InsertAnywhere: Geometrically Grounded and Optics-Aware Video Object Insertion

Hoiyeong Jin[1⋆], Hyojin Jang[1⋆], Junha Hyung[1⋆], Jeongho Kim[1⋆], Kinam Kim[1], Dongjin Kim[1], Huijin Choi[2], Hyeonji Kim[2], and Jaegul Choo[1]

[1] KAIST AI
[2] SK Telecom
{hy.jin, wkdgywlsrud, rlawjdghek, sharpeeee, kinamplify, dj_kim, jchoo}@kaist.ac.kr
{astehelen, hyeonji}@sk.com

Fig. 1: InsertAnywhere achieves realistic video object insertion by 4D scene understanding and optics-aware video synthesis. From a single user-specified 3D placement, our method seamlessly propagates the reference object to produce geometrically and photometrically consistent insertions across complex camera motions, occlusions, and viewpoints, while naturally synthesizing optical effects such as shadows and reflections.

Abstract. Recent advances in diffusion models have enabled impressive video editing capabilities, yet production-grade Video Object Insertion (VOI) remains challenging due to inadequate 4D scene understanding and a lack of proper optical interactions, such as shadows and reflections. To address these limitations, we present InsertAnywhere, a comprehensive VOI framework that achieves geometrically grounded object placement and optics-aware video synthesis. Our approach first leverages a 4D-aware mask generation module that allows users to anchor an object's 3D pose in a single frame. The framework automatically propagates this placement across the video, accurately handling local scene dynamics and occlusions. To synthesize realistic physical lighting interactions, we introduce Optics-Aware Representation Alignment, a novel

⋆ Authors contributed equally to this work.

strategy that utilizes an extended mask to guide feature extraction, enabling optical effects to seamlessly extend beyond the inserted object's boundary. Finally, to overcome the lack of training data for such phenomena, we construct and open-source ROSE++, a specialized quadruplet dataset tailored for the supervised learning of optical effects. Extensive experiments demonstrate that InsertAnywhere produces geometrically plausible and photometrically realistic insertions in complex real-world scenarios, significantly outperforming existing research and commercial generative tools.

1 Introduction

Recent advances in diffusion-based generative models have driven significant progress in user-controllable video editing, propelling Video Object Insertion (VOI) into a prominent role for applications like content creation, advertising, and film post-production. The goal of VOI is to seamlessly integrate new objects into existing scenes while maintaining strict spatial, temporal, and photometric consistency. While recent research [2, 11, 12, 21, 23] and commercial video generation tools such as Kling [22] and Pika-Pro [18] have made strides in temporally coherent object insertion, existing models still fall short in two critical aspects: 4D-aware object placement and optics-aware video synthesis. These limitations severely restrict their ability to achieve production-grade quality.

Achieving accurate object placement requires both user-guided control and a robust 4D geometric understanding of the scene. Since a single 2D reference image lacks scale and depth context, users must be able to explicitly specify the object's initial position, size, and pose. However, manually annotating this across all video frames is highly impractical. Therefore, a robust framework must automatically propagate the initial placement throughout the sequence while gracefully handling complex real-world dynamics—such as moving support surfaces and occlusions caused by foreground elements—as demonstrated in Fig. 1 and Fig. 9.

Beyond placement, the generative model must faithfully synthesize the object's appearance alongside local optical variations induced by the insertion, such as cast shadows and reflections. While mask-free VOI models [3, 11, 12, 15, 29] can hallucinate these effects, they often fail to preserve the unedited background regions. Conversely, mask-conditioned models [8, 23, 28] rigidly confine edits within the provided boundary, struggling to synthesize optical interactions that naturally extend into the surrounding scene. A major bottleneck in solving this is the lack of open-source training data; learning these effects requires a dataset containing a paired tuple of [source video, target video, object mask video, reference image], where the target video explicitly contains the optically integrated object.

To address these challenges, we introduce InsertAnywhere, a comprehensive VOI framework that combines 4D-aware object placement with optics-aware video generation. First, our 4D-aware mask generation module reconstructs the input video into a 4D scene representation. Through an intuitive user interface,

users can anchor the object's 3D pose and scale in a single reference frame. Our module then automatically propagates this placement temporally, accurately tracking local scene dynamics (e.g., an object moving synchronously with the luggage cart it rests upon) while preserving occlusion boundaries. The 3D object is then reprojected into 2D to extract a temporally coherent mask video, which serves as the spatial condition for the subsequent video generation stage.

For optics-aware video synthesis, we construct and open-source ROSE++, a synthetic dataset specifically tailored to train models on complex optical interactions. We demonstrate that models fine-tuned on ROSE++ generalize highly effectively to real-world VOI tasks. To maximize visual fidelity, our generation pipeline leverages the strong priors of image-based insertion models via first-frame anchoring. Furthermore, we propose Optics-Aware Representation Alignment to enhance photometric consistency. By aligning the intermediate features of an extended mask with those of the primary input mask, it enables mask-conditioned video diffusion models to synthesize realistic shadows and reflections that seamlessly extend beyond the immediate boundaries of the inserted object.

