VOID: Video Object and Interaction Deletion

Saman Motamed[1,2]*, William Harvey[1], Benjamin Klein[1], Luc Van Gool[2], Zhuoning Yuan[1], and Ta-Ying Cheng[1]**

[1] Netflix
[2] INSAIT, Sofia University "St. Kliment Ohridski"

void-model.github.io

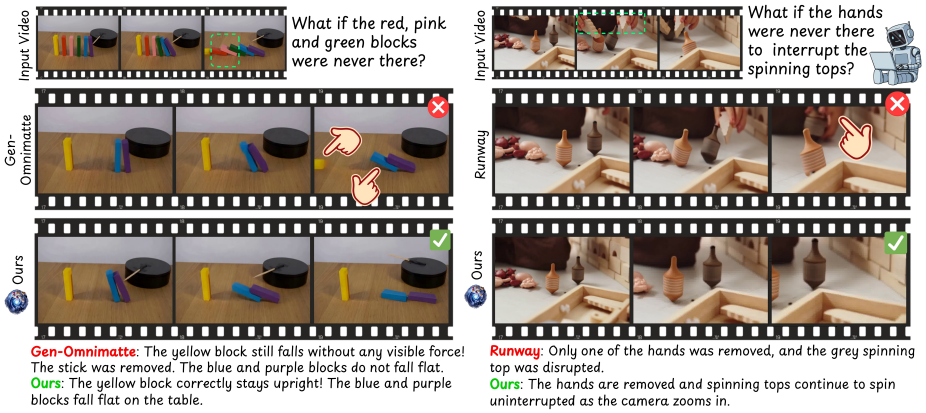

Fig. 1: Removing an object and its interactions can require rewriting the entire scene. On the left, when the middle three blocks are removed, VOID correctly models the domino effect halting so that the yellow block never falls. On the right, when the hands are removed, VOID correctly models the spinning tops continuing without interruption.

Abstract. Existing video object removal methods excel at inpainting content "behind" the object and correcting appearance-level artifacts such as shadows and reflections. However, when the removed object has more significant interactions, such as collisions with other objects, current models fail to correct them and produce implausible results. We present VOID, a video object removal framework designed to perform physically-plausible inpainting in these complex scenarios. To train the model, we generate a new paired dataset of counterfactual object removals using Kubric and HUMOTO, where removing an object requires altering downstream physical interactions. During inference, a vision-language model identifies regions of the scene affected by the removed object. These regions are then used to guide a video diffusion model that generates physically consistent counterfactual outcomes. Experiments on both synthetic and real data show that our approach better preserves

* Work done during an internship at Netflix.
** Project lead.

consistent scene dynamics after object removal compared to prior video object removal methods. We hope this framework sheds light on how to make video editing models better simulators of the world through high-level causal reasoning.

Keywords: Video object removal · Video generation · Plausible video editing

1 Introduction

Videos capture the complex causal dynamics of our physical world. The presence of an object exerts a diverse range of effects on its environment, ranging from photometric phenomena like shadows and reflections to kinetic events like collisions. This spatiotemporal entanglement makes the seemingly simple task of "video object removal" non-trivial for video generation and editing models. It requires the model to first imagine "What would happen if this object was removed?" and then synthesize a realistic video of its answer. Consider the line of collapsing domino tiles shown on the left of Fig. 1. If we use a video inpainting model to remove the middle tiles, the later tiles keep falling, which is a physically impossible scenario. A realistic solution requires the model to reason that without the middle tiles, tiles appearing later in the chain must remain standing. Doing so requires an editing model to simulate the world through high-level causal reasoning and not rely solely on low-level visual features. Mastering this capability will benefit film visual effects and make advanced video editing accessible to non-experts.

Video decomposition methods [17, 22, 32, 33] aim to decompose videos into layers, such that each object has an associated layer containing its effects disentangled from those of other objects in the scene. These methods excel at extracting effects where an object creates shadows, reflections, or other distinct visual entities. However, they are not capable of disentangling interactions where one object affects another, such as by breaking or moving it. Alternatively, a text-guided video editing model [14, 31] could remove an object and its effects following a user prompt. However, current diffusion models often lack the physical intuition to determine what should happen with the object removed, and it is often impossible to specify all effects precisely with a text prompt.

To this end, we propose an extension of video object removal to more dynamic scenarios. These require not only removing a specified object, but also modeling how its removal affects other objects in the scene. We then present VOID, a framework to elicit this high-level causal reasoning from a video diffusion model.

VOID is built on advances along three axes: data construction, training strategy, and inference optimization. To create data pairs that capture dynamic changes when an object is removed, we repurpose the Kubric simulation and rendering engine [9] and the HUMOTO human motion capture dataset [23]. For training, we propose two improvements over prior work [17]: (i) "quadmask" conditioning that explicitly identifies regions of each frame that may change after the object is removed, and (ii) a video appearance refiner applied in a second

pass to remove artifacts like unwanted object morphing. During inference, we generate quadmasks with vision-language models (VLMs), leveraging their world knowledge to expand a simple object mask into richer pixel-space guidance.

