PHYSLAYER: LEARNING OBJECT–EFFECT SEPARATION THROUGH PHYSICS-GROUNDED LAYER DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent diffusion-based approaches have advanced image layer decomposition, enabling flexible extraction and editing of individual image components. However, complex natural scenes remain challenging when a target is partially occluded, covers previously invisible content, or induces effects such as shadows and reflections. Existing representations often entangle these effects with the foreground or background, while natural 2D imagery rarely provides aligned supervision for the complete target, the target-free scene, and their interactions. We propose PhysLayer, a target-conditioned framework that decomposes a composite image, given a target bounding box, into three complementary components: an amodal foreground RGBA layer, a clean target-removed background, and a signed residual capturing target-induced scene interactions. To supervise this formulation, we construct PhysLayer-30K through controllable 3D scene manipulation and counterfactual rendering, providing matched foreground, background, visibility, and interaction annotations from the same scene. A joint diffusion transformer predicts the three layers within a shared denoising process using layer-aware spatial encoding, targeted foreground and background supervision, and residual attention alignment. We evaluate PhysLayer on foreground matting, object removal, and layer decomposition. Compared with the box-conditioned RevealLayer baseline, PhysLayer improves object-removal PSNR by 2.51 dB on OBER-Test and reduces background LPIPS by 47.4% on PhysLayer-Test. It also achieves alpha SoftIoU scores of 0.9333 on RefMatte-RW100 and 0.945 on PhysLayer-Test, demonstrating improved foreground recovery and target-free scene reconstruction.

1 INTRODUCTION

Generative image models have improved image synthesis and editing (Wu et al., 2025). Manipulating individual objects while preserving the surrounding scene requires structured, independently editable layers (Yin et al.). Recovering these layers is challenging because a target can hide background content, be occluded by other objects, and induce shadows or reflections beyond its silhouette.

Foreground and background layers alone can leave object appearance and scene effects entangled. Effects retained in the background leave traces of the removed target, while effects absorbed into the foreground tie its appearance to the original scene. We therefore use three explicit outputs: an amodal foreground RGBA layer, a clean target-removed background, and a signed interaction residual. The residual represents target-induced effects rather than serving only as an auxiliary training signal. Figure 1 illustrates this distinction.

A central challenge is obtaining matched supervision for all three components. Composing known layers offers explicit targets but limited realistic object-scene interactions (Li et al., 2023; Chen et al., 2025); inferring layers from natural images using segmentation, matting, or generative completion preserves image statistics but relies on inferred hidden content (Chen et al., 2026; Wang et al., 2026). Neither approach readily provides the complete target, the target-free scene, and a separate interaction reference under the same scene configuration, especially when foreground and background content are both occluded.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

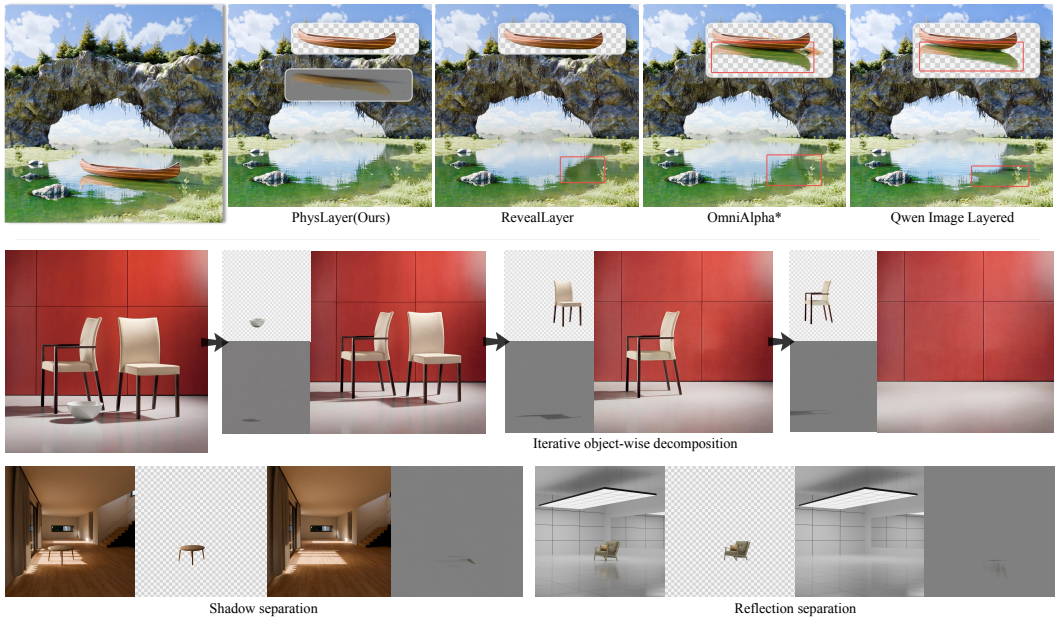

Figure 1: Layer decomposition with PhysLayer. Top: comparison with existing layer decomposition methods. Middle: iterative object-wise decomposition. Bottom: representative residual effects, including cast shadows and reflections.