Our contributions are summarized as follows:

- We propose a geometrically grounded mask generation module that propagates object masks across all frames via a 4D scene representation. Based on a single user-specified placement, it produces reliable, temporally coherent masks even under complex camera motions and occlusions.
- We introduce an optics-aware video generation strategy, incorporating a first-frame anchoring technique and Optics-Aware Representation Alignment, which enables the model to synthesize physical lighting interactions, such as shadows and reflections, that extend beyond the object mask.
- We construct and open-source ROSE++, a specialized dataset for video object insertion comprising quadruplets of source videos, target videos, mask sequences, and reference images, explicitly tailored for the supervised training of optical effects.
- We demonstrate that **InsertAnywhere** achieves geometrically plausible and photometrically realistic object insertions across diverse real-world scenarios, significantly outperforming existing commercial and research-based generative tools.

2 Related Work

Mask-Free Video Object Insertion Mask-free models [3, 11, 12, 15, 29] typically drive insertion using text prompts and an edited first frame. Crucially, the absence of spatial mask conditioning deprives users of the ability to explicitly control the precise location, scale, and boundaries of the inserted object. While their unbounded generation region theoretically allows them to hallucinate natural optical interactions like shadows, this lack of strict spatial constraints frequently degrades unedited background regions. Furthermore, without explicit geometric guidance, these methods struggle to maintain temporal consistency and fail under complex, dynamic occlusion patterns.

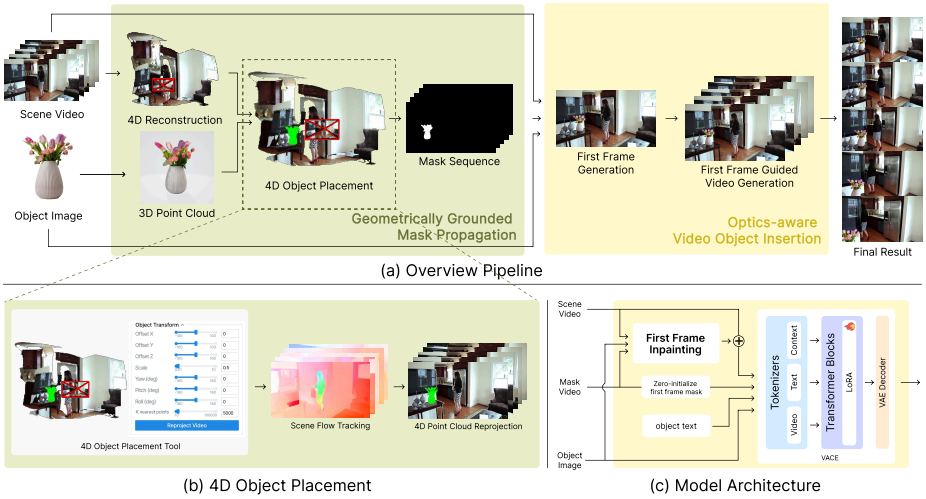

Fig. 2: The InsertAnywhere architecture. Our framework first utilizes 4D scene reconstruction and scene flow tracking to propagate a single user-specified 3D placement into a geometrically grounded mask sequence (a, b). This mask then conditions an optics-aware diffusion model (c), which leverages first-frame anchoring and LoRA fine-tuning to seamlessly synthesize the inserted object and its surrounding optical effects.

Mask-Conditioned Video Object Insertion To strictly preserve the background, mask-conditioned models [4,8,21,23,28] formulate VOI as localized video inpainting governed by an explicit spatial mask. However, this rigid confinement prevents them from synthesizing optical interactions (e.g., cast shadows and reflections) that naturally extend into the surrounding scene. Additionally, these methods typically lack 4D geometric understanding, resulting in brittle and incoherent insertions when the target object interacts with moving support surfaces or becomes occluded.

In contrast to both paradigms, **InsertAnywhere** achieves production-grade VOI by automatically generating a 4D-aware mask sequence for geometrically grounded, occlusion-robust placement, while simultaneously employing an optics-aware synthesis strategy to render photorealistic lighting interactions that naturally expand beyond the primary object boundary.

3 Method

3.1 Overview

Video Object Insertion (VOI) is the task of seamlessly integrating a novel object into an existing video sequence. Formally, given a source video $\mathcal{V}_{src} = \{I_1, I_2, \ldots, I_T\}$, a reference object image I_{ref}, and a user-defined initial 3D placement, the goal is to generate a photorealistic output video $\hat{\mathcal{V}}$ where the object is geometrically and photometrically embedded into the scene.

To achieve production-grade VOI, a framework must overcome two critical limitations present in prior works: the lack of robust 4D scene understanding for controllable object placement, and the rigid, tightly-bound nature of standard mask-conditioned generative models that fail to synthesize scene-wide optical interactions.

To simultaneously address both spatial controllability and photometric realism, we propose **InsertAnywhere**. As depicted in Figure 2, our two-stage framework consists of:

- **Geometrically Grounded Mask Propagation (Section 3.2):** Based on a user-specified object placement within a 3D-reconstructed scene, the object is automatically propagated and reprojected using the underlying 4D spatio-temporal dynamics. This yields a reliable, temporally consistent mask sequence $\{M_t\}_{t=1}^{T}$ that accurately handles complex camera motions and occlusions.
- **Optics-Aware Video Synthesis (Section 3.3):** Given the propagated mask sequence, source video, and reference image, our diffusion-based generation pipeline synthesizes the final photorealistic video. By leveraging our ROSE++ dataset and a novel Optics-Aware Representation Alignment strategy, the model seamlessly expands object-induced optical effects (e.g., shadows and reflections) beyond the strict boundaries of the primary mask.

