PhysRemover: A Unified Framework for Physically Realistic Object Removal

Dongxu Yue[1*] Weixuan Jin[1,*] Yuhang Yu[2] Bo Li[2†] Qi Wen[2] Xinrui Chen[1]
Jinwei Chen[2] Tianyi Zheng[2] Hao Zhang[2] Zhihai He[3,†] Chun Yuan[1,†]
[1]Shenzhen International Graduate School, Tsinghua University
[2]vivo [3]Southern University of Science and Technology
ydx25@mails.tsinghua.edu.cn, yuanc@sz.tsinghua.edu.cn

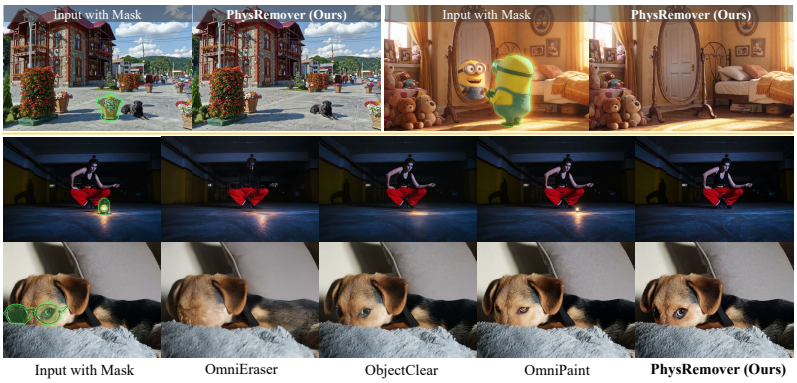

Figure 1: **Physically Realistic Object Removal.** The proposed **PhysRemover** addresses a challenging dual objective: seamlessly eliminating objects and their associated physical side effects, as well as intricate transparent materials, while faithfully preserving the underlying background semantics.

Abstract

While diffusion models have revolutionized image generation and editing, achieving physically realistic object removal remains a challenge. Physically realistic object removal posits that if an object ceases to exist, all its directly associated transient physical effects must also vanish. However, existing methods either confine the removal strictly to the masked region or limit "side effects" to simple shadows and reflections, ignoring complex interactions like caustics, emitted light, and volumetric media. Furthermore, they struggle with transparent objects, destroying the underlying background semantics. To bridge these gaps, we formalize the task of physically realistic object removal and propose **PhysRemover**, a unified framework addressing both the elimination of objects with their physical effects and the handling of transparent materials. Specifically, we incorporate an *In-Context Contrastive Guidance* mechanism guided by both masked and unmasked image, coupled with a *Learnable Removal Trigger*. This design effectively resolves the conflict between erasing the object and its physical side effects, and preserving the underlying background. Furthermore, we introduce **PhyTrace**, a large-scale image dataset capturing diverse physical effects, along with two benchmarks tailored for the proposed task of physically realistic object removal. Additionally, we develop a novel metric, **PhysRM**, for evaluating object removal quality, establishing a new standard for physically realistic image editing. Extensive experiments demonstrate that PhysRemover effectively eliminates objects and their physical side effects while faithfully preserving the background when removing transparent objects, outperforming existing methods in these challenging settings.

*Equal contribution.
†Corresponding authors: Bo Li, Zhihai He, and Chun Yuan.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Recent advances in generative models [32, 15] have demonstrated remarkable capabilities not only in image generation but also in high-fidelity image editing [13, 35, 1]. Among real-world applications, object removal has emerged as a pivotal task, enabling users to erase unwanted elements Traditional approaches typically follow an image inpainting pipeline, which focuses on suppressing the visual presence of the target object within a masked region. However, these methods often ignore the associated physical effects exerted by the object on the scene (*e.g.*, shadows and reflections).

While recent works [39, 38] have attempted to address object removal with its side effects, our analysis reveals that these methods rely on a constrained definition of "effects", limiting them to cast shadows and reflections. However, the physical effects of an object extend far beyond these basic interactions. They encompass a wide range of physically associated phenomena, such as the illumination cast by a light source or the volumetric steam generated by fire. Furthermore, existing methods face a dilemma when dealing with transparent objects: **none of them can eliminate the transparent object while preserving the underlying background**, as shown in Fig. 1.

Despite object removal being a long-standing task, three challenges remain unresolved: (1) Holistic Elimination: Removing not just the object but its physically associated side effects; (2) Transparency Handling: Preserving the background semantics originally visible through transparent objects; and (3) Hallucination-free Restoration: Reconstructing complex backgrounds without introducing unwanted artifacts or altering original details. As shown in Fig. 1, even SOTA models fall short of these goals. This failure stems partly from the limited definition of physical effects and partly from the condition paradigm, where the input mask fully occludes the target, preventing the model from perceiving the surrounding context of the object and the background behind transparent surfaces.

In addition, the absence of large-scale, publicly available datasets for physically realistic object removal limits further progress in the field. Early datasets relied on naive "copy-paste" heuristics [20, 26] or pseudo-ground-truths from inpainting models [43], ignoring physically realistic interactions. While recent efforts [39, 45] have collected real-world data considering shadows and reflections, these datasets are constrained by small scales, limited effect categories, and a lack of public accessibility.

To bridge these gaps, we first formulate a rigorous definition of **Physically Realistic Object Removal**: *If an object ceases to exist, all its directly transient physical effects must also vanish, while the authentic background content behind transparent objects must be faithfully preserved.* These effects include not only shadows and reflections but also caustics, emitted light, and volumetric participating media. Guided by this definition, we propose **PhysRemover**, a unified diffusion-based object removal framework. We first identify conflicts, including the dilemma between eliminating side effects and preserving the background outside the mask for physically realistic removal, as well as the struggle to erase the masked area while retaining background within the mask for transparent objects. To address these challenges, we design a novel architecture that incorporates both the masked image and the unmasked image as a contrastive pair. It preserves the background semantics in an in-context manner, enabling physical side-effect elimination and transparent object removal for the first time.

To address the lack of publicly available datasets and benchmarks, we introduce **PhyTrace**, a high-quality and large-scale dataset, followed by the establishment of two benchmarks and a novel metric **PhysRM** tailored for the proposed task. PhyTrace features all kinds of foregrounds and backgrounds with diverse physical effects, and comprises 10,000 training samples. All samples are provided at 1 Megapixel resolution, and we will make this dataset publicly available to facilitate community research. In summary, our contributions are as follows:

- **Task Definition:** We analyze the limitations of existing object removal methods in handling physical side effects, and propose this new task of physically realistic object removal.
- **PhysRemover:** A unified framework that introduces a novel *In-Context Contrastive Guidance* mechanism and a *Learnable Removal Trigger*. To the best of our knowledge, it is the first model to achieve both physical side-effect elimination and transparent object removal.
- **PhyTrace:** A large-scale, high-quality dataset featuring diverse physical effects to advance research in physically consistent editing. The dataset comprises 10,000 high-resolution training samples, accompanied by two benchmarks specifically tailored for the proposed task.
- **PhyRM:** A novel evaluation metric that comprehensively assesses the removal of objects and their associated physical side effects, demonstrating a high correlation with human perception.

2 Related Work

Image Inpainting. Image inpainting, a fundamental task in computer vision, aims to reconstruct missing regions within a given mask [37, 44, 49, 12]. Early approaches predominantly relied on GANs [9, 14] which struggle to generate diverse and high-fidelity textures in complex scenes [17, 34]. Recent breakthroughs in diffusion models [32, 15] have revolutionized generative editing. Notably, state-of-the-art frameworks like SDXL-Inpainting [30] and FLUX.1-Fill [22] utilize powerful generative priors to deliver superior photorealism. However, these approaches operate under a "hole-filling" assumption. Although visual coherence is maintained within the mask, the complex semantic interaction between the target object and the global scene is frequently neglected. Consequently, these models often fail to identify and remove physically associated side effects (*e.g.*, shadows, reflections), leading to inconsistent results where the object is gone, but its physical footprint remains.

General Image Editing. Beyond specific inpainting, general-purpose image editing models driven by text instructions have gained significant traction recently [13, 1, 35, 46]. Methods like Emu Edit [35] and MGIE [7] utilize Multimodal Large Language Models (MLLMs) to interpret complex user instructions for diverse editing tasks. Similarly, instruction-based approaches like MagicQuill [27] and DreamOmni3 [40] introduce auxiliary inputs or unified frameworks to enhance controllability and user interaction. While these methods offer impressive versatility, they often lack the precision required for physically realistic object removal, especially when targeting a specific instance among multiple similar objects. General editing models tend to struggle with the trade-off between instruction adherence and background preservation, leading to issues such as over-editing (*e.g.*, removing all similar objects) or under-editing (*e.g.*, copy the input image).

Object Removal. Object removal can be viewed as a specialized sub-task of inpainting that demands a higher level of physical consistency. Early approaches adapted inpainting models tailored for removing content within the masked region. However, these methods struggle to address the physical side effects of the target object that extend beyond the mask [51, 36, 19, 23, 20, 4, 26]. Recent works [39, 38, 48, 45, 50] have attempted to address object removal with its side effects. For instance, ObjectDrop [39] explored constructing real-world counterfactual image pairs. Following ObjectDrop, LayerDecomp [42] and OmniPaint [45] adopt costly camera-capture data. However, their broader adoption is hindered by the limited scale of the collected data and the lack of public availability. Alternatively, GeoRemover [50] constructs a two-stage pipeline, but it struggles with partial removal tasks and incurs substantial computational overhead. To mitigate this, OmniEraser [38] attempts to construct paired frames by extracting samples from fixed-viewpoint videos. More recently, ObjectClear [48] seeks to scale up data by combining real images with synthetic samples. Nevertheless, despite these efforts, existing methods remain constrained by limiting "side effects" to merely shadows and reflections, while failing to effectively achieve physically realistic side-effect elimination and transparent object removal. In contrast, the proposed PhysRemover achieves both holistic side-effect elimination and transparent object removal within a unified framework.

3 Methods

3.1 Preliminary

We employ Conditional Flow Matching (CFM) [25] to learn a time-dependent velocity field v_θ that transports a source distribution p_0 to the data distribution p_1. Specifically, the network is trained to regress the target velocity defining the probability path. Following recent advancements [5], we adopt a linear interpolation path between $\mathbf{z}_0 \sim p_0$ and $\mathbf{z}_1 \sim p_1$:

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1, \quad t \in [0,1], \tag{1}$$

where $\mathbf{z}_t$ represents the noisy latent at timestep t. The ground-truth velocity field $u_t(\mathbf{z}_t|\mathbf{z}_1, \mathbf{z}_0)$ for this optimal transport path is simply the displacement vector $\mathbf{z}_1 - \mathbf{z}_0$. Consequently, the training objective is to minimize the expected mean squared error:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left[\left\|v_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{z}_0)\right\|^2\right], \tag{2}$$

where $\mathbf{c}$ denotes the condition. To parameterize the velocity field v_θ, we leverage a Diffusion Transformer (DiT) architecture [5], which treats multiple modalities as a unified sequence of tokens.

