e,p\}} \lambda_F^\nu \mathcal{R}(\ell^v; \nu) \\ &+ \lambda_{\mathrm{edit}}\big[\mathcal{R}(\ell^y; m) + \mathcal{R}(\ell^y; e)\big] + \lambda_{\mathrm{keep}}\mathcal{R}(\ell^x; p) \\ &+ \lambda_T\big[\mathcal{T}(m) + \mathcal{T}(e) + \lambda_F^p \mathcal{T}(p)\big].\end{aligned} \tag{18}$$

Here $\omega(\sigma)$ is the backbone scheduler's training weight. Velocity and temporal differences match the edited target. Endpoint losses match z^y in object and changed regions and z^x in preserved regions. Background preservation retains region area through full-volume normalization:

$$\mathcal{L}_{\mathrm{bg}} = \frac{1}{|\Omega|} \sum_{r \in \Omega} p_r \left(\eta_v \ell_r^v + \eta_x \ell_r^x \right). \tag{19}$$

Unlike the regional flow term, this background term does not apply $\omega(\sigma)$. The point-wise auxiliary uses the fixed mask denominator in Eq. (6). Neither endpoint objective applies an additional noise multiplier. Their noise dependence follows from

$$\hat{z}^y - z^y = \sigma(u - v_\theta), \tag{20}$$

so evidence-dependent allocation supplements the dense endpoint constraints rather than replacing them.

C Implementation settings

Architecture and conditioning. We use Wan2.1-Fun InP 1.3B with 30 transformer layers, 12 heads, width 1536, and a $1 \times 2 \times 2$ patch kernel. Each 16-channel latent stream is zero-padded to the frozen projection's 36 input channels. Type-Isolated Attention applies in every self-attention block. Rank-32 LoRA adapters update the query, key, value, and output projections of self- and cross-attention and both feed-forward linear layers. Three learned type embeddings distinguish the inputs. The backbone weights, VAE, and patch projection remain frozen.

Training and inference use the fixed instruction "Remove the object indicated by the white mask while preserving the background and motion." The tokenizer and text encoder provide c_{txt} through the backbone's text-conditioning pathway.

Optimization. Training clips contain 89 frames at 832×480. The reported model accumulates 13,500 optimization updates. We use AdamW with weight decay 0.01. The combined-objective optimization uses four devices with one video per device and gradient accumulation one. Region-wise supervision uses the warmup defined in Eq. (7). The evidence branch provides detached supervision and is omitted at inference.

Inference. Reported PROVE-M evaluation uses 30 sampling steps, guidance scale 1.0, scheduler shift 5.0, and seed 42. For a descending schedule $1 = \sigma_J > \cdots > \sigma_0 = 0$, sampling begins with Gaussian target noise and follows

$$z_{\sigma_{j-1}} = z_{\sigma_j} + (\sigma_{j-1} - \sigma_j) v_\theta(z_{\sigma_j}, \sigma_j; z^x, z^m, c_{\mathrm{txt}}), \qquad \widehat{Y} = \mathcal{D}(z_{\sigma_0}). \tag{21}$$

Here $\mathcal{D}$ is the frozen VAE decoder. The original-video and mask latents remain fixed conditioning inputs throughout sampling.

D Additional experimental details

D.1 Human evaluation

Participants and presentation. We collect evaluations from 40 researchers in artificial intelligence and computer vision. We use 30 in-the-wild clips collected from Pexels, with identical clips and removal masks for all methods. For each clip, we pair TIE with each of four baselines, ProPainter, EffectErase, SVOR, and UnderEraser, in separate anonymous A/B comparisons. This yields 120 comparisons, distributed across four questionnaires with one comparison per clip in each. As illustrated in Figure 8, each question presents three synchronized views: the original video on the left, with the removal target highlighted by a translucent green mask, and anonymous outputs A and B in the middle and on the right. Method names are hidden, and TIE appears equally often in positions A and B within each questionnaire and across each baseline's comparisons to reduce identity and position bias.