To address this gap, we introduce PhysLayer, a physics-grounded framework built on matched three-layer supervision. Its data foundation, PhysLayer-30K, is constructed through controllable 3D scene manipulation and counterfactual rendering. Matched scene states provide aligned composite, foreground, background, and visibility references, from which we derive target-specific interaction supervision. This grounds joint learning in a shared scene configuration rather than independently inferred 2D pseudo-labels.

Building on this supervision, Phys-MMDiT jointly generates the three layers from a composite image and a target bounding box. Layer-aware spatial encoding preserves global coordinates for target-region tokens within a shared diffusion process. Targeted objectives use the matched annotations to supervise complete foreground appearance and alpha, background completion, and effect removal, while effect-aware residual attention alignment associates interaction-related semantics with affected regions.

We train with PhysLayer-30K and external layered data and evaluate on matting, object removal, and layer decomposition. Foreground and background comparisons, residual recovery measurements, and qualitative examples assess the resulting decomposition on synthetic and natural images.

Our contributions are summarized as follows:

- We construct PhysLayer-30K, providing matched amodal foreground, target-removed background, visibility, and derived interaction supervision through controllable counterfactual rendering for target-level three-layer decomposition.
- We formulate target-level object–effect separation using three explicit outputs for complete target appearance, target-free scene content, and target-induced effects, accounting for occlusion of both foreground and background.
- We develop Phys-MMDiT to jointly learn the three-layer decomposition from matched supervision, and validate the framework across matting, object removal, and layer decomposition benchmarks.

2 Related Work

Image Matting and Object Removal. Image matting estimates alpha mattes for fine-grained foreground extraction, with recent methods extending to point-guided, box-guided, or text-guided settings. MAM (Li et al., 2024) predicts continuous alpha mattes from SAM features (Kirillov et al., 2023), while OmniAlpha (Yu et al., 2025a) formulates target extraction as generative RGBA prediction. In parallel, object removal focuses on recovering the scene after removing a specified target; representative methods include PowerPaint (Zhuang et al., 2024), Attentive Eraser (Sun et al., 2025), and ObjectClear (Zhao et al., 2026), with the latter additionally addressing associated effects such as shadows and reflections. These tasks primarily recover either the target or the target-free scene, but do not explicitly retain both together with a separate representation of target-induced interactions. PhysLayer instead models all three within a unified target-level decomposition framework.

Image Layer Decomposition. Image layer decomposition recovers independently editable components from a composite image. Qwen-Image-Layered (Yin et al.) predicts multiple RGBA layers, while RefLayer (Chen et al., 2026) and RevealLayer (Wang et al., 2026) focus on prompt-conditioned or box-conditioned amodal layer recovery under occlusion. LayerDecomp (Yang et al., 2025) further incorporates visual effects into transparent foreground layers, whereas LiWi (He et al., 2026) introduces an explicit shadow layer. A remaining challenge is to disentangle object appearance from scene interactions under reliable supervision. Existing datasets often rely on manually authored layers or model-based completion of hidden content. PhysLayer instead uses controllable 3D rendering to provide aligned amodal foregrounds, target-removed backgrounds, visibility annotations, and target-specific interaction residuals from the same scene configuration.

3 Method

To achieve target-level decomposition into complete object appearance, target-free scene content, and object-induced interactions, we construct PhysLayer-30K through a physics-grounded data pipeline and develop Phys-MMDiT, a joint diffusion model that predicts the three layers within a shared denoising process.

3.1 Problem Formulation

Given a composite image I, a bounding box b specifying the target object, and layer-wise textual conditioning s, PhysLayer predicts three aligned layers:

$$\mathcal{D}(I, b; s) = \{F, B, R\}, \tag{1}$$

where F is the amodal foreground RGBA layer, B is the clean background after removing the target, and R is a signed interaction residual.

The foreground recovers the complete target, including occluded regions, while the background reconstructs the scene content revealed after target removal. The residual captures target-induced scene effects that should belong to neither layer, such as cast shadows and reflections. In this way, PhysLayer decomposes a queried object into its appearance, the underlying scene, and its interaction with the environment.

The visibility-aware compositing operator is

$$\mathcal{C}(F, B, R; V) = P_F \odot V + B \odot (1 - A \odot V) + R, \tag{2}$$

where $F = (P_F, A)$, P_F is the premultiplied foreground color, A is the amodal alpha matte, and $V \in [0, 1]^{H \times W}$ denotes target visibility.

