Physics-Aware Video Instance Removal Benchmark

Zirui Li[1] **Xinghao Chen**[2] **Lingyu Jiang**[1] **Dengzhe Hou**[1]
Fangzhou Lin[1,3] **Kazunori Yamada**[1] **Xiangbo Gao**[3] **Zhengzhong Tu**[3,*]

[1]Tohoku University [2]University of Washington [3]Texas A&M University

li.zirui.r7@dc.tohoku.ac.jp

[*]Corresponding author.

Abstract

Video Instance Removal (VIR) *requires removing target objects while maintaining background integrity and physical consistency, such as specular reflections and illumination interactions. Despite advancements in text-guided editing, current benchmarks primarily assess visual plausibility, often overlooking the* physical causalities*—such as lingering shadows—triggered by object removal. We introduce the* ***Physics-Aware Video Instance Removal (PVIR)*** *benchmark, featuring 95 high-quality videos annotated with instance-accurate masks and removal prompts. PVIR is partitioned into* Simple *and* Hard *subsets, the latter explicitly targeting complex physical interactions. We evaluate four representative methods—PISCO-Removal, UniVideo, DiffuEraser, and CoCoCo—using a decoupled human evaluation protocol across three dimensions to isolate semantic, visual, and spatial failures:* instruction following*,* rendering quality*, and* edit exclusivity*. Our results show that* ***PISCO-Removal*** *and* ***UniVideo*** *achieve state-of-the-art performance, while* ***DiffuEraser*** *frequently introduces blurring artifacts and* ***CoCoCo*** *struggles significantly with instruction following. The persistent performance drop on the* Hard *subset highlights the ongoing challenge of recovering complex physical side effects.*

1. Introduction

Removing a target instance from a video is a foundational editing capability with direct utility in post-production, privacy protection, robotics simulation, and synthetic data curation. Compared with image object removal, video instance removal requires consistency over time and consistency with scene physics. When an object is removed, not only should the object disappear, but its side effects should be updated as well [15, 16]: reflections in windows, mirror appearances, indirect occlusion patterns, and local illumination cues should all remain plausible. These requirements expose a major gap between qualitative demos and reliable, comparable evaluation.

Recent video editing and inpainting systems have improved temporal coherence and controllability [12, 25, 30]. At the same time, methods designed for side-effect-aware removal demonstrate that physics-aware editing is now an explicit research target [15]. Despite these algorithmic advances, the field lacks a common yardstick: existing evaluations are typically performed on private subsets, inconsistent prompts, and varying resolution constraints, leading to a ”closed-world” comparison that obscures true progress. As a result, it remains difficult to answer basic questions: Which model follows removal instructions best? Which model preserves visual quality over time? Which model minimizes unintended edits outside the target region?

To address this gap, we introduce **Physics-Aware Video Instance Removal Benchmark**, a task-focused benchmark that standardizes data, protocols, and evaluation. Our benchmark contains 95 high-quality videos with per-video target segmentation and removal prompts. We explicitly partition data into *Simple* and *Hard* subsets, where *Hard* clips include stronger interactions with real-world physics (e.g., specular reflections, mirror appearance, and pronounced scene coupling).

A second key contribution is a decoupled human evaluation protocol. Instead of a single holistic score, we assess three independent dimensions: (1) **Instruction Following**, i.e., whether the target is correctly removed; (2) **Rendering Quality**, i.e., whether the inpainted result is temporally stable and visually plausible; (3) **Edit Exclusivity**, i.e., whether non-target content remains unchanged. Each dimension uses a 1–4 rubric with explicit criteria, enabling interpretable diagnosis rather than one-number ranking.

We evaluate four representative models under a unified setting: CoCoCo [30], UniVideo [25], DiffuEraser [12], and PISCO-Removal [3]. For PISCO-Removal, we evaluate the variant trained in the PISCO paper [3] using the ROSE dataset [15]; this version supports more demanding

practical configurations, including 720p resolution, portrait orientations, and sequences up to 120 frames. This cross-model comparison is designed to expose the trade-off between general-purpose inpainting stability and task-specific physical fidelity.

Contributions. Our contributions are summarized as follow:

- We present a new benchmark for *physics-aware video instance removal*, with 95 high-quality videos, instance-level masks, and removal prompts.
- We define a decoupled, interpretable human evaluation protocol with three independent 1–4 metrics and an explicit aggregation rule.
- We provide a unified benchmark of four representative methods and establish analysis protocols for overall, per-difficulty, and failure-mode evaluation.