3.2 Geometrically Grounded Mask Propagation

This stage generates a highly controllable, geometrically grounded mask sequence based on the source video and a reference object. The pipeline executes three main steps: 4D scene reconstruction, user-controlled object placement, and scene flow-based temporal propagation. This ensures that the resulting mask sequence accurately aligns with the scene's geometry, camera motion, and temporal continuity.

4D Scene Reconstruction. We first reconstruct a temporally consistent 4D scene representation from the input monocular video by building upon the Uni4D framework [26]. By orchestrating multiple pretrained vision foundation models including depth estimation [17], optical flow [9], camera pose estimation, and segmentation [5, 10, 13], this stage infers the underlying scene geometry and dynamics. Specifically, the 4D representation consists of spatio-temporal point clouds explicitly decoupled into static components for global structure and dynamic components for moving entities. This decoupling allows our framework to handle complex real-world occlusions and perspective changes with high geometric fidelity.

User-Controlled Object Placement. To ensure precise spatial control, the target object is introduced into the 4D scene as a standalone rigid entity. First, the reference object image I_{ref} is lifted into a local 3D point cloud y using TRELLIS [25], an image-to-3D generation model. Through an interactive interface, users align this object point cloud with the reconstructed scene space by

applying a rigid transformation:

$$y_s = s_{obj} R_{obj} y + t_{obj} \tag{1}$$

where the rotation R_{obj}, translation t_{obj}, and global scale factor s_{obj} are intuitively adjusted. This interface provides real-time visualization within the anchor frame s, allowing users to verify the object's perspective and scale relative to the supporting scene geometry before committing to video generation.

Scene Flow-Based Object Propagation. To maintain physical realism in dynamic scenarios (e.g., an object resting on a moving luggage cart), the object's trajectory must synchronize with local scene dynamics. Simply fixing the initial 3D pose in world coordinates often results in physically implausible drifting relative to moving surfaces. To model these physical interactions, we propagate the object motion using a K-Nearest Neighbor (KNN) scene flow aggregation strategy.

We first estimate the dense 2D optical flow of scene points around the object using SEA-RAFT [24]. The depth value of each displaced pixel is back-projected into 3D world coordinates using the estimated camera intrinsics and extrinsics, yielding the 3D displacement Δp_i^t for each tracked scene point at frame t:

$$\Delta p_i^t = p_{i,world}^{t+1} - p_{i,world}^{t} \tag{2}$$

For each frame, we select the K-nearest scene points to the object centroid in 3D space and aggregate their displacement vectors to estimate the object's motion:

$$\Delta p_{obj}^t = \frac{1}{K} \sum_{k \in \text{TopK}} \Delta p_k^t \tag{3}$$

This aggregated displacement drives the object to faithfully follow the dominant movement of nearby scene surfaces while suppressing noisy flow estimates. To propagate the object in 4D space, the 3D coordinates y'_t are sequentially updated:

$$y'_t = y'_{t-1} + \Delta p_{obj}^{t-1} \tag{4}$$

where the sequence is initialized with $y'_s = y_s$ from Equation (1).

Camera-Aligned Reprojection. The updated 3D object points y'_t are projected onto the image plane of frame t using the estimated camera intrinsics K and extrinsics $P_t = [R_t|t_t]$:

$$\begin{bmatrix} u_{j,t} \\ v_{j,t} \\ 1 \end{bmatrix} \sim K \left(R_t y'_{j,t} + t_t \right) \tag{5}$$

By projecting and rasterizing all visible points, we obtain the object's silhouette for each frame. This reprojection naturally accounts for camera motion, parallax, and occlusion, producing geometrically consistent renderings from real camera viewpoints. The projected silhouettes are further refined using SAM 2 [20] to yield a temporally coherent binary mask sequence $\{M_t\}_{t=1}^{T}$, serving as the foundational spatial condition for synthesis.