We gather a new benchmark of videos with diverse and complex interactions, comprising synthetic and real-world data. Extensive studies involving perceptual metrics, user studies, and VLM-as-a-judge demonstrate compelling results from VOID against both video inpainting methods and general video editing models. We further see surprising generalizations to unseen effects: VOID models a balloon floating up when the person holding it is removed, despite there being no floating objects in its training data; VOID also prevents food inside a blender from moving when the person turning it on disappears, despite there being no blenders or electrical devices in its finetuning data. These extrapolations demonstrate that VOID does not just recall simple visual cues from its training data, but applies high-level reasoning and world knowledge from the VLM and underlying video diffusion model to video editing. As such, it is likely to benefit as more capable generative models become available.

In summary, we have three contributions. First, we investigate the current pitfalls of physics-aware object removal when the removed object has complex interactions with the rest of the scene. Second, we introduce VOID, a framework that tackles these problems from three perspectives: data curation, training strategy, and inference time VLM-guided scene analysis. Finally, we present extensive evaluations on previous and new benchmarks that show the superiority of our model in disentangling and removing a wide range of object effects. We believe that this work illuminates an interesting and underexplored direction for video generation and world modeling research.

2 Related Work

Video Generation and Editing. The breakthrough of diffusion and flow-matching methods has led to large advancements in video generation models. Notable examples include closed-source models like Veo 3 [7] and Runway [31] and open-source models such as WAN [35], VACE [14], CogVideo [38], and LTX-2 [11]. Much of the stunning performances of these models came from the large video-text datasets driving forward the performance, allowing them to extend to various editing controls including text and sketches. However, since they are learning from unstructured data, they often create pixel-perfect yet physically implausible scenes, especially for tasks that require extensive reasoning (e.g., removing an object).

Several models aim to improve reasoning via VLMs [12, 16, 39]. However, these models either work in a specific domain (e.g., LangDriveCtrl [12] works purely for driving scenes) or only solve simplee reasoning tasks like grounding (Video-Repair [16]) and segmentation (Veggie [39]). VOID takes a leap towards applying VLM reasoning for complex video editing tasks, where we need to synthesize counterfactual scenarios in which an object is removed.

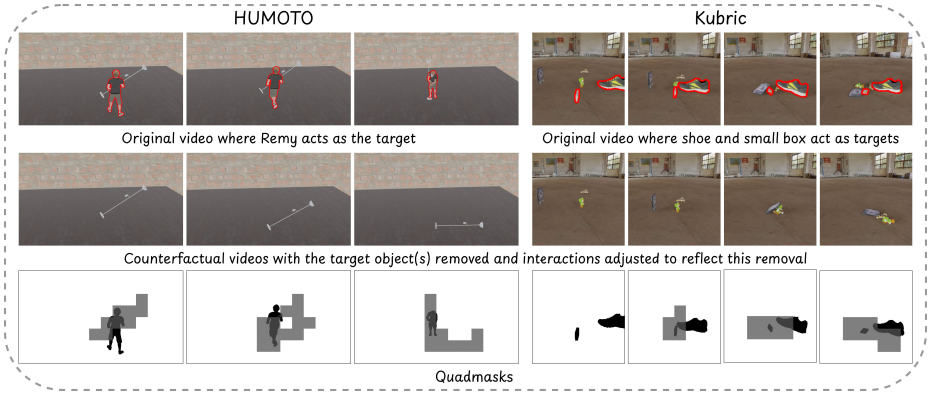

Fig. 2: Counterfactual supervision examples. Top: videos $\mathbf{V}$ where O is outlined in red. Bottom: re-simulated counterfactuals $\hat{\mathbf{V}}$ generated without O. In Kubric scenes, downstream motion changes when the initiating object is removed. In HUMOTO scenes, supported objects transition naturally under gravity.

Video Decomposition and Effect Removal. The problem of decomposing a video into RGBA layers was significantly advanced by Omnimatte [22], which introduced a self-supervised framework to associate subjects with their "effects" (e.g., shadows and reflections). While OmnimatteRF [20] extended this by modelling the static background with 3D radiance fields, these methods remain fundamentally reconstructive, focusing on uncovering existing background pixels rather than synthesizing new content. More recently, Generative Omnimatte [17] integrated priors from a video inpainting model, using a trimask setup to decompose images into object-specific layers and effects. OmnimatteZero [32] introduced training-free extraction of effects and objects via attention maps.

There are also a plethora of recently-released video inpainting methods. Propainter [42] proposed a dual-domain propagation from image and feature side for better inpainting. DiffuEraser [18] integrated flow-based pixel propagation with transformer-based generation to better restore textures and objects. AVID [41] and FDM [8] proposed sampling pipelines to extend video inpainting lengths. Minimax-Remover [43] is an object removal approach with a more efficient model architecture followed by distilling a remover on human annotations. ROSE [24] proposed an effect removal inpainting framework focusing on photometric effects such as shadows, reflections, light, and translucency, while ObjectWiper [15] presented a training-free method targeting these effects. While these methods all handle some associated photometric effects of the removed object, they cannot model complex physical interactions.

