AGILE: Hand-object Interaction Reconstruction from Video via Agentic Generation

JIN-CHUAN SHI*, Zhejiang University, China
BINHONG YE*, Zhejiang University, China
TAO LIU, Zhejiang University, China
XIAOYANG LIU, Zhejiang University, China
YANGJINHUI XU, Zhejiang University, China
JUNZHE HE, Zhejiang University, China
ZEJU LI, Zhejiang University, China
HAO CHEN†, Zhejiang University, China
CHUNHUA SHEN, Zhejiang University, China and Zhejiang University of Technology, China



Fig. 1. **High-Fidelity Hand-Object Reconstruction from Video.** We present AGILE, a framework that reconstructs simulation-ready interaction sequences from monocular video. By leveraging agentic generative priors, AGILE robustly recovers watertight geometry, realistic textures, and precise 6D poses for diverse objects, ranging from thin structures (scissors, pen) to complex topologies (game controller), even under severe hand occlusion and rapid manipulation.

Reconstructing dynamic hand-object interactions from monocular videos is critical for dexterous manipulation data collection and creating realistic digital twins for robotics and VR. However, current methods face two prohibitive barriers: (1) reliance on neural rendering often yields fragmented, non-simulation-ready geometries under heavy occlusion, and (2) dependence on brittle Structure-from-Motion (SfM) initialization leads to frequent failures on in-the-wild footage. To overcome these limitations, we introduce AGILE, a robust framework that shifts the paradigm from *reconstruction* to *agentic generation* for interaction learning. First, we employ an agentic pipeline where a Vision-Language Model (VLM) guides a generative model to synthesize a complete, watertight object mesh with high-fidelity texture, independent of video occlusions. Second, bypassing fragile SfM entirely, we propose a robust *anchor-and-track* strategy. We initialize the object pose at a single interaction onset frame using a foundation model and propagate it temporally by leveraging the strong visual similarity between our generated asset and video observations. Finally, a contact-aware optimization integrates semantic, geometric, and interaction stability constraints to enforce physical plausibility. Extensive experiments on HO3D, DexYCB, ARCTIC,

*Equal contribution.
†Corresponding author.

Authors' Contact Information: Jin-Chuan Shi, Zhejiang University, Hangzhou, China, jinchuanshi@zju.edu.cn; Binhong Ye, Zhejiang University, Hangzhou, China, 22421273@zju.edu.cn; Tao Liu, Zhejiang University, Hangzhou, China, 3220103422@zju.edu.cn; Xiaoyang Liu, Zhejiang University, Hangzhou, China, liuxiaoyang@zju.edu.cn; Yangjinhui Xu, Zhejiang University, Hangzhou, China, 3230105918@zju.edu.cn; Junzhe He, Zhejiang University, Hangzhou, China, hejunzhe@zju.edu.cn; Zeju Li, Zhejiang University, Hangzhou, China, zejuli@zju.edu.cn; Hao Chen, Zhejiang University, Hangzhou, China, haochen.cad@zju.edu.cn; Chunhua Shen, Zhejiang University, Hangzhou, China and Zhejiang University of Technology, Hangzhou, China, chunhuashen@zju.edu.cn.

This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGGRAPH Conference Papers '26, Los Angeles, CA, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2554-8/26/07
https://doi.org/10.1145/3799902.3811134

and in-the-wild videos reveal that AGILE outperforms baselines in global geometric accuracy while demonstrating exceptional robustness on challenging sequences where prior arts frequently collapse. By prioritizing physical validity, our method produces simulation-ready assets validated via real-to-sim retargeting for robotic applications. Project page: agile-hoi.github.io.

CCS Concepts: • **Computing methodologies → Reconstruction**; *Mesh geometry models.*

ACM Reference Format:
Jin-Chuan Shi, Binhong Ye, Tao Liu, Xiaoyang Liu, Yangjinhui Xu, Junzhe He, Zeju Li, Hao Chen, and Chunhua Shen. 2026. AGILE: Hand-object Interaction Reconstruction from Video via Agentic Generation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3799902.3811134

1 Introduction

Reconstructing dynamic hand-object interactions (HOI) from monocular video is a pivotal challenge in computer vision and graphics. The ultimate aspiration is to create high-fidelity digital twins of these interactions, assets that are not only visually coherent but also geometrically explicit and physically plausible for downstream tasks such as robotic imitation learning and physics-based simulation. While the internet offers a vast repository of in-the-wild interaction videos, converting this data into simulation-ready 3D assets remains an unresolved challenge. We identify two fundamental barriers in existing pipelines that hinder scalability and reliability.