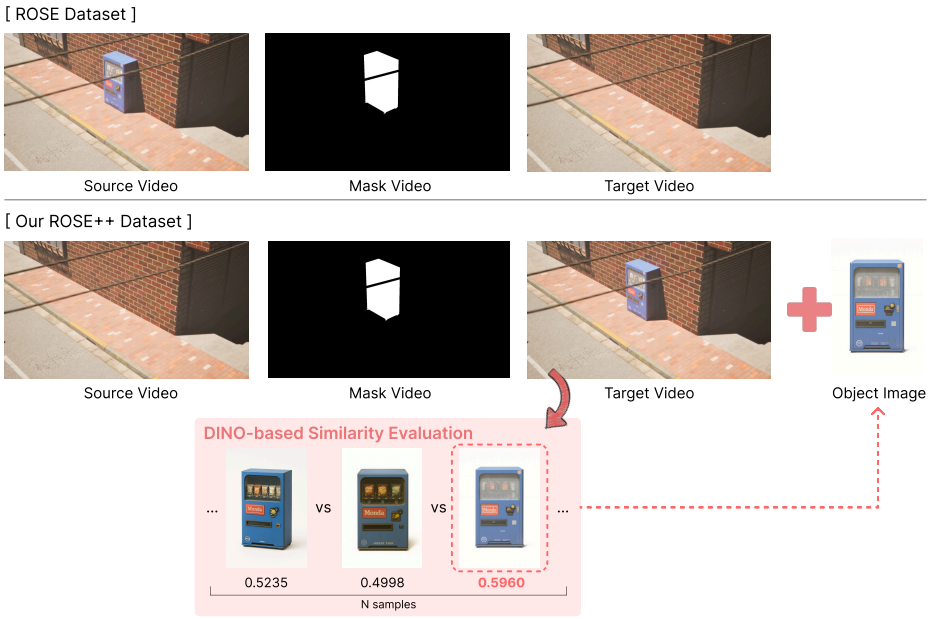

Fig. 3: ROSE++ dataset generation pipeline. Reference object images are synthesized using an image editing model [1] and selected through a DINO-guided rejection sampling pipeline to ensure high appearance fidelity.

3.3 Optics-Aware Video Synthesis

Large-scale mask-conditioned video generation models (e.g., VACE [8]) provide an efficient foundation for object insertion, as they are trained to strongly adhere to the given spatial condition. However, their localized nature strictly confines generation within the provided mask, leaving the surrounding region untouched. Consequently, they fail to synthesize scene-wide optical interactions, such as cast shadows and reflections, that naturally extend beyond the object's strict boundary.

ROSE++ Dataset Construction. To supervise the generation of these physical lighting effects, we construct ROSE++, a specialized dataset derived from the ROSE object removal benchmark [14]. We reorganize the dataset's structure by treating the object-removed video as the source $\mathcal{V}_{src}$ and the object-present video as the target $\mathcal{V}_{tgt}$. This inherently embeds insertion-induced illumination changes into the supervision signal, as shadows and lighting variations absent in the source safely appear in the target.

Since the original dataset lacks isolated reference images, we generate them using an image editing model [1] coupled with a rejection sampling pipeline. Multi-view object crops f_j are extracted from $\mathcal{V}_{tgt}$ using the ground-truth masks and provided to the editing model to synthesize candidate reference images on white backgrounds, with scene-dependent lighting cues removed. This ensures

that a model trained with ROSE++ infers illumination from the target scene rather than relying on copy-and-paste shortcuts from the reference image. To filter out candidate images that deviate from the true object's appearance, we employ a rejection sampling strategy. The generated candidates are ranked using a DINO-based similarity metric [16] against the cropped frames:

$$s_k = \frac{1}{N} \sum_{j=1}^{N} \text{sim}\left(\Phi_{\text{DINO}}(\hat{o}_k), \Phi_{\text{DINO}}(f_j)\right) \tag{6}$$

where $\hat{o}_k$ denotes the k-th generated candidate, $\Phi_{\text{DINO}}(\cdot)$ extracts the feature embeddings, and $\text{sim}(\cdot, \cdot)$ represents cosine similarity. The highest-scoring candidate is selected as the final reference image.

Optics-Aware Representation Alignment. To encourage the generative model to synthesize optical variations beyond the primary object mask, we propose the representation alignment technique. As illustrated in Fig. 4(a), we compute an **optics-aware extended mask** (M_{ext}) by extracting the pixel-wise difference between $\mathcal{V}_{src}$ and $\mathcal{V}_{tgt}$. After thresholding and morphological post-processing, this mask successfully captures both the object region and its associated photometric footprint (i.e., shadows and reflections), such that $M \subset M_{\text{ext}}$.

During training, both the mask and the extended-mask are tokenized and passed through a shared diffusion backbone (Fig. 4(b)). We extract intermediate features from multiple transformer blocks respectively and enforce alignment between the mask features (F_l) and the extended-mask features (F_l^{ext}):

$$\mathcal{L}_{align} = \sum_{l} ||F_l - \text{sg}(F_l^{\text{ext}})||_2^2 \tag{7}$$

Crucially, a stop-gradient operator (sg) is applied to the extended-mask branch, allowing it to act as a fixed, optics-aware teacher. This alignment forces the standard fine-mask input to anticipate and generate extended optical variations, even during inference when only the tightly bound fine mask is provided.

First-Frame Anchoring and LoRA Fine-Tuning. We fine-tune the pre-trained mask-conditioned video diffusion model [8] on ROSE++ using LoRA adapters [6], preserving its core generative prior while adapting it to optics-aware object insertion. To further maximize visual fidelity, we employ an optional but highly effective first-frame anchoring technique. Because image-level object insertion benefits from highly mature generative priors compared to the more complex VOI task, we first synthesize a high-quality edited image using a dedicated image insertion model [27]. We then provide this synthesized image to our video diffusion model as an explicit first-frame conditioning signal. This establishes highly reliable object appearance and local illumination cues, which are naturally propagated to subsequent frames, ensuring robust temporal consistency in color, texture, and lighting throughout the final synthesized video.