3.2 PhysLayer-30K Dataset

To provide aligned supervision for the target layers, we construct PhysLayer-30K through controllable 3D scene composition and counterfactual rendering with Blender Cycles.

As illustrated in Figure 2, we prepare background scenes and a foreground asset pool from 3D resources including Poly Haven and BlenderKit. Candidate insertion regions are identified in each

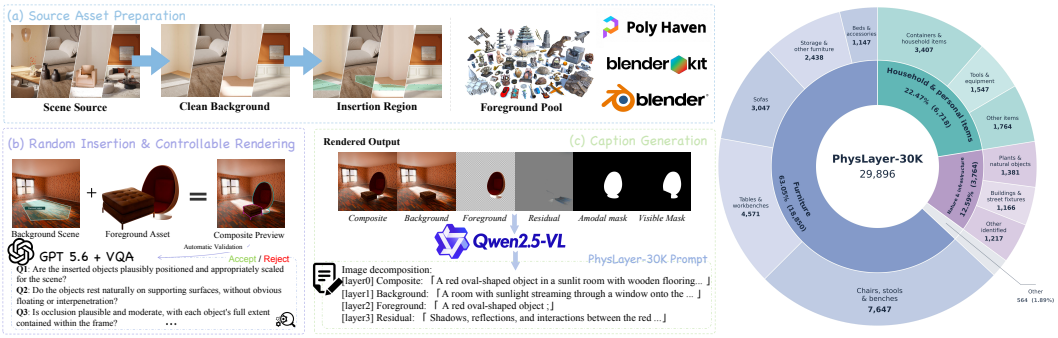

Figure 2: Dataset construction pipeline of PhysLayer-30K. Controllable scene composition and counterfactual rendering are performed with Blender Cycles to obtain matched rendering references and derived interaction annotations for foreground, background, visibility, and object–scene interactions.

scene, and foreground assets are randomly placed with varied pose, scale, and location. Preview-based VQA filtering rejects implausible configurations before final rendering.

For each valid configuration, matched scene states are rendered under the same camera and illumination to obtain the composite I, target-free background B, amodal foreground F, amodal alpha A^{gt}, and visible alpha $A^{\mathrm{gt}}_{\mathrm{vis}}$. Each sample provides the aligned supervision $\mathcal{S} = (I, B, F, R, A^{\mathrm{gt}}, A^{\mathrm{gt}}_{\mathrm{vis}})$. The dataset annotations $M_{\mathrm{amodal}} \equiv A^{\mathrm{gt}}$ and $M_{\mathrm{vis}} \equiv A^{\mathrm{gt}}_{\mathrm{vis}}$ are soft-alpha references in $[0, 1]$, rather than binary masks. Qwen2.5-VL (Bai et al., 2025) additionally generates layer-wise semantic descriptions used during training.

Using the matched composite and background, we construct the interaction target outside the visible object:

$$R^{\mathrm{gt}} = S_R \odot \left[(I - B^{\mathrm{gt}}) \odot (1 - A^{\mathrm{gt}}_{\mathrm{vis}})\right], \tag{3}$$

where $A^{\mathrm{gt}}_{\mathrm{vis}}$ is the visible alpha and S_R is soft interaction support derived from the magnitude of the unsuppressed difference to reduce weak responses. The support-filtered interaction target provides supervision for effect separation, while Eq. 2 defines a soft reconstruction consistency constraint rather than an exact identity for the reference layers.

Training combines 29,896 self-rendered samples with 10,304 samples from RevealLayer, yielding 40,200 training records. PhysLayer-30K refers to the self-rendered dataset; the RevealLayer samples provide external layered supervision. Auxiliary objectives require their corresponding annotations.

3.3 Joint PhysLayer Diffusion

PhysLayer jointly generates the three target layers in a shared latent diffusion process. We use the frozen OmniAlpha RGBA-VAE to encode all image layers into a common latent space, and Qwen2.5-VL to provide image-text context h_s. The signed residual is stored in image range as $E_R = (R + 1)/2$, where 0.5 represents zero interaction.

In addition to the clean input latent z_I, the model jointly denoises four generation slots $k \in \{C, B, F, R\}$ during training and sampling. The auxiliary composite C receives scene-level flow supervision; the task outputs are $B/F/R$. All slots use one shared output projection. The clean input remains fixed, and C is distinct from the image recomposed from the three task outputs.