Evaluation instructions. Participants first inspect the original video, then watch both outputs in full, replaying them as needed. They assess two criteria independently. *Removal completeness* concerns removal of the designated object and its associated visible effects, including effects beyond

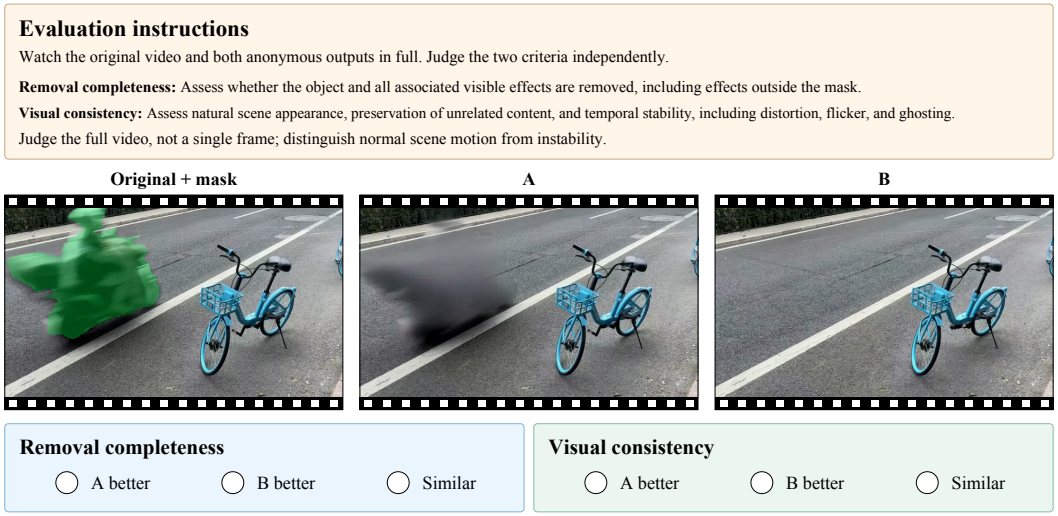

Figure 8: **Overview of human study.** Participants compare anonymous video outputs independently for removal completeness and visual consistency.

Table 5: **Auxiliary supervision on PROVE-M.** Rows a and TIE match Table 4. Uniform point-wise sets $S_r = 1$ in variant a.

Variant	PSNR↑	SSIM↑	LPIPS↓	ReMOVE↑	CFD↓	RC-S↑	RC-T↓
a (Calibrated weights)	28.7706	0.90823	0.0895	0.8849	0.2992	0.5285	0.253
Uniform point-wise	28.7727	0.90822	0.0895	**0.885**	**0.2931**	0.529	**0.249**
TIE (Full supervision)	**28.9711**	**0.9099**	**0.0871**	0.8849	0.2971	**0.5294**	**0.249**

the marked region. *Visual consistency* concerns natural scene appearance, preservation of unrelated content, and temporal stability, including blur, distortion, implausible textures, hallucinations, flicker, and ghosting. Participants judge the entire video, distinguish normal camera or scene motion from instability, and assess plausibility without guessing the hidden ground-truth background. Each criterion offers A better, B better, or similar, and choices may differ between criteria.

Responses are pooled separately for each criterion, retaining ties and excluding unanswered items.

D.2 Additional point-wise supervision ablation

We extend Table 4 with a uniform point-wise auxiliary under the same initialization, data, and evaluation settings. Starting from variant a, we set $S_r = 1$, replacing calibrated weights $m_r S_r$ with m_r. The region-wise auxiliary remains disabled, and dense reconstruction and background preservation are unchanged. This comparison isolates point-wise reweighting without region-wise supervision. Table 5 includes a and TIE from the main table as references. TIE achieves better reconstruction fidelity and perceptual quality than both point-wise-only auxiliary configurations, supporting the complementary supervision design.

D.3 Cross-frame input sensitivity

Figure 1 compares cross-frame input sensitivity under matched settings. At step 15 of a 30-step denoising trajectory, we fix the noisy target and differentiate aligned regions in three consecutive RGB frames decoded from the predicted clean endpoint with respect to the original-video latent, through the VAE decoder and DiT. Four shared Rademacher projections per frame estimate squared gradient energy, summed over latent channels and averaged across projections and frames.

We normalize each map by its total energy across source latent positions and display an earlier latent slice over a representative input frame with a shared logarithmic mapping. This temporally com-



Figure 9: **More video object removal results from TIE.**

pressed slice represents source-latent sensitivity, not sensitivity to an individual RGB input frame. The maps characterize a single-step prediction rather than the full denoising trajectory. Reduced masked-region responses with retained boundary context suggest more selective contextual use; normalized maps compare spatial distributions, not absolute sensitivity or evidence reliability.