Scope and current status. This paper focuses on benchmark construction and standardized evaluation rather than proposing a new removal architecture. Our analysis uncovers a "performance ceiling" where even state-of-the-art models fail to resolve secondary physical interactions, revealing that the primary bottleneck in VIR has shifted from temporal flickering to physical incoherence.

2. Related Works

Evolution of Video Inpainting and Instance Removal. Modern video object removal pipelines inherit core spatial advances from image inpainting [17, 22, 28] and have evolved through flow-guided propagation [2, 13] and transformer-based temporal modeling [14, 29] to ensure long-horizon consistency. Recent diffusion-based formulations, such as DiffuEraser [12], further enhance realism under complex motion, while ROSE [15] specifically formalizes physical interactions like reflections as first-order constraints. However, while these methods focus on the architectural challenge of texture propagation and physical coupling, our work shifts the focus toward providing a standardized evaluation framework to quantify how well these models actually recover such complex physical causalities.

Text-guided Video Generation and Editing. The field has shifted from general video generation [7, 20] to highly controllable editing pipelines [5, 11, 26]. Recent large-scale models like UniVideo [25] and CoCoCo [30] leverage strong generative priors to achieve high-quality restoration and better workflow compatibility. Despite their impressive zero-shot capabilities, these generative models often struggle with "semantic leakage" or fail to strictly respect local masking constraints in instance removal tasks. Unlike these generative frameworks that prioritize visual plausibility, our benchmark emphasizes the decoupling of instruction following and spatial exclusivity to expose these specific failure modes.

Benchmarking and Perceptual Assessment. While general-purpose video benchmarks [1, 4, 8, 24] and visual metrics like FID [6] and FVD [23] have improved coverage, they often fail to capture high-level physical logic, such as lingering shadows or inconsistent illumination. Existing data infrastructures from segmentation [19, 27] provide annotation principles but do not isolate the entangled dimensions of target compliance and background preservation. Our PVIR benchmark addresses this gap by introducing a physics-aware dataset and a decoupled human evaluation protocol with a 1–4 scoring rubric, providing a more interpretable assessment than conventional automatic metrics.

3. Dataset: Physics-Aware Video Instance Removal

Design goal. The dataset is designed for one core task: remove a designated instance from a real video while preserving non-target content and maintaining physically plausible side effects. To enable robust benchmarking, we prioritize high visual quality, diverse scenes, and explicit interaction complexity.

Scale and split. The benchmark comprises 95 high-quality sequences, ensuring a balanced distribution between foundational removal tasks and advanced physics-aware challenges. We organize them into two subsets: **Simple** and **Hard**. Simple videos (57 videos) usually contain objects with simpler geometry and weaker coupling to scene physics. They serve as a baseline to evaluate a model's fundamental ability to maintain spatial-temporal coherence in the absence of complex physics. Hard videos (38 videos) include stronger physics interactions, such as mirror reflections, specular highlights, and complex motion-appearance coupling. These cases require the model to not only fill the disoccluded pixels but also maintain physical causality by updating or removing secondary side effects, such as reflections and shadows, that are anchored to the target. Each video includes two annotations:

- **Target segmentation mask**: a high-quality instance mask indicating the object to remove.
- **Removal prompt**: a natural-language instruction that unambiguously refers to the target instance.

These two elements define a standardized input interface for all benchmarked models.

3.1. Collection and Annotation Pipeline

Data sourcing and pre-filtering. Candidate videos are collected from diverse real-world scenes and filtered by minimal requirements on spatial quality, temporal smoothness, and compression artifacts. To avoid overly synthetic bias, we prioritize clips with natural motion and illumination variation. Our primary sources include the **Inter4k** dataset [21], selected for its ultra-high-definition (UHD) clarity and high frame rates, and the **DAVIS2016** dataset [18], chosen for its diversity in object-to-background interactions. Specifically, we curate 45 sequences from Inter4k to ensure high-fidelity rendering assessment, and 50 sequences from DAVIS2016 to leverage its complex motion patterns. All source videos are either licensed under Creative Commons (CC) or are explicitly permitted for non-commercial research use, ensuring a clear and ethical release metadata.