For each output latent $z_0^{(k)}$, flow matching (Lipman et al., 2022) defines

$$z_\sigma^{(k)} = (1 - \sigma) z_0^{(k)} + \sigma \epsilon^{(k)}, \qquad u^{(k)} = \epsilon^{(k)} - z_0^{(k)}, \quad \epsilon^{(k)} \sim \mathcal{N}(0, \mathbf{I}). \tag{4}$$

Phys-MMDiT (Esser et al., 2024) is initialized from Qwen-Image-Edit-2511 (Yin et al.). The pretrained Transformer, text encoder, and RGBA-VAE are frozen; LoRA modules and the shared output projection are trainable. Rather than generating each layer independently, a shared Transformer

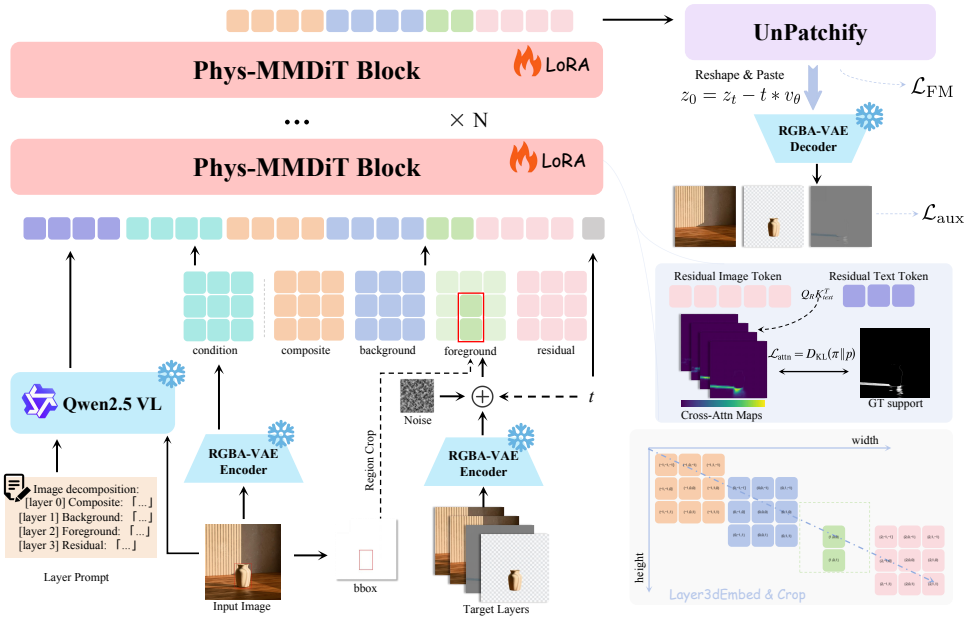

Figure 3: The framework of PhysLayer. PhysLayer jointly decomposes a target into an amodal foreground, a target-removed background, and a signed interaction residual. Layer-aware spatial encoding preserves global scene alignment, while effect-aware attention supervision aligns residual responses with interaction support.

jointly predicts their flow velocities:

$$\left\{v_\theta^{(k)}\right\}_k = \Phi_\theta\left(z_I,\, \{z_\sigma^{(k)}\}_k,\, h_s,\, \sigma\right). \tag{5}$$

Joint processing allows each output to access both the observed scene context and the evolving representations of the other layers.

Layer-Aware Spatial Encoding. The background and residual span the full image, whereas foreground generation is restricted to the queried object region. Cropping foreground tokens naively would discard their original scene coordinates. We instead assign each image token a layer-aware coordinate

$$p_i^{(k)} = (\ell_k, x_i, y_i), \qquad \Omega_F = \Omega_b, \quad \Omega_C = \Omega_B = \Omega_R = \Omega, \tag{6}$$

where ℓ_k identifies the layer, (x_i, y_i) denotes its position in the original image grid, Ω is the full token grid, and Ω_b is the target-box region.

Only foreground generation tokens are selected within the target box; the input composite retains full scene context. Training boxes cover the amodal target extent, while the spatial query bounds the foreground token region at inference. Selected tokens retain their original global coordinates through layer-aware rotary positional encoding (Su et al., 2024), preserving alignment among the outputs.

3.4 Training Objectives

Validity-Aware Flow Matching. Let m_k indicate whether a slot's target is valid, and $\mathcal{E}^{(k)}$ its velocity MSE averaged over channels and valid spatial elements. We optimize

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\frac{\sum_{k\in\{C,B,F,R\}} m_k w_k \mathcal{E}^{(k)}}{\sum_{k\in\{C,B,F,R\}} m_k w_k}\right]. \tag{7}$$

The foreground domain is the target-box region; other slots span the full canvas. Missing labels disable corresponding direct supervision rather than supplying zero interaction. For image-space losses, we recover

$$\hat{z}_0^{(k)} = z_\sigma^{(k)} - \sigma v_\theta^{(k)} \tag{8}$$

and decode $B/F/R$ through the frozen RGBA-VAE, preserving gradients to the estimated latents. Below, $\langle X \rangle_M$ denotes channel and spatial averaging weighted by soft mask M and normalized by its mass; Ω denotes the full canvas. We use $\rho_\epsilon(e) = \sqrt{e^2 + \epsilon^2}$.