D.4 Temporal discrepancy visualization

Revealing temporal inconsistencies. Figure 6 complements individual output frames by localizing temporal discrepancies that may appear as ghosting or unstable content during playback. All methods use the same input masks, evaluated regions, and display settings.

Measuring changes between adjacent outputs. For adjacent outputs, we extract DINO features from a crop around their mask union. Within overlapping local windows, Gaussian-kernel maximum mean discrepancy compares feature distributions at sites covered by both masks. Windows with fewer than five selected tokens are skipped. The window size and stride scale with the feature-grid resolution as $\max(1, \lfloor 16q \rfloor)$ and $\max(1, \lfloor 8q \rfloor)$, where $q = \max(H_f, W_f)/64$. We average overlapping scores, resize the map to the crop, and restrict it to the mask intersection.

Mapping discrepancies over time. We accumulate maps over up to 18 adjacent frame pairs and divide each pixel's sum by the number of pairs whose mask intersection covers it. Their union defines the displayed temporal support. All methods share a linear color scale capped at the pooled 95th percentile across methods and frames within each video. Brighter regions indicate larger temporal feature discrepancies, localizing unstable residuals discussed in the main text. Yellow contours mark the current mask, magenta contours mark temporal support, and translucent earlier object positions provide motion context. Pixels outside this support are dimmed rather than assigned a zero discrepancy. The maps use image coordinates without motion compensation and serve as spatial diagnostics, rather than the aggregate RC-T score or a dedicated ghosting detector.

E Additional visual results

Figure 9 presents additional object removal results from TIE. Figures 10–13 extend the qualitative comparisons in Figure 5 with more frames. All sequences are sampled at fixed temporal intervals. The supplementary archive also provides video demonstrations. We encourage readers to view these videos to better assess removal quality and temporal coherence during playback.