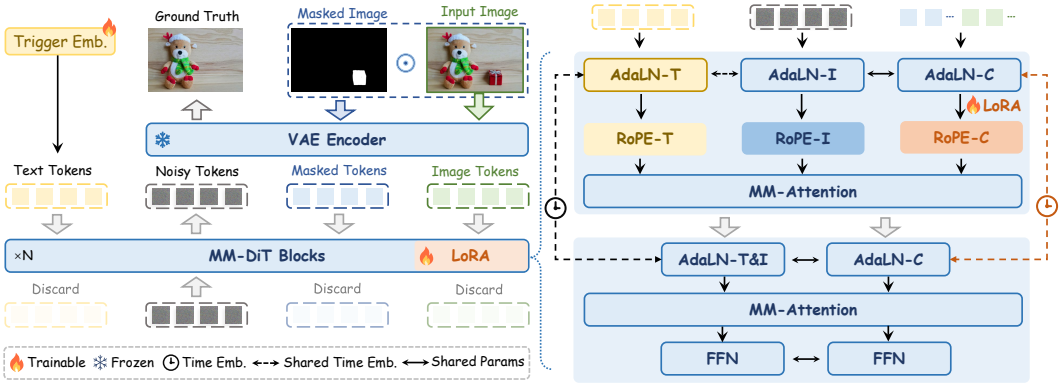

Figure 2: **The overview of PhysRemover pipeline.** *Left:* The proposed architecture and the Learnable Removal Trigger. *Right:* Details of the proposed In-Context Contrastive Guidance.

3.2 Task Definition

We formulate **Physically Realistic Object Removal** not merely as the inverse process of physical scene composition. Given an observed image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ and a binary mask $\mathbf{m}$ indicating the spatial footprint of the target object $\mathcal{O}$, we model the observed scene as a superposition of the underlying clean background $\mathbf{x}_{\text{bg}}$ and the physical presence of the object. The forward image formation process can be generalized via a physical composition function $\mathcal{C}$:

$$\mathbf{x} = \mathcal{C}(\mathbf{x}_{\text{bg}}, \mathbf{x}_{\text{obj}}, \mathcal{E}_{\text{phy}}), \tag{3}$$

where $\mathbf{x}_{\text{obj}}$ denotes the intrinsic visual appearance of $\mathcal{O}$ strictly bounded by the mask $\mathbf{m}$, and $\mathcal{E}_{\text{phy}}$ represents the unbounded physical interactions induced by $\mathcal{O}$ (*e.g.*, cast shadows, reflections, and refractive distortions from transparent materials). Crucially, while conventional removal tasks implicitly assume the image is perfectly pristine outside the mask, the presence of $\mathcal{E}_{\text{phy}}$ explicitly models how the physical influence of an object extends globally across the entire image domain.

Counterfactual Objective. Consequently, our primary goal is to estimate the *counterfactual state* of the scene, which involves reconstructing how the environment would appear had the object $\mathcal{O}$ never been placed there. We seek a mapping function $\mathcal{F}_\theta$ to synthesize the clean background:

$$\hat{\mathbf{x}}_{\text{bg}} = \mathcal{F}_\theta(\mathbf{x}, \mathbf{m}), \tag{4}$$

where the optimal estimation $\hat{\mathbf{x}}_{\text{bg}}$ must satisfy: (1) **Holistic Decoupling**: It must eliminate not only the object semantics within $\mathbf{m}$ but also the distributed physical footprints $\mathcal{E}_{\text{phy}}$ across the entire domain. (2) **Contextual Fidelity**: It must faithfully reconstruct the occluded or optically distorted background details, preserving the global illumination coherence and structural integrity.

3.3 Model Design

For physically realistic object removal, existing approaches adhere to a "mask-and-inpaint" paradigm. This paradigm discards all visual content within the masked region, preventing the model from perceiving the physical attributes (*e.g.*, transparency, specularity) of the target object. In contrast, we propose a novel conditioning paradigm, named *In-Context Contrastive Guidance (ICCG)*, which retains both the masked and unmasked images as in-context references, employing spatially-aligned positional embedding and counterfactual timestep scheduling. Furthermore, to better activate the model's priors for the object removal task, we introduce a *Learnable Removal Trigger (LRT)*. Deviating from the base DiT architecture that conditions solely on textual prompts, we first extend the framework to a multi-modal input stream, as shown in Fig. 2. Crucially, we encode the visual conditions using the VAE [21] from the model, eliminating the need for external encoders.

In-Context Contrastive Guidance. To enable physically realistic removal, the model requires simultaneous knowledge of surrounding context of the target object and its precise localization. ICCG addresses this challenge by formulating the unmasked and masked images as contrastive guidance pairs via in-context conditioning. This mechanism incorporates three key components: *(1) in-context contrastive pairs, (2) spatially-aligned RoPE,* and *(3) counterfactual timestep scheduling*.

(1) In-context contrastive pairs. Fundamentally, physically realistic object removal presents an inherent paradox: it requires eliminating the object semantics within the mask $\mathbf{m}$ and the physical side effects $\mathcal{E}_{\text{phy}}$ in its complement region $\overline{\mathbf{m}}$, while preserving the surrounding background context. What makes transparent object removal particularly challenging is the strong coupling across both spatial domains: the entanglement between the exterior background $\mathbf{x}_{\text{bg}}$ and $\mathcal{E}_{\text{phy}}$ within $\overline{\mathbf{m}}$, as well as the visual blending between the interior background $\mathbf{x}_{\text{bg}}$ and the object $\mathbf{x}_{\text{obj}}$ within $\mathbf{m}$. To address this, we explore a new strategy to decouple $\mathbf{x}_{\text{bg}}$ from $\mathbf{x}_{\text{obj}}$ within the model. Specifically, we reconstruct the input sequence by retaining *both* the masked image and the unmasked image as distinct conditions: $[\mathbf{c}_{\text{txt}}, \mathbf{z}_t, \mathbf{c}_{\text{cond}}]$, where $\mathbf{c}_{\text{txt}}$, $\mathbf{z}_t$, and $\mathbf{c}_{\text{cond}} = [\mathcal{E}(\mathbf{x}_{\text{masked}}), \mathcal{E}(\mathbf{x})]$ denote text tokens, noisy latent tokens, and condition tokens, respectively, and $\mathcal{E}(\cdot)$ is the visual encoder.

(2) Spatially-aligned RoPE. To preserve background details, the noisy latent tokens and condition tokens must maintain structural consistency. However, since the processing of these tokens shares network parameters, employing identical positional encodings would inadvertently cause the target object to be retained in the final output. To address this, we propose SA-RoPE, which shares coordinate values between the condition positional encoding and the noisy latent token positional encoding but computes them separately. Specifically, we assign identical 2D spatial coordinates to the noisy latent and both conditions: $p_i^{\text{img}} = p_i^{\text{masked}} = p_i^{\text{orig}} = (r_i, c_i)$. Crucially, RoPE is applied separately to each stream before concatenation into the joint attention sequence:

$$\begin{cases} \hat{\mathbf{Q}} = \left[\ \mathrm{RoPE}(\mathbf{Q}^{\text{txt,img}},\, \mathbf{P}^{\text{txt,img}})\ ;\ \mathrm{RoPE}(\mathbf{Q}^{\text{cond}},\, \mathbf{P}^{\text{cond}})\ \right], \\ \hat{\mathbf{K}} = \left[\ \mathrm{RoPE}(\mathbf{K}^{\text{txt,img}},\, \mathbf{P}^{\text{txt,img}})\ ;\ \mathrm{RoPE}(\mathbf{K}^{\text{cond}},\, \mathbf{P}^{\text{cond}})\ \right], \end{cases} \tag{5}$$

where $\mathbf{P}^{\text{txt,img}}, \mathbf{P}^{\text{cond}} \in \mathbb{R}^{N\times 2}$ are the 2D coordinate matrices, and $[\,\cdot\ ;\ \cdot\,]$ denotes sequence-wise concatenation. SA-RoPE ensures that cross-stream attention is governed solely by the absolute spatial coordinates, encouraging pixel-wise comparison between $\mathcal{E}(\mathbf{x}_{\text{masked}})$ and $\mathcal{E}(\mathbf{x})$.

(3) Counterfactual timestep scheduling. In DiT, time embeddings inform the network of the current noise level and modulate feature normalization via Adaptive Layer Norm (AdaLN). Naively assigning the same timestep t to all streams implies that the conditions carry the same noise level as the generation target $\mathbf{z}_t$. This assumption, however, is factually incorrect, as the conditions are inherently clean signals. Consequently, while this approach can achieve the removal of general objects, more challenging transparent objects often cannot be effectively eliminated due to the necessity of retaining the majority of semantics within the masked region. To address this issue, we instead adopt a counterfactual scheduling strategy that explicitly separates the temporal identity of each stream:

$$\mathbf{e}^{\text{img}} = \mathrm{AdaLN}\big(\mathbf{z}_t,\ \tau(t)\big), \quad \mathbf{e}^{\text{cond}} = \mathrm{AdaLN}\big(\mathbf{c}_{\text{cond}},\ \tau(0)\big), \tag{6}$$

where $\tau(\cdot)$ denotes the timestep embedding network that maps a scalar timestep to a modulation vector. By fixing the condition timestep at $t{=}0$, we convey a counterfactual semantic that the conditions represent a fully denoised state. This allows the model to easily distinguish the noisy target to be generated at $\tau(t)$ from the clean reference at $\tau(0)$, requiring no additional parameters.

Learnable Removal Trigger. Relying solely on text prompts (*e.g.*, "remove the object") can be ambiguous and inconsistent. To address this, we introduce a learnable removal trigger that replaces the text embedding with a set of optimizable tokens $\mathbf{c}_{\text{task}} \in \mathbb{R}^{K\times d}$, trained via a *pivot-based optimization* strategy. To accommodate the dual-encoder architecture of our base model, we define separate learnable tokens for the T5 and CLIP branches. In the first training stage, only the T5 tokens are optimized, allowing the model to learn a general inpainting instruction. In the second stage, we initialize the CLIP tokens via a pivot derived from the target objects in the training set by:

$$\mathbf{c}_{\text{task}}^{\text{CLIP}} \leftarrow \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},\mathbf{m})\in\mathcal{D}} \mathrm{CLIP}_{\text{vis}}(\mathbf{x} \odot \mathbf{m}), \tag{7}$$

where $\mathrm{CLIP}_{\text{vis}}(\cdot)$ denotes the CLIP visual encoder and $\mathbf{x} \odot \mathbf{m}$ isolates the target object region. Starting from this pivot, the CLIP tokens are then optimized alongside the model weights, guiding the attention mechanism to focus on the target object and its associated physical effects. This pivot-based initialization provides a strong inductive bias toward object-centric semantics, improving convergence and removal precision compared to random or text-derived initialization.

Training Pipeline. Our training pipeline is structured into two distinct stages to ensure robust convergence and generalization. Inspired by OmniPaint [45], we first adapt the FLUX.1-dev model

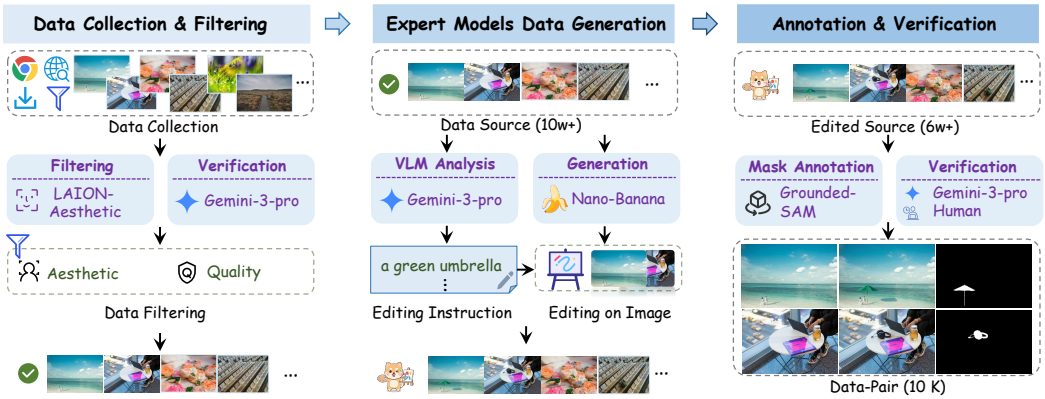

Figure 3: **PhyTrace dataset construction pipeline.** The triplet data is constructed through three processes, yielding accurately annotated pairs with plausible physical side effects.

to the general inpainting domain using the LAION dataset [33]. At this stage, we synthesize training pairs by applying random masks to the images. To ensure computational efficiency, we employ LoRA [16] for lightweight fine-tuning. The model is optimized using a standard self-supervised CFM loss, where the goal is to reconstruct the original image $\mathbf{z}_1$ within the masked region:

$$\mathcal{L}_{\text{S1}} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left[\left\|v_\theta(\mathbf{z}_t, t, \mathbf{c}_1) - (\mathbf{z}_1 - \mathbf{z}_0)\right\|^2\right], \tag{8}$$

where $\mathbf{c}_1 = [\mathbf{c}_{\text{task}}, \mathcal{E}(\mathbf{x}_{\text{masked}})]$. This stage equips the model with structure-completion capabilities.