Foreground Supervision. Premultiplied foreground color is supervised within the amodal support, and alpha over the full canvas:

$$\mathcal{L}_{\text{fg}}^{\text{aux}} = \lambda_{\text{rgb}} \left\langle \rho_\epsilon(\widehat{P}_F - P_F^{\text{gt}}) \right\rangle_{A^{\text{gt}}} + \lambda_\alpha \left\langle \rho_\epsilon(\widehat{A} - A^{\text{gt}}) \right\rangle_{\Omega}. \tag{9}$$

Alpha weights the color error without multiplying premultiplied color again. Support normalization prevents transparent regions from diluting appearance supervision, including for occluded target content. Full-canvas alpha supervision constrains target coverage and spurious foreground outside it.

Background Supervision. Background recovery must complete the target-occupied region and remove its surrounding effects. We use $O = A_{\text{vis}}^{\text{gt}}$ and an effect proxy E_{bg} from the composite–background difference excluding the visible object. This proxy is distinct from residual support S_R and residual encoding E_R. With channel-averaged squared background velocity error e_B and latent-grid weights O_ℓ, E_ℓ, regional flow supervision is

$$\mathcal{L}_{\text{bg}}^{\text{reg}} = \tfrac{1}{2}\langle e_B \rangle_{O_\ell} + \tfrac{1}{2}\langle e_B \rangle_{E_\ell}. \tag{10}$$

The background auxiliary objective combines regional flow, photometric error on $U = \text{clip}(O + E_{\text{bg}}, 0, 1)$, and contextual-crop LPIPS (Zhang et al., 2018):

$$\begin{aligned} \mathcal{L}_{\text{bg}}^{\text{aux}} = \lambda_{\text{reg}} \mathcal{L}_{\text{bg}}^{\text{reg}} + \lambda_{\text{pix}} \left\langle \rho_\epsilon(\widehat{B} - B^{\text{gt}}) \right\rangle_U \\ + \lambda_{\text{perc}} \,\text{LPIPS}\left(\mathcal{T}_O(\widehat{B}), \mathcal{T}_O(B^{\text{gt}})\right). \end{aligned} \tag{11}$$

$\mathcal{T}_O$ applies the same expanded visible-object crop, resize, and normalization to both images, retaining surrounding pixels. Object-region terms supervise completion; effect-region terms supervise interaction removal. Full-canvas background flow remains active.

Effect-Aware Residual Attention Alignment. To associate sparse interactions with affected regions, we supervise responses from residual queries to effect-related tokens in the residual description. For valid task-text keys $\mathcal{T}$ and effect tokens $\mathcal{T}_{\text{eff}} \subset \mathcal{T}$, define

$$r_i = \frac{1}{H_a} \sum_{h=1}^{H_a} \sum_{j \in \mathcal{T}_{\text{eff}}} \left[\text{softmax}_{j' \in \mathcal{T}} \left(\frac{(q_{ih}^R)^\top k_{j'h}^{\text{text}}}{\sqrt{d}} \right) \right]_j. \tag{12}$$

Spatially normalizing r and the resized reference support S_R gives p and π, respectively. We minimize

$$\mathcal{L}_{\text{attn}} = D_{\text{KL}}(\pi \| p), \tag{13}$$

averaged over selected blocks. This text-normalized response supplies supervision while preserving the model's joint image–text attention.

Cross-Layer Consistency. Using the visible and amodal reference alpha to obtain V^{gt}, we encourage consistency through Eq. 2:

$$\begin{aligned} \widehat{I}_{\text{rec}} &= \mathcal{C}(\widehat{F}, \widehat{B}, \widehat{R}; V^{\text{gt}}), \\ \mathcal{L}_{\text{rec}} &= \left\langle \rho_\epsilon(\widehat{I}_{\text{rec}} - I) \right\rangle_\Omega. \end{aligned} \tag{14}$$

Visibility keeps completed but occluded foreground behind its occluders. The auxiliary composite C does not enter this reconstruction. Residual direct supervision $\mathcal{L}_{\text{res}}^{\text{aux}}$ combines full-canvas and support-weighted errors in encoded image space. The total objective is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{fg}}^{\text{aux}} + \mathcal{L}_{\text{bg}}^{\text{aux}} + \mathcal{L}_{\text{res}}^{\text{aux}} \\ + \lambda_{\text{attn}} \mathcal{L}_{\text{attn}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}. \end{aligned} \tag{15}$$

Auxiliary groups include their internal weights and require valid annotations. External samples retain $C/B/F$ flow and foreground auxiliaries, but lack the visibility and effect annotations needed for background auxiliaries, residual supervision, and reconstruction.

At inference, PhysLayer uses I, b, and s without reference visibility or interaction support, and recovers $\widehat{R} = 2\widehat{E}_R - 1$.