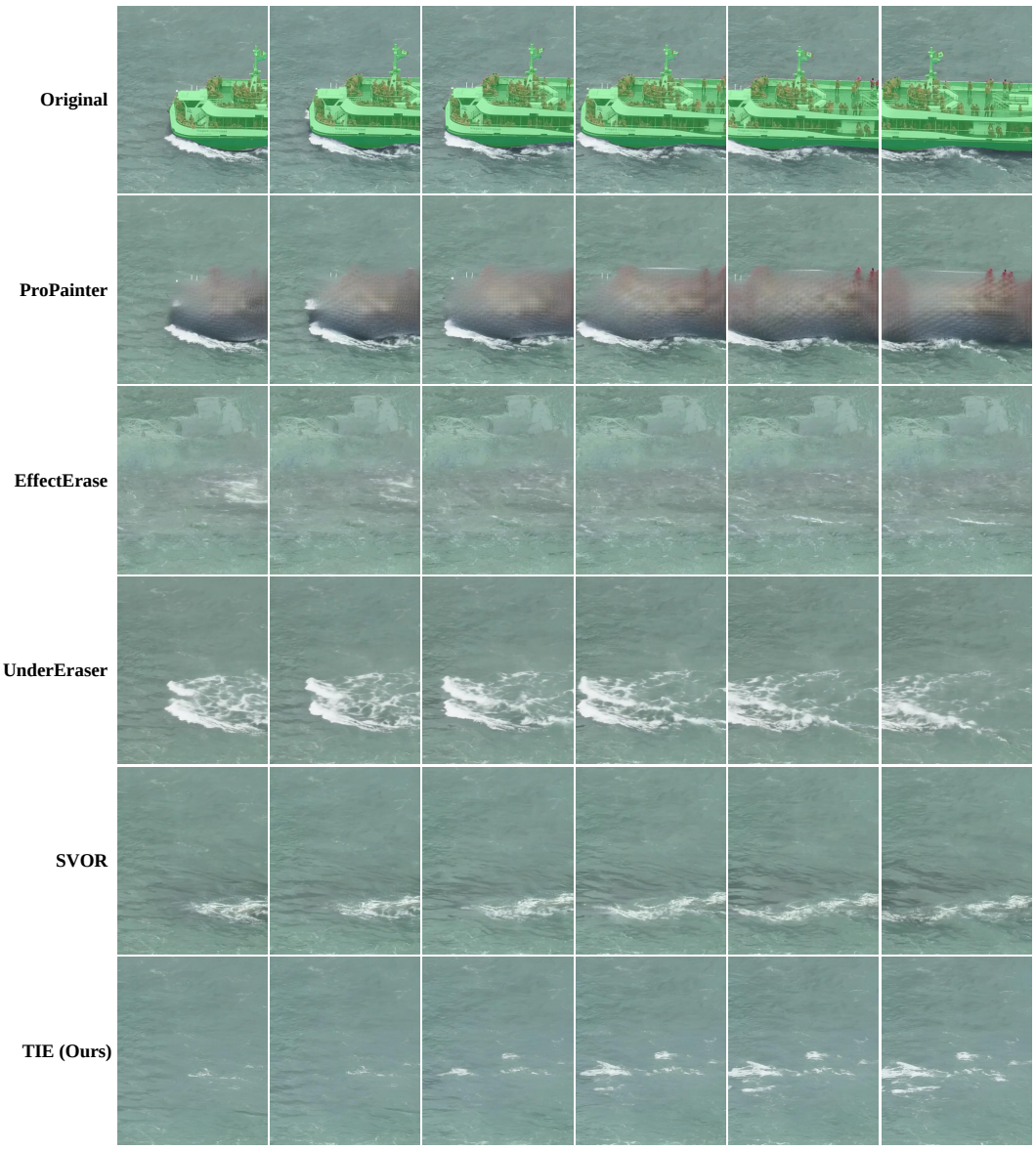

Figure 10: **More frames of qualitative comparison.**

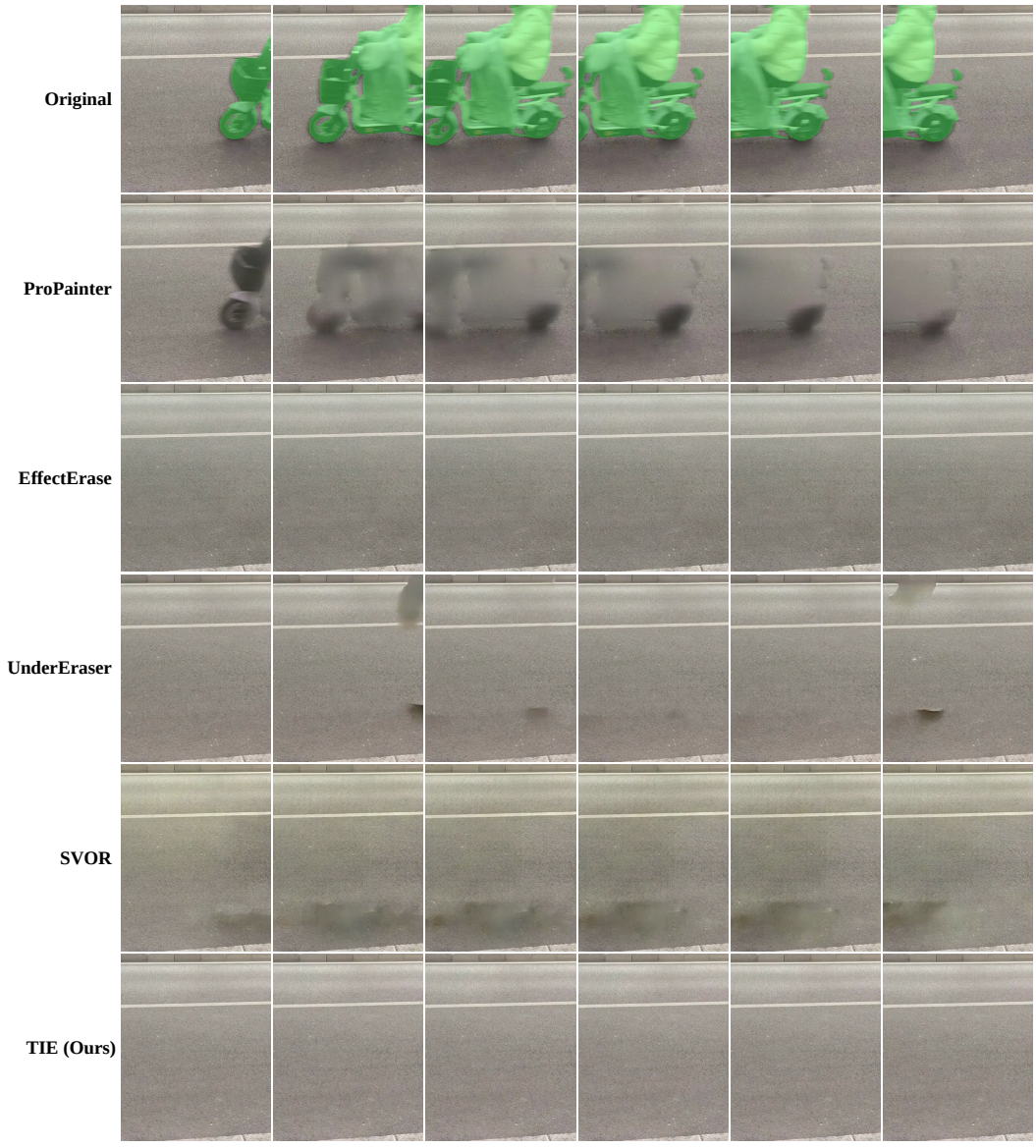