Mask annotation workflow. To ensure pixel-level precision, we employ a multi-stage annotation pipeline. Annotators first utilize the Segment Anything Model 2 (SAM 2) [10] to generate initial object tracks across each sequence. By providing sparse point or box prompts on keyframes, SAM 2's memory-based propagation produces coarse masks for the entire video. Subsequently, annotators perform manual frame-level refinement to clean up boundaries, particularly in challenging cases involving motion blur or occlusion. To ensure temporal smoothness, we conduct a final quality-control pass focusing on reducing "shape jitter" (flickering boundaries). Any masks exhibiting temporal instability are manually corrected or re-propagated using flow-based consistency checks. This hybrid workflow combines the efficiency of foundation models with the rigor of human verification, yielding the instance-accurate masks required for high-fidelity removal.

Prompt writing workflow. Prompts are designed to be specific, concise, and target-disambiguating. To ensure consistency across the benchmark, we adopt a structured template: `[Action] + [Target Attributes] + [Spatial/Contextual Qualifiers]`. For instance, a PVIR prompt specifies "Remove the silver sedan parked under the flickering streetlamp" instead of a generic command, providing unique semantic grounding that minimizes instruction ambiguity.

About the language policy and disambiguation, all prompts are authored in English using standard descriptive vocabulary. Ambiguous references—such as multiple similar objects in a single frame—are strictly disallowed unless unique qualifiers (e.g., "the person on the far left") are included. For the *Hard* subset, prompts are intentionally augmented with physical context, such as "including its reflection on the water surface," to explicitly signal the expected physics-aware behavior. Each prompt undergoes a cross-verification pass by a second annotator to ensure that the textual description uniquely identifies the instance masked in the ground truth.

3.2. Difficulty Taxonomy

Simple subset. Simple clips typically contain targets with limited appearance variation and weak side-effect coupling. Examples include matte surfaces, stable backgrounds, and short-term motion without heavy occlusion.

Hard subset. Hard clips emphasize complex light-surface coupling and geometric reconstruction challenges.

Table 1. **Summary of the PVIR Benchmark Dataset.** The benchmark comprises 95 high-definition videos, categorized into Simple and Hard subsets based on the complexity of physical interactions (e.g., reflections, shadows, and dynamic fluid wakes).

Subset	#Videos	Avg. Frames	Key Properties
Simple	57	81	weaker interaction and simpler geometry
Hard	38	81	reflection/mirror/specular and stronger coupling
Total	95	81	mixed scenes and motions

4. Benchmark Protocol

Task definition. Given an input video V, a target mask sequence M, and a removal prompt P, the model outputs an edited video $\hat{V}$ where the target instance is removed. The benchmark requires: (1) complete target removal alongside its associated physical derivatives, (2) high-quality temporally consistent rendering, (3) minimal unintended changes outside the edited region.

Evaluated models. We evaluate four representative methods: **CoCoCo** [30], **UniVideo** [25], **DiffuEraser** [12], and **PISCO-Removal** [3], where PISCO-Removal is a WACE [9]-like model fine-tuned with the ROSE dataset [15] that support 720p, portrait videos, and up to 120 frames [3].

4.1. Unified Inference Setup

Input/output interface normalization. Each model receives the same semantic input triplet (V, M, P) and outputs a completed video sequence. We standardize video decoding, prompt formatting, and export codecs to minimize evaluation noise from non-model factors.

Figure 1. **Qualitative comparison on the PVIR benchmark.** Each row presents a specific instance removal scenario with its corresponding textual prompt. As a unified model, **UniVideo** demonstrates strong physics-awareness, successfully removing coupled side effects like ground shadows (e.g., rows b and c); however, it suffers from severe semantic hallucinations, occasionally generating unprompted artifacts to fill the void (e.g., the distorted figure generated in row d). **PISCO-Removal** consistently achieves clean erasure of both the target and its physical side effects while maintaining high background fidelity. **DiffuEraser** reliably masks the correct instance but frequently leaves unnatural residual shadows and spatial blurring (rows c and d). Finally, **CoCoCo** struggles significantly with instruction following, often leaving obvious "ghosting" silhouettes or failing to remove the object entirely (e.g., the white duck in row a).