Table 1: Foreground matting results on RefMatte-RW100 and PhysLayer-Test. On PhysLayer-Test, MSE, MAD, and SAD are evaluated within the target bounding box for all methods.

Guidance	Method	RefMatte-RW100				PhysLayer-Test			
		MSE ↓	MAD ↓	SAD ↓	SoftIoU ↑	MSE ↓	MAD ↓	SAD ↓	SoftIoU ↑
Mask	SAM2	0.0587	0.0611	107.21	0.8769	0.0478	0.0656	4.39	0.8719
	SAM3	0.0406	0.043	72.21	0.9062	0.0514	0.0698	4.51	0.8670
	MAM	0.0267	0.0369	62.13	0.8912	0.0568	0.0962	4.80	0.8214
Box	RevealLayer	0.0136	0.0158	27.89	0.9234	0.0183	0.035	1.86	0.9158
	PhysLayer	**0.0122**	**0.0142**	**24.79**	**0.9333**	**0.0127**	**0.0224**	**1.37**	**0.9450**

Table 2: Object removal results on OBER-Test and PhysLayer-Test. For ObjectClear*, the reported metrics exclude the post-processing step involving background blending with the original image.

Guidance	Method	OBER-Test				PhysLayer-Test			
		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Mask	RORem	24.83	0.7578	0.1120	62.58	27.28	0.8580	0.0829	31.24
	AttentiveEraser	26.40	0.7716	0.1251	54.06	29.57	0.9310	0.0801	26.64
	ObjectClear*	27.57	0.7671	0.0840	31.06	28.83	0.8640	0.0650	26.14
	OmniPaint	29.05	0.8736	0.0521	20.70	30.51	0.9200	0.0519	18.38
Box	RevealLayer	30.42	0.9204	0.0669	27.08	29.56	0.9166	0.0961	18.84
	PhysLayer	**32.93**	**0.9513**	**0.0369**	**19.18**	**33.66**	**0.9495**	**0.0505**	**9.91**

4 EXPERIMENTS

We evaluate PhysLayer from three complementary perspectives: foreground matting, object removal, and layer decomposition.

4.1 EXPERIMENTAL SETUP

Benchmarks. For foreground matting, we evaluate on RefMatte-RW100 (Chen et al., 2026) and PhysLayer-Test. Object removal is evaluated on OBER-Test (Zhao et al., 2026) and PhysLayer-Test. Layer decomposition is evaluated on PhysLayer-Test, which provides aligned ground truth for amodal foreground, target-free background, and interaction residual.

Baselines. For foreground matting, we compare PhysLayer with SAM2 (Ravi et al., 2025), SAM3 (Carion et al., 2026), MAM, and RevealLayer. For object removal, we include mask-guided methods RORem (Li et al., 2025), AttentiveEraser, ObjectClear*, and OmniPaint (Yu et al., 2025b), together with the box-guided RevealLayer baseline. For layer decomposition, we compare against Qwen-Image-Layered, OmniAlpha*, and RevealLayer where corresponding outputs are available. Since removal methods use different spatial guidance, mask-guided and box-guided approaches are distinguished in the corresponding table.

Implementation Details. PhysLayer is initialized from Qwen-Image-Edit-2511 and uses the OmniAlpha RGBA-VAE for layered image encoding and decoding. We freeze the pretrained text encoder, VAE, and Transformer weights, and optimize rank-64 LoRA modules together with the shared output projection. All main and ablation models are trained for 10,000 steps on 8 NVIDIA H200 GPUs at 1024×1024 resolution with a global batch size of 8, using AdamW with a learning rate of 1×10^{-4}. The reconstruction and residual-attention losses use weights of 0.05 and 0.01, respectively. At inference, we use 28 sampling steps with classifier-free guidance scale 4.0.

Table 3: Layer decomposition results on PhysLayer-Test. OmniAlpha* combines its referring matting output as foreground and object removal output as background.

Method	Background					Foreground		
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	RGB L1 ↓	PSNR ↑	FID ↓	SoftIoU ↑
Qwen-Image-Layered	21.32	0.8380	0.2310	94.10	0.0540	19.54	119.53	0.5760
OmniAlpha*	<u>31.43</u>	0.8980	0.1503	<u>12.58</u>	<u>0.0164</u>	<u>33.17</u>	37.63	0.8533
RevealLayer	29.56	<u>0.9166</u>	<u>0.0961</u>	18.84	0.0227	32.84	<u>26.01</u>	<u>0.9158</u>
PhysLayer	**33.66**	**0.9495**	**0.0505**	**9.91**	**0.0129**	**38.74**	**16.02**	**0.9450**

Table 4: Ablation of interaction modeling on PhysLayer-Test. Residual metrics are evaluated within the ground-truth interaction support.