In the second stage, we build upon the initialized LoRA weights and transition to our PhyTrace dataset. Here, we introduce two critical modifications to the architecture and objective. First, we replace the masked condition $\mathcal{E}(\mathbf{x}_{\text{masked}})$ with our in-context contrastive pairs $\mathbf{c}_{\text{cond}} = [\mathcal{E}(\mathbf{x}_{\text{masked}}), \mathcal{E}(\mathbf{x})]$ and append the learnable task tokens $\mathbf{c}_{\text{task}}$. Second, the training paradigm shifts from self-supervised training to **counterfactual background estimation**. The ground-truth target $\mathbf{z}_1$ is no longer the original image, but corresponds to the latent representation of the clean background $\mathbf{z}_{\text{bg}} = \mathcal{E}(\mathbf{x}_{\text{bg}})$:

$$\mathcal{L}_{\text{S2}} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_{\text{bg}}}\left[\left\|v_\theta(\mathbf{z}_t, t, \mathbf{c}_2) - (\mathbf{z}_{\text{bg}} - \mathbf{z}_0)\right\|^2\right], \tag{9}$$

where $\mathbf{c}_2 = [\mathbf{c}_{\text{task}}, \mathbf{c}_{\text{cond}}]$. In this stage, we employ various mask augmentation strategies to ensure robustness against potentially inaccurate masks during inference (details in Appendix A.2).

3.4 PhyTrace Dataset

As discussed in Sec. 1, existing datasets fail to provide counterfactual data pairs featuring physical side effects and transparent object removal, thereby restricting physically realistic object removal. To address this limitation, we introduce PhyTrace, a high-quality and large-scale dataset comprising 10,000 accurately annotated and rigorously verified data pairs. Figure 3 illustrates the construction pipeline of the PhyTrace dataset, which consists of three main stages. First, we collect a massive number of raw images and perform quality filtering via an aesthetic classifier and a Vision-Language Model (VLM), retaining only high-quality images with a 1-megapixel (1M) resolution or higher. Second, we employ expert models to analyze the images and determine the specific objects and locations for editing based on the unique characteristics of each image, subsequently generating the edited image pairs. Finally, mask annotations are generated for the edited pairs using the editing instructions from the previous stage. Furthermore, to ensure data quality, we introduce a dual-verification mechanism, which combines VLM assessment with human review, to guarantee that all constructed data pairs adhere to physical laws and are free from over-editing. **Detailed construction pipeline and implementation details are provided in Appendix A.3.**

4 Experiments

4.1 Experimental Setup

Evaluation Benchmarks. Given the absence of publicly available datasets and benchmarks, we propose two new benchmarks to evaluate the performance of the methods in physically realistic object

Table 1: **Quantitative comparison with baselines.** "Side Effect" indicates whether the method considers physical side effects, and "Transparent Object" denotes the capability of handling transparency.

Method	Side Effect	Transparent Object	PhyTrace-Test						PhyTrace-TP		
			FID ↓	CMMD ↓	LPIPS ↓	PhysRM-PER ↑	PhysRM-CB ↑	PhysRM-ORS ↑	$PSNR_{tp}$ ↑	$LPIPS_{tp}$ ↓	$DINO_{tp}$ ↓
MAT	✘	✘	42.485	0.242	0.155	1.564	1.216	1.370	10.872	0.029	0.644
SD-Inpainting	✘	✘	43.817	0.426	0.197	2.102	1.788	2.530	14.430	0.023	0.449
FLUX-Fill	✘	✘	40.446	0.178	0.110	1.952	1.664	1.774	14.263	0.025	0.454
CLIPAway	✘	✘	43.290	0.435	0.200	2.164	1.620	2.430	14.130	0.025	0.432
PowerPaint	✘	✘	55.150	0.350	0.120	1.344	1.038	1.072	8.533	0.038	0.715
FreeCompose	✘	✘	35.360	0.456	0.196	2.660	2.124	3.642	16.455	0.018	0.386
Attentive Eraser	✘	✘	25.412	0.203	0.109	2.654	2.592	4.294	17.091	0.016	0.305
RORem	✘	✘	33.135	0.277	0.199	2.648	2.626	4.164	18.280	0.017	0.278
SmartEraser	✘	✘	34.168	0.245	0.164	2.476	2.434	3.368	16.272	0.018	0.380
ObjectClear	✔	✘	20.317	0.200	0.118	3.360	3.664	4.696	19.551	0.013	0.194
OmniPaint	✔	✘	22.001	0.149	0.107	3.334	3.936	4.452	19.331	0.012	0.184
GeoRemover	✔	✘	23.620	0.180	0.120	3.492	3.740	4.488	19.531	0.011	0.231
OmniEraser	✔	✘	51.893	0.382	0.227	3.142	2.532	4.386	16.306	0.017	0.379
PhysRemover	✔	✔	**17.402**	**0.106**	**0.105**	**4.528**	**4.470**	**4.808**	**20.882**	**0.009**	**0.134**

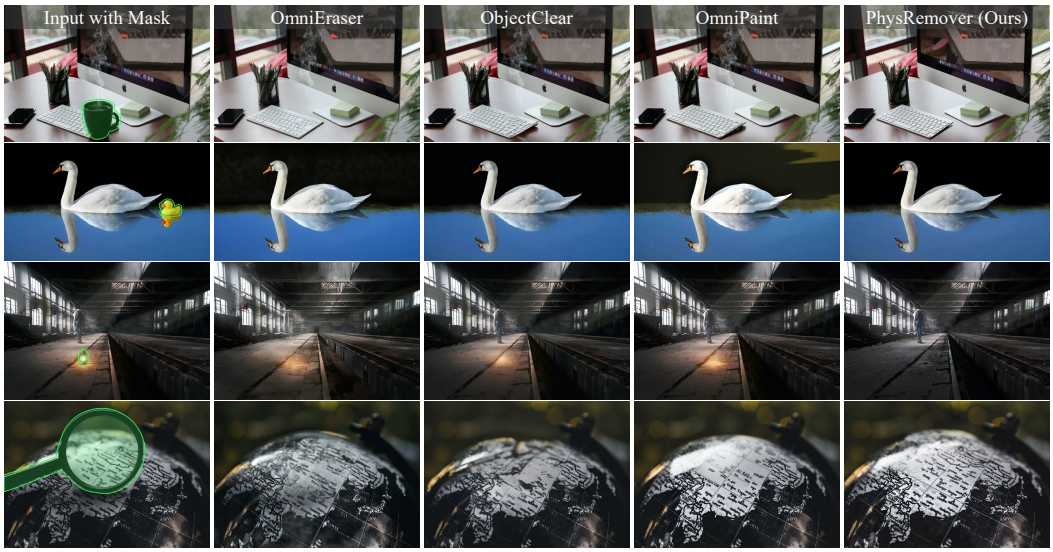

Figure 4: **Qualitative comparison of physically realistic object removal.** Comparison results of our PhysRemover and state-of-the-art methods across challenging scenarios: removing objects with physical side effects, and erasing transparent objects while preserving backgrounds

removal from two perspectives: (1) ***PhyTrace-Test***: We construct the *PhyTrace-Test* benchmark to evaluate the physically realistic object removal performance. We split our collected PhyTrace dataset into train and test sets. For the test set, we select 500 high-quality image pairs that feature diverse physical interactions. (**2**) ***PhyTrace-TP***: To further investigate the performance in transparent object removal, we introduce *PhyTrace-TP*, a specialized subset consisting of 150 images from the test set that exclusively feature transparent or semi-transparent objects.

Evaluation Metrics. To comprehensively evaluate the performance across the two proposed benchmarks, we employ tailored sets of metrics for each. For the **PhyTrace-Test** benchmark, which evaluates general object and physical side effect removal, we report FID [14] and CMMD [18] to assess the global fidelity and perceptual quality of the generated images, along with LPIPS [47] for perceptual similarity. Furthermore, to accurately evaluate the effectiveness of physically realistic removal, we introduce **PhysRM**, a novel VLM-based evaluation metric powered by Gemini-3-Flash. It assesses the removal quality on a 1-5 scale (5 being the best) across three dimensions: (1) *Physical Effects Removal (PER)* assesses whether all transient physical side effects associated with the target are effectively eliminated; (2) *Completion & Blending (CB)* evaluates the naturalness and visual coherence of the infilled region; and (3) *Object Removal Success (ORS)* ensures the complete erasure of the target and penalizes object hallucination. For the **PhyTrace-TP** benchmark, which focuses on transparent objects, we aim to assess the restoration of the background content behind the objects. To

Table 2: **Ablation studies.** We incrementally ablate each proposed component. Each row removes one design choice from the full model while keeping all others intact.

Setting	Components				PhyTrace-Test			PhyTrace-TP		
	Contrastive Pairs	SA-RoPE	Counter. Timestep	Learnable Trigger	FID ↓	LPIPS ↓	PhysRM ↑	PSNR_{tp} ↑	LPIPS_{tp} ↓	DINO_{tp} ↓
$Cond_{\text{fill}}$	✘	✔	✔	✔	50.150	0.155	2.784	18.227	0.010	0.482
$Cond_{\text{replace}}$	✘	✔	✔	✔	87.513	0.192	1.898	9.536	0.021	0.730
w/o SA-RoPE	✔	✘	✔	✔	21.835	0.112	4.223	19.104	0.012	0.198
w/o Counter. Timestep	✔	✔	✘	✔	23.271	0.116	4.068	18.563	0.013	0.221
w/o Learnable Trigger	✔	✔	✔	✘	29.340	0.120	3.831	19.845	0.011	0.213
PhysRemover (Full)	✔	✔	✔	✔	**17.402**	**0.105**	**4.602**	**20.882**	**0.009**	**0.134**

this end, we employ PSNR_{tp}, LPIPS_{tp}, and DINO_{tp} [29], which are computed within the masked regions to measure the consistency and accuracy of the recovered background. **For reproducibility, the implementation details and justifications for PhysRM are provided in Appendix A.4.**

Baselines. We benchmark PhysRemover against a wide range of state-of-the-art approaches, which can be categorized into: (1) General inpainting models, including MAT [24], SD-Inpainting [32], and FLUX-Inpainting [22]; and (2) Specialized object removal models, covering CLIPAway [4], PowerPaint [51], FreeCompose [3], Attentive Eraser [36], RORem [23], ObjectClear [48], OmniPaint [45], GeoRemover [50], SmartEraser [20], and OmniEraser [38].

