BeyondMasks: Evaluating Causal and Physical Consistency in Video Object Removal

Yiğit Ekin[1†], Enes Sanli[2,4†], Aykut Erdem[2,4], Erkut Erdem[3,4], and Aysegul Dundar[1]

[1] Bilkent University, Department of Computer Engineering, Ankara, Türkiye
[2] Koç University, Department of Computer Engineering, Istanbul, Türkiye
[3] Hacettepe University, Department of AI and Data Engineering, Ankara, Türkiye
[4] KUIS AI Center, Istanbul, Türkiye
yigit.ekin@bilkent.edu.tr

Project Page · Code · Dataset URL

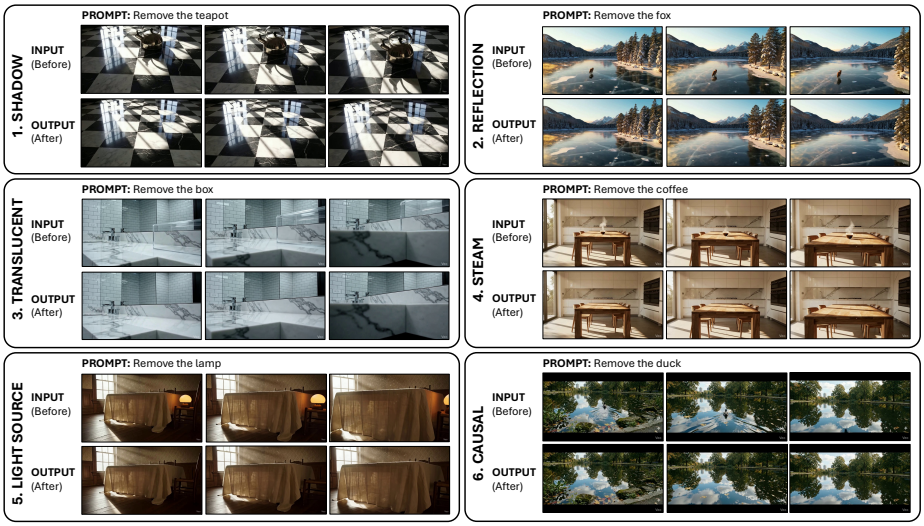

Fig. 1: Video object removal requires eliminating both objects and their physical after-effects. Objects interact with their environment through shadows, reflections, translucency, illumination changes, volumetric effects (e.g., steam), and dynamic physical traces. Removing the object alone does not restore the scene if these effects persist. BeyondMasks introduces a benchmark to evaluate such causal object-environment interactions using temporally aligned video pairs, enabling systematic assessment of whether methods recover the scene as if the object had never been present.

Abstract. Recent advances in generative video models have significantly improved visual realism in video object removal, yet evaluation protocols still focus on masked-region fidelity, treating removal as local inpainting. In real scenes, object removal is a causal intervention:

† Denotes equal contribution.

eliminating an object also requires removing its induced physical effects, such as shadows, reflections, illumination changes, translucency, and dynamic traces. Existing benchmarks lack aligned clean references or remain limited to simplified synthetic settings, preventing systematic evaluation of causal consistency. We introduce *BeyondMasks*, a paired benchmark for causally consistent video object removal, consisting of temporally aligned synthetic and real-world video pairs with clean background references. The dataset spans diverse photometric, geometric, volumetric, and dynamic interactions, and supports both mask-based and instruction-driven editing. We further propose *CORE*, a structured vision–language model-based evaluation protocol that jointly measures object disappearance and after-effect consistency, aligning more closely with human judgments than existing metrics. Benchmarking state-of-the-art methods reveals systematic failures in removing secondary physical effects despite high masked-region fidelity, exposing a gap between visual plausibility and causal correctness. BeyondMasks reframes video object removal as causal scene consistency rather than local reconstruction and provides a unified framework for its evaluation.

Keywords: Video object removal with after effect· Benchmark dataset · Paired video evaluation · Vision-language models

1 Introduction

Video object removal aims to edit a video such that a selected object appears as if it were never present. With the rise of large-scale generative video models [5, 20, 24, 25], recent systems achieve striking perceptual realism, enabling increasingly sophisticated scene manipulation. Modern video foundation models further integrate instruction-driven editing [9] and open-world reasoning, positioning object removal [14] as a core capability within general-purpose video generation pipelines. However, evaluation protocols have not kept pace with model capabilities. Most existing benchmarks assess reconstruction fidelity only within the masked region, implicitly framing object removal as a localized inpainting problem.