Variant	Background Level				Foreground Level			Residual Level	
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	FID ↓	SoftIoU ↑	PSNR_m ↑	SSIM_m ↑
w/o R_{attn}	32.51	0.944	0.0534	12.28	37.23	18.01	0.942	22.64	0.7289
w/o R_{layer}	32.91	0.944	0.056	13.45	38.47	22.94	0.941	–	–
PhysLayer	**33.66**	**0.9495**	**0.0505**	**9.91**	**38.74**	**16.02**	**0.945**	**24.41**	**0.7931**

4.2 Comparisons with State-of-the-Art

4.2.1 Foreground Matting

Table 1 evaluates foreground alpha prediction on RefMatte-RW100 and PhysLayer-Test. On RefMatte-RW100, PhysLayer consistently improves over the reproduced RevealLayer baseline, reducing SAD from 27.89 to 24.79 and increasing SoftIoU from 0.9234 to 0.9333. The gains are more pronounced on PhysLayer-Test, where PhysLayer improves SoftIoU from 0.9158 to 0.945 while also reducing the target-region MSE, MAD, and SAD.

4.2.2 Object Removal

We evaluate target-free scene recovery under both mask- and box-guided settings on OBER-Test and PhysLayer-Test. PhysLayer performs strongly on both benchmarks despite relying only on box guidance. On OBER-Test, it reaches 32.93 dB PSNR, exceeding the box-guided RevealLayer baseline by 2.51 dB, while also achieving the best SSIM and LPIPS among the compared methods. A similar trend is observed on PhysLayer-Test, where PhysLayer obtains 33.66 dB PSNR and reduces FID to 9.91.

4.2.3 Layer Decomposition

PhysLayer-Test. We jointly evaluate foreground recovery and target-free scene reconstruction on PhysLayer-Test. As shown in Table 3, PhysLayer improves both foreground and background recovery. Compared with RevealLayer, foreground PSNR increases from 32.84 to 38.74 dB and background LPIPS decreases from 0.0961 to 0.0505. PhysLayer also outperforms the task-composed OmniAlpha* baseline. The following qualitative examples illustrate target-associated effect separation.

Qualitative Results. Figure 4 compares target extraction and scene recovery. In the chair example, the adjacent ottoman belongs to the remaining scene: the displayed OmniAlpha* and Qwen-Image-Layered foregrounds also include it, whereas PhysLayer isolates the chair and retains the ottoman in the background. The car example highlights local scene preservation, including tree shadows and pavement markings; the displayed Qwen-Image-Layered background retains the car. The staircase example shows a small target and differences around its cast shadow. PhysLayer's gray residual insets separately represent target-associated changes. These examples illustrate the roles of the three outputs.

Table 5: Ablation of training objectives on PhysLayer-Test.

Variant	Background Level				Foreground Level		
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	FID ↓	SoftIoU ↑
w/o $\mathcal{L}_{\text{rec}}$	33.39	0.943	0.053	12.42	37.60	27.66	0.9293
w/o $\mathcal{L}_{\text{bg}}^{\text{aux}}$	32.38	0.939	0.0585	17.05	37.21	17.45	0.942
w/o $\mathcal{L}_{\text{fg}}^{\text{aux}}$	32.98	0.946	0.0526	11.83	36.46	21.36	0.934
PhysLayer	**33.66**	**0.9495**	**0.0505**	**9.91**	**38.74**	**16.02**	**0.945**

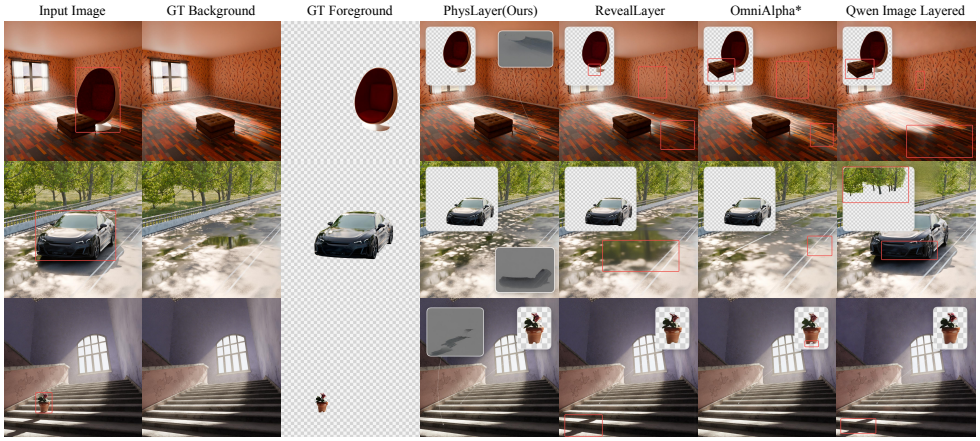

Figure 4: Layer decomposition on PhysLayer-Test. Prior methods may incompletely separate the target from its scene effects and can also introduce residual artifacts or unwanted content.