3 Approach

We start from an input video $\mathbf{V} = \{I_t\}_{t=1}^{T}$ and a mask sequence $\mathbf{M}_o = \{m_t\}_{t=1}^{T}$ identifying one or more target objects to remove, O. Our objective is to learn a

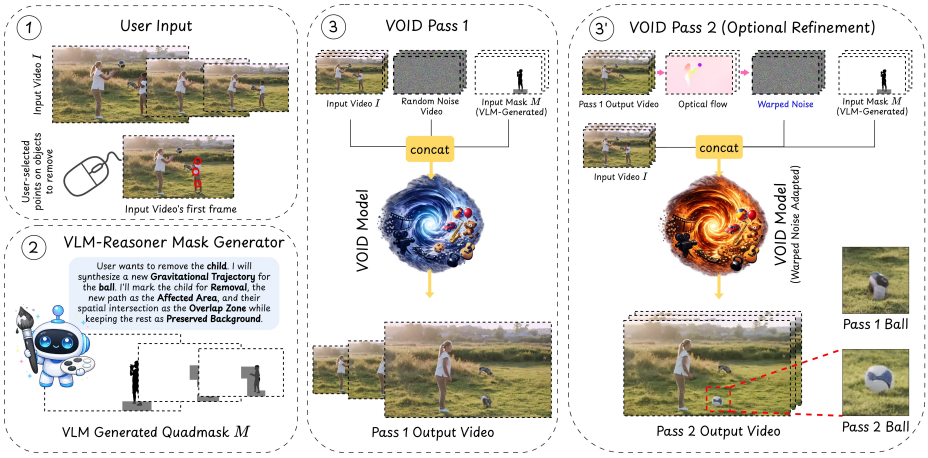

Fig. 3: VOID: Interaction-Aware Counterfactual Video Generation. A user provides an input video and clicks on an object to mask it for removal. A VLM-based pipeline expands the mask to identify other areas that will be affected. VOID's first pass then predicts a counterfactual trajectory. The optional second pass stabilizes object deformation using flow-warped noise derived from the initially predicted motion.

model f that generates a counterfactual video $\hat{\mathbf{V}}$, in which O and all induced interactions are removed:

$$\hat{\mathbf{V}} = f(\mathbf{V}, \mathbf{M}_o). \tag{1}$$

In general, the interactions can be complex. Removing a support can cause something to fall. Removing an obstacle can prevent a collision. To work well in these settings, f cannot rely on spatial hole filling but must conceptualize how the scene would have evolved in the target object's absence. It should then *(i)* eliminate the target object, *(ii)* regenerate regions affected through potentially complex relationships, and *(iii)* preserve unaffected regions.

3.1 Counterfactual Dataset Supervision

Training this model requires a dataset of counterfactual video pairs $(\mathbf{V}, \hat{\mathbf{V}})$ with and without object O, respectively. Video pairs used by existing video inpainting and omnimatte datasets [17, 24, 29, 37] focus mostly on photometric effects such as shadows and lack supervision for the removal of objects that physically affect other objects in the scene. We generate new counterfactual pairs to address this gap with physics-based simulations from Kubric [9] and human motion capture data from HUMOTO [23]. For both datasets, we randomize camera trajectories and focal zoom during rendering to help with the disentanglement of object effects and camera trajectories. We show in Fig. 6 that training on these two synthetic datasets enables extensive generalization to real-world domains.

Rigid-body dynamics (Kubric). The diverse set of objects offered by Kubric is ideal for simulating collisions, falling, and structural dependencies. We create videos $\mathbf{V}$ by sampling initial conditions of multiple objects with varying initial positions and velocities, and simulating the interactions over time. We then define one or more of the objects to be O. The counterfactual video $\hat{\mathbf{V}}$ is created by removing O while keeping initial conditions for all other objects the same, and re-simulating the scene. This yields a new, physically consistent alternative set of interactions. We generate ∼1900 videos pairs in this manner.

Articulated interactions (HUMOTO). We use HUMOTO, a 4D motion-capture dataset of human-object interactions, to collect articulated interaction data. In these sequences, O corresponds to the human performing diverse activities. $\mathbf{V}$ and $\hat{\mathbf{V}}$ are created by passes over the simulation and rendering engine with and without the human. These videos teach the model how to perform object removal in videos containing dynamic manipulations. We randomize the textures of the objects in the scene, the background wall and the human, and generate ∼4500 video pairs.

3.2 Interaction-Aware Quadmask Conditioning

To provide further guidance over the binary object mask $\mathbf{M}_o$, Lee et al. [17] proposed a trimask which distinguishes between three image regions: the object to be removed (black), the area affected by its removal (light gray), and areas which should stay the same (white). Their setup, however, creates two ambiguities.

The first is that they highlight almost the entire region of each image frame in light gray and mark only specific objects as white. Their model therefore learns that it typically needs to modify only a small portion of the light gray mask to remove the effects. We provide stronger guidance by focusing the light gray region closely on where effects take place. While generating data, we use the rendering engines to determine these regions. We then gridify the regions to better match our inference time procedure described in Sec. 3.6.

The second ambiguity occurs when there is overlap between the object to remove and the area with dynamic effects. In the example in Fig. 3, we want to remove a child and once they are removed, the ball should fall to the ground instead of them catching it. Consider what value the trimask should have around the boy's upper body while he catches the ball. Should it be black because the boy is being removed? Or should it be light gray because the "effect" of the removal is that the ball should now continue its trajectory and pass through this area? To resolve these ambiguities, we extend the trimask to a quadmask $\mathbf{M}_q$ with a fourth color (dark grey) that describes overlap between (i) the object to be removed and (ii) other parts of the scene that are affected. See Fig. 2 for examples.