In real scenes, object removal is fundamentally a *causal intervention*. Objects do not merely occupy pixels; they interact with their environment across space and time. As illustrated in Fig. 1, they cast shadows, alter illumination, appear in reflections, scatter light through translucent materials, and induce dynamic traces such as footprints or dust displacements. Removing an object therefore requires eliminating both the object itself and the downstream physical effects it induces. Evaluating masked-region reconstruction alone overlooks this broader causal footprint and fails to measure whether the scene has been physically restored.

Current benchmarks are ill-suited for this purpose. Widely used video datasets such as DAVIS [15] and YouTube-VOS [23] lack paired clean backgrounds, preventing full removal verification. More recent benchmarks like ROSE-Bench [14]

introduce synthetic paired data to study specific physical side effects. While representing important progress, these datasets remain limited in interaction diversity and real-world variability, and typically focus on isolated photometric effects. They do not systematically evaluate dynamic object-environment interactions, nor do they naturally support instruction-driven editing pipelines increasingly used in modern video foundation models.

Motivated by these limitations, we advocate a causally grounded formulation of video object removal. Under this view, the task requires disentangling the target object from the scene and undoing its induced environmental changes consistently over time. This perspective shifts evaluation from local reconstruction quality toward object-environment consistency and temporal physical reasoning. To enable this paradigm, we introduce *BeyondMasks*, a diagnostically focused benchmark for evaluating causal and physical consistency in video object removal. BeyondMasks contains a collection of temporally aligned paired videos, including synthetic sequences generated through controlled editing and real-world tripod-captured scenes. Rather than maximizing scale, we prioritize precise causal control, physical diversity, and clean reference alignment, enabling systematic evaluation of object-environment disentanglement. Each sequence provides a clean reference background, per-frame object masks, and instruction prompts, supporting both mask-based and text-guided editing scenarios.

Beyond paired pixel-level evaluation, we propose CORE (Causal Object Removal Evaluation) score, a vision-language model (VLM)-based evaluation metric that jointly measures object disappearance and after-effect consistency. We validate this protocol through human studies and inter-model agreement analysis, demonstrating its reliability and alignment with perceptual judgments.

Extensive benchmarking of state-of-the-art mask-based and instruction-driven models reveals systematic failures in removing secondary physical effects, even when masked-region fidelity remains high. These findings expose a substantial gap between visual plausibility and causal scene consistency, highlighting the need for evaluation frameworks that capture physical reasoning rather than local reconstruction alone.

The main contributions of this paper can be summarized as follows:

- We introduce a causally grounded formulation of video object removal, reframing evaluation around object-environment consistency rather than masked-region reconstruction.
- We construct *BeyondMasks*, a paired benchmark of 180 synthetic and real-world video sequences with aligned clean backgrounds, masks, and instruction prompts, spanning diverse causal interactions. It more than doubles the scale of prior paired video object removal benchmarks.
- We propose and validate *CORE*, a structured VLM-based evaluation framework that jointly assesses object disappearance and the causal consistency of induced physical after-effects.
- We benchmark contemporary mask-based and instruction-driven video object removal models and reveal systematic failures in handling dynamic and interaction-driven physical effects.

Table 1: Comparison of video object removal benchmarks. Unlike prior datasets, BeyondMasks provides temporally aligned paired ground truth while explicitly modeling diverse causal after-effects in both synthetic and real-world scenes.

Dataset	Size	Paired GT	After-effects	Causal effects	Avg. video length	Source
DAVIS	90	✗	✗	✗	∼3 secs	Real
YouTube-VOS	508	✗	✗	✗	∼4.5 secs	Real
HQVI	10	✓	✗	✗	∼3.5 secs	Real
ROSE-Bench	60	✓	✓	✗	∼3 secs	Synthetic
Ours	180	✓	✓	✓	∼5.5 secs	Both

2 Related Work

2.1 Video Object Removal Benchmarks

Early evaluations of video object removal rely on segmentation datasets such as DAVIS [15] and YouTube-VOS [23]. Although these provide high-quality masks, they lack paired clean background videos in which both the object and its induced effects are absent. As a result, evaluation is limited to masked-region reconstruction, implicitly treating removal as local inpainting. HQVI [4] addresses this by constructing paired ground-truth videos via alpha compositing of foreground objects onto real backgrounds. While it provides temporally aligned pairs and occlusion-aware annotations, the compositing process assumes no secondary environmental effects, keeping evaluation focused on spatial reconstruction rather than object-environment disentanglement. ROSE-Bench [14] introduces synthetic paired videos generated via 3D rendering and models side effects such as shadows, reflections, mirror interactions, light emission, and translucency[5]. Despite advancing physically grounded evaluation, synthetic benchmarks remain limited in interaction diversity and environmental variability, primarily emphasizing isolated photometric effects. They do not systematically capture dynamic interactions such as deformation, fluid perturbations, particle traces, or scene rearrangement under realistic motion, nor do they support instruction-driven editing. Table 1 summarizes the differences between prior benchmarks and BeyondMasks. In contrast, BeyondMasks provides temporally aligned paired videos across synthetic and real-world scenes, explicitly modeling diverse causal after-effects and dynamic interactions, enabling systematic evaluation beyond masked-region fidelity.