The first barrier lies in the inherent limitations of *reconstruction-based geometry*. State-of-the-art methods typically rely on neural rendering techniques, such as NeRF [Mildenhall et al. 2021] or 3D Gaussian Splatting [Kerbl et al. 2023], to optimize scene representations [Fan et al. 2024b; On et al. 2025]. However, these methods fundamentally depend on multi-view consistency. In HOI scenarios, severe hand occlusion frequently violates this premise, leading to fragmented, noisy, or non-watertight geometries. Such representations, while visually passable from trained viewpoints, are ill-suited for physics engines that require clean, explicit topology.

The second, and more critical, bottleneck is the *brittleness of pose initialization*. Most prevailing approaches [Fan et al. 2024b; Wang et al. 2025a] are anchored on Structure-from-Motion (SfM) pipelines like COLMAP [Schönberger and Frahm 2016] to estimate initial camera and object poses. SfM is notoriously fragile in dynamic scenes featuring textureless objects, rapid motions, or significant occlusions. When SfM fails to register frames, a common occurrence in real-world footage, the entire pipeline collapses. This reliance on classic geometric matching prevents current methods from achieving reliable performance on complex benchmarks [Chao et al. 2021].

Recent generative advancements offer potential but lack robustness for high-fidelity HOI. MagicHOI [Wang et al. 2025a] leverages diffusion priors but produces over-smoothed meshes and retains a dependency on brittle SfM. Single-view methods like SAM3D [Team et al. 2025] suffer from occlusion-induced information loss, yielding coarse geometries and poor textures that destabilize subsequent pose optimization. To overcome these barriers, it is essential to aggregate multi-view cues from the video. However, since standard 3D generators require canonical cameras unavailable in dynamic footage, we must resort to 2D generative priors to synthesize missing perspectives. Yet, the stochastic nature of 2D diffusion models frequently introduces hallucinations inconsistent with the actual object in the video. Consequently, reliably harnessing these priors requires an intelligent mechanism capable of identifying informative frames and strictly filtering generated views for consistency—a capability missing in current end-to-end pipelines.

To bridge this gap, we introduce AGILE, a robust framework that shifts the paradigm from *reconstruction* to *agentic generation* for interaction learning. First, we employ an agentic pipeline where a Vision-Language Model (VLM) acts as an intelligent supervisor. The VLM selects informative keyframes to guide multi-view synthesis and rigorously filters the outputs via rejection sampling to ensure consistency. Crucially, the pipeline incorporates a *texture refinement* phase that recovers high-frequency details from the video observations. This yields a watertight mesh with hyper-realistic appearance—a prerequisite that enables our subsequent foundation models to reliably initialize pose and optimize semantic alignment.

Leveraging this high-quality asset, we propose a robust *anchor-and-track* optimization strategy that eliminates the need for brittle SfM. We utilize a foundation model to initialize the object pose only at a single interaction onset frame. For the remainder of the video, we perform online optimization by propagating the pose temporally. By capitalizing on the strong visual similarity between our generated textured mesh and the video observations, we drive this alignment using semantic feature loss and mask constraints, while simultaneously enforcing physical contact consistency. This approach proves significantly more stable than relying on noisy per-frame predictions or fragile SfM.

Comprehensive evaluations across benchmarks (HO3D, DexYCB) and diverse in-the-wild videos demonstrate the transformative impact of our approach. AGILE achieves state-of-the-art performance across all datasets, establishing a new standard for both accuracy and robustness. While prior methods suffer from high failure rates (up to 75%) on complex sequences, AGILE maintains a 100% success rate. Our analysis reveals a critical trade-off: while some prior methods optimize for local proximity—often allowing physical interpenetration—our approach prioritizes global geometric accuracy and non-penetration. Consequently, AGILE delivers digital twins that are validated via a real-to-sim retargeting pipeline, successfully driving dynamic interactions in a physics simulator.