4.2 Qualitative Comparisons

We present qualitative comparison results in Fig. 1 and Fig. 4. To demonstrate the effectiveness of our proposed method, we select baselines that explicitly account for the removal of objects and their associated side effects. Our evaluation covers three distinct categories of physical interactions: Radiometric Surface Interactions, Specular & Refractive Light Transport, and Volumetric Participating Media. As illustrated in the second row of Fig. 1 and the third row of Fig. 4, when removing a light-emitting object, all baselines fail to eliminate the diffuse reflections on the ground. While OmniEraser attempts to address this, it tends to over-erase illumination and fails to correctly model the intrinsic correlation between the target object and the radiometric surface. For Specular & Refractive Light Transport, we evaluate mirror reflections (Row 2, Fig. 4). OmniEraser and ObjectClear successfully remove both the object and its reflection. However, OmniPaint induces noticeable unwanted artifacts in the background. Regarding Volumetric Participating Media (Row 1, Fig. 4), the removal of a cup containing hot water dictates that the associated steam diffusing into the atmosphere must also vanish. Existing baselines ignore this physical association between the target and the environment. In contrast, PhysRemover achieves the best performance across all categories.

Beyond side effects, eliminating transparent objects while preserving the physical reality behind them is a critical aspect of this task. As shown in Fig. 1 and Fig. 4, baselines adhere to the traditional "mask-and-inpaint" paradigm: they erase the target object and hallucinate the missing content based on the surrounding context. This leads to the loss of details visible through the transparent surface. In contrast, PhysRemover successfully removes transparent objects while precisely restoring the details of the underlying background, even in scenarios with highly complex and challenging contexts. For instance, as shown in Row 4 of Fig. 4, PhysRemover effectively eliminates the magnifying glass and its shadow while perfectly preserving the intricate text and lines in the background. **Additional comparisons are detailed in Appendix A.5, featuring comparisons against existing baselines in Figs. 7, 8, and 9, alongside *in-the-wild results in* Fig. 11**.

4.3 Quantitative Comparisons

We compare the proposed PhysRemover against state-of-the-art baselines on the two introduced benchmarks: *PhyTrace-Test* and *PhyTrace-TP*, as summarized in Tab. 1.

General Object and Side Effect Removal. Evaluated on the *PhyTrace-Test* benchmark, PhysRemover achieves state-of-the-art performance across all evaluation metrics. In terms of global generation quality, our method secures the best scores on FID and CMMD, indicating that the infilled regions are seamlessly integrated with the surroundings, closely mimicking natural image statistics. More importantly, regarding removal effectiveness, PhysRemover significantly outperforms all baselines on LPIPS and the comprehensive VLM-based PhysRM metrics. The superior scores in



Figure 5: Qualitative ablation on different conditioning strategies, where Condition Type 1 denotes the $Cond_{\text{fill}}$ setting and Condition Type 2 denotes the $Cond_{\text{replace}}$ setting.

PhysRM-PER, PhysRM-CB, and PhysRM-ORS demonstrate exceptional capability to completely erase target objects and their transient physical side effects of our model, ensuring natural blending without introducing unwanted object hallucinations.

Transparent Object Removal. Furthermore, we evaluate the capability of transparent object removal on the *PhyTrace-TP* benchmark. As shown in Tab. 1, the proposed PhysRemover achieves the highest PSNR_{tp} alongside the lowest LPIPS_{tp} and DINO_{tp} scores. Since these metrics are specifically computed within the masked regions, this quantitative superiority confirms the effectiveness of PhysRemover in not only eradicating transparent occlusions but also precisely recovering the structural consistency and fine-grained details of the blocked background.

4.4 Ablation Study

We ablate each component to validate its individual contribution, as summarized in Tab. 2.

Effectiveness of In-Context Contrastive Guidance. We first evaluate the conditioning strategy by comparing against two baselines. $Cond_{\text{fill}}$ follows the standard inpainting paradigm, conditioning solely on the masked image. Without access to the original scene content, the model is forced to hallucinate the background behind transparent objects, resulting in significant semantic degradation. $Cond_{\text{replace}}$ attempts to provide richer context by replacing the masked image with the unmasked original. However, as illustrated in Fig. 5, this induces structural leakage. We further validate the two mechanisms within ICCG. Removing SA-RoPE and reverting to a naïve shared positional encoding weakens the pixel-wise spatial correspondence between streams, degrading both generation fidelity and transparent object restoration. Similarly, disabling the counterfactual timestep scheduling impairs the model's capacity to distinguish clean references from noisy latents, leading to the most pronounced decline on transparent object metrics. Notably, in challenging scenarios such as removing transparent objects while preserving the extensive and complex underlying background, SA-RoPE tends to retain the entire structure. Conversely, the introduction of the counterfactual timestep enables precise object elimination alongside faithful background preservation, as shown in Fig. 5.

Effectiveness of Learnable Removal Trigger. We evaluate the necessity of the learnable removal trigger by replacing the optimized tokens $\mathbf{c}_{\text{task}}$ with the fixed natural language prompt "remove the object". As evidenced in Tab. 2, the fixed prompt yields notably inferior performance across all metrics. We attribute this to its inability to encode the nuanced semantics of physical eradication, which encompasses both side-effect elimination and transparency-aware background preservation and is acquired by the learnable tokens through end-to-end optimization alongside the model weights.

5 Conclusion

In this paper, we introduce a new task: **Physically Realistic Object Removal**. This task posits that if an object ceases to exist, all its associated transient physical effects must also vanish. To address this, we propose **PhysRemover**, the first unified framework to achieve both physically realistic side-effect elimination and transparent object removal enabled by the proposed *In-Context Contrastive Guidance* and *Learnable Removal Trigger*. Furthermore, we construct **PhyTrace**, a large-scale, high-quality dataset capturing diverse physical effects, along with two benchmarks for the task: *PhyTrace-Test* and *PhyTrace-TP*. Additionally, we develop a novel metric, **PhysRM**, for evaluating object removal quality, establishing a new standard for physically realistic image editing. Extensive experiments demonstrate that PhysRemover achieves state-of-the-art performance in physically realistic object removal. Finally, PhysRemover is naturally extensible to removal in artistic styles or unconstrained in-the-wild scenarios, offering practical value to the broader community.

Acknowledgments and Disclosure of Funding

This work was supported by the SSTIC Grants under Nos. KJZD20230923115106012, KJZD20230923114916032, and GJHZ20240218113604008, and by the National Natural Science Foundation of China under Grant No. 62331014. The authors declare no competing interests.

References

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

[2] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. TOM-Net: Learning transparent object matting from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Zhekai Chen, Wen Wang, Zhen Yang, Zeqing Yuan, Hao Chen, and Chunhua Shen. Freecompose: Generic zero-shot image composition with diffusion prior. In *European Conference on Computer Vision*, pages 70–87. Springer, 2024.

[4] Yiğit Ekin, Ahmet Burak Yildirim, Erdem Eren Çağlar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models. *Advances in Neural Information Processing Systems*, 37:17572–17601, 2024.

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[6] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10571–10580, 2021.

[7] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

[8] Yang Fu, Yike Zheng, Ziyun Dai, and Henghui Ding. EffectErase: Joint video object removal and insertion for high-quality effect erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[9] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[10] Google. A new era of intelligence with gemini 3. `https://blog.google/products-and-platforms/products/gemini/gemini-3/`, 2025.

[11] Google DeepMind. Nano banana. `https://deepmind.google/models/gemini-image/`, 2025.

[12] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14134–14143, 2021.

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

[14] 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. *Advances in neural information processing systems*, 30, 2017.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[17] Jitesh Jain, Yuqian Zhou, Ning Yu, and Humphrey Shi. Keys to better image inpainting: Structure and texture go hand in hand. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 208–217, 2023.

[18] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.

[19] Yueru Jia, Yuhui Yuan, Aosong Cheng, Chuke Wang, Ji Li, Huizhu Jia, and Shanghang Zhang. Designedit: Multi-layered latent decomposition and fusion for unified & accurate image editing. *arXiv preprint arXiv:2403.14487*, 2024.

[20] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. *arXiv preprint arXiv:2501.08279*, 2025.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[23] Ruibin Li, Tao Yang, Song Guo, and Lei Zhang. Rorem: Training a robust object remover with human-in-the-loop. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14024–14035, 2025.

[24] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[25] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[26] Yi Liu, Hao Zhou, Benlei Cui, Wenxiang Shang, and Ran Lin. Erase diffusion: Empowering object removal through calibrating diffusion pathways. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2418–2427, 2025.

[27] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13072–13082, 2025.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[31] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[34] Pourya Shamsolmoali, Masoumeh Zareapoor, and Eric Granger. Transinpaint: Transformer-based image inpainting with context adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 849–858, 2023.

[35] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.

[36] Wenhao Sun, Xue-Mei Dong, Benlei Cui, and Jingqun Tang. Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20734–20742, 2025.

[37] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

[38] Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, and Zhanyu Ma. Omnieraser: Remove objects and their effects in images with paired video-frame data. *arXiv preprint arXiv:2501.07397*, 2025. URL `https://arxiv.org/abs/2501.07397`.

[39] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*, pages 112–129. Springer, 2024.

[40] Bin Xia, Bohao Peng, Jiyang Liu, Sitong Wu, Jingyao Li, Junjia Huang, Xu Zhao, Yitong Wang, Ruihang Chu, Bei Yu, et al. Dreamomni3: Scribble-based editing and generation. *arXiv preprint arXiv:2512.22525*, 2025.

[41] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023.

[42] Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7643–7653, 2025.

[43] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023.

[44] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019.

[45] Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. *arXiv preprint arXiv:2503.08677*, 2025.

[46] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[48] Jixin Zhao, Shangchen Zhou, Zhouxia Wang, Peiqing Yang, and Chen Change Loy. Objectclear: Complete object removal via object-effect attention. *arXiv preprint arXiv:2505.22636*, 2025.

[49] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion. *arXiv preprint arXiv:2306.04632*, 2023.

[50] Zixin Zhu, Haoxiang Li, Xuelu Feng, He Wu, Chunming Qiao, and Junsong Yuan. Georemover: Removing objects and their causal visual artifacts. *arXiv preprint arXiv:2509.18538*, 2025.

[51] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer, 2024.

A Appendix

A.1 Implementation Details.

We implement our framework based on the state-of-the-art text-to-image diffusion model, FLUX.1-dev [22]. To achieve parameter-efficient fine-tuning, we inject LoRA [16] layers into the transformer blocks with a rank of 4. A selective scaling strategy is applied where the LoRA scale is set to 0 for non-condition tokens, ensuring the preservation of the base model's original generative priors [41]. Our experiments are conducted on $2\times$ NVIDIA A100 (80GB) GPUs. We utilize the AdamW optimizer [28] with a learning rate of 1×10^{-4} and a total batch size of 8. The training process is divided into two stages: the first stage is trained for $25,000$ steps, followed by the second stage for $15,000$ steps.

A.2 Mask Augmentation Strategy

In practice, user-provided masks are often imprecise—they may over-extend beyond the object boundary or miss portions of the target. To ensure robustness against such inaccuracies during inference, we apply stochastic mask augmentation during Stage 2 training. Following the probability distribution inspired by OmniPaint [45], we augment the training masks with the following strategies:

- **Original (20%):** The mask is used without modification.
- **Dilation (10%):** A random morphological dilation (kernel size 5–20) is applied, simulating over-masking scenarios where users mark beyond the object boundary.
- **Erosion (20%):** A random morphological erosion (kernel size 5–20) is applied, simulating under-masking where object edges are missed.
- **Shape Addition (10%):** Small geometric shapes (squares or circles, size 4–12 pixels) are drawn in white near the object boundary, simulating accidental brush strokes.
- **Shape Removal (40%):** Small geometric shapes are drawn in black *inside* the object region, creating holes in the mask that simulate incomplete annotation. This is the most frequent augmentation, as incomplete masks are the most common failure mode in practice.