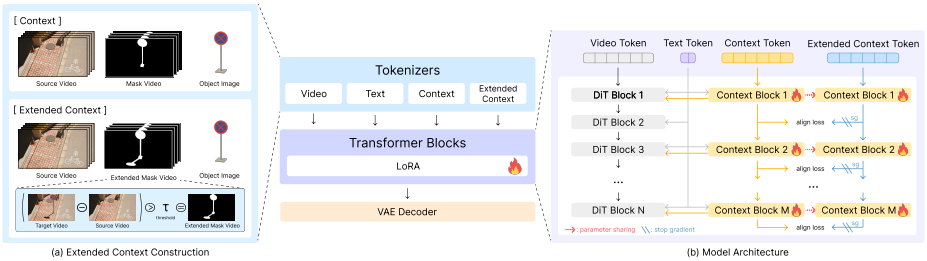

Fig. 4: (a) We extract an extended mask containing the object and its optical footprint. (b) By explicitly aligning the intermediate representations of the standard fine mask to those of the extended mask via an alignment loss, the model learns to synthesize physical lighting effects (e.g., cast shadows) that extend beyond the strict object mask boundary.

Type	Method	Subject Consistency		VBench			
		CLIP-I ↑	DINO-I ↑	Background Consistency ↑	Subject Consistency ↑	Motion Smoothness ↑	Imaging Quality ↑
Commercial	Pika-Pro [18]	0.4940	0.3856	0.9080	0.8720	0.9889	0.6546
	Kling [22]	0.6349	0.5028	0.9335	0.9494	**0.9940**	0.7069
Open Source	AnyV2V [11]	0.7033	0.2217	0.8884	0.8699	0.9795	0.5973
	ReVideo [15]	0.7385	0.3651	0.9391	0.9403	0.9906	0.6526
	Señorita [29]	0.7499	0.3982	0.9262	0.9266	0.9902	0.6333
	$\text{VACE}_{\text{our mask}}$ [8]	0.7368	0.5060	0.9011	0.8855	0.9887	0.6046
	Ours	**0.8132**	**0.5669**	**0.9503**	**0.9534**	0.9925	**0.7473**

Table 1: Quantitative comparisons with baseline methods.

4 Experiments

4.1 Experimental Setup

Evaluation Dataset. We introduce VOIBench, a new benchmark designed to rigorously assess video object insertion. VOIBench spans three axes that jointly stress geometric and photometric robustness: *scene* (indoor / outdoor / natural), *occlusion* (none / dynamic / initially-occluded), and *lighting* (daylight / indoor / low-light), giving the benchmark broad coverage and evaluative power. The dataset consists of 200 video clips covering a broad spectrum of settings, such as indoor scenes, outdoor environments, and natural landscapes, with each video containing a pair of objects. We crawled contextually relevant objects for each scene.

Baselines and Evaluation Metrics. We evaluate against top commercial tools (Pika-Pro [18], Kling [22]) and representative open-source models (AnyV2V [11], ReVideo [15], Señorita [29]), as well as a VACE-based variant ($\text{VACE}_{\text{our mask}}$), conditioned on our generated mask, to ensure fair VOI evaluation. We also report ablation results to isolate the impact of our proposed components.

We assess three key aspects: (1) *Subject Consistency* via CLIP-I [19] and DINO-I [16]; (2) *Video Quality* using VBench [7] (Image Quality, Background/

Fig. 5: Qualitative comparisons with baseline methods. Best viewed when zoomed in.

Subject Consistency, Motion Smoothness); and (3) *Multi-View Consistency* [7] to gauge how reliably the object is preserved across viewpoint changes and dynamic occlusions.

4.2 Qualitative Results

Fig. 5 qualitatively highlights two key factors: 1) a 4D-aware mask for occlusion-robust insertion and 2) Optics-Aware Representation Alignment for realistic illumination effects. Without a 4D-aware mask, Kling [22] and Señorita [29] fail to maintain a stable separation between the insertion region and the rest of the scene, leading to background drift and deformation when a dynamic occluder (the moving tube) crosses the target area. In particular, even with a detailed prompt specifying the insertion location ("Add a paper boat to the right rear of the river in the video background."), Kling often misinterprets the scene and transforms the tube into a paper boat. Señorita, which propagates edits from an edited first frame, degrades over time and cannot sustain consistent insertions under occlusions. In contrast, methods equipped with a 4D-aware mask (VACE [8] and ours) preserve both the tube geometry and the original background, producing consistent insertions. However, VACE lacks optics-aware modeling, making the inserted boat appear pasted with implausible reflections and weak cast shadows.

4.3 Quantitative Results

To assess subject consistency, we uniformly sample 10 frames from each generated video and measure how closely they match the reference subject using

Method	Subject Consistency		VBench			
	CLIP-I ↑	DINO-I ↑	Background Consistency ↑	Subject Consistency ↑	Motion Smoothness ↑	Imaging Quality ↑
VACE $_{-,\text{ our 4D mask}}$ [8]	0.7368	0.5060	0.9011	0.8855	0.9887	0.6046
VACE$_{\text{ROSE++, bbox mask}}$ [8]	0.6383	0.4856	0.9026	0.8818	0.9864	0.5824
VACE$_{\text{ROSE++, our 4D mask}}$ [8]	0.7485	0.5103	0.9035	0.8895	0.9888	0.6583
Ours	**0.8132**	**0.5669**	**0.9503**	**0.9534**	**0.9925**	**0.7473**

Table 2: Fair comparison against VACE fine-tuned on ROSE++ with our 4D-aware mask and with a coarse bounding-box mask, evaluated on the VOIBench. Both fine-tuned VACE variants underperform our full model.