Figure 11: **More frames of qualitative comparison.**

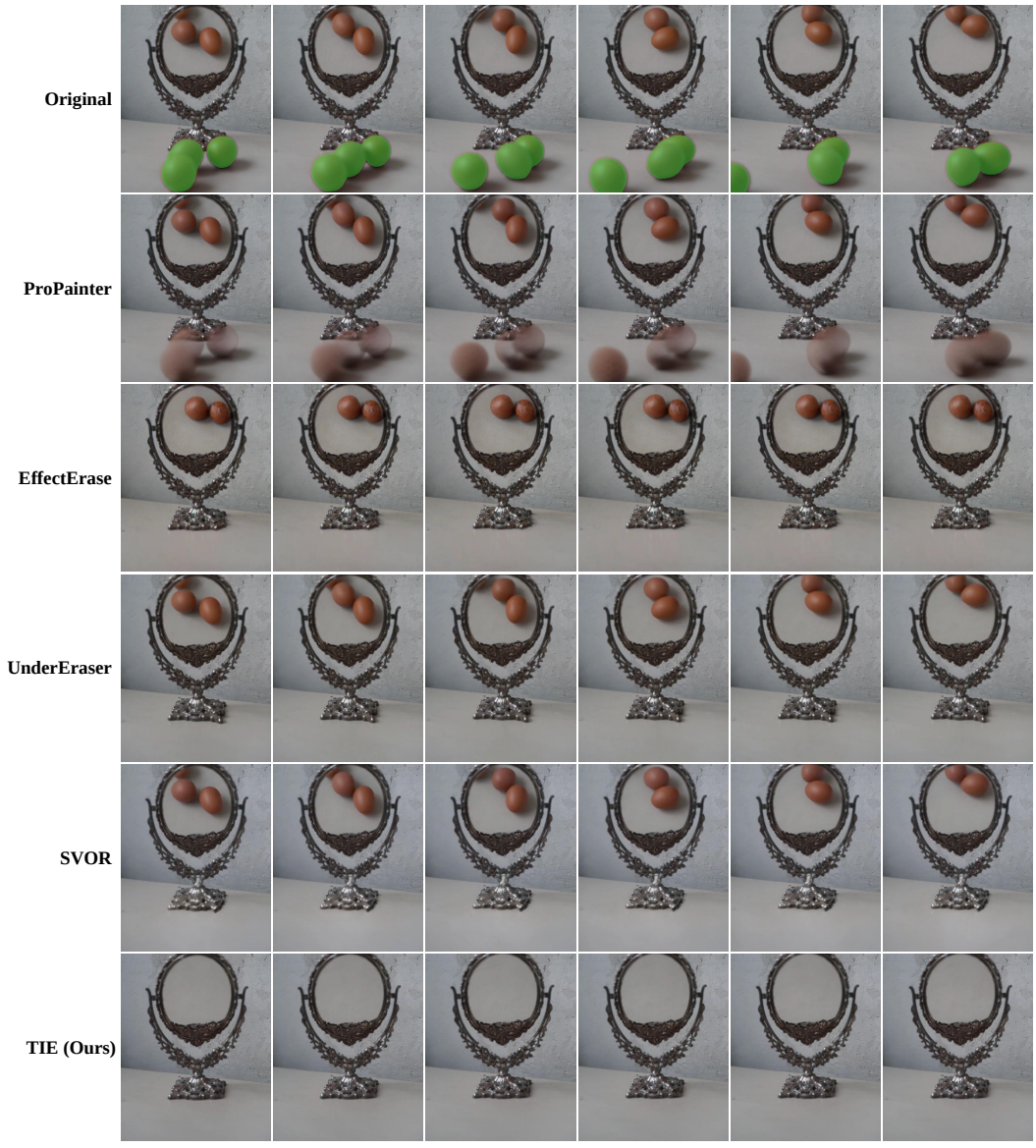

Figure 12: **More