Our core contribution is a paradigm shift from reconstruction to *agentic generation* for HOI, realized through the following designs. First, and central to our novelty, we propose the first agentic HOI pipeline that integrates VLM-guided quality assessment with generative models, enabling the production of high-fidelity, watertight meshes independent of video occlusions. Second, we introduce a robust anchor-and-track optimization strategy that eliminates the dependency on brittle SfM by anchoring pose initialization at a single contact frame and propagating it via semantic and geometric alignment. Third, extensive experiments across single-hand benchmarks, bimanual interactions, and in-the-wild sequences, together with real-to-sim retargeting on a dexterous robotic hand, validate that AGILE achieves state-of-the-art geometric accuracy, exceptional robustness, and produces simulation-ready digital twins.

2 Related Work

3D Hand Reconstruction. Monocular 3D hand reconstruction has evolved from fitting parametric models like MANO [Romero et al. 2022] to images [Boukhayma et al. 2019; Pavlakos et al. 2024; Zhang et al. 2019; Zhou et al. 2020], towards recovering dynamic motion from video. While recent temporal approaches [Yu et al. 2025; Zhang et al. 2025] regress 4D trajectories, they often sacrifice the per-frame geometric precision essential for contact analysis. In our framework, we prioritize geometric fidelity to ensure accurate interaction modeling. Consequently, we leverage state-of-the-art estimators like WiLoR [Potamias et al. 2025] to provide robust, high-fidelity initial hand meshes.

Generative 3D Object Reconstruction. The field is shifting from multi-view optimization (e.g., NeRF [Mildenhall et al. 2021], Gaussian Splatting [Kerbl et al. 2023])—which requires dense, unoccluded coverage rarely found in interaction videos—to generative inference. Large Reconstruction Models (LRMs) [Hong et al. 2023; Tang et al. 2024; Team 2025; Xu et al. 2024] and diffusion-based methods [Liu et al. 2023; Long et al. 2024] now generate assets from sparse inputs. However, single-view methods like SAM3D [Team et al. 2025], while capable of end-to-end inference, inherently lack the capacity to aggregate multi-view cues from video. This limitation prevents them from resolving severe occlusions or maintaining texture consistency across views. We address this by proposing an agentic framework that intelligently fuses video evidence with generative priors, synthesizing watertight, textured assets that strictly align with video observations.

Hand-Object Interaction (HOI) Reconstruction. Reconstructing dynamic HOI is complicated by severe mutual occlusion. Early template-based [Corona et al. 2020; Fan et al. 2024a; Tekin et al. 2019; Yang et al. 2021] or depth-dependent methods limit in-the-wild applicability. *Reconstruction-based methods* like HOLD [Fan et al. 2024b] and BIGS [On et al. 2025] optimize implicit representations but yield fragmented, non-physical geometries under occlusion. Furthermore, their reliance on brittle Structure-from-Motion (SfM) initialization (e.g., COLMAP [Schönberger and Frahm 2016]) leads to frequent failures on textureless or moving objects. *Generative approaches* attempt to hallucinate missing regions; MagicHOI [Wang et al. 2025a] integrates diffusion priors with NeRF but relies on Score Distillation Sampling (SDS), often producing over-smoothed meshes lacking simulation-ready detail. Crucially, it retains the fragility of SfM initialization. In contrast, AGILE eliminates the SfM bottleneck via a robust anchor-and-track strategy, delivering high-fidelity, simulation-ready digital twins.

3 Method

Given a monocular video of a hand-object interaction captured by a fixed camera, our goal is to reconstruct the 4D trajectory of both the hand and the object, yielding high-fidelity, simulation-ready 3D assets, as illustrated in Figure 3. Unlike previous methods [Fan et al. 2024b; On et al. 2025] that rely on neural rendering and fragile Structure-from-Motion (SfM) initialization, we propose AGILE, a generative framework that shifts the paradigm from *reconstruction* to *agentic generation* for hand-object interaction reconstruction.

3.1 Agentic Textured Object Generation

A primary challenge in reconstructing hand-object interactions from monocular video is the severe occlusion of the object by the user's hand. While neural rendering [Kerbl et al. 2023; Mildenhall et al. 2021] excels at view synthesis, it struggles to recover complete geometry from such partially visible data. Similarly, single-view generative methods like SAM3D [Team et al. 2025] inherently suffer from information loss, yielding coarse geometries and low-fidelity textures that destabilize subsequent pose optimization. To overcome these limitations, it is essential to aggregate multi-view cues from the video using 2D generative priors to synthesize missing perspectives. However, the stochastic nature of diffusion models frequently introduces hallucinations that deviate from the actual object in the video. Reliably harnessing these priors therefore requires an intelligent mechanism capable of identifying the most informative reference frames and strictly filtering generated views for consistency—a capability missing in standard pipelines. To address this, we propose an agentic framework where a VLM supervisor acts as this intelligent critic, bridging the gap between noisy generative priors and rigorous video evidence.