4.3 Ablation Studies

Interaction Modeling. We examine the role of explicit target–scene interaction modeling. PhysLayer directly recovers interaction residuals with PSNR_m 24.41 and SSIM_m 0.7931; removing R_{attn} reduces them to 22.64 and 0.7289, while also degrading background and foreground FID from 9.91 to 12.28 and from 16.02 to 18.01. Removing R_{layer} further increases these FIDs to 13.45 and 22.94; since this variant also disables reconstruction consistency, it reflects the benefit of the complete interaction modeling configuration rather than the residual layer alone. These results support explicit interaction recovery and its spatial supervision.

Training Objectives. We ablate the training objectives while keeping the three-layer formulation fixed. Removing $\mathcal{L}_{\text{rec}}$ disables only reconstruction consistency and notably degrades foreground quality, increasing FID from 16.02 to 27.66 and reducing Alpha SoftIoU from 0.9450 to 0.9293. The layer-specific auxiliary losses show complementary effects: removing $\mathcal{L}_{\text{bg}}^{\text{aux}}$ degrades background quality (FID 9.91→17.05), while removing $\mathcal{L}_{\text{fg}}^{\text{aux}}$ mainly affects the foreground (PSNR 38.74→36.46, SoftIoU 0.945→0.934). These results support both reconstruction consistency and targeted layer-specific supervision.

5 Conclusion

We presented PhysLayer, a target-conditioned framework that decomposes a scene into an amodal foreground, a target-free background, and an explicit target-induced interaction residual. PhysLayer-30K provides aligned supervision for this formulation through controlled scene manipulation and counterfactual rendering. Experiments across matting, object removal, and layered decomposition validate both interaction recovery and its complementary benefit to foreground-background reconstruction.

6 AI USE STATEMENT

We used GPT-5.6 Sol to assist with dataset quality review, GPT-6 Astra for proofreading and light copy-editing, and Codex for code maintenance and implementation assistance. The research ideas, methodological design, and scientific conclusions were developed by the human authors. All technical content, scientific claims, mathematical formulations, algorithms, experimental design, dataset processing decisions, figures, and reported evaluations were authored and verified by the human authors. LLMs were not used to generate research ideas, data, experimental results, or peer reviews, and did not contribute content at the level of a co-author. All AI-assisted suggestions and code changes were manually inspected and validated, with the authors making the final decision on whether to accept or reject them.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Fangyi Chen, Yaojie Shen, Lu Xu, Ye Yuan, Shu Zhang, Yulei Niu, and Longyin Wen. Referring layer decomposition. *arXiv preprint arXiv:2602.19358*, 2026.

Junwen Chen, Heyang Jiang, Keming Wu, Ji Li, Chao Zhang, Keiji Yanai, Dong Chen, and Yuhui Yuan. Prismlayers: Open data for high-quality multi-layer transparent image generative models. 2025.

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.

Yu He, Fang Li, Haoyang Tong, Lichen Ma, Xinyuan Shan, Jingling Fu, Dong Chen, Luohang Liu, Junshi Huang, and Yan Li. Liwi: Layering in the wild. *arXiv preprint arXiv:2605.14552*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1775–1785. IEEE, 2024.

Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22448–22457. IEEE, 2023.

Ruibin Li, Tao Yang, Song Guo, and Lei Zhang. Rorem: Training a robust object remover with human-in-the-loop. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14024–14035. IEEE, 2025.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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, pp. 20734–20742, 2025.

Binhao Wang, Shihao Zhao, Bo Cheng, Qiuyu Ji, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Dawei Leng, and Yuhui Yin. Reveallayer: Disentangling hidden and visible layers via occlusion-aware image decomposition. *arXiv preprint arXiv:2605.11818*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7643–7653. IEEE, 2025.

Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, et al. Qwen-image-layered: Towards inherent editability via layer decomposition, 2025. *URL https://arxiv. org/abs/2512.15603*.

Hao Yu, Jinglin Wang, Jiabo Zhan, Rui Chen, Zile Wang, Huaisong Zhang, Hongyu Li, Xinrui Chen, Yongxian Wei, and Chun Yuan. Omnialpha: Aligning transparency-aware generation via multi-task unified reinforcement learning. *arXiv preprint arXiv:2511.20211*, 2025a.

Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17324–17334. IEEE, 2025b.

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

Jixin Zhao, Zhouxia Wang, Peiqing Yang, and Shangchen Zhou. Precise object and effect removal with adaptive target-aware attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19370–19379, 2026.

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*, pp. 195–211. Springer, 2024.