frames of qualitative comparison.**

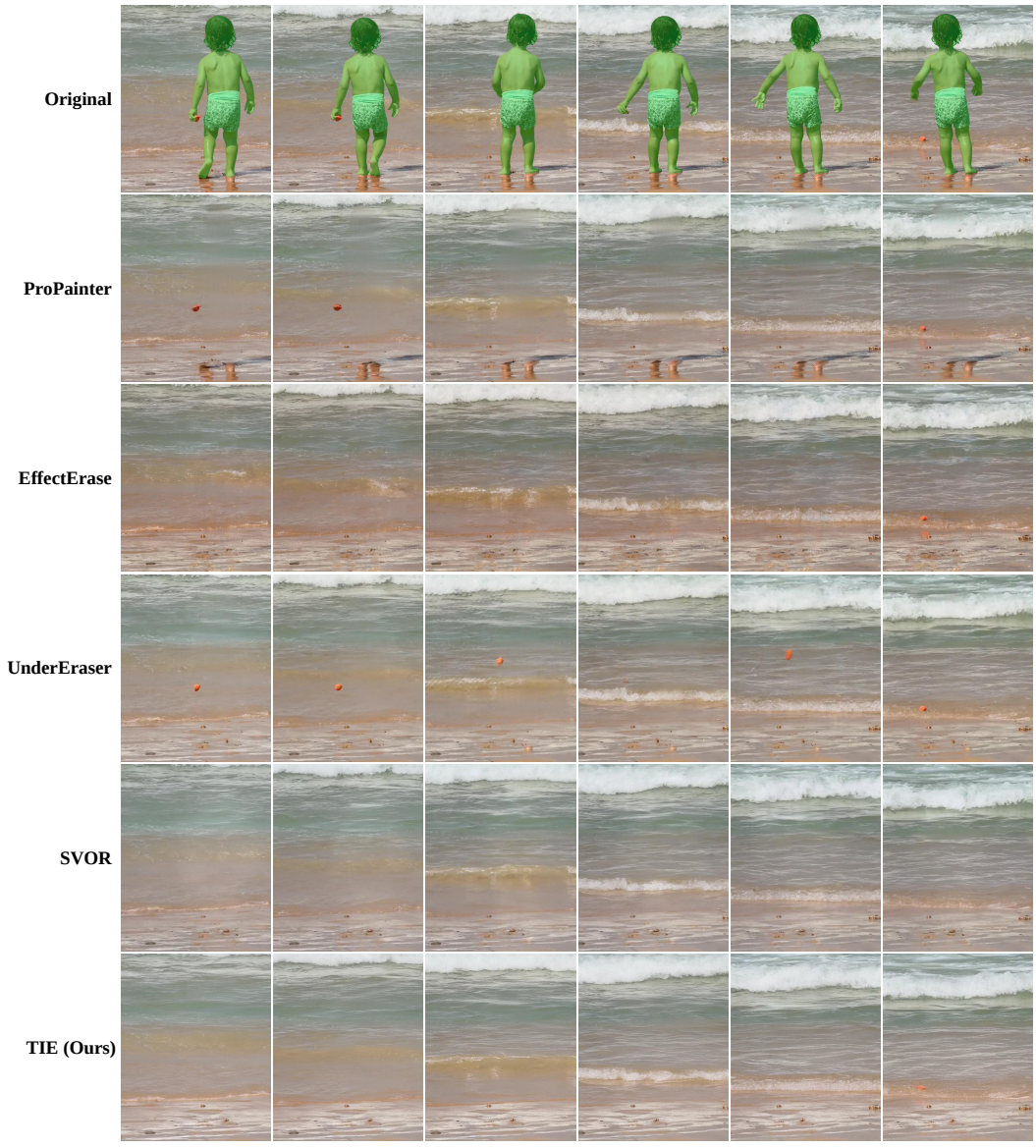

Figure 13: **More frames of qualitative comparison.**