subject-masked regions only. As shown in Table 1, InsertAnywhere achieves the highest CLIP-I [19] and DINO-I [16] scores among all compared methods, demonstrating superior identity preservation. Moreover, InsertAnywhere attains the best overall VBench score, with clear advantages in Background Consistency, Subject Consistency, and Imaging Quality. Overall, the results suggest that our model not only provides a reliable 4D-aware mask for subject control but also improves full-video quality without sacrificing visual fidelity. In contrast, Kling and Pika-Pro often introduce undesirable changes to the original background during insertion. To verify that our gains stem from our design rather than from data exposure, we additionally fine-tune VACE on ROSE++ with our 4D-aware mask and with a coarse bounding-box mask (Table 2). Both variants underperform our full model, confirming that the improvement comes from our optics-aware design rather than merely from training on ROSE++. Notably, our method is best on five of the six metrics; the only exception is Motion Smoothness, where we closely match Kling (0.9925 vs. 0.9940), a gap within VBench's per-clip noise.

4.4 Ablation Study

We evaluate the effectiveness of our proposed components through a progressive ablation study, with qualitative and quantitative results detailed in Fig. 6 and Tab. 3. First, introducing our 4D-aware mask sequence successfully resolves dynamic occlusion failures (e.g., orange box, row 2), though basic object fidelity remains low. Next, incorporating first-frame anchoring significantly enhances identity preservation, yet temporal inconsistencies in shape and color still emerge post-occlusion. While subsequent LoRA fine-tuning on the ROSE++ dataset successfully stabilizes temporal consistency and appearance, the network still struggles to hallucinate external optical interactions. Finally, integrating our Optics-Aware Representation Alignment ($\mathcal{L}_{align}$) naturally synthesizes insertion-induced lighting effects, such as shadows. Combining all components yields geometrically accurate and photometrically realistic insertions with high object fidelity across the entire sequence.

Effect of Alignment Loss on Optics-Aware Rendering. Table 4 (right) reports quantitative comparisons for rendering quality using LPIPS/PSNR/SSIM against the ground truth. To enable a paired evaluation with ground truth,

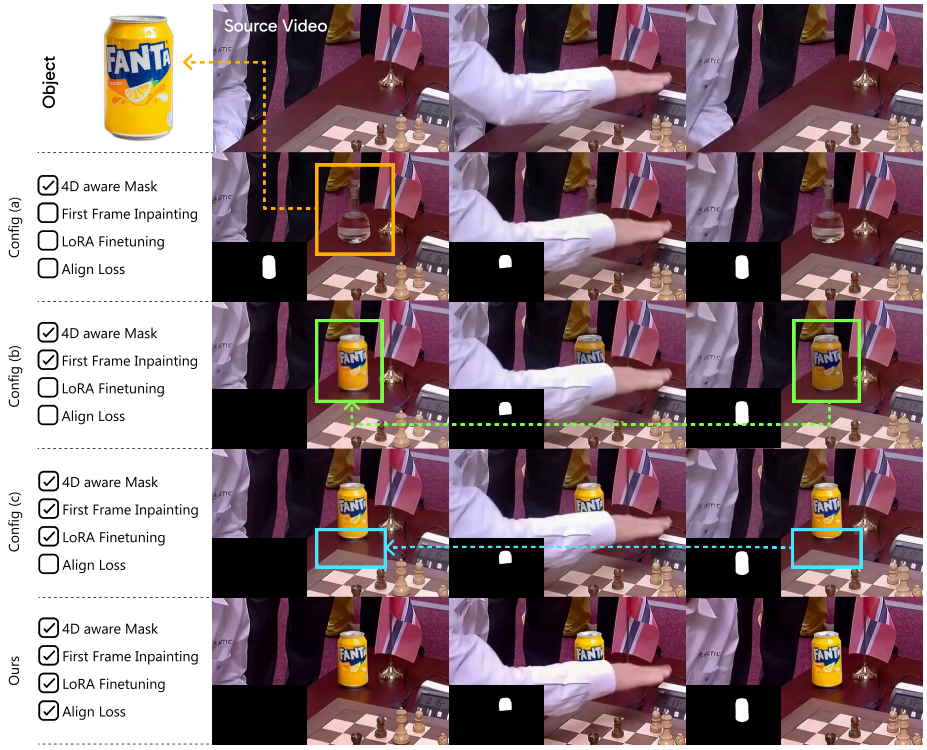

Fig. 6: Qualitative ablation study. Sequentially adding our proposed components progressively improves occlusion handling, object fidelity, and realistic optics-aware generation. Best viewed zoomed in.