VLM-Guided Multi-View Synthesis. As illustrated in Figure 2, our pipeline operates through a cascade of generative models steered by a VLM agent. Instead of relying on a single frame, the VLM first selects N informative keyframes (typically 1–4) from the input video to maximize viewpoint coverage. These frames prompt an image generation model to synthesize orthogonal views of the object. To ensure these hallucinations faithfully reflect the video content, we introduce a VLM-based critic. The VLM evaluates the consistency between generated views and original video frames, scoring them on geometry, texture, and material correspondence. Through a rejection sampling process, candidates falling below a strict consistency threshold are discarded and re-generated. This ensures that the input for 3D lifting is not only multi-view consistent but also faithful to the real-world observation. Implementation details are provided in the Supplementary Material.

3D Lifting and Mesh Refinement. The validated multi-view images are processed by a feed-forward 3D generation model [Team 2025] to produce an initial mesh. While geometrically plausible, raw outputs often suffer from irregular topology and blurry textures. To address this, we first apply automated retopology and UV unwrapping to create a clean, lightweight mesh optimized for physics simulation. Subsequently, we perform an *agentic texture refinement* step. The initial texture is enhanced using an image-to-image editing model conditioned on the evaluated high-resolution multi-view images to recover details.Crucially, as shown in Figure 2, this process is also supervised by the VLM agent. The VLM critic evaluates the refined texture against the multi-views via rejection sampling, ensuring strict visual fidelity and discarding any hallucinated artifacts. This rigorous quality control is motivated by two downstream requirements: (1) Foundation models used for pose initialization (e.g., FoundationPose [Wen et al. 2024]) rely heavily on sharp textual features for robust matching; and (2) our subsequent optimization

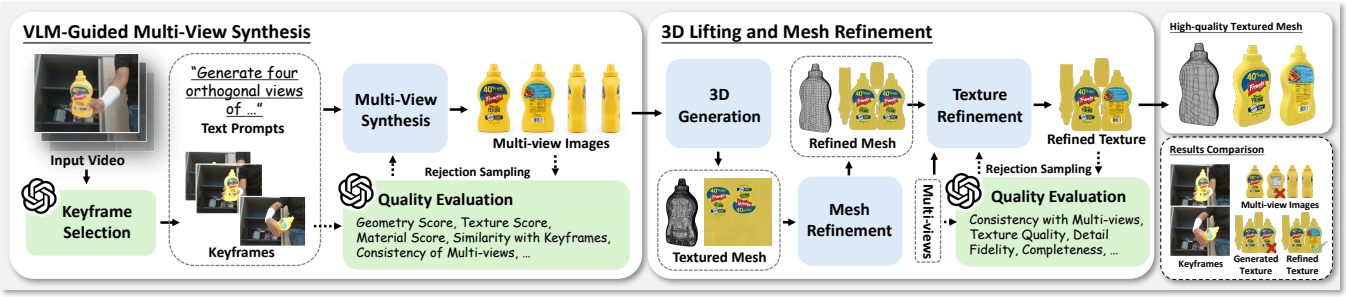


Fig. 2. **Pipeline for Agentic Textured Object Generation.** A VLM agent first selects informative keyframes from the input video to guide multi-view synthesis. To ensure consistency, a VLM-based critic filters the generated views via rejection sampling. The validated images are then lifted to 3D, followed by automated topology optimization and texture refinement. As highlighted in the bottom-right comparison, this refinement step significantly enhances texture fidelity against the evaluated multi-views, yielding a high-quality, simulation-ready asset.

utilizes DINO features, which demand high visual fidelity to maintain accuracy. This refinement yields a simulation-ready asset with photo-realistic appearance and explicit topology.

Our agentic framework offers distinct advantages over traditional baselines. First, it enables fully automated asset acquisition from arbitrary videos, effectively bypassing the failure modes of photogrammetry under heavy occlusion. Second, by continually aggregating temporal cues via VLM supervision, we ensure exceptional visual consistency between the generated asset and the video. Finally, unlike implicit representations, our approach directly produces clean, watertight meshes with high-frequency textures, providing a robust initialization that is critical for the stability of subsequent interaction tracking and physics simulation.