For the shape-based augmentations, we sample positions adaptively: additions are placed along the object's outer boundary band (obtained via dilation minus the original mask), while removals are sampled from the interior foreground pixels. This ensures that the augmented masks remain geometrically plausible rather than randomly distributed.

A.3 Details of PhysTrace Dataset

Image Filtering & VLM Analysis. We start by harvesting raw images from internet sources. To ensure high visual fidelity, we first filter out low-quality samples, retaining only images with resolutions exceeding 1 megapixel and high aesthetic scores assessed by the LAION aesthetic predictor [33]. Subsequently, we employ Gemini-3-Pro [10] to perform fine-grained visual analysis on the filtered images, including: (1) Quality Check: Detecting and discarding images with watermarks, subtitles, or text overlays. (2) Attribute Profiling: Analyzing scene types, artistic styles, and critically, the presence of specific physical properties. Based on these attributes, we prompt Gemini to suggest a context-appropriate object name to be inserted, along with a detailed editing instruction. Crucially, this instruction includes explicit descriptions of physical side effects to guide the subsequent generation.

Generative Insertion. We utilize NanoBanana [11], a state-of-the-art instruction-based image editing model, as our generative engine. Initially, we attempted a direct removal approach, instructing the model to remove a specific target from an image. However, this led to two critical failures: the model frequently altered unrelated background regions, and in multi-object scenes it struggled to identify the correct target. To overcome these limitations, we adopt an inverse generation strategy. We treat the collected real image as the clean background $\mathbf{x}_{\text{bg}}$ and instruct the model to *insert* the specified object with its physical effects into the scene to generate $\mathbf{x}$.

Mask Annotation. To obtain precise object masks $\mathbf{m}$, we implement an automated segmentation pipeline. Leveraging the generated image $\mathbf{x}$ and the corresponding object name, we deploy Grounded SAM [31] to derive the initial object mask. To further resolve ambiguities particularly where SAM

might segment the wrong instance in multi-object scenes, we compute the difference map between the edited image $\mathbf{x}$ and the background $\mathbf{x}_{\text{bg}}$. We then refine the mask to ensure accurate localization.

Post-processing & Verification. A common artifact in instruction-based editing models is the introduction of global pixel shifts, where the generated image $\mathbf{x}$ is slightly spatially misaligned compared to $\mathbf{x}_{\text{bg}}$. Even minor shifts can lead to significant errors in training. We address this by calculating the optimal translation vector that minimizes the pixel error in the unmasked regions and register the images to perfect alignment. Finally, we employ Gemini-3-Pro again as a verifier to assess the consistency between the instruction and the result, filtering out failed edits. This rigorous pipeline yields a final set of 10,000 high-quality, physically consistent training triplets. We provide randomly sampled visualization of the PhyTrace dataset in Fig. 14.

Physical-Effect Categories. Using the construction metadata and target-effect descriptions, we assign each annotated triplet to one mutually exclusive primary physical-effect category. Table 3 summarizes the sample counts and proportions. The categories cover cast shadows, reflections, refraction and transparency, emitted illumination, volumetric media, and other physical interactions.

Table 3: **Distribution of primary physical-effect categories in PhyTrace.** Each triplet is assigned to one primary category. Percentages are rounded to one decimal place.

Primary physical-effect category	Samples	Fraction
Cast shadows	3,062	30.2%
Reflections	1,527	15.0%
Refraction, transparency, and caustics	2,301	22.7%
Emitted illumination	946	9.3%
Volumetric media (smoke, steam, dust, and vapor)	1,627	16.0%
Other physical interactions	689	6.8%
Total	**10,152**	**100.0%**

A.4 PhysRM

Existing evaluation metrics for object removal (*e.g.*, FID, LPIPS) primarily measure global image quality or pixel-level similarity, but fail to capture whether physically associated side effects have been properly eliminated. To address this gap, we introduce **PhysRM**, a VLM-based evaluation framework powered by Gemini-3-Flash [10].

Evaluation Protocol. For each test sample, we construct two composite images: (1) a *full-image composite* providing global context, and (2) a *cropped-region composite* enabling fine-grained local inspection. Each composite is a 2×2 grid containing the original image, the edited result, the ground-truth background, and the binary mask. The VLM evaluates three dimensions on a 1–5 integer scale:

- **Physical Effects Removal (PER):** Whether the physical effects directly caused by the masked object (*e.g.*, reflections, shadows, emitted light, vapor) have been successfully removed.
- **Completion & Blending (CB):** Whether the masked target region is reasonably completed with background content and blends naturally with surrounding areas.
- **Object Removal Success (ORS):** Whether the object inside the mask is fully removed without generating new similar object-like content.

Scoring Criteria. Each dimension is scored independently: 5 = excellent (no obvious problem), 4 = good (slight issue), 3 = moderate, 2 = poor (obvious issue), 1 = very poor (severe problem). The evaluator is instructed to be conservative and assign extreme scores only when evidence is visually clear. The complete evaluation prompt is provided in Tab. 4.

Justification. To validate that PhysRM faithfully reflects human perception, we conduct a human evaluation study. We randomly sample 100 images from the PhyTrace-Test benchmark, covering diverse physical effects and transparent objects. For each image, we collect removal results from 5 representative methods, including PhysRemover and 4 baselines, yielding 500 image-result pairs. We recruit 20 human raters from diverse academic backgrounds. Each rater independently scores all

three dimensions (PER, CB, ORS) on the same 1–5 scale used by PhysRM. To ensure consistency, raters undergo a calibration session with 10 example images before the formal evaluation.

We measure inter-rater reliability using Krippendorff's α, obtaining $\alpha = 0.74$ (PER), $\alpha = 0.71$ (CB), and $\alpha = 0.78$ (ORS), indicating substantial agreement. We then compute the per-image mean human score and correlate it with PhysRM predictions using PLCC (Pearson Linear Correlation Coefficient) and SROCC (Spearman Rank-Order Correlation Coefficient). The MainScore is defined as $\frac{\text{PLCC}+\text{SROCC}}{2}$. As shown in Fig. 6, PhysRM achieves MainScores of 0.865 (PER), 0.799 (CB), and 0.910 (ORS), with 95% confidence intervals of ± 0.032, ± 0.041, and ± 0.025, respectively.

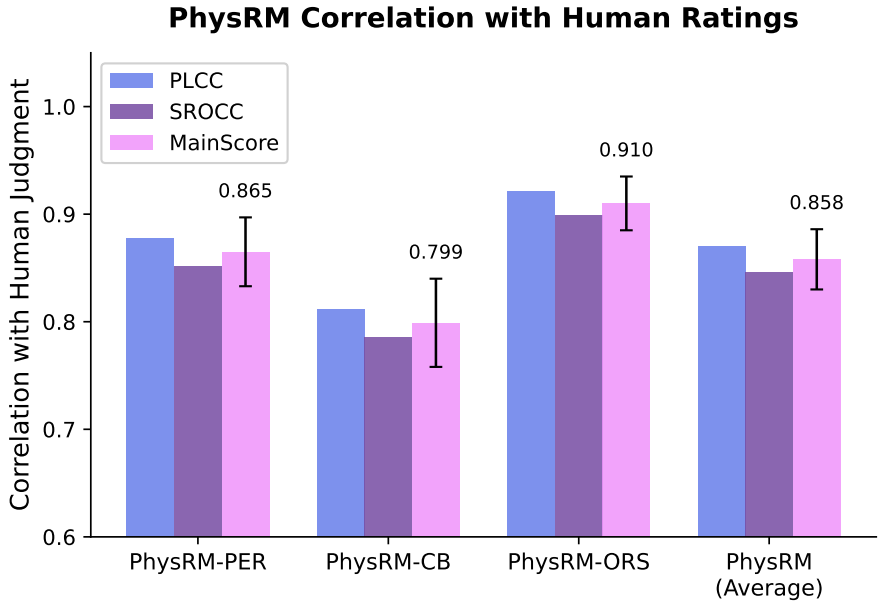

Figure 6: **Correlation between metrics and human ratings.** PhysRM achieves substantially higher PLCC, SROCC, and MainScore compared to existing metrics (ReMOVE, CFD), demonstrating strong alignment with human perception. Error bars indicate 95% confidence intervals.

A.5 Additional Comparisons

Baseline Configurations. Table 5 summarizes the baseline configurations used in our comparisons. All methods receive the input image and an object-only mask. We report the positive text or task input, sampling steps, guidance scale, and inference resolution for each method. For generic text-guided inpainting models, the prompt explicitly requests object removal and an empty scene. N/A denotes a setting that is not applicable. GeoRemover uses two consecutive 28-step stages, resulting in 56 denoising steps in total.

Table 4: **PhysRM evaluation prompt.** The complete prompt used to query Gemini-3-Flash for scoring object removal quality.

You are evaluating object removal quality in an image editing task. I will provide 2 composite images. Both are 2×2 grids formed by stitching four images together. For both, interpret the four quadrants in reading order (left to right, top to bottom): Top-left: BEFORE image; Top-right: EDIT image; Bottom-left: GT image; Bottom-right: MASK image. Image 1 is the full-image composite. Image 2 is the cropped-region composite. The MASK indicates the object that should be removed. Use Image 1 for global context and Image 2 for local inspection.

Evaluate the following 3 items (integer scores 1–5):
1. physical_effects: Evaluate whether the physical effects directly caused by the masked object have been successfully removed in the EDIT image. This includes only clearly visible effects such as: reflection, smoke/vapor/mist, emitted light or direct illumination, and cast shadow caused by the object. Only count effects that are clearly and directly caused by the masked object. 5 = removed almost perfectly; 4 = most removed with slight traces; 3 = partially removed with noticeable traces; 2 = poorly removed with obvious traces; 1 = almost not removed at all.
2. target_completion_and_blending: Evaluate whether the masked target region is reasonably completed with background content and blends naturally and seamlessly with surrounding areas. 5 = completed very naturally and blends seamlessly; 4 = completed well with slight blending issues; 3 = moderately completed with noticeable issues; 2 = poorly completed with obvious problems; 1 = completed very badly.
3. object_removed_no_new_similar_object: Evaluate whether the object inside the mask is removed, and whether no new similar object-like content is generated inside the masked region. 5 = fully removed with no new similar object; 4 = almost fully removed with minor residue; 3 = partially removed or mildly similar content appears; 2 = poorly removed or obvious similar content appears; 1 = object clearly remains or a new similar object is generated.

Important: Judge only the masked object and its directly related area. Do not score overall image beauty. Be conservative and only assign very high or very low scores when the evidence is visually clear. When full-image and cropped-region evidence differ, use Image 2 for local detail and Image 1 for global context. Return JSON only: `{"physical_effects": int, "target_completion_and_blending": int, "object_removed_no_new_similar_object": int}`.

Table 5: **Baseline inference configurations.** Text entries specify the positive prompt or task input. N/A (mask-only) indicates that no runtime text instruction is used.