Method	Subject Consistency		VBench				
	CLIP-I ↑	DINO-I ↑	Background Consistency ↑	Subject Consistency ↑	Motion Smoothness ↑	Imaging Quality ↑	Multi-View Consistency ↑
Config (a)	0.7532	0.3861	0.9232	0.9380	0.9913	0.6298	0.5308
Config (b)	0.7880	0.5135	0.9175	0.9290	0.9911	0.6318	0.5436
Config (c)	0.8122	**0.5678**	0.9429	0.9520	0.9916	0.7101	0.5857
Ours	**0.8132**	0.5669	**0.9503**	**0.9534**	**0.9925**	**0.7473**	**0.5865**

Table 3: Quantitative results of the ablation study.

we hold out 25 samples from ROSE++ as a test set and compute all metrics on this split. We compare VACE, our ROSE++ LoRA fine-tuned model without the alignment loss (Ours $_{\text{w/o}\mathcal{L}_{\text{align}}}$), and the full model trained with the proposed alignment loss (Ours). We evaluate results under two complementary regions: Object region covers only the precise foreground area of the inserted object, while Optics region isolates the area affected by the object-induced optical effects such as shadows, specular highlights, and contact-dependent illumination changes. Using Object Mask, the two Ours variants achieve similar

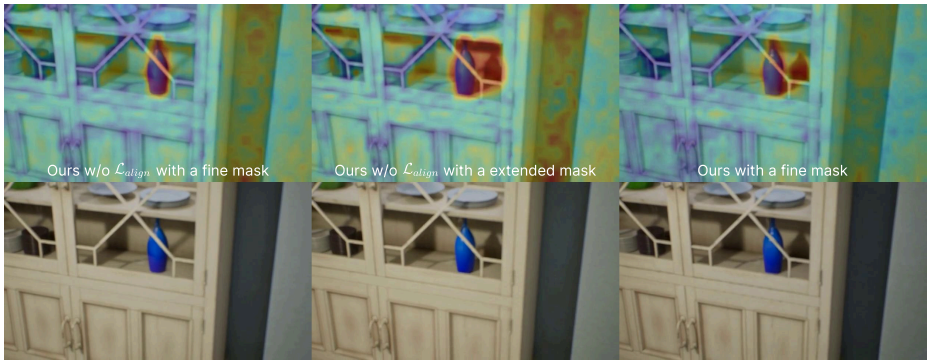

Fig. 7: Impact of Optics-Aware Representation Alignment. Attention maps for (a) a baseline model (w/o $\mathcal{L}_{align}$) with a fine mask, (b) a baseline model with an extended mask, and (c) our model with a fine mask. Despite training on the same ROSE++ dataset, the baseline (a) remains strictly confined to the inserted object. In contrast, our aligned model (c) exhibits strong external activations closely matching (b). This confirms that our alignment strategy effectively trains the network to synthesize surrounding optical effects (e.g., shadows) without requiring an extended input mask.

performance, indicating that the object appearance fidelity is largely insensitive to $\mathcal{L}_{align}$. In contrast, under Optics Mask, Ours yields clear gains with higher PSNR/SSIM and lower LPIPS, demonstrating that $\mathcal{L}_{align}$ effectively improves the preservation and consistency of optics-aware effects. We further verify that $\mathcal{L}_{align}$ does not degrade fidelity in the unedited region: across the held-out ROSE++ split, the unedited-region PSNR is nearly identical with and without $\mathcal{L}_{align}$ ($\text{VACE}_{\text{our mask}}$: 33.27, $\text{Ours}_{\text{w/o } \mathcal{L}_{align}}$: 33.31, Ours: 33.30), while the optics-region PSNR rises by +7.3 dB. The mild color drift occasionally observed thus reflects sample-to-sample variation rather than a systematic effect of $\mathcal{L}_{align}$. Qualitative results in Fig. 8 corroborate this trend: in both the desk-top mouse and the floating tube examples, VACE shows noticeable appearance mismatch to the reference and reduced overall image quality, whereas both Ours variants maintain comparable object quality, with Ours producing more realistic and coherent specular highlights and shadows than $\text{Ours}_{\text{w/o } \mathcal{L}_{\text{align}}}$.

Fig. 7 visualizes the network's attention maps during generation. We compare three configurations: (a) a baseline model trained without $\mathcal{L}_{align}$ conditioned on a fine mask, (b) baseline model conditioned on an extended mask, and (c) our full model conditioned on a fine mask. Despite being fine-tuned on the same ROSE++ dataset, the attention of the baseline model (a) remains strictly confined to the inserted object's boundaries. In contrast, when provided only the tightly-bound fine mask, our aligned model (c) exhibits strong external activations that closely mirror the extended mask conditions of (b). This visual evidence confirms that our alignment strategy effectively teaches the network to anticipate and synthesize surrounding optical effects, such as cast shadows, without requiring an explicitly expanded mask during inference.

Fig. 8: Qualitative ablation for Optics-Aware Representation Alignment.

Methods	Multi-View Consistency ↑
$\text{Ours}_{\text{random}}$	0.5295
Ours	**0.5865**

Methods	Object Mask			Optics Mask		
	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑
$\text{VACE}_{\text{our mask}}$ [8]	0.0250	21.1980	0.9795	0.0130	19.8345	0.9877
Ours $_{\text{w/o } \mathcal{L}_{\text{align}}}$	**0.0137**	**26.0282**	0.9843	0.0121	19.9065	0.9883
Ours	0.0161	25.8785	**0.9872**	**0.0089**	**27.2362**	**0.9920**

Table 4: Left: Multi-view Consistency evaluation. Models trained with random-frame references vs. our ROSE++ reference. Right: Quantitative evaluation of shadow rendering quality on object and optics masks, comparing VACE, $\text{Ours}_{\text{w/o } \mathcal{L}_{\text{align}}}$, and Ours.