3.3 Backbone Initialization and Counterfactual Generation

We propose VOID, a model built upon the CogVideoX diffusion transformer backbone [38] and initialized from the weights released with Generative Omn-

imatte [17]. This initialization provides a strong prior for layered object–effect disentanglement under trimask guidance. We finetune it with quadmask conditioning on the counterfactual video pairs described previously. This teaches the model mask semantics and re-enables the underlying video model's native capacity for physically plausible trajectory synthesis, transforming it from a layered removal model to a dynamic counterfactual rewriting model.

3.4 Pass 1: Counterfactual Trajectory Synthesis

In its first pass, VOID generates an initial counterfactual prediction:

$$\hat{\mathbf{V}}_{\text{p1}} = \text{VOID}(\mathbf{z}, \mathbf{V}, \mathbf{M}_q), \tag{2}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes the Gaussian diffusion noise, $\mathbf{V}$ is the input video sequence, and $\hat{\mathbf{V}}_{\text{p1}}$ is VOID's first-pass prediction of how the scene would evolve in the absence of the target object. This pass typically captures broadly correct hypotheses around motion, such as previously supported objects entering free-fall and previously-obstructed objects continuing their motion. However, we find that the objects undergoing newly synthesized motion can exhibit structural deformation.

3.5 Pass 2: Flow-Warped Noise Stabilization

Why deformation occurs. Video diffusion models, especially relatively lightweight models like the 5 billion parameter CogVideoX model we build on, struggle to maintain temporal coherence when generating complex, motion-heavy videos [3, 26]. Prior work mitigates this issue through large-scale tracking supervision or explicit motion conditioning [3, 10, 13, 21, 27]. In the simple object removal case where only photometric effects need to be corrected, the input video provides similarly strong constraints on the generated motion. For example, if we need to remove a shadow from a surface, the motion and geometry of the surface in the output video should be the same as in the input. In our more complex settings, the diffusion model often needs to generate new motion. We find that this leads to bending, stretching, or structural drift of the objects undergoing changed motion, similar to the artifacts seen from running the CogVideoX image-to-video model without motion guidance. We now describe how to resolve these artifacts in the object removal setting without changing the base model.

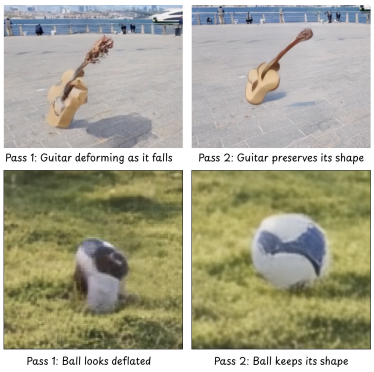

Fig. 4: Frames from generated videos featuring a guitar entering free-fall and a thrown ball following a new trajectory. Pass 1 (left) produces correct counterfactual trajectories but exhibits structural deformation. Pass 2 (right) better preserves object rigidity by using motion-aligned warped noise.

Second pass to fix deformation Go-with-the-Flow [1] observed that using temporally correlated noise based on predicted motion trajectories can encourage the diffusion model to denoise consistently along those trajectories. We follow Go-with-the-Flow [1] to derive warped noise from the optical flow field of our first-pass output $\hat{\mathbf{V}}_{\text{p1}}$, and then use it as input to a second pass as:

$$\hat{\mathbf{V}} = \text{VOID}_{\text{warp}}(\mathbf{z}_{\text{warp}}, \mathbf{V}, \mathbf{M}_q), \tag{3}$$

where $\text{VOID}_{\text{warp}}$ is a warped noise variant of VOID. It is trained with the same data and quadmask conditioning but with flow-aligned noise derived from each training target $\hat{\mathbf{V}}$.

This second pass is not always required so we trigger it only when object removal is predicted to cause substantial dynamic reconfiguration. The same VLM used to create a quadmask additionally classifies whether removal induces significant object motion (e.g., free-fall or trajectory change). We trigger the second pass only when such dynamics are detected.

3.6 VLM-Guided Quadmask Generation at Inference Time

At inference time, we start from an input video $\mathbf{V}$ and user-provided binary object mask $\mathbf{M}_o$. To run VOID, we need to first infer the affected region and use it to create the quadmask $\mathbf{M}_q$. Doing so requires reasoning about counterfactual dependencies and object dynamics, so we use a VLM [5, 19, 25]. We start by inputting $\mathbf{V}$ and $\mathbf{M}_o$ to the VLM and using it to produce a list of descriptions of objects that are affected by the removed object. We use Segment Anything 3 [2] to get a mask $\mathbf{M}_a^{\text{orig}}$ covering all objects in the list. Since the affected objects may be in different places in the counterfactual scenario, we also need to predict their counterfactual positions to fully capture the changes between $\mathbf{V}$ and $\hat{\mathbf{V}}$. We therefore feed $\mathbf{M}_a^{\text{orig}}$ into the VLM and use it to predict the mask sequence describing the positions of these objects in the counterfactual scenario. This is done by overlaying a coarse spatial grid on the input video and asking it to list which cells in each frame may contain effects. This gives us a block-structured mask $\mathbf{M}_a^{\text{count}}$ describing where the affected objects in $\mathbf{M}_a^{\text{orig}}$ go in the counterfactual scenario. We combine the masks to get the final affected area mask $\mathbf{M}_a := \mathbf{M}_a^{\text{orig}} \vee \mathbf{M}_a^{\text{count}}$. We finally compute the quadmask $\mathbf{M}_q$ by setting it to black for pixels only in $\mathbf{M}_o$; dark grey for pixels where $\mathbf{M}_o$ and $\mathbf{M}_a$ overlap; light grey for pixels only in $\mathbf{M}_a$; and white everywhere else.