Method	Positive text/task input	Steps	Guidance	Test resolution
MAT	N/A (mask-only)	N/A	N/A	512×512
SD-Inpainting	"remove the object, an empty scene."	50	7.5	512×512
FLUX-Fill	"remove the object, an empty scene."	28	30.0	1024×1024
CLIPAway	N/A (mask-only)	50	7.5	512×512
PowerPaint	P_{ctxt}; "empty scene blur"	45	10.0	640×640
FreeCompose	"something in some place" $\rightarrow$ "some place"	100	7.5	512×512
Attentive Eraser	N/A (mask-only)	28	9.0	1024×1024
RORem	"4K, high quality, photorealistic, realistic"	50	1.0	512×512
SmartEraser	"Remove the instance of object"	50	1.5	512×512
OmniEraser	"There is nothing here."	28	3.5	1024×1024
ObjectClear	"remove the instance of object"	20	2.5	512×512
OmniPaint	N/A (mask-only)	28	3.5	1024×1024
GeoRemover	"A beautiful scene"	56	30	1024×1024

Evaluation on External Benchmarks. To evaluate generalization beyond PhyTrace, we evaluate the same PhysRemover checkpoint on OmniPaint-Bench and RemovalBench without benchmark-specific fine-tuning. Table 6 reports the quantitative comparisons. PhysRemover achieves the best results on all five metrics on OmniPaint-Bench and the best FID, LPIPS, and PSNR on RemovalBench. These results support its generalization beyond the PhyTrace evaluation setting.

Table 6: **Quantitative comparison on external object-removal benchmarks.** O- and R- denote OmniPaint-Bench and RemovalBench, respectively. The same PhysRemover checkpoint is used without benchmark-specific fine-tuning.

Method	O-FID ↓	O-LPIPS ↓	O-CMMD ↓	O-PSNR ↑	O-DINO ↑	R-FID ↓	R-LPIPS ↓	R-CMMD ↓	R-PSNR ↑	R-DINO ↑
SD-Inpaint	131.265	0.325	0.279	18.703	0.601	164.486	0.396	0.470	20.232	0.616
FLUX-Fill	111.545	0.206	0.311	19.724	0.623	154.596	0.262	0.463	20.337	0.645
CLIPAway	104.557	0.302	0.287	19.494	0.679	140.710	0.399	0.387	20.410	0.685
PowerPaint	104.881	0.263	0.205	19.751	0.672	154.214	0.306	0.424	20.917	0.674
RORem	58.063	0.276	0.139	22.307	0.834	77.067	0.377	0.180	23.280	0.821
FreeCompose	82.152	0.282	0.192	21.120	0.777	116.149	0.380	0.350	22.190	0.773
AttentiveEraser	59.499	0.183	0.097	22.308	0.842	64.022	0.197	0.205	23.726	0.855
SmartEraser	81.709	0.327	0.385	20.713	0.745	91.600	0.391	0.250	22.051	0.770
OmniEraser	72.667	0.257	0.110	21.243	0.796	65.320	0.291	0.369	23.201	0.848
OmniPaint	45.529	0.157	0.046	23.459	0.884	31.657	0.211	0.032	24.676	0.941
ObjectClear	43.789	0.230	0.061	23.519	0.884	28.328	0.287	0.033	25.812	0.938
GeoRemover	42.263	0.164	0.131	23.806	0.881	31.800	0.244	0.186	25.137	0.918
PhysRemover	36.948	0.131	0.032	31.607	0.905	27.115	0.175	0.082	29.621	0.938

Comparison with Traditional Removal Methods. To further demonstrate the superior performance of our proposed **PhysRemover**, we conduct qualitative comparisons with a broad range of traditional removal approaches, including MAT, SD-Inpainting, FLUX-Fill, CLIPAway, PowerPaint, FreeCompose, Attentive Eraser, RORem, and SmartEraser. The comparative results are presented in Fig. 7, 8 and 9.

The results reveal limitations in these traditional methods. They fail to completely eradicate the target object or inadvertently hallucinate new, semantically similar entities in the masked region, let alone address the physical side effects associated with the target. Furthermore, we observe an unusual discrepancy: despite yielding visually inferior removal results (*e.g.*, incomplete removal or artifacts), some baselines still achieve competitive scores on standard evaluation metrics. This phenomenon highlights the insufficiency of current metrics in faithfully evaluating physically realistic object removal. We intend to explore more accurate and effective evaluation protocols for this task in future work.

Extended Comparisons with SOTA Methods. In Fig. 10, we provide extended comparisons between **PhysRemover** and state-of-the-art approaches, including ObjectClear, OmniPaint, GeoRemover, and OmniEraser. While these methods are explicitly designed to account for associated physical side effects, their scope is predominantly confined to simple shadows and reflections. It is worth noting that, with the exception of ObjectClear, all methods operate at a native resolution of 1024×1024. ObjectClear performs removal at a lower resolution of 512×512, followed by post-processing upscaling. This pipeline inevitably leads to a noticeable loss of high-frequency details—a degradation that, similar to the issue mentioned above, is not effectively penalized by existing quantitative metrics. Regarding GeoRemover, it employs a two-stage pipeline that cascades two FLUX-fill models. Despite its massive computational overhead, it fails to generate physically realistic removal results consistent with the environmental lighting and physics.

As illustrated in Fig. 10, baselines fail to correctly associate the target object with its corresponding physical effects, including illumination changes and volumetric smoke. Furthermore, regarding transparent objects, ObjectClear and OmniEraser tend to exhibit *excessive erasure*, thereby destroying the underlying semantic structure of the background. Meanwhile, GeoRemover, while attempting to remove transparent objects, frequently neglects concurrent side effects, such as cast shadows. In contrast, **PhysRemover** achieves superior physical consistency, simultaneously accomplishing the elimination of objects along with their side effects and the precise removal of transparent objects with clearer background preservation.

A.6 Additional Qualitative Results

To demonstrate the robust generalization capability of our proposed method, we present a broader set of qualitative results in this section. Specifically, we provide additional samples from diverse in-the-wild scenarios and evaluate performance on stylized artistic images. Furthermore, we highlight the practical utility of PhysRemover by showcasing its ability to remove arbitrarily specified objects from complex scenes.

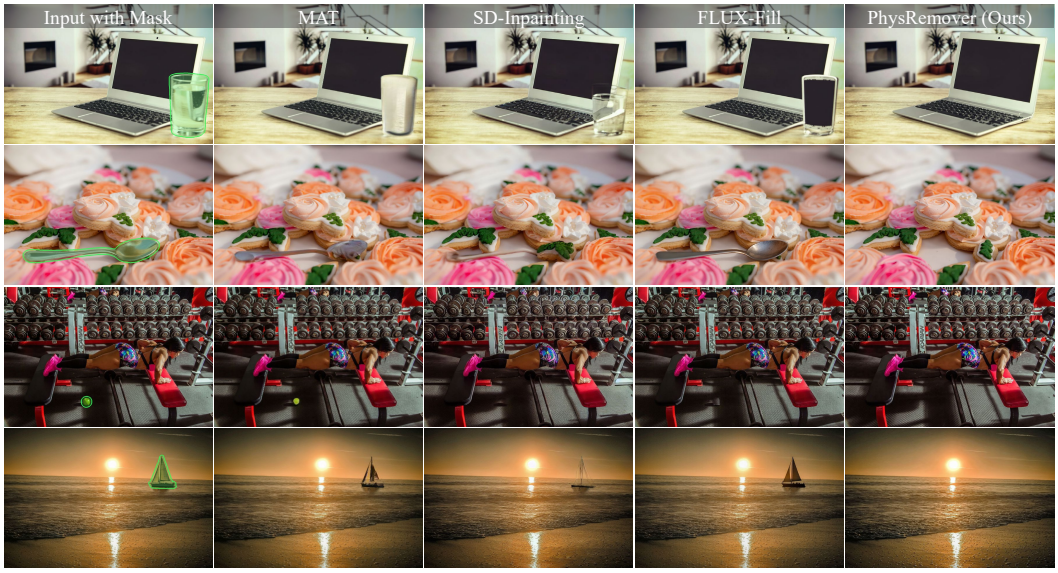

Figure 7: **Additional Qualitative Comparison with Traditional Removal Methods.**

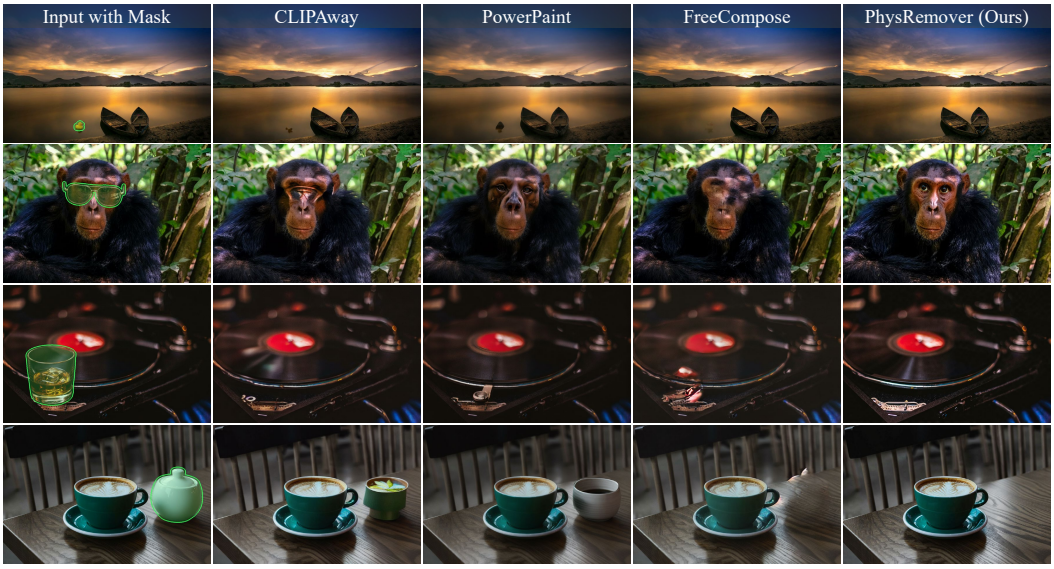

Figure 8: **Additional Qualitative Comparison with Traditional Removal Methods.**

More In-the-Wild Results. As illustrated in Fig. 11, we present extensive results on real-world images characterized by complex lighting conditions and diverse semantic contexts. These examples further validate that our model can effectively generalize to unseen data, accurately handling intricate shadows, reflections, and transparency.

Performance on Artistic Styles. As shown in Fig. 12, we evaluate the model's performance on stylized images, such as oil paintings and digital art. The results demonstrate that PhysRemover not only removes the target object and its physical effects but also seamlessly preserves the original artistic style and texture patterns of the background, avoiding any style degradation or inconsistency.

Arbitrary Object Removal. To verify the flexibility and practical value of our method, we demonstrate the removal of arbitrary specified objects within a scene. As shown in Fig. 13, whether the target is a prominent foreground entity or a subtle background element, our method consistently eliminates the specified target along with its associated side effects, proving its versatility for general image editing tasks.