Method-specific adaptation policy. To eliminate performance bias stemming from disparate input constraints, all evaluated baselines are unified under a standardized configuration: 720p resolution at 81 frames. Unlike previous benchmarks that often resort to downsampling or temporal truncation, our protocol ensures that every model is tested at its maximum practical capacity. We utilize the officially released checkpoints for each method, ensuring that any observed performance gaps are intrinsic to the models' architectures rather than artifacts of suboptimal parameter tuning. This high-resolution, long-duration setup serves as a rigorous stress test for temporal-physical consistency in Video Instance Removal.

4.2. Human Evaluation Protocol

Decoupled scoring principle. All three dimensions are scored *independently*; a score in one dimension must not influence another. Each dimension uses a 1–4 ordinal rubric, where higher is better.

(1) Instruction Following (IF). Core question: Does the edited video correctly satisfy the removal instruction? For this benchmark, judges check whether the specified target instance is removed, and whether visible remnants (edges, fragments, obvious traces) remain.

(2) Rendering Quality (RQ). Core question: Is the filled content visually plausible over space and time? Judges inspect naturalness, sharpness, temporal stability (flicker/jitter), and physical plausibility of motion/appearance.

(3) Edit Exclusivity (EE). Core question: Did the model only perform the requested removal? Judges verify that non-target regions preserve original content, including background structures, lighting appearance, and unrelated objects.

Aggregation. Let $s_d^{(i,r)} \in \{1, 2, 3, 4\}$ denote the score for video i, rater r, and dimension $d \in \{\text{IF}, \text{RQ}, \text{EE}\}$. We first average over raters:

$$\bar{s}_d^{(i)} = \frac{1}{R} \sum_{r=1}^{R} s_d^{(i,r)}. \tag{1}$$

Then report per-dimension dataset mean:

$$S_d = \frac{1}{N} \sum_{i=1}^{N} \bar{s}_d^{(i)}. \tag{2}$$

The aggregated benchmark score is

$$S_{\text{overall}} = \frac{1}{3}(S_{\text{IF}} + S_{\text{RQ}} + S_{\text{EE}}). \tag{3}$$

To ensure unbiased assessment, videos are assigned to raters using a balanced, randomized sampling strategy. Each video-model pair is evaluated by at least 2 independent raters, and the presentation order of the four methods is shuffled for every trial to eliminate model-specific or sequential bias.

4.3. Reporting Protocol

Primary Metrics. We report the mean scores for Instruction Following (IF), Rendering Quality (RQ), and Edit Exclusivity (EE) across the entire benchmark. An Overall Score is computed as the unweighted arithmetic mean of these three dimensions, providing a holistic measure of instance removal performance.

Split-wise Analysis. To isolate model robustness under varying physical complexities, we additionally report performance partitioned by the Simple and Hard subsets. This granular reporting highlights the "performance decay" models experience when transitioning from basic scenarios to those with strong physical coupling (e.g., reflections and wakes).

Uncertainty and Significance. To ensure the reliability of our rankings, we report 95% Confidence Intervals (CIs) for all primary metrics. For pairwise model comparisons, we employ Bootstrap Significance Testing (with $N = 10,000$ iterations). Our analysis confirms that the performance superiority of PISCO-Removal and UniVideo is statistically significant ($p < 0.05$) across all dimensions, particularly on the Hard subset where simpler baselines suffer from severe physical artifacts.

5. Experiments

5.1. Experimental Setup

Benchmark setting. We evaluate all methods on the full 95-video benchmark and report: (1) overall scores, (2) split-wise scores on Simple and Hard subsets, (3) qualitative failure analysis. Unless otherwise noted, all scores are human ratings following Sec. 4.2.

Models. We benchmark CoCoCo [30], UniVideo [25], DiffuEraser [12], and PISCO-Removal [3, 15].

Implementation Details. All experiments are conducted on a workstation with three **NVIDIA A100 (80GB)** GPUs. To ensure a high-fidelity evaluation, we process all sequences at a native **720p (81 frames)** resolution. The inference latency varies significantly across baselines: **UniVideo** requires ∼3.5 hours per video, while **PISCO-Removal**, **DiffuEraser**, and **CoCoCo** complete in ∼30 minutes. For models with limited temporal receptive fields (e.g., CoCoCo), we apply a sliding window inference with a **4-frame overlap** to maintain coherence. All outputs are exported in a lossless format to avoid secondary compression artifacts during human evaluation.