4 Results

We test on two datasets. The first comprises 75 real-world videos involving object manipulation, support removal, collisions, articulated interactions, and shadow/reflection removal. The second is synthetic and consists of 30 Kubric and HUMOTO test videos combined with existing synthetic object removal datasets.

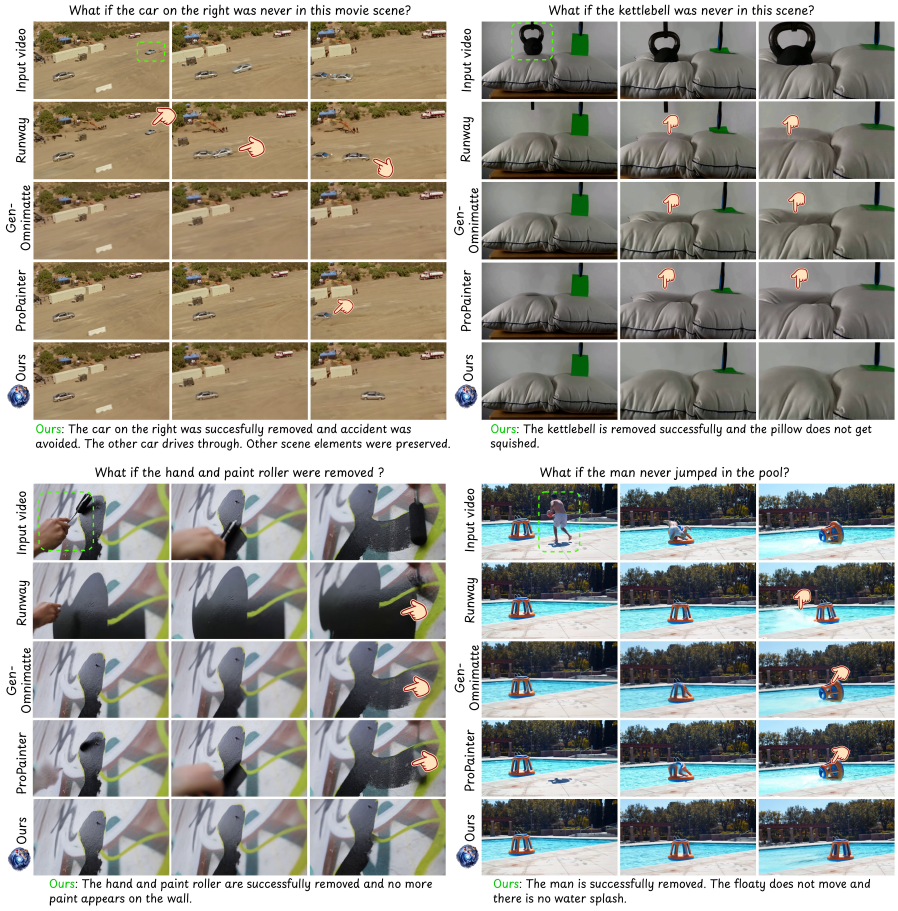

Fig. 5: Qualitative comparisons on real-world videos. VOID maintains object structure and produces plausible motion over time, while the baselines exhibit deformation (the kettlebell on the pillow, and the floaty deforming), incomplete removal (two cars crashing), or implausible outputs (paint appearing after the roller is removed).

4.1 Experimental Details

For each real-world video, a user specifies a primary object via sparse clicks, which are converted into the binary object mask $\mathbf{M}_o$ using Segment Anything 2 [30]. We convert the binary mask into a quadmask with the VLM-based pipeline in Sec. 3.6. All results in the main paper use Gemini 3 Pro as the VLM in this pipeline; we also report scores with GPT-5.2 and Qwen-3.5 VL in the appendix.

For fair comparison, each baseline is evaluated using its preferred conditioning format: binary masks for ProPainter, DiffuEraser, ROSE, and MiniMax-

Remover; trimasks for Generative Omnimatte; and natural-language editing prompts for Runway (Aleph), a commercial video editing system.

Since Runway is a text-guided editor rather than a mask-conditioned inpainting model, we explicitly describe both *(i)* the object to remove and *(ii)* the expected scene evolution after removal (e.g., "remove the person and ensure the held object falls naturally"). This makes the counterfactual requirement explicit while allowing each model to operate under its intended interface. We did not compare with Object-Wiper [15] as code is unavailable, nor OmnimatteZero [32] due to an acknowledged issue with their released code at the time of writing.

4.2 Real-World Counterfactual Comparisons

Since there are no ground truth counterfactuals for real-world videos, we evaluate with a human preference study, three VLM judges on fine-grained criteria, and several qualitative comparisons.