2.2 Evaluation of Physical and Causal Consistency

Video editing is typically evaluated using reference-based metrics such as PSNR, SSIM [21], and LPIPS [27], which measure pixel similarity to ground truth. While effective for reconstruction quality, these metrics do not capture whether

[5] The authors also describe a real-world split generated via copy-and-paste compositing onto background videos, which is not publicly available at the time of writing.

object-induced effects persist or whether dynamic physical interactions remain temporally consistent. Perceptual measures such as FVD [19] assess realism but also remain insensitive to intervention correctness. At the image level, task-specific metrics such as ReMOVE [3] and CFD [26] incorporate object-aware signals to penalize residual foreground traces and assess semantic consistency. However, they operate on single frames and do not address temporal coherence or dynamic object–environment interactions. Recent benchmarks employ pre-trained vision-language models (VLMs) as automated evaluators. VBench [7,28], VideoPhy [1,2], and LLM-based judge frameworks [12] use structured prompting to assess semantic alignment and motion realism in generated videos. Yet these approaches evaluate realism without aligned references and are not designed for intervention-based editing. Causal object removal requires comparison against temporally aligned clean backgrounds to verify both object disappearance and elimination of induced physical effects. We address this through CORE, a structured VLM-based evaluation protocol that leverages paired references and explicitly reasons about object removal and residual after-effects. CORE enables scalable evaluation of causal consistency beyond masked-region fidelity.

3 BeyondMasks: Benchmark Design and Construction

We introduce *BeyondMasks*, a benchmark for evaluating video object removal under a causal intervention framework. While the recent ROSE-Bench [14] dataset incorporates selected physical side effects in synthetic settings, prior benchmarks largely assess removal through masked-region reconstruction or effect-specific realism. We instead formalize object removal as an intervention on a scene-generating process, aiming to recover the interventional scene state in which both the object and its induced environmental changes are absent, so that the resulting video corresponds to the scene as if the object had never been present.

Let S_t denote the latent physical state of a scene at time t, and let O denote the presence of a target object. We conceptualize each observed frame as arising from a rendering process $V_t = R(S_t, O)$, where a scene state together with the object configuration is mapped to pixel observations. Under an intervention in which the object is removed (i.e., $O = \varnothing$), the same scene state is rendered without the object, yielding the clean background frame $V_t^{bg} = R(S_t, \varnothing)$. This formulation highlights that object removal is not merely masked reconstruction, but a counterfactual rendering problem: recovering the scene as it would appear under the intervention $O = \varnothing$. Importantly, O influences V_t not only through direct appearance (object pixels), but also indirectly through environmental interactions that alter S_t, including shadows, reflections, illumination shifts, surface deformation, and dynamic perturbations. True object removal therefore requires approximating the interventional distribution $R(S_t, \varnothing)$ rather than merely hallucinating plausible background content within a masked region.

BeyondMasks implements this formulation through paired, temporally aligned intervention data, as illustrated in Fig. 2. Each sample consists of: (i) an object-present sequence V^{obj} generated under $O \neq \varnothing$, and (ii) a clean reference se-

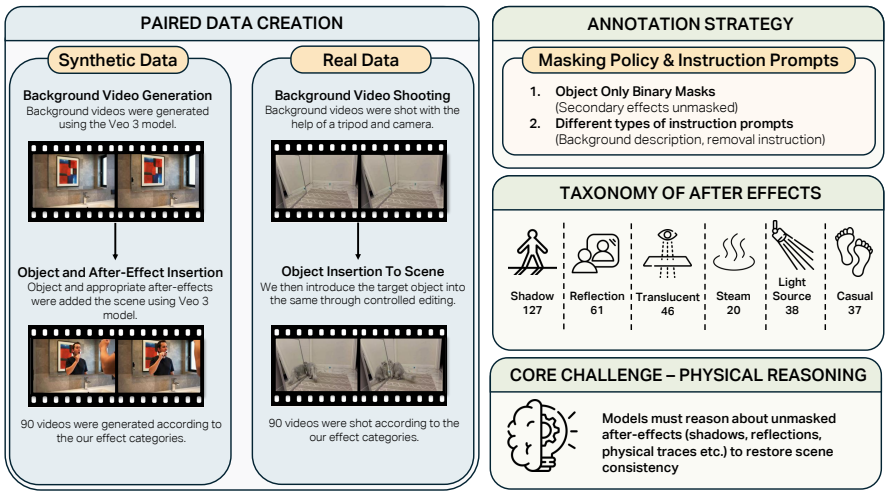

Fig. 2: Overview of the BeyondMasks construction pipeline. Top: Paired intervention-based generation, where an object-present video is derived from a clean background video, ensuring temporal alignment. Bottom: Representative categories of object-induced causal after-effects, including illumination, reflection, volumetric, and dynamic interaction effects.

quence V^{bg} generated under $O = \varnothing$. Because the two sequences differ only by this controlled intervention, evaluation directly measures whether a removal model successfully recovers the interventional scene state.