5.2. Main Results

Overall comparison. Tab. 3 summarizes the overall performance. We observe a clear performance hierarchy: **PISCO-Removal** and **UniVideo** consistently define the state-of-the-art across all metrics, forming a high-fidelity tier. In contrast, while **DiffuEraser** provides competitive efficiency, it suffers from a significant "fidelity gap" compared to the leaders. The most striking finding is the universal performance decay on the *Hard* subset, where even the top-performing models struggle to maintain physical consistency, highlighting the diagnostic value of our benchmark.

Instruction following. This dimension evaluates the model's ability to ground textual instructions into pixel-level removal, where we observe a clear trade-off between semantic autonomy and execution reliability. **UniVideo**

Table 2. **Decoupled Human Evaluation Rubric.** Each of the three dimensions—Instruction Following (IF), Rendering Quality (RQ), and Edit Exclusivity (EE)—is scored independently on a 1–4 scale, enabling granular diagnosis of model failure modes.

Score	Instruction Following	Rendering Quality	Edit Exclusivity
Score 1	*Target not removed or unrelated edit*	*Severe artifacts; unusable output*	*Uncontrolled edits across scene*
Score 2	*Partial removal with obvious residual traces*	*Obvious distortion/flicker; poor temporal consistency*	*Multiple non-target regions altered*
Score 3	*Target mostly removed with minor artifacts*	*Moderate quality degradation but viewable result*	*Minor non-target changes but structure mostly preserved*
Score 4	*Target precisely removed with no obvious residuals*	*High visual quality and stable temporal consistency*	*Non-target regions preserved with only negligible differences*

Table 3. Performance comparison on the comprehensive benchmark (95 videos), including overall results and the Simple/Hard split breakdown. Metric scores range from 1 to 4, where higher values indicate superior performance. **Dark green** and light green denote the best and second-best results within each subset, respectively.

Subset	Method	Instruction Following ↑	Rendering Quality ↑	Edit Exclusivity ↑	Overall ↑
Overall	CoCoCo	1.60	1.84	3.07	2.17
	UniVideo	3.06	**3.45**	3.53	3.35
	DiffuEraser	2.89	2.63	3.52	3.01
	PISCO-Removal	**3.62**	3.28	**3.58**	**3.49**
Simple	CoCoCo	1.75	1.98	3.33	2.35
	UniVideo	3.21	**3.34**	3.53	3.36
	DiffuEraser	2.73	2.73	**3.73**	3.06
	PISCO-Removal	**3.75**	3.32	3.57	**3.55**
Hard	CoCoCo	1.52	1.76	2.92	2.07
	UniVideo	2.96	**3.53**	3.53	3.34
	DiffuEraser	3.00	2.56	3.38	2.98
	PISCO-Removal	**3.45**	3.23	**3.59**	**3.42**

and **PISCO-Removal**, as representative end-to-end architectures, demonstrate superior visual integration in the majority of cases; however, they exhibit occasional "grounding drift" where the target subject is either ignored or only partially erased. This suggests that while their latent semantic alignment is powerful, it can occasionally fail to localize the instance accurately amidst cluttered backgrounds. In contrast, **DiffuEraser** achieves the highest reliability in complete removal across the benchmark. Since its pipeline is explicitly constrained by the input mask, it bypasses the semantic localization errors inherent in end-to-end models, ensuring the designated regions are always processed. Notably, **CoCoCo** consistently fails this dimension; as a general-purpose inpainting model, it lacks the specialized inductive bias for large-scale instance removal, often producing "ghosting" artifacts that retain the original object's silhouette or failing to initiate the removal command altogether in favor of local texture synthesis.

Rendering quality. This dimension evaluates visual fidelity and temporal stability, where we observe a stark contrast in spatial resolution and coherence. **UniVideo** and **PISCO-Removal** consistently produce the most visually pleasing results, maintaining high-frequency textures that blend seamlessly with the original background. While they exhibit occasional flickering in complex dynamic scenes, their overall rendering remains stable at 720p. In contrast, **DiffuEraser** suffers from pervasive *spatial blurring* within the inpainted regions. Despite its ability to reliably remove the target, the synthesized textures often lack the sharpness of the surrounding environment, creating a noticeable "patchwork" effect that disrupts the scene's visual harmony. **CoCoCo** performs the worst in this category, frequently generating severe "ghosting" artifacts—where remnants of the original object reappear as semi-transparent textures—and significant temporal flickering. These failures suggest that while general inpainting models can fill small holes, they lack the structural priors necessary to reconstruct large-scale, 81-frame backgrounds with the requi-

site physical and temporal plausibility.