3.2 Initialization of Pose and Scale

To enable robust tracking without relying on brittle Structure-from-Motion pipelines, we establish a consistent metric initialization for both the hand and object. First, we extract essential scene information, including segmentation masks, camera intrinsics, and estimated metric depth. Second, we recover the metric geometry of the hand, comprising its 6D pose, global scale, and 2D joint projections. Finally, we estimate the object's global scale across the sequence and initialize its pose at the interaction onset frame.

Data Preprocessing. For each input video frame, we extract a set of key components. Specifically, we employ the monocular metric depth estimator MoGe-2 [Wang et al. 2025b] to obtain the camera intrinsic matrix $\mathbf{K}$ and a per-frame metric depth map $\mathbf{D} \in \mathbb{R}^{H\times W}$. Simultaneously, we use SAM2 [Ravi et al. 2024] to acquire precise segmentation masks for both the hand ($\mathbf{M}_h$) and the object ($\mathbf{M}_o$). Given our static camera assumption, the camera coordinate system serves as the world frame, with extrinsic parameters fixed as the identity matrix. These preprocessed priors form the foundational inputs for all subsequent stages.

Metric Hand Initialization. We utilize the preprocessed data to initialize a physically plausible hand model. We first employ the off-the-shelf estimator WiLoR [Potamias et al. 2025] to predict the MANO [Romero et al. 2022] parameters for each frame, including shape $\boldsymbol{\beta}$, pose $\boldsymbol{\theta}$, global rotation $\mathbf{R}_h \in SO(3)$, and the corresponding 2D keypoint annotations $\mathbf{J}_{2D} \in \mathbb{R}^{21\times 2}$. Since WiLoR provides reliable rotation estimates, we fix $\mathbf{R}_h$ and focus on recovering the missing metric scale and translation. To determine the global scale $s_h \in \mathbb{R}^+$, we unproject the masked hand pixels into 3D point clouds using the metric depth map $\mathbf{D}$. We align the MANO mesh to these depth observations using a constrained Iterative Closest Point (ICP) algorithm [Chetverikov et al. 2002]. In this step, we hold $\mathbf{R}_h$ fixed and optimize for s_h and a temporary translation to minimize the model-to-scan distance. Finally, to ensure accurate image alignment, we fix s_h, $\mathbf{R}_h$, and the intrinsics $\mathbf{K}$, and solve for the per-frame translation $\mathbf{T}_h \in \mathbb{R}^3$ via a Perspective-n-Point (PnP) approach, utilizing the 3D model joints and the predicted 2D keypoints $\mathbf{J}_{2D}$.

Object Pose and Scale Estimation. Once the canonical object mesh is created (Sec. 3.1), we align it to the video observations using a scale-first strategy. We first estimate the object's global metric scale s_o by applying the same constrained ICP algorithm used for the hand, aligning the generated mesh with the unprojected object point clouds across all available frames. Next, we identify the *interaction onset frame* (IOF), defined as the moment the object mask $\mathbf{M}_o$ exhibits significant displacement. Finally, to determine the initial object pose $[\mathbf{R}_o^{IOF}, \mathbf{T}_o^{IOF}]$, we apply FoundationPose [Wen et al. 2024] at the IOF. This estimator takes the RGB frame, the metric depth map, and the generated mesh pre-scaled by s_o as input, ensuring the estimated pose is metrically consistent with the scene.

3.3 Hand-Object Interaction Optimization

To ensure pixel-level alignment and physical consistency, we employ a bi-directional online optimization strategy anchored at the *interaction onset frame (IOF)* (identified in Sec. 3.2). Starting from the IOF, we propagate the optimization frame-by-frame towards both the start and end of the video sequence. At the IOF, we jointly optimize the hand and object poses along with the object's anisotropic scale $\mathbf{s}_o$ to establish a reliable canonical geometry. For all subsequent frames, $\mathbf{s}_o$ is frozen, and we perform the following two-step optimization per frame:

Step 1: Hand Translation Refinement. For the current frame, we first refine the hand trajectory. Relying on the robust orientation