3.1 Synthetic Paired Generation

The synthetic subset contains 90 scenes generated using Veo3's [22] video editing capabilities. For each scene, we first generate a background-only video without the target object. This sequence serves as the clean reference. We then introduce the target object into the same scene through controlled editing, producing the object-present video. Since the latter is derived from the former, temporal alignment is preserved. The insertion process naturally induces physically coherent interactions with the environment, including cast shadows, illumination changes, reflections, translucency effects, and scene-dependent dynamic traces. Because the background sequence remains unchanged apart from the object insertion, the paired design ensures that the only differences between videos correspond to the object and its causal footprint. This controlled construction enables both pixel-level comparison and higher-level semantic assessment. All synthetic pairs undergo manual review to verify physical plausibility, temporal coherence, and consistency between object presence and environmental interaction.

3.2 Real-World Paired Capture

To complement synthetic data, BeyondMasks includes 90 real-world paired sequences captured with a tripod-stabilized setup. For each scene, we record a short video with the object present and immediately capture a second video after removing the object, keeping camera position or settings fixed. This back-to-back protocol preserves geometry, lighting, and scene composition, ensuring that differences between paired videos arise solely from object presence and its associated effects. These real-world captures introduce natural texture complexity, subtle lighting variation, and realistic interaction patterns that are difficult to model synthetically, improving the realism and diversity of the benchmark.

3.3 Mask Annotation Pipeline

For each object-present sequence, we provide temporally consistent binary masks to support mask-based removal methods. Masks are generated through a semi-automatic pipeline that combines interactive segmentation using SAM2 [16] with temporal propagation via CoTracker 3 [10]. Annotators refine the initial-frame segmentation using positive and negative point prompts to obtain precise object boundaries, after which masks are propagated across frames and periodically corrected to maintain temporal coherence. Each sequence undergoes cross-review by a second annotator to ensure segmentation accuracy and consistency. Importantly, masks include only the target object and explicitly exclude its after-effects. Shadows, reflections, illumination changes, and other induced phenomena are not masked. This design choice reflects our causal formulation: models must infer and remove secondary effects through physical reasoning rather than relying on explicit supervision.

3.4 Taxonomy of Causal After-Effects

A central component of BeyondMasks is its taxonomy of object-induced environmental interactions. Rather than categorizing scenes by object type, we group samples according to the underlying physical mechanisms through which objects causally influence their surroundings. This mechanism-driven taxonomy ensures that evaluation reflects distinct modes of object-environment entanglement.

Shadow. Objects occlude or redirect illumination, producing cast shadows and local radiance changes. Successful removal requires restoring a shadow-consistent illumination field, not merely deleting object pixels.

Reflection. Objects may appear indirectly through specular or diffuse reflections on mirrors, glass, or water. Removal must eliminate both direct and reflected evidence while preserving surface appearance and temporal consistency.

Translucent effects. Semi-transparent objects (e.g., glass) partially reveal the background while modulating light transport. These cases require recovering background structure consistent with partial visibility and attenuation.

Steam/scattering. Volumetric media such as steam or smoke introduce spatially diffuse, time-varying traces. Removal demands undoing these effects while maintaining coherent scene dynamics and background continuity.

Causal physical effects. Some objects alter scene geometry or material state through motion or contact, yielding effects that are not visually attached to the object yet remain causally dependent on it. Examples include water ripples, dust trails, displaced objects, and surface deformation such as footprints.

Light source effects. When the object emits light, its removal requires compensating for localized illumination changes and restoring consistent exposure and shading.

No explicit after-effect. A small subset of sequences exhibits minimal secondary interaction, serving as controls for evaluating pure object removal with limited environmental entanglement.

3.5 Instruction Prompts

To support instruction-driven editing, each sample is associated with one or more validated text prompts specifying the removal target while emphasizing natural scene restoration. Prompts are constructed using structured templates to ensure object specificity and clarity, and are manually verified for consistency across both mask-based and text-conditioned paradigms.