Edit exclusivity. This dimension assesses the models' precision in localized editing, specifically focusing on whether the transformation "leaks" into non-target background regions. Overall, all four baselines demonstrate a respectable ability to preserve the surrounding scene context. **UniVideo** and **PISCO-Removal** exhibit high spatial fidelity, maintaining the original pixel values of the static background with minimal drift. However, **DiffuEraser** occasionally suffers from *blurring leakage*, where the spatial smoothing intended for the erased region inadvertently spreads beyond the mask boundaries into the neighboring textures. This creates a subtle but perceptible halo of reduced sharpness in the background, a phenomenon that is particularly visible at 720p. For the *Hard* subset involving complex reflections, we observe that models often struggle to disentangle the target instance from its environmental "side effects," sometimes leading to unintended modifications of the global lighting or shadow maps in areas that should remain untouched.

5.3. Simple vs. Hard Split Analysis

The split-wise analysis isolates model robustness against physics-coupled interactions. As detailed in Tab. 3, we observe a noticeable performance degradation across multiple dimensions when transitioning from the *Simple* to the *Hard* subset. Looking beyond the aggregate scores, the multi-dimensional breakdown reveals that the most significant drops typically occur in **Instruction Following** and **Rendering Quality**. For instance, top-tier models like **PISCO-Removal** experience a drop in IF from 3.75 on the Simple set to 3.45 on the Hard set, illustrating the increased difficulty of executing clean removals amidst complex physical constraints. While all four baselines can reliably erase an isolated object against a static background in the *Simple* set, they frequently struggle to maintain structural consistency when the task scales in physical complexity.

The interaction types involving **specular reflections** and **dynamic wakes/fluid ripples** prove to be the most challenging. In many *Hard* cases, even when the primary instance is successfully erased by leaders like **UniVideo** or **PISCO-Removal**, its corresponding physical "side effects"—such as a moving shadow on a textured wall or a mirror reflection on a car's surface—remain stubbornly visible. This severely impacts the **Edit Exclusivity** scores for methods like **DiffuEraser**, which drops from 3.73 (Simple) to 3.38 (Hard). Its local blurring strategy fails to properly propagate the background's global illumination, leading to physically implausible "ghost reflections." These findings underscore that current Video Instance Removal (VIR) models generally treat removal as a 2D texture-filling task rather than a 3D-aware scene reconstruction, highlighting a critical direction for future physics-augmented research.

5.4. Cross-Metric Trade-off Analysis

To analyze whether methods sacrifice one dimension for another, we examine the pairwise relationships in Fig. 2. This visualization exposes the inherent tension between a model's "semantic fluidity" and its adherence to local constraints.

High-Fidelity Clustering. **PISCO-Removal** and **UniVideo** (blue and green) exhibit tight clustering in the (4,4) quadrants of the IF-RQ and IF-EE plots, defining a high-fidelity tier that balances task completion with background preservation. However, a clear shift toward lower RQ scores is evident for the **Hard** subset ($\times$), confirming that complex physical coupling remains the primary performance bottleneck even for top-tier models.

Reliability vs. Precision. **DiffuEraser** (red) achieves high IF scores (Score 3–4) due to its mask-guided nature, yet shows significant dispersion on the RQ and EE axes. This reflects a trade-off where reliable object removal often introduces spatial blurring or "halo" artifacts in non-target regions. Conversely, **CoCoCo** (purple) is heavily skewed toward low IF scores (Score 1–2), often failing to execute the removal command while maintaining high EE scores simply by leaving the scene unedited.

Metric Correlations. Statistical analysis reveals that IF and RQ are moderately correlated ($r \approx 0.65$), suggesting that models with better instruction grounding typically synthesize more plausible textures. However, the weak correlation between RQ and EE in lower-performing models underscores the necessity of our decoupled protocol: visual plausibility alone does not guarantee that the rest of the scene remains untouched.

5.5. Discussion

Why decoupled evaluation matters. Traditional VIR evaluation often relies on a single visual plausibility score, which masks critical failure modes. Our three-axis protocol reveals that a model can score highly on *instruction following* while simultaneously failing *edit exclusivity*, or produce visually plausible textures while completely missing the target removal. By decoupling these dimensions, we expose the inherent trade-offs in current architectures—such as the balance between the strict spatial constraints of mask-guided models like **DiffuEraser** and the semantic fluidity of end-to-end generative models like **UniVideo**. This granular mapping is essential for diagnosing whether a model's failure stems from poor instruction grounding, low rendering fidelity, or unintended scene drift.