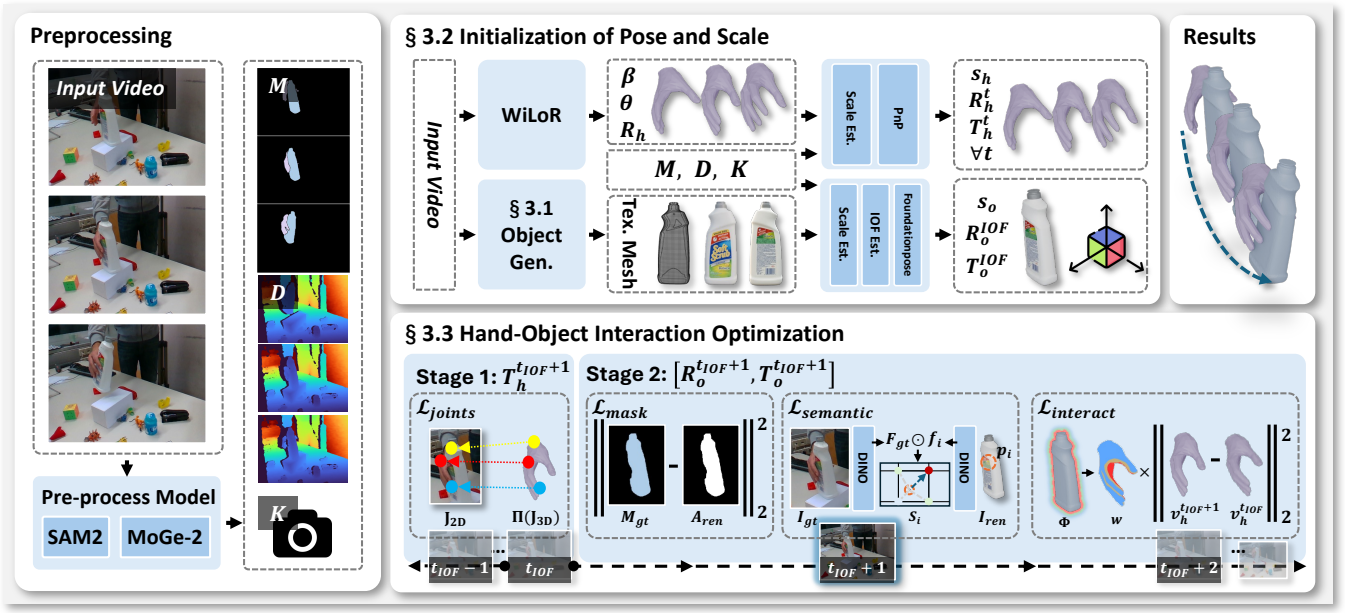


Fig. 3. **Pipeline of AGILE.** Our framework processes the input video in three phases: (1) **Agentic Generation (§3.1):** A VLM-guided loop extracts keyframes and supervises the synthesis of a watertight, textured object mesh $\mathcal{M}_o$, utilizing rejection sampling to ensure visual fidelity. (2) **SfM-Free Initialization (§3.2):** We decouple metric scale and pose. The hand is initialized via WiLoR, while the object pose is anchored at the Interaction Onset Frame (IOF) using a foundation model. (3) **Contact-Aware Optimization (§3.3):** A bi-directional tracking process refines the trajectories. We stabilize the hand via geometric alignment and track the object using semantic ($\mathcal{L}_{dino}$) and interaction constraints ($\mathcal{L}_{interact}$) to ensure physical plausibility.

estimates from the initialization stage, we fix the hand rotation $\mathbf{R}_h$ and scale s_h. We exclusively optimize the translation $\mathbf{T}_h$ to align the mesh with 2D observations. This step is driven solely by the *joint reprojection loss* ($\mathcal{L}_{\text{joint}}$), which minimizes the Euclidean distance between projected 3D joints and 2D detections. This strictly geometric update ensures the hand serves as a stable anchor for the object.

Step 2: Interaction-Aware Object Tracking. With the refined hand pose fixed, we optimize the object's rigid pose $(\mathbf{R}_o, \mathbf{T}_o)$. To handle rapid motion or occlusion, the object pose is initialized using the result from the preceding processed frame. During this stage, we minimize a composite objective function to ensure visual and physical fidelity:

$$\mathcal{L}_{\text{obj}} = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{dino}}\mathcal{L}_{\text{dino}} + \lambda_{\text{interact}}\mathcal{L}_{\text{interact}}. \tag{1}$$

Here, $\mathcal{L}_{\text{mask}}$ enforces silhouette alignment, and $\mathcal{L}_{\text{dino}}$ maintains semantic feature consistency using DINOv3 [Siméoni et al. 2025] to mitigate texture ambiguity. Crucially, $\mathcal{L}_{\text{interact}}$ penalizes interpenetration and encourages surface attraction, ensuring the object remains tightly locked to the hand in a physically plausible manner.