4 CORE: Causal Object Removal Evaluation

While BeyondMasks enables paired pixel-level comparison, reconstruction metrics alone are insufficient for evaluating interventional correctness. Pixel-based measures such as PSNR, SSIM, and LPIPS penalize any deviation from the ground-truth background, even when the reconstructed scene is physically plausible and free of residual object traces. As illustrated in Fig. 3, methods with nearly identical LPIPS scores can differ substantially in whether object-induced physical effects are properly removed. In object removal, multiple visually coherent reconstructions may exist for unseen regions, making strict pixel identity an unreliable proxy for causal success. Under our causal formulation, successful removal corresponds to recovering the interventional scene state $R(S_t, \varnothing)$, that is, the scene that would be observed if the object and its induced environmental changes were absent. Deviations from this target state arise at two distinct levels: (i) *residual object presence*, where visible traces of the object remain, and (ii) *residual causal effects*, where shadows, reflections, illumination shifts, or dynamic perturbations persist despite the object's removal. To explicitly measure these dimensions, we introduce *CORE (Causal Object Removal Evaluation)*, a structured VLM-based assessment framework illustrated in Fig. 4.

CORE evaluates removal using three temporally aligned inputs: the object-present video V, the aligned clean background V^{bg}, and the edited output $\hat{V}^{bg}$ produced by the evaluated method.

Fig. 3: Pixel metrics fail to capture causal correctness. Although LPIPS scores are similar across methods, their ability to remove object-induced effects (e.g., shadows or reflections) differs substantially. CORE separates these failure modes by evaluating object removal (CORE-OS) and elimination of causal after-effects (CORE-AES). ↓ better for LPIPS; ↑ better for CORE scores.

Rather than measuring pixel correspondence, CORE assesses whether the edited video is semantically and physically consistent with the interventional reference in which the object and its induced environmental changes are absent. The evaluation is implemented via a pretrained vision-language model operating under a fixed, structured prompt that enforces comparative reasoning between $\hat{V}^{bg}$ and V^{bg}. The prompt explicitly guides the model to reason about two intervention-specific criteria: (i) whether the object has been fully removed, and (ii) whether its causal after-effects have been eliminated. These two dimensions correspond to complementary failure modes in object removal. Residual object traces may remain even if the background appears plausible, and conversely, shadows or dynamic perturbations may persist despite apparent object disappearance. To disentangle these behaviors, CORE decomposes evaluation into two scores:

- **ObjectScore:** completeness of object removal and plausibility of background restoration;
- **AfterEffectScore:** elimination of object-induced physical traces, including shadows, reflections, illumination changes, and dynamic perturbations.

We implement CORE using Gemini 3.1 [6]. Each evaluation call receives the three videos along with structured metadata specifying the target object and relevant after-effect categories. The VLM performs structured comparative reasoning: identifying object traces in the object-present video V, inferring expected

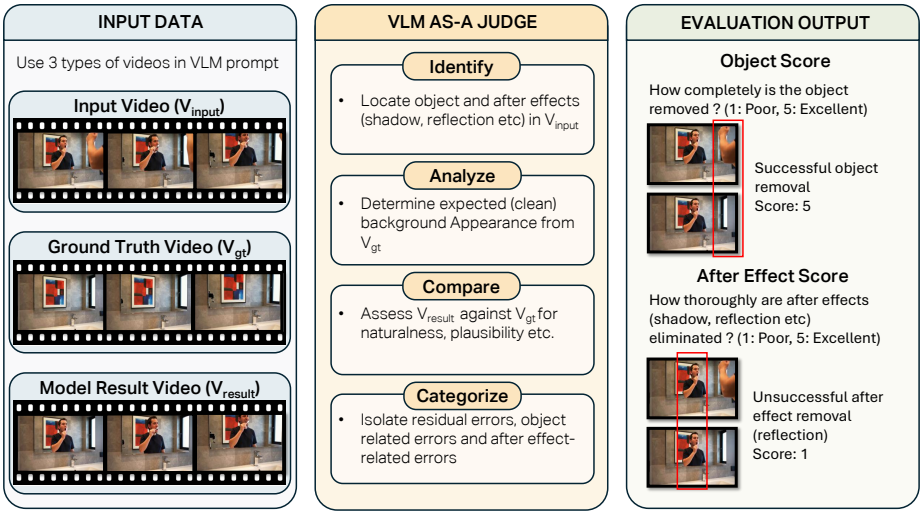

Fig. 4: Overview of CORE (Causal Object Removal Evaluation). Given an edited video $\hat{V}^{bg}$ and its temporally aligned clean reference V^{bg}, CORE performs structured VLM-based comparison to assess interventional correctness. The evaluation jointly measures (i) object disappearance and (ii) consistency of object-induced physical after-effects, such as shadows, reflections, and dynamic perturbations. By conditioning semantic reasoning on aligned references, CORE moves beyond masked-region fidelity toward evaluation of causal and physical consistency.

appearance from the aligned clean background V^{bg}, comparing the edited output $\hat{V}^{bg}$ against V^{bg} under a naturalness criterion, and categorizing residual errors as object-related or aftereffect-related. The scoring rubric explicitly rewards semantically coherent background reconstruction, even when pixelwise differences from GT exist, while penalizing residual object evidence, persistent after-effects, or newly hallucinated foreground content. By conditioning semantic reasoning on aligned references, CORE transforms VLM-based scoring from observational realism assessment into structured intervention verification. The full prompt and rubric are provided in the supplementary material.