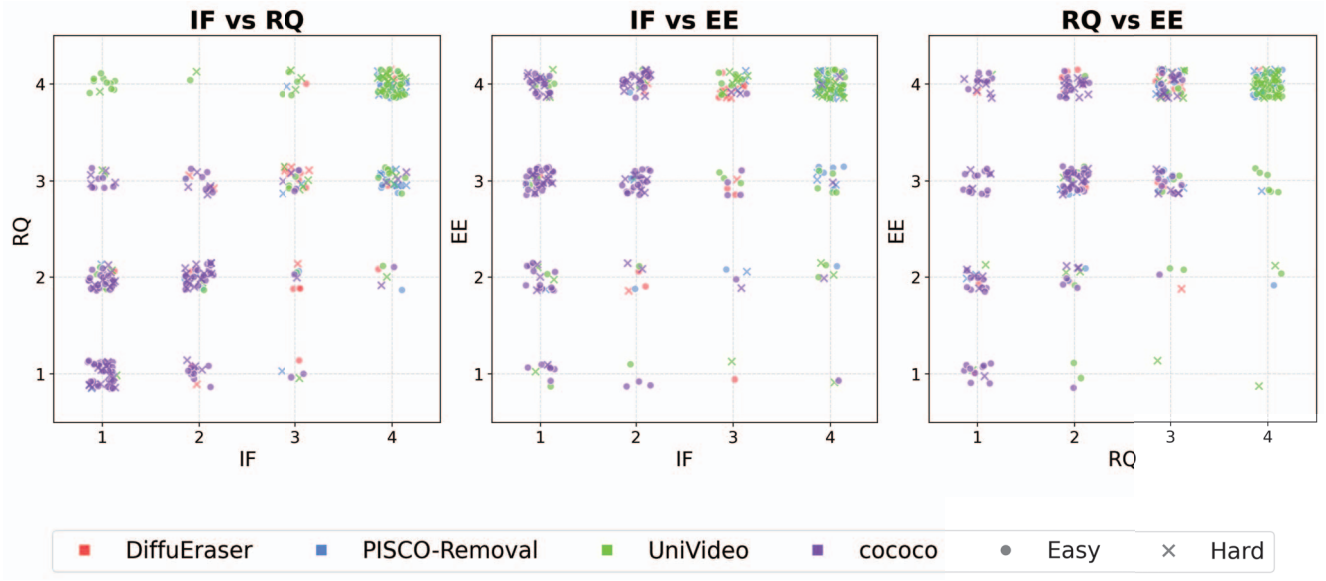

Figure 2. **Cross-metric trade-off analysis across Instruction Following (IF), Rendering Quality (RQ), and Edit Exclusivity (EE).** Each point represents a single video-model pair. Markers distinguish between the *Simple* (•) and *Hard* (×) subsets. The tight clustering in the top-right quadrant for **PISCO-Removal** and **UniVideo** indicates high-fidelity, balanced performance. Conversely, the wide dispersion of **DiffuEraser** and **CoCoCo** reveals inherent trade-offs between reliable target removal and background preservation.

Current limitations. Current benchmark evaluations remain heavily human-centered, making them both cost-sensitive and difficult to scale. While we provide a comprehensive human study, the development of robust automatic metrics for physics-aware side effects—such as detecting residual reflections, lingering shadows, or fluid inconsistencies—remains highly challenging. Existing feature-space proxies fail to capture this high-level physical logic, indicating an urgent need for automated, physics-aware evaluation models.

Future directions. Our empirical results highlight a clear paradigm shift: large-scale generative models significantly outperform smaller, traditional inpainting networks. Furthermore, within the regime of large models, specialized architectures fine-tuned for precise editing demonstrate noticeable superiority over unified, general-purpose video models in handling strict spatial constraints. This suggests that while foundational priors are necessary, they are not sufficient for instance-level physical accuracy. Consequently, future advancements in video instance removal should focus on two key pillars: the continued scaling of large video foundation models, and the rigorous curation of high-quality, domain-specific datasets designed to inject precise physical and spatial constraints into these models.