Human Preference Study. We conduct a user study with 25 participants to measure perceptual realism and physical plausibility of counterfactual edits. For each participant, we randomly sample 5 out of the 75 real-world scenarios, resulting in 125 total comparisons. For each video, participants saw the original input and outputs of all seven models in randomized order. They are asked to select the video that best reflects how the scene should *realistically* appear after the specified object is removed, considering visual quality, temporal consistency, blending, realism of scene evolution, and absence of artifacts. An example of the user interface for the user study is provided in the appendix.

Model	Win %
VOID (ours)	**64.8**
Runway	18.4
Gen-Omni.	11.2
DiffuEraser	4.0
ROSE	1.6
MiniMax-Rem.	0.0
ProPainter	0.0

Table 1: Human preferences on real-world edits. 25 participants each evaluated 5 scenarios.

Table 1 summarizes the results. VOID is selected **64.8%** of the time, substantially outperforming all baselines, including the closed-source Runway model which required additional text guidance on what should happen. Models optimized for traditional inpainting (e.g., ProPainter) receive few or no selections, showing that they are not automatically capable of interaction-aware synthesis.

VLM-as-a-judge evaluation. To complement human evaluation with more fine-grained criteria, we employ three VLMs (Gemini 3 Pro, GPT-5.2, and Qwen 3.5-32B) as automated judges [4,36]. Each judge scores outputs across six criteria (0–5 per category; total 30): "Interaction & Physics", "Object Removal", "Background & Artifacts", "Temporal Consistency", "Preservation", and "Sharpness". Table 2 reports the full results. Across all three judges, VOID achieves the highest total score. The overall ranking is broadly consistent across judges and aligns with the human preference study: VOID is ranked first, Runway second, and Generative Omnimatte third in most cases. The strongest and most consistent gains appear in "Interaction & Physics", which directly evaluates whether the

Fig. 6: Generalization on various object interactions. VOID removes the target object and resolves downstream physical consequences (e.g., released objects fall; prevented collisions do not occur; shadows/reflections disappear).

scene updates causally after removal. VOID correctly simulates the intuitive physics while achieving visual quality at least on par with the Runway general video editing model.

Qualitative comparisons. Fig. 5 presents qualitative comparisons on 4 real-world videos. We present 3 representative baselines: Runway for text-based video

Table 2: VLM-as-a-judge evaluation on real-world videos. Each criterion is scored in $[0, 5]$. Best per judge and column is green; second best is orange.

Judge	Model	Int.Phys↑	Obj.Rem↑	Bg.Art↑	Temp↑	Pres↑	Sharp↑	**Total↑**
Gemini-3 Pro	ProPainter	0.82	3.81	2.64	3.64	4.81	3.35	19.05
	DiffuEraser	1.48	4.43	3.53	4.12	4.84	4.13	22.53
	MiniMax-Remover	1.94	4.47	3.70	4.30	4.80	4.24	23.46
	ROSE	2.25	4.77	3.86	4.26	4.92	4.17	24.22
	Gen-Omnimatte	2.30	4.75	3.81	4.32	4.92	4.23	24.34
	Runway	2.61	4.62	4.16	4.49	4.82	4.35	25.05
	Ours	3.66	4.82	4.10	4.44	4.88	4.22	26.13
GPT-5.2	ProPainter	0.76	2.91	2.11	2.44	3.13	3.20	14.55
	MiniMax-Remover	0.81	3.49	2.51	2.71	3.17	3.67	16.31
	DiffuEraser	0.97	3.53	2.67	2.79	3.32	3.64	17.00
	ROSE	1.39	4.13	2.95	2.93	3.51	3.67	18.59
	Gen-Omnimatte	1.33	4.05	3.25	3.31	3.56	3.73	19.23
	Runway	1.85	3.59	3.60	3.48	3.75	3.89	20.21
	Ours	3.19	4.35	3.48	3.88	4.41	3.81	23.16
Qwen3.5-32B	ProPainter	1.93	4.43	2.96	3.89	5.00	4.00	22.21
	MiniMax-Remover	1.68	4.60	3.12	4.03	5.00	4.03	22.45
	Runway	2.03	4.25	3.36	4.07	4.95	4.05	22.79
	ROSE	1.75	4.45	3.41	4.17	4.93	4.04	22.96
	DiffuEraser	2.19	4.75	3.04	4.00	5.00	4.00	23.04
	Gen-Omnimatte	2.19	4.89	3.16	4.08	5.00	4.05	23.44
	Ours	2.64	4.75	3.55	4.24	5.00	4.12	24.49

editing, Gen-Omnimatte for effect removal inpainting, and Propainter for traditional video inpainting without effect removal. Additional results from other baselines are presented in the appendix. All baselines present various types of failure cases, such as not removing anything or removing more than intended (two-car crashing example) or creating physically implausible scenes (pillow squished without kettlebell, floaty falling without a collision, and paint still appearing after the paint roller was removed). VOID exhibits high generalization capabilities across all examples, performing accurate object and effect removal while ensuring the scene remains physically plausible and artifact-free.