Validation of CORE. To assess alignment with human judgment, we conduct a user study in which 8 annotators independently evaluate results from 5 methods across 20 videos using the same scoring rubric. The Pearson correlation between CORE scores and the mean human ratings is 0.615 for *ObjectScore* and 0.733 for *AfterEffectScore*. For reference, human inter-rater agreement yields Pearson correlations of 0.693 and 0.785 for the same dimensions, respectively. These results indicate that CORE closely aligns with human perception and approaches the upper bound defined by human agreement.

Table 2: Quantitative comparison on our benchmark. We evaluate representative video object removal methods using standard reconstruction metrics (PSNR, SSIM, LPIPS), the video realism metric FVD, and our VLM-based CORE evaluation. **Type** indicates the supervision required by each method: **T** denotes text-instruction-based editing, **M** denotes mask-guided removal, and **T+M** indicates methods that support both inputs. **Bold** indicates the best score.

Method	Type	PSNR↑	SSIM↑	LPIPS↓	FVD↓	CORE-OS↑	CORE-AES↑
Lucy Edit (arXiv 2025) [18]	T	18.5884	0.7239	0.2758	480.08	1.443	1.551
VACE (ICCV 2025) [9]	T+M	19.9932	0.7856	0.2077	418.79	1.799	1.698
CoCoCo (AAAI 2025) [31]	T+M	20.8565	0.7287	0.2404	381.84	1.986	2.043
Gen. Omnimatte (CVPR 2025) [11]	T+M	23.8106	0.8175	0.1968	250.76	3.771	**3.021**
ProPainter (ICCV 2023) [29]	M	23.5979	0.8429	0.1635	225.28	3.063	2.368
DiffuEraser (arXiv 2025) [13]	M	**25.1880**	**0.8769**	**0.1124**	**164.26**	3.494	2.625
MiniMax (NeurIPS 2025) [30]	M	23.1954	0.8218	0.1682	203.64	3.631	2.553
ROSE (NeurIPS 2025) [14]	M	23.8062	0.8173	0.1659	316.04	**3.784**	2.920
OmnimatteZero (SIGGRAPH 2025) [17]	M	22.9306	0.8127	0.1638	227.24	3.190	2.499

5 Experimental Analysis

5.1 Experimental Setup

We benchmark nine recent video object removal and video editing methods spanning mask-based (M), text-guided (T), and hybrid (T+M) paradigms. These include classical mask-guided inpainting approaches such as ProPainter [29] and DiffuEraser [13], layer-based decomposition models including Generative Omnimatte [11] and OmnimatteZero [17], and recent large-scale editing systems such as VACE [9], Lucy Edit [18], CoCoCo [31], MiniMax [30], and ROSE [14]. Together, these methods represent the dominant paradigms currently used for video object removal and instruction-driven video editing. All methods are evaluated using publicly released implementations and default configurations at the native resolution of BeyondMasks. For each method, we report standard reference-based metrics, PSNR, SSIM, LPIPS, computed against the paired GT backgrounds, as well as FVD for temporal quality and our proposed VLM-based CORE ObjectScore (CORE-OS) and AfterEffectScore (CORE-AES).

5.2 Quantitative Results

Table 2 summarizes the results. Across standard reconstruction metrics, several methods achieve comparable PSNR, SSIM, and LPIPS scores, indicating similar pixel-level fidelity with respect to the clean reference videos. However, these metrics do not reliably reflect removal correctness. Methods that achieve strong reconstruction scores frequently leave residual object evidence or fail to remove induced physical effects such as shadows or reflections. This discrepancy is particularly evident when comparing reconstruction metrics with CORE

Table 3: Average performance across after-effect categories. Results are averaged across the top-performing six models for each after-effect type. Higher values are better for PSNR, SSIM, CORE-OS, and CORE-AES, while lower values indicate better performance for LPIPS.