Joint Reprojection Loss ($\mathcal{L}_{joint}$). To spatially anchor the hand pose using reliable 2D cues, we introduce a joint reprojection loss. Let $\Pi(\cdot)$ denote the perspective projection function given the camera intrinsics. We constrain the projected positions of the 3D hand joints $\mathbf{J}_{3D}$ to match the preprocessed 2D keypoint annotations $\mathbf{J}_{2D} \in \mathbb{R}^{21\times 2}$. The loss is defined as the mean squared error (MSE):

$$\mathcal{L}_{\text{joint}} = \frac{1}{N_j}\sum_{i=1}^{N_j} \|\Pi(\mathbf{J}_{3D}^{(i)}) - \mathbf{J}_{2D}^{(i)}\|_2^2. \tag{2}$$

This term ensures the estimated hand geometry remains consistent with the detected anatomical landmarks on the image plane.

Mask Alignment Loss ($\mathcal{L}_{mask}$). To ensure the reconstructed geometry aligns with the visual evidence, we employ a mask alignment loss. This term measures the discrepancy between the rendered alpha silhouette $\mathbf{A}_{\text{ren}}$ and the ground-truth segmentation mask $\mathbf{M}_{\text{gt}}$. We formulate this as the L_2 distance between the two maps:

$$\mathcal{L}_{\text{mask}} = \|\mathbf{A}_{\text{ren}} - \mathbf{M}_{\text{gt}}\|_2^2. \tag{3}$$

By minimizing this difference, the optimization drives the rendered mesh to tightly fit the observed pixel-level silhouette, effectively handling occlusion boundaries.

Semantic Feature Loss ($\mathcal{L}_{dino}$). To mitigate tracking ambiguities caused by partial occlusions or textureless regions, we leverage deep semantic features from a pre-trained Vision Transformer, specifically DINOv3 [Siméoni et al. 2025]. Unlike pixel-wise photometric losses, semantic features provide robust correspondences invariant to local lighting changes.

We establish explicit 3D-to-2D semantic correspondences through a sampling-based approach. Periodically, we sample a set of N_p visible points $\mathcal{P}_{can} = \{\mathbf{p}_i\}_{i=1}^{N_p}$ on the canonical object surface. For each point $\mathbf{p}_i$, we compute a *feature similarity map* $\mathbf{S}_i \in \mathbb{R}^{H_f\times W_f}$

by correlating its feature vector (extracted from the rendered view) against the dense feature map of the target ground-truth image $\mathbf{F}_{gt}$. This similarity map $\mathbf{S}_i$ encodes the likelihood of the point $\mathbf{p}_i$ corresponding to each spatial location in the target image.

During optimization, we transform these canonical points into the current camera frame using the estimated object pose $(\mathbf{R}_o, \mathbf{T}_o)$ and scale $\mathbf{s}_o$, and project them to obtain 2D coordinates $\mathbf{u}_i = \Pi(\mathbf{R}_o(\mathbf{s}_o \odot \mathbf{p}_i) + \mathbf{T}_o)$. The loss is formulated to maximize the semantic similarity at the projected locations, weighted by an occlusion mask to ignore unreliable regions:

$$\mathcal{L}_{\text{dino}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathcal{M}_{occ}(\mathbf{u}_i) \cdot \text{sample}(\mathbf{S}_i, \mathbf{u}_i), \tag{4}$$

where $\mathcal{V}$ denotes the set of currently visible sampled points, $\mathcal{M}_{occ}$ is the binary occlusion mask (0 for occluded regions), and sample$(\cdot)$ represents bilinear interpolation of the similarity map at coordinates $\mathbf{u}_i$. By maximizing this similarity, the optimization drives the object pose to align its semantic parts with the corresponding regions in the video frame.

Interaction Stability Loss ($\mathcal{L}_{interact}$). Given the severe occlusion of the object during manipulation, strictly geometric cues from the image are often insufficient. To resolve this, we impose a physical prior that the relative spatial configuration between the hand and the object remains stable during a grasp, drawing on similar contact stability assumptions explored in prior works [Fan et al. 2024b; Wang et al. 2025a]. This enforces the assumption that the hand and object move approximately as a rigid aggregate during interaction.