Benchmark extension roadmap. As an evolving platform, the PVIR roadmap includes expanding the dataset to encompass greater category diversity, extended sequences (e.g., 10+ seconds) to stress-test long-term temporal consistency, and more complex physical phenomena such as multi-object interactions and fluid/smoke dynamics.

6. Conclusion

In this work, we introduced the Physics-Aware Video Instance Removal (PVIR) benchmark, a dedicated evaluation infrastructure designed to bridge the gap between visual plausibility and physical consistency in video editing. Our benchmark contributes three key pillars: a high-quality 95-video dataset with dense annotations, a Simple/Hard difficulty split driven by complex physical interactions, and a decoupled evaluation protocol spanning instruction following, rendering quality, and edit exclusivity. By benchmarking four representative models under a unified, high-resolution inference setup, we revealed a significant performance hierarchy and a universal "physics blindness" in current state-of-the-art methods.

Crucially, our analysis demonstrates that while existing models excel in static background synthesis, they suffer from severe performance when encountering optical reflections, shadows, and fluid interactions in our Hard subset. This finding underscores that current video instance removal is still predominantly treated as a 2D texture-filling task rather than a 3D-aware physical reconstruction. We offer PVIR as a rigorous yardstick to encourage the community to move beyond surface-level aesthetic metrics and toward physically grounded video intelligence.

References

[1] Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. Ivebench: Modern benchmark suite for instruction-guided video editing assessment. *arXiv preprint arXiv:2510.11647*, 2025. 2

[2] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Fgvc: Flow-guided video completion. In *CVPR*, 2019. 2

[3] Xiangbo Gao, Renjie Li, Xinghao Chen, Yuheng Wu, Suofei Feng, Qing Yin, and Zhengzhong Tu. Pisco: Precise video instance insertion with sparse control. *arXiv preprint arXiv:2602.08277*, 2026. 1, 3, 5

[4] Xiangbo Gao, Mingyang Wu, Siyuan Yang, Jiongze Yu, Pardis Taghavi, Fangzhou Lin, and Zhengzhong Tu. The pulse of motion: Measuring physical frame rate from visual dynamics. *arXiv preprint arXiv:2603.14375*, 2026. 2

[5] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2

[7] Jonathan Ho, William Chan, and Pieter Abbeel. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[8] Weizhe Huang, Xiaofeng Liu, Yifan Wang, Xin Li, et al. Vbench++: Comprehensive and versatile benchmark for video understanding and generation. *arXiv preprint*, 2024. 2

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

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[11] Xuan Li, Yujie Wang, Chuhang Zhang, Bin Zhao, and Ying Shan. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 2

[12] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*, 2025. 1, 2, 3, 5

[13] Zhen Li, Cheng Xie, Weidi Zhang, Yebin Liu, Qi Tian, and Ying Shan. Towards an end-to-end framework for flow-guided video inpainting. In *CVPR*, 2022. 2

[14] 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*, pages 14040–14049, 2021. 2

[15] 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. 1, 2, 3, 5

[16] Saman Motamed, William Harvey, Benjamin Klein, Luc Van Gool, Zhuoning Yuan, and Ta-Ying Cheng. Void: Video object and interaction deletion. *arXiv preprint arXiv:2604.02296*, 2026. 1

[17] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[18] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 3

[19] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[21] Alexandros Stergiou and Ronald Poppe. Adapool: Exponential adaptive pooling for information-retaining downsampling. 2021. 3

[22] Roman Suvorov, Ekaterina Logacheva, Anton Mashikhin, Mikhail Melnikov, Mikhail Kaigorodov, Sergey Yudin, Denis Davydov, Anastasia Molchanova, Artem Malkov, Alexander Ilin, et al. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, 2022. 2

[23] Thomas Unterthiner, Bernhard Nessler, Guenter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2

[24] Jiayu Wang, Yicheng Zhang, Kai Liu, Xin Sun, Lijuan Ye, Yunchao Wei, et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2024. 2

[25] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025. 1, 2, 3, 5

[26] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Z. Lei, Yuchao Gu, Bolei Huo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2023. 2

[27] Ning Xu, Linjie Yang, Yuchen Fan, Dingkang Yue, Yuchen Liang, James Yang, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 2

[28] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 2

[29] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10477–10486, 2023. 2

[30] Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Rong Xiao, Kam-Fai Wong, and Lei Zhang. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11067–11076, 2025. 1, 2, 3, 5