After-effect	PSNR↑	SSIM↑	LPIPS↓	CORE-OS↑	CORE-AES↑
Shadow	23.6870	0.8298	0.1621	3.498	2.520
Reflection	23.9842	0.8558	0.1580	3.332	2.446
Light source	22.4233	0.8201	0.1849	3.554	2.248
Steam	22.8375	0.8466	0.1624	3.236	2.032
Translucent	22.9326	0.8156	0.2162	3.256	2.701
Causal	22.7912	0.8163	0.1867	3.486	2.141

scores. For example, DiffuEraser attains the highest PSNR and SSIM, yet exhibits noticeably lower AfterEffectScore than several competing methods, indicating incomplete removal of object-induced effects. Conversely, ROSE achieves the highest CORE-OS despite only moderate reconstruction metrics, suggesting more effective elimination of object traces even when pixel similarity is comparable. These results highlight a fundamental limitation of conventional evaluation: pixel-based fidelity does not necessarily correspond to causal correctness. By explicitly separating object disappearance (CORE-OS) from elimination of physical after-effects (CORE-AES), CORE reveals failure modes that remain hidden under traditional reconstruction metrics. This analysis confirms that realistic video object removal requires reasoning about object-environment interactions rather than merely reconstructing masked regions.

To further analyze performance across different physical phenomena, Table 3 reports results aggregated by after-effect category. The breakdown reveals substantial variation in difficulty across interaction types. While shadows and reflections are handled relatively well on average, volumetric effects such as steam and dynamic interactions (e.g., footprints or surface disturbances) exhibit significantly lower CORE-AES scores. These results indicate that secondary physical processes extending beyond the masked region remain challenging for current removal methods. Overall, the analysis confirms that realistic object removal requires modeling object-environment interactions rather than relying solely on local reconstruction.

5.3 Qualitative Analysis

Fig. 5 presents representative examples comparing several state-of-the-art methods. While most approaches successfully remove the target objects, many fail to eliminate object-induced effects. Shadows, reflections, and illumination changes often persist even after the object disappears, resulting in physically inconsistent

Fig. 5: Qualitative comparison across six after-effect categories. Each category shows two temporal frames.

scenes. These residual traces are particularly evident in cases involving reflective surfaces, cast shadows, and lighting interactions.

After-Effect Removal is the Dominant Challenge. Across the examples in Fig. 5, removing the object itself is often easier than eliminating the physical effects it induces in the environment. While most methods successfully erase the primary object, secondary phenomena such as shadows, reflections, and illumination changes frequently persist. These effects are spatially diffuse and may extend beyond the masked region, making them difficult to address with standard

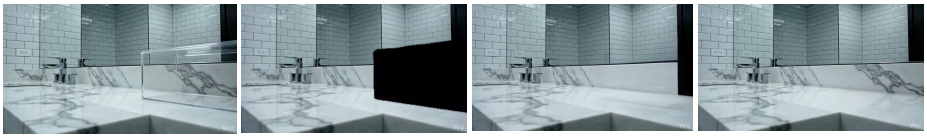

Fig. 6: Context loss in mask-based video object removal. Many mask-based pipelines zero out the masked region prior to inference, eliminating potentially informative visual evidence (e.g., background partially visible through translucent structures). This removal of observable context forces the model to hallucinate missing content rather than recover available information, resulting in degraded restoration and incomplete causal consistency relative to the ground truth.

inpainting strategies. This observation highlights the importance of evaluating object removal as a causal intervention rather than a purely local reconstruction task.

Mask-Based vs. Text-Guided Removal. Qualitative differences also emerge between mask-based and instruction-driven approaches. Mask-based methods typically achieve strong object removal due to explicit spatial supervision, but often leave residual environmental effects such as shadows or reflections. In contrast, text-guided approaches offer greater editing flexibility but show inconsistent removal of fine-grained physical traces and may introduce hallucinated content (see Supplementary for visual examples). These observations suggest that current systems, regardless of supervision modality, struggle to disentangle objects from the physical processes they induce in the surrounding environment.

Context Loss in Mask-Based Inpainting. A structural limitation of many mask-based pipelines is the removal or replacement of masked regions before inference. This operation discards potentially informative visual evidence, particularly for translucent or semi-transparent objects where background content remains partially observable. As illustrated in Fig. 6, eliminating these signals forces models to hallucinate missing content rather than recover available context. Similar observations have been reported in image-based removal [8]. In video, temporal redundancy provides additional opportunities to exploit such information across frames, suggesting that future removal systems should incorporate context-aware reasoning rather than relying solely on masked inpainting.