Generalizations to unseen effects. Figure 6 shows generations by VOID on samples from our real-world dataset involving effects unseen in training. To our delight, VOID is frequently able to extrapolate to these new types of physical interactions. It disentangles complex motions, such as a Jenga tower being simultaneously pushed by a hand and a cat, and a bowling ball hitting multiple bowling pins. It infers physics effects not present in the training dataset, such as a balloon floating up after the holder is removed, and a blender not turning on when the person pressing it is removed. Finally, it remains robust to removing the reflection of the Big Ben tower, letting the stick fall when the dog chewing it is removed, and corrects the ball rolling trajectory when the ducky obstacle is removed. The diverse set of effects are strong indicators that VOID learns to leverage the intuitive physics reasoning of the VLM and video diffusion base

Table 3: Synthetic benchmark evaluation on 10 classic shadow/reflection removal cases and 30 dynamic interaction cases (Kubric + HUMOTO) capturing a wide range of effects. All metrics measure fidelity to the ground-truth counterfactual targets.

Model	PSNR↑	LPIPS↓	DreamSim↓	DINOv2↑	FVD↓	VLM-Judge↑
MiniMax-Remover	29.96	0.11	0.09	0.91	448.43	22.83
ProPainter	30.48	0.10	0.10	0.89	471.13	21.38
DiffuEraser	30.11	0.12	0.10	0.89	496.61	21.30
ROSE	29.21	0.13	0.11	0.89	480.18	21.62
Gen-Omnimatte	29.44	0.12	0.12	0.87	437.88	20.40
Runway	26.68	0.11	0.15	0.85	442.76	21.67
Ours	31.49	0.12	0.07	0.92	260.31	25.10

model in a general manner, letting it excel on tasks far from the synthetic data we use to train it. We provide more video examples in the appendix.

4.3 Synthetic Dataset Comparisons

To compute metrics requiring ground-truth counterfactual targets, we take a synthetic benchmark of 10 videos focusing on object/shadow/reflection removal used by prior work [17] and add another 30 dynamic counterfactual cases from Kubric and HUMOTO that captures a wider range of object interactions. These includes altered collision outcomes and released objects entering free fall. These videos were held-out from our training dataset.

We follow previous work in reporting pixel-based metric PSNR and perceptual metric LPIPS [40]. However, with the new dataset introducing more diverse sets of effects, we also add in the more recent frame-wise perceptual metrics DreamSim [6] and DINOv2 [28], as well as video metric FVD [34]. These can better capture intricate and semantically high-level effects. We also include a VLM-Judge evaluation by Gemini 3 Pro, which is the closest aligned with the human evaluation on our real-world dataset. We modify our VLM-judge protocol on this dataset by showing the judge the ground-truth counterfactual video in addition to the model output. The same six-category scoring protocol (0–5 per criterion; total 30) is applied as in the real-world setting, but judges now assess fidelity relative to the true counterfactual outcome.

VOID achieves the strongest performance across all metrics except LPIPS. Note that this frame-wise LPIPS metric is sensitive to local translations, and therefore can penalize counterfactual effects being generated in slightly incorrect regions. For example, if we remove a person holding a stick, a model accurately portraying a stick falling but at the wrong speed may be penalized more than a model that removes the stick altogether The largest margin between VOID and baselines appears in the FVD and the VLM-judge metrics, which are the two comprehensive video-level metrics we report. This is strongly supportive of our claim that VOID excels at producing physically plausible and semantically coherent videos.

Table 4: Ablation study evaluated by VLM judge on 75 real-world test cases.

Model (Dataset Size)	Int.Phys↑	Obj.Rem↑	Bg.Art↑	Temp↑	Pres↑	Sharp↑	**Total**↑
Kubric-Only (1200)	2.63	4.06	2.33	3.46	4.34	3.54	20.36
HUMOTO-Only (1200)	2.50	4.22	2.36	3.33	4.30	3.41	20.12
Both Datasets (1200)	3.04	4.30	2.31	3.87	4.45	3.96	21.93
Gen-Omni. Mask (Full)	3.30	4.73	3.04	4.04	4.22	4.06	23.39
VOID (Full)	3.66	4.82	4.10	4.44	4.88	4.22	26.13

4.4 Ablation

Table 4 presents ablations on our training data and quadmask strategy. To best capture model robustness and generalization, all variants are evaluated on the same 75 real-world test cases using Gemini 3 Pro as a VLM judge. See the appendix for a further ablation on VOID's second pass.

Data composition. To analyze the effect of our datasets, we train ablations with three alternative datasets: Kubric-Only is trained on a 1200 sample subset of Kubric; HUMOTO-Only is trained on a 1200 sample subset of HUMOTO; and Both Datasets is trained on another 1200 samples split equally between Kubric and HUMOTO. In Tab. 4 we see that Kubric-Only and HUMOTO-Only both underperform Both-Datasets, meaning the diversity we get by mixing Kubric and HUMOTO data is beneficial even when the dataset size is held constant.

Masking strategy. We train another ablation that uses less detailed trimasks, similar to Generative Omnimatte [17], so that we can drop the VLM-guided mask generation pipeline. These masks are simply black wherever the object to be removed is and light gray everywhere else, meaning that there are no constraints on what parts of the video the diffusion model can change. Table 4 shows that this ablation degrades performance across all categories, confirming the importance of our detailed masks and our mask generation pipeline.