Effectiveness of ROSE++ We evaluate whether the generated reference images in ROSE++ successfully prevent copy-and-paste artifacts. To do so, we compare our model against a baseline trained with reference objects trivially cropped from the target video. As reported in Table 4 (left), our method achieves substantially higher multi-view consistency. This confirms that our dataset construction forces the network to genuinely infer 4D geometry and lighting, rather than relying on direct image copying.

5 Conclusion

We introduced **InsertAnywhere**, a novel framework for production-grade Video Object Insertion (VOI). By combining Geometrically Grounded Mask Propagation with an optics-aware diffusion model, our approach simultaneously guarantees precise spatial control and photorealistic rendering. From a single 3D anchor, our method automatically handles complex camera motions and dynamic occlusions. Furthermore, by leveraging our custom ROSE++ dataset and Optics-Aware Representation Alignment, the network successfully synthesizes object-induced physical lighting effects, such as shadows, outside the immediate

Fig. 9: Effectiveness of scene flow-based object propagation in Section 3.2. Unlike the static mask (second row), our proposed propagation ensures temporally consistent and 4D-aware alignment with moving objects (third row).

mask boundary. Extensive evaluations confirm that InsertAnywhere significantly outperforms existing baselines in both geometric consistency and photometric realism, paving the way for advanced practical applications in virtual product placement and visual effects.

Limitations. Even when the propagated mask is not perfectly pixel-accurate, our pipeline re-synthesizes regions beyond the boundary and absorbs small errors into a plausible output. Nonetheless, complex physics-aware interactions (*e.g.*, dropping a heavy object onto a soft sofa) lie outside our current scope and are left for future work.

Acknowledgement

This work was supported by and carried out in close collaboration with SK Telecom (SKT). This research was supported by Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports, and Tourism (MCST) in 2024 (Project Name: Development of Technology for Convergence Performance Planning and Production Platform to Revitalize the Production of Convergence Performance by Traditional Artist Dance Music, Project Number: RS-2024-00398536, Contribution Rate: 30%). This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)). This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263841, Development of a Real-time Multimodal Framework for Comprehensive Deepfake Detection Incorporating Common Sense Error Analysis).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, C., Shao, Z., Zhang, G., Liang, D., Yang, J., Zhang, Z., Guo, Y., Zhong, C., Qiu, Y., Wang, Z., et al.: Anything in any scene: Photorealistic video object insertion. arXiv preprint arXiv:2401.17509 (2024)
3. Chen, J., Li, X., Bai, X., Ma, T., Zhang, P., Chen, Z., Li, G., Liu, L., Zhao, S., Li, B., et al.: Omniinsert: Mask-free video insertion of any reference via diffusion transformer models. arXiv preprint arXiv:2509.17627 (2025)
4. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6593–6602 (2024)
5. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1316–1326 (2023)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
7. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., et al.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. arXiv preprint arXiv:2411.13503 (2024)
8. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598 (2025)
9. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6013–6022 (2025)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
11. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. Transactions on Machine Learning Research (2024)
12. Liu, S., Wang, T., Wang, J.H., Liu, Q., Zhang, Z., Lee, J.Y., Li, Y., Yu, B., Lin, Z., Kim, S.Y., et al.: Generative video propagation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17712–17722 (2025)
13. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
14. Miao, C., Feng, Y., Zeng, J., Gao, Z., Liu, H., Yan, Y., Qi, D., Chen, X., Wang, B., Zhao, H.: Rose: Remove objects with side effects in videos. arXiv preprint arXiv:2508.18633 (2025)
15. Mou, C., Cao, M., Wang, X., Zhang, Z., Shan, Y., Zhang, J.: Revideo: Remake a video with motion and content control. Advances in Neural Information Processing Systems **37**, 18481–18505 (2024)
16. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

17. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. arXiv preprint arXiv:2502.20110 (2025)
18. Pika: Pika additions. https://pika.art/pikadditions (2025), accessed: 2025-11-14
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
20. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
21. Saini, N., Bodla, N., Shrivastava, A., Ravichandran, A., Zhang, X., Shrivastava, A., Singh, B.: Invi: Object insertion in videos using off-the-shelf diffusion models. arXiv preprint arXiv:2407.10958 (2024)
22. Team, K.A.: Klingai (2025), https://app.klingai.com/global/multimodal-to-video/add-object/new
23. Tu, Y., Luo, H., Chen, X., Ji, S., Bai, X., Zhao, H.: Videoanydoor: High-fidelity video object insertion with precise motion control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
24. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: European Conference on Computer Vision. pp. 36–54. Springer (2024)
25. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
26. Yao, D.Y., Zhai, A.J., Wang, S.: Uni4d: Unifying visual foundation models for 4d modeling from a single video. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1116–1126 (2025)
27. Yu, Y., Zeng, Z., Zheng, H., Luo, J.: Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. arXiv preprint arXiv:2503.08677 (2025)
28. Zhao, Q., Ma, Z., Zhou, P.: Dreaminsert: Zero-shot image-to-video object insertion from a single image. arXiv preprint arXiv:2503.10342 (2025)
29. Zi, B., Ruan, P., Chen, M., Qi, X., Hao, S., Zhao, S., Huang, Y., Liang, B., Xiao, R., Wong, K.F.: Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. In: NeurIPS D&B (2025)