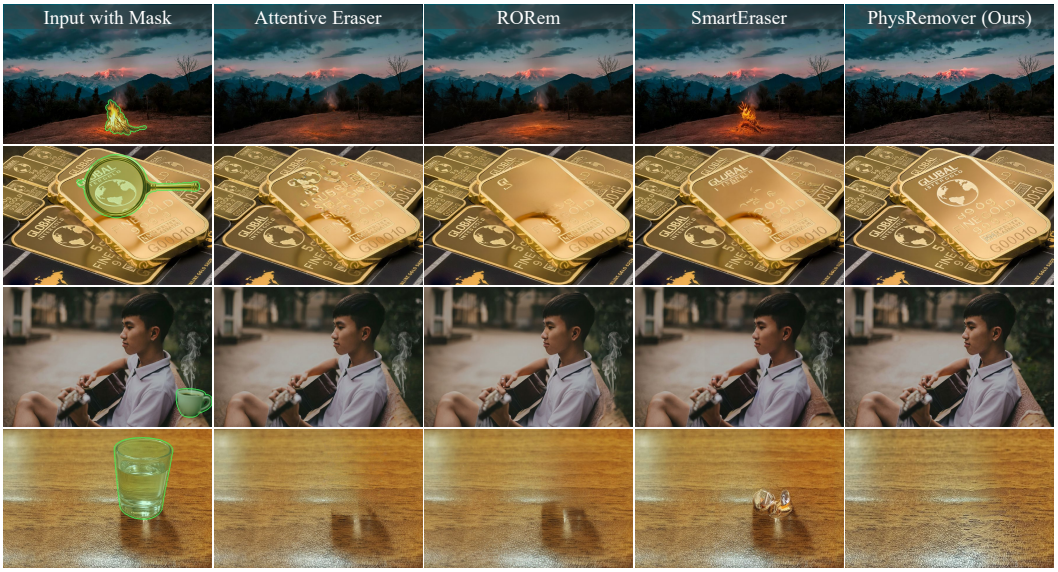

Figure 9: **Additional Qualitative Comparison with Traditional Removal Methods.**

A.7 Additional Related Work

Transparent Object Matting and Shadow Removal. Transparent object matting, exemplified by TOM-Net [2], estimates representations such as attenuation and refractive flow for compositing transparent objects onto new backgrounds. Shadow-removal methods [6] instead recover shadow-free appearance while preserving scene content. These tasks are related to physically realistic object removal but have different outputs: neither directly removes a specified object together with its spatially distributed effects and reconstructs the counterfactual background. PhysRemover jointly addresses these requirements, including background recovery behind transparent objects.

Video Object and Effect Removal. EffectErase [8] studies video object removal through joint removal–insertion learning and introduces VOR, a paired-video dataset covering occlusion, shadow, lighting, reflection, and deformation. It models spatiotemporal object–effect correlations to support consistent removal across frames. Our work addresses the complementary single-image setting, focusing on removing a specified object and its transient effects while recovering background content behind transparent objects. The two methods therefore address related objectives under different temporal and input settings.

A.8 Limitations

As with all latent diffusion-based editing methods, PhysRemover operates in the compressed latent space of a pretrained VAE. While the VAE enables efficient high-resolution processing, the encode-decode cycle inevitably introduces subtle information loss in high-frequency details that are imperceptible to the human eye but may be detected by pixel-level metrics. This is a fundamental limitation shared by all VAE-based generative frameworks and is not specific to our method. Future work may explore higher-fidelity autoencoders or hybrid approaches to further mitigate this artifact.