5 Conclusion

We present VOID, an object removal framework that generates the counterfactual video corresponding to an object is removed. VOID is built upon two new paired datasets of counterfactual object removal videos derived from the Kubric engine and HUMOTO dataset. We also present a VLM-guided quadmask generation pipeline to guide VOID into generating physics-informed counterfactual videos. Through extensive evaluations against inpainting and text-guided video model baselines on synthetic and real-world data, we show that VOID excels at modeling complex dynamics which can follow on from object removal. It also generalizes to a broad range of scenarios far from our training data. VOID is a strong starting point for future research to continue transferring strong world modeling capabilities to the video editing domain.

Limitations and future work. Despite the various generalization capabilities VOID exhibits, there are still certain domain gaps we observe, such as when test videos have the cameras at an unusual angle or too close to the object. Future work could obtain better training datasets beyond rendering engines. The generated video lengths are still in the range of a few seconds, and resolutions could be further improved.

References

1. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., Ryoo, M., Debevec, P., Yu, N.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In: CVPR (2025), licensed under Modified Apache 2.0 with special crediting requirement
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719
3. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. arXiv preprint arXiv:2502.02492 (2025)
4. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: Forty-first International Conference on Machine Learning (2024)
5. Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., Wang, Y.: Physbench: Benchmarking and enhancing vision-language models for physical world understanding. arXiv preprint arXiv:2501.16411 (2025)
6. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
7. Google DeepMind: Veo 3 technical report. Tech. rep., Google (2025), https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf
8. Green, D., Harvey, W., Naderiparizi, S., Niedoba, M., Liu, Y., Liang, X., Lavington, J., Zhang, K., Lioutas, V., Dabiri, S., et al.: Semantically consistent video inpainting with conditional diffusion models. arXiv preprint arXiv:2405.00251 (2024)
9. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3761 (2022)
10. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
11. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
12. He, Y., Pittaluga, F., Jiang, Z., Zwicker, M., Chandraker, M., Tasneem, Z.: Langdrivectrl: Natural language controllable driving scene editing with multi-modal agents. arXiv preprint arXiv:2512.17445 (2025)
13. Jeong, H., Huang, C.H.P., Ye, J.C., Mitra, N.J., Ceylan, D.: Track4gen: Teaching video diffusion models to track points improves video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7276–7287 (2025)
14. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598 (2025)

15. Kushwaha, S.S., Nag, S., Tian, Y., Kulkarni, K.: Object-wiper: Training-free object and associated effect removal in videos. arXiv preprint arXiv:2601.06391 (2026)
16. Lee, D., Yoon, J., Cho, J., Bansal, M.: Videorepair: Improving text-to-video generation via misalignment evaluation and localized refinement. arXiv preprint arXiv:2411.15115 (2024)
17. 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: arXiv preprint arXiv:2411.16683 (2024)
18. Li, X., Xue, H., Ren, P., Bo, L.: Diffueraser: A diffusion model for video inpainting. arXiv preprint arXiv:2501.10018 (2025)
19. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A survey of state of the art large vision language models: Benchmark evaluations and challenges (2025)
20. Lin, G., Gao, C., Huang, J.B., Kim, C., Wang, Y., Zwicker, M., Saraf, A.: Omnimatterf: Robust omnimatte with 3d background modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23471–23480 (2023)
21. Liu, Z., Yanev, A., Mahmood, A., Nikolov, I., Motamed, S., Zheng, W.S., Wang, X., Van Gool, L., Paudel, D.P.: Intragen: Trajectory-controlled video generation for object interactions. arXiv preprint arXiv:2411.16804 (2024)
22. Lu, E., Cole, F., Freeman, W.T., Dekel, T., Zisserman, A., Rubinstein, M.: Omnimatte: Associating objects and their effects in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
23. Lu, J., Huang, C.H.P., Bhattacharya, U., Huang, Q., Zhou, Y.: Humoto: A 4d dataset of mocap human object interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10886–10897 (2025)
24. Miao, C., Feng, Y., Zeng, J., Gao, Z., Liu, H., Yan, Y., Qi, D., Chen, X., Wang, B., Zhao, H.: Rose: Remove objects with side effects in videos. arXiv preprint arXiv:2508.18633 (2025)
25. Motamed, S., Chen, M., Van Gool, L., Laina, I.: Travl: A recipe for making video-language models better judges of physics implausibility. arXiv preprint arXiv:2510.07550 (2025)
26. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? arXiv preprint arXiv:2501.09038 (2025)
27. Motamed, S., Van Gansbeke, W., Van Gool, L.: Investigating the effectiveness of cross-attention to unlock zero-shot editing of text-to-video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 7406–7415 (June 2024)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
29. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
30. 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. arXiv preprint arXiv:2408.00714 (2024)
31. Runway Research: Runway gen-4: Advanced video generation model. https://runwayml.com (2025), accessed: 2026-02-24
32. 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) (2025)

33. Shrivastava, G., Lim, S.N., Shrivastava, A.: Video decomposition prior: Editing videos layer by layer. In: The Twelfth International Conference on Learning Representations (2024)
34. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
35. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
36. Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., Li, C.: Llava-critic: Learning to evaluate multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13618–13628 (2025)
37. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
38. 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)
39. 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)
40. 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)
41. Zhang, Z., Wu, B., Wang, X., Luo, Y., Zhang, L., Zhao, Y., Vajda, P., Metaxas, D., Yu, L.: Avid: Any-length video inpainting with diffusion model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7162–7172 (2024)
42. 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) (2023)
43. 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)