6 Conclusion

We introduced *BeyondMasks*, a benchmark for evaluating video object removal from a causal intervention perspective. Unlike existing evaluation protocols that primarily assess masked-region reconstruction, BeyondMasks enables systematic evaluation of whether removal methods eliminate both the object and the physical effects it induces in the surrounding environment. By providing temporally aligned video pairs with clean background references and diverse interaction

categories, the benchmark allows controlled analysis of object-environment disentanglement in realistic editing scenarios. Using this benchmark, we evaluated a diverse set of recent video object removal and video editing methods and identified a consistent gap between high reconstruction fidelity and true causal correctness. While many approaches successfully remove the visible object, they frequently leave behind residual shadows, reflections, illumination changes, or other secondary physical traces. To address limitations of existing evaluation protocols, we introduced *CORE*, a VLM-based evaluation framework that separately measures object disappearance and the removal of object-induced after-effects. Our experiments show that conventional pixel-based metrics often fail to capture these failure modes, whereas CORE provides a more faithful assessment of causal consistency and aligns more closely with human perception.

Acknowledgements

This work was partly supported by the KUIS AI Center Research Awards to ES, and TÜBİTAK-2247-A Program Award (No. 123C550) to AE. We thank all the reviewers for their valuable comments.

References

1. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=9D2QvO1uWj, Accessed: 1 July 2026
2. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=HA8KSQW7SO, Accessed: 1 July 2026
3. Chandrasekar, A., Chakrabarty, G., Bardhan, J., Hebbalaguppe, R., AP, P.: Remove: A reference-free metric for object erasure. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 7901–7910 (June 2024)
4. Cho, S., Oh, S.W., Lee, S., Lee, J.Y.: Elevating flow-guided video inpainting with reference generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025). pp. 2527–2535 (2025)
5. Dharmarajan, K., Huang, W., Wu, J., Fei-Fei, L., Zhang, R.: Dream2flow: Bridging video generation and open-world manipulation with 3d object flow. arXiv preprint arXiv:2512.24766 (2025)
6. Gemini Team, Google DeepMind: Gemini 3.1 pro - model card (February 2026), https://deepmind.google/models/model-cards/gemini-3-1-pro/, Accessed: 1 July 2026
7. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21807–21818 (June 2024)
8. Jiang, L., Wang, Z., Bao, J., Zhou, W., Chen, D., Shi, L., Chen, D., Li, H.: Smarteraser: Remove anything from images using masked-region guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24452–24462 (2025)
9. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: 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)
10. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6013–6022 (2025)
11. Lee, Y.C., Lu, E., Rumbley, S., Geyer, M., Huang, J.B., Dekel, T., Cole, F.: Generative omnimatte: Learning to decompose video into layers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12522–12532 (2025)
12. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llms-as-judges: A comprehensive survey on llm-based evaluation methods (2024), https://arxiv.org/abs/2412.05579, Accessed: 1 July 2026
13. Li, X., Xue, H., Ren, P., Bo, L.: Diffueraser: A diffusion model for video inpainting (2025), https://arxiv.org/abs/2501.10018, Accessed: 1 July 2026
14. Miao, C., Feng, Y., Zeng, J., Gao, Z., Hantang, L., Yan, Y., Qi, D., Chen, X., Wang, B., Zhao, H.: ROSE: Remove objects with side effects in videos. In: NeurIPS (2025), https://openreview.net/forum?id=xTWWKMxY1x, Accessed: 1 July 2026
15. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (June 2016)
16. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025)
17. Samuel, D., Levy, M., Darshan, N., Chechik, G., Ben-Ari, R.: Omnimattezero: Fast training-free omnimatte with pre-trained video diffusion models. In: SIGGRAPH Asia 2025 Conference Papers. SA Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025)
18. Team, D.: Lucy edit: Open-weight text-guided video editing (2025)
19. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: A new metric for video generation (2019), https://openreview.net/forum?id=rylgEULtdN, Accessed: 1 July 2026
20. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
22. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
23. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: ECCV (September 2018)

24. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
25. Yu, S., Liu, D., Ma, Z., Hong, Y., Zhou, Y., Tan, H., Chai, J., Bansal, M.: Veggie: Instructional editing and reasoning video concepts with grounded generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15147–15158 (2025)
26. Yu, Y., Zeng, Z., Zheng, H., Luo, J.: Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17324–17334 (October 2025)
27. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
28. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness (2025), https://arxiv.org/abs/2503.21755, Accessed: 1 July 2026
29. Zhou, S., Li, C., Chan, K.C., Loy, C.C.: Propainter: Improving propagation and transformer for video inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10477–10486 (October 2023)
30. Zi, B., Peng, W., Qi, X., Wang, J., Zhao, S., Xiao, R., Wong, K.F.: Minimax-remover: Taming bad noise helps video object removal. arXiv preprint arXiv:2505.24873 (2025)
31. Zi, B., Zhao, S., Qi, X., Wang, J., Shi, Y., Chen, Q., Liang, B., Xiao, R., Wong, K.F., Zhang, L.: Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 11067–11076 (2025)