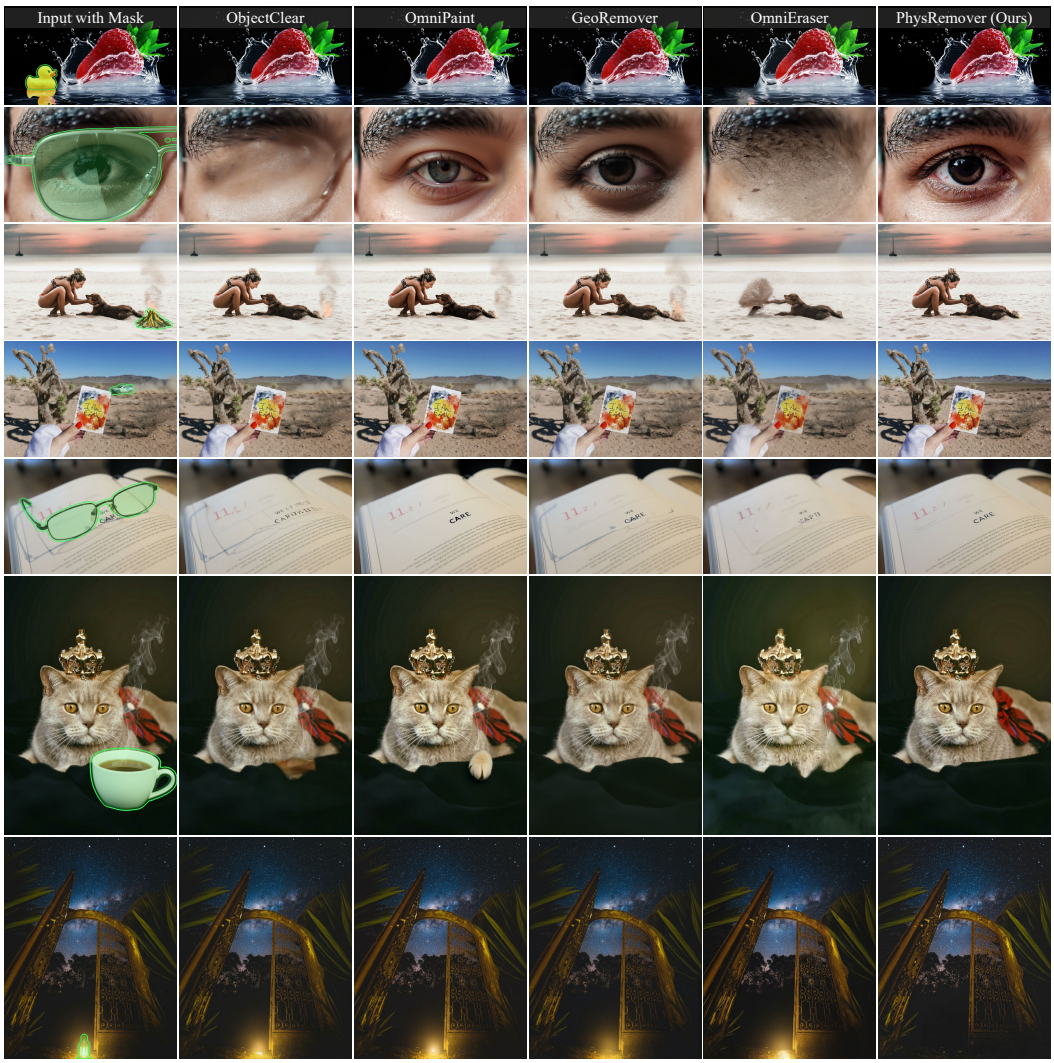

Figure 10: **Additional Qualitative Comparison Results.**

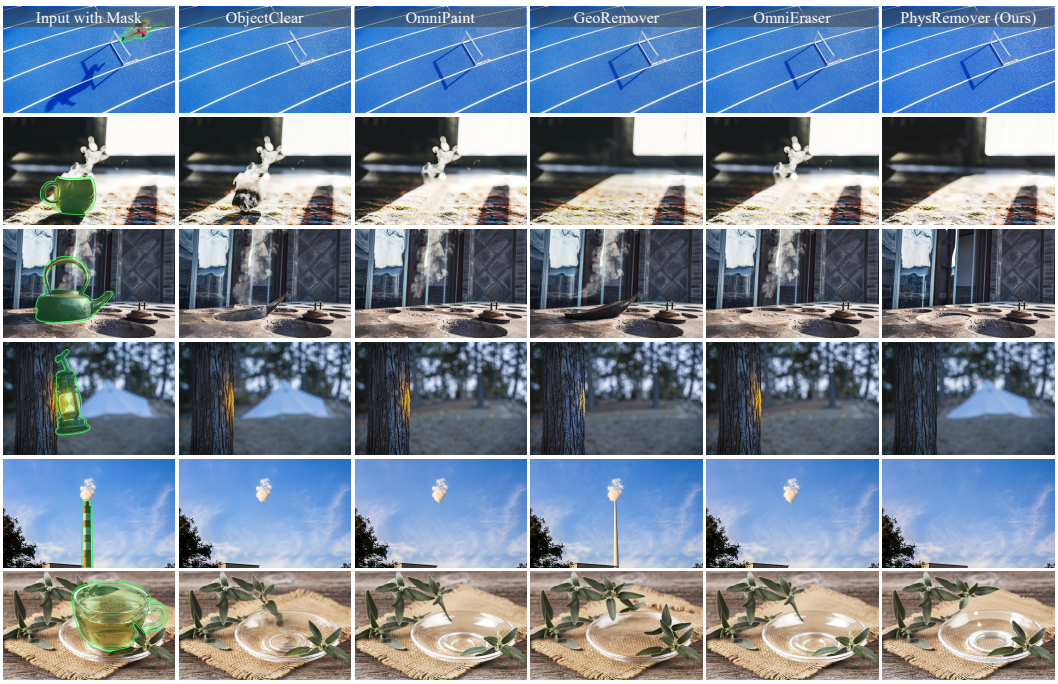

Figure 11: **Additional Qualitative Results for In-the-Wild Results.**

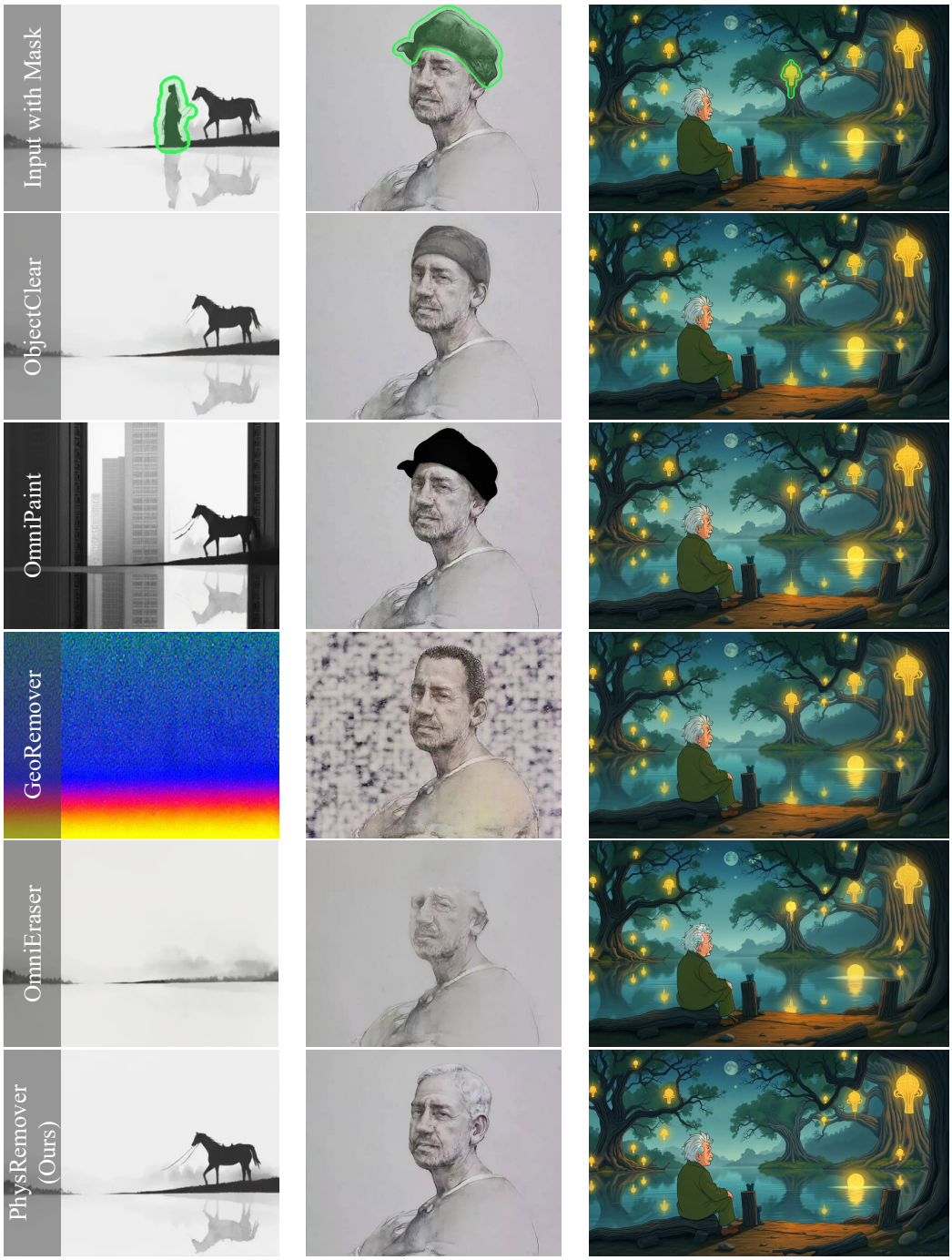

Figure 12: **Additional Qualitative Results on Artistic Styles.**

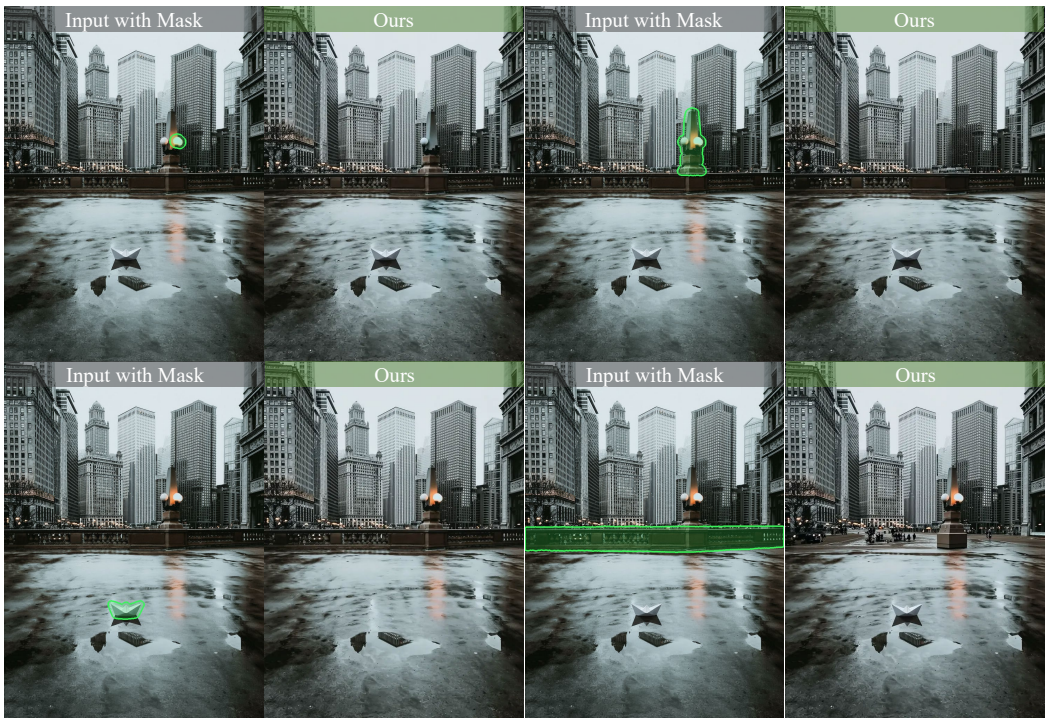

Figure 13: **Additional Qualitative Results on Arbitrary Object Removal.**

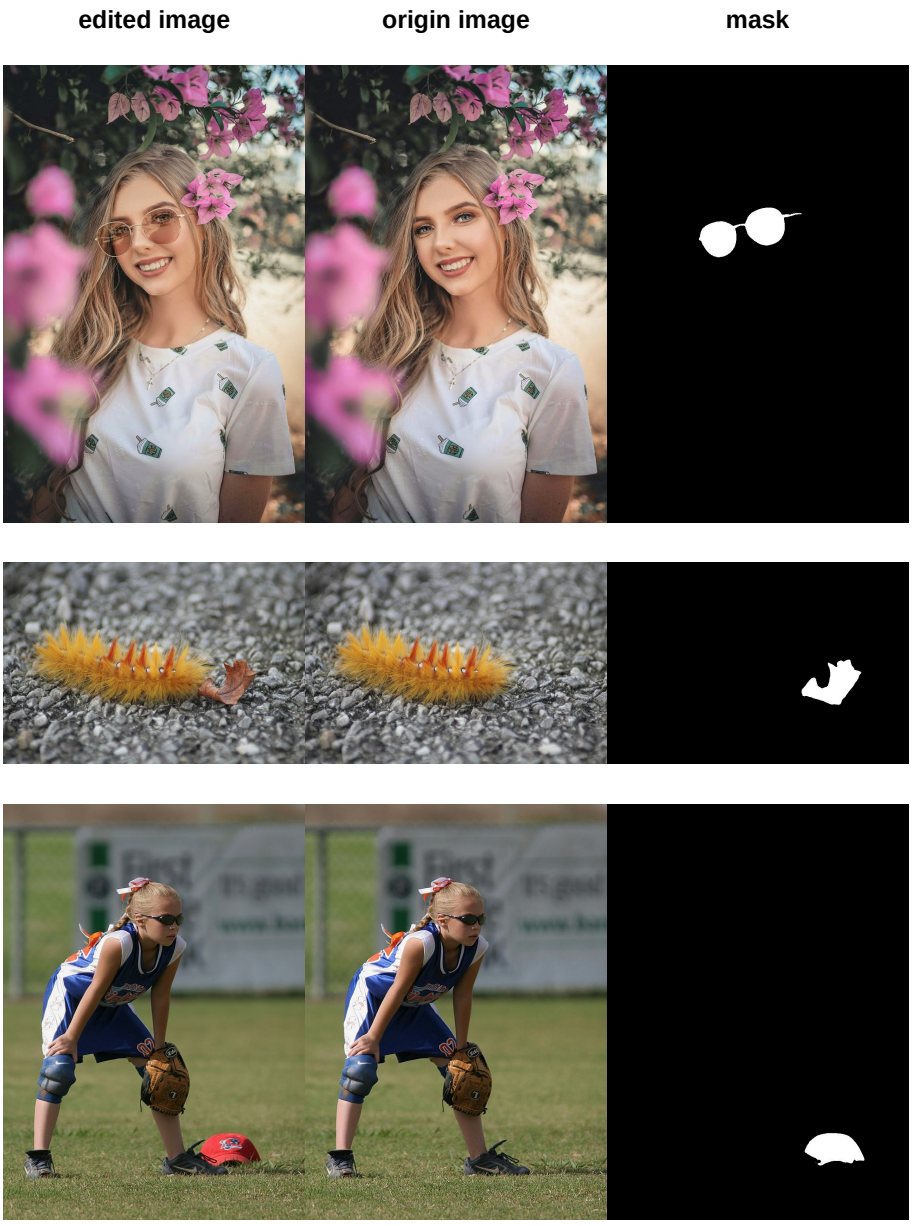

Figure 14: **Randomly sampled examples from the PhyTrace dataset.** Each triplet consists of the original image $\mathbf{x}$ (with object and its physical effects), the clean background ground truth $\mathbf{x}_{\text{bg}}$, and the object mask $\mathbf{m}$. The dataset covers diverse physical interactions including cast shadows, reflections, emitted illumination, volumetric media, and transparent objects.

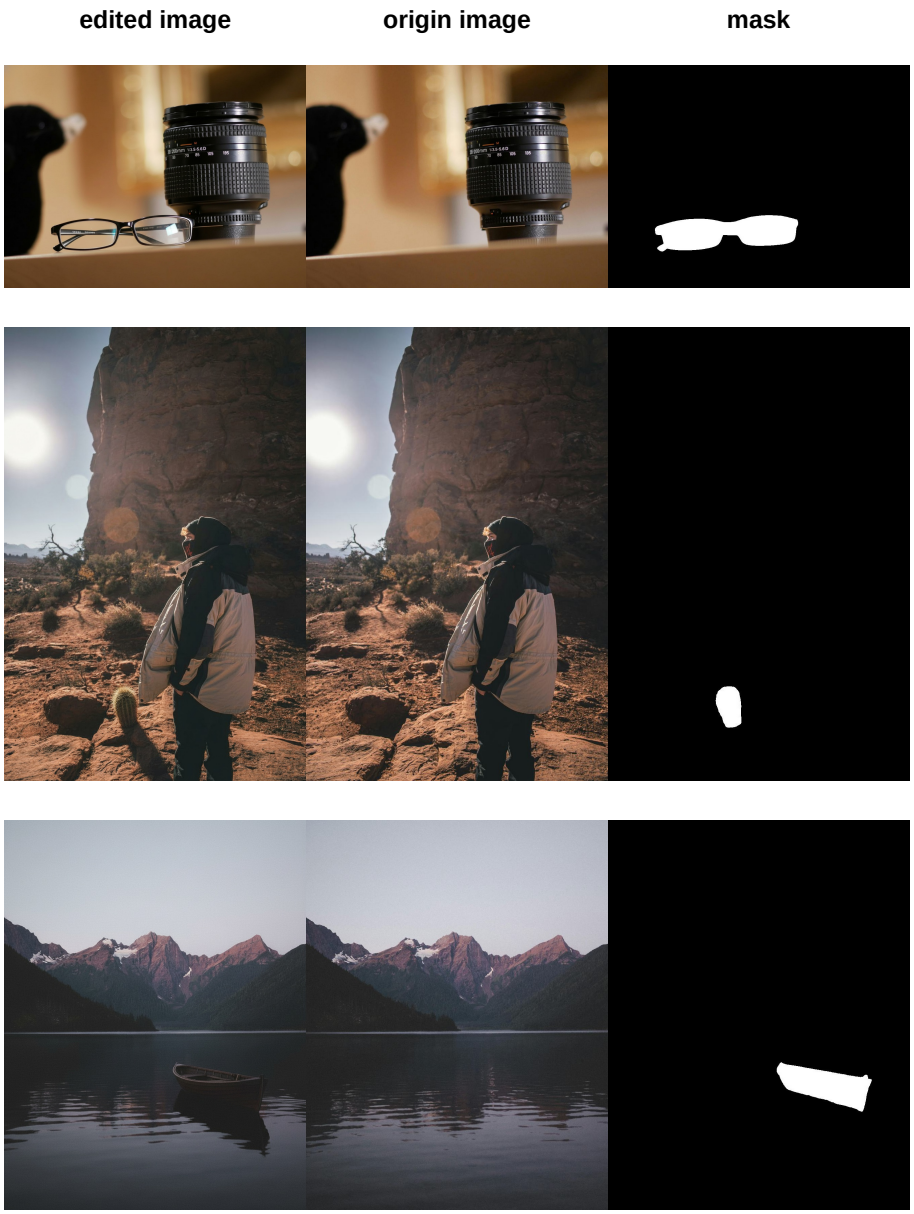

Figure 15: **Randomly sampled examples from the PhyTrace dataset.** Each triplet consists of the original image $\mathbf{x}$ (with object and its physical effects), the clean background ground truth $\mathbf{x}_{\text{bg}}$, and the object mask $\mathbf{m}$. The dataset covers diverse physical interactions including cast shadows, reflections, emitted illumination, volumetric media, and transparent objects.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper contains sufficient details for reproducing the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In the supplementary material, we have provided the proposed benchmark, along with the calculation code for the proposed VLM-based metrics. The full codebase, pre-trained models, and the complete dataset are currently being cleaned and documented. We commit to making them fully open-source and publicly available upon the acceptance to facilitate reproducibility.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides the settings and details of the method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following previous methods, we did not report the error bars. But in order to calculate the metrics, we used the same seed and averaged the 4 generated results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide the sufficient information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research is conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All collected data are released under the CC 4.0 public license.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, they are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, they are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.