Specifically, we formulate a interaction stability loss in the object's canonical coordinate system. Let $\mathbf{V}_h^{(t)} \in \mathbb{R}^{N\times 3}$ denote the hand vertices in the camera frame at time t, and let $(\mathbf{R}_o^{(t)}, \mathbf{T}_o^{(t)})$ be the estimated object pose. We map the current hand vertices into the current object's local frame via the inverse transformation:

$$\tilde{\mathbf{v}}_{h,i}^{(t)} = (\mathbf{R}_o^{(t)})^\top (\mathbf{v}_{h,i}^{(t)} - \mathbf{T}_o^{(t)}), \tag{5}$$

where $\mathbf{v}_{h,i}^{(t)}$ is the i-th vertex of the hand. We then minimize the displacement of these local coordinates relative to their positions in the previous tracked frame $(t-1)$.

Crucially, to ensure the constraint applies only to the effective grasping regions, we compute a vertex-wise weight map $\mathbf{w} \in \mathbb{R}^N$ based on the object's Signed Distance Function (SDF), denoted as $\Phi(\cdot)$. We calculate the distance of the previous hand vertices to the object surface: $d_i = \max(0, \Phi(\tilde{\mathbf{v}}_{h,i}^{(t-1)}))$, and derive the weights via a soft gating function $w_i = 1 - \tanh(\sigma \cdot d_i)$, where σ is a scaling factor controlling the sensitivity of the fall-off. This formulation assigns high weights ($w_i \approx 1$) to vertices in close proximity to the object surface while suppressing the influence of non-interacting fingers. The final loss is defined as:

$$\mathcal{L}_{\text{interact}} = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot \|\tilde{\mathbf{v}}_{h,i}^{(t)} - \tilde{\mathbf{v}}_{h,i}^{(t-1)}\|_2. \tag{6}$$

Minimizing this term effectively "locks" the object to the grasping hand parts, preventing unnatural sliding or jitter when visual features are ambiguous.

In summary, this optimization framework prioritizes reconstruction fidelity and physical consistency. By directly optimizing explicit 3D meshes, our method allows for the seamless integration of 2D visual supervision with 3D interaction constraints, which is often difficult for implicit representations. This design effectively resolves tracking ambiguities caused by severe hand occlusions. Furthermore, our decoupled, interaction-aware strategy ensures that the reconstructed motion is not only pixel-aligned but also physically grounded, preserving the high-quality geometry and texture assets generated in the previous stage.

4 Experiments

Datasets. We evaluate AGILE on both standard benchmarks and challenging in-the-wild sequences. For quantitative evaluation, we utilize the **DexYCB** dataset [Chao et al. 2021], selecting 20 diverse trajectories that cover a wide range of object categories and interaction types. To assess robustness under severe occlusion, we employ the **HO3D-v3** dataset [Hampali et al. 2020], testing on 18 sequences following the protocol of HOLD [Fan et al. 2024b]. To evaluate generalization to bimanual interactions, we further include the **ARCTIC** dataset [Fan et al. 2023], adopting its rigid-object subset following the HOLD protocol. Furthermore, to demonstrate generalization capabilities, we curate an **In-the-Wild** dataset. This collection integrates sequences from prior work [Fan et al. 2024b] with self-captured footage, specifically targeting objects with complex geometries and intricate manipulation patterns to verify robustness in arbitrary settings.

Baselines. We benchmark our framework against state-of-the-art approaches in monocular hand-object reconstruction. Specifically, we compare with **HOLD** [Fan et al. 2024b], which reconstructs interactions via implicit neural rendering, and **MagicHOI** [Wang et al. 2025a], a recent method leveraging diffusion-based priors for geometry generation. Detailed hyperparameters, implementation details, and runtime analysis are provided in the Supp.

Metrics. We evaluate system performance focusing on geometric fidelity, interaction plausibility, and robustness. (1) *Hand Accuracy*: We report the Root-relative Mean Per-Joint Position Error (MPJPE) in millimeters. (2) *Object Geometry*: Following the protocol of HOLD [Fan et al. 2024b], we evaluate reconstruction quality using Chamfer Distance (CD) in squared centimeters (cm^2) and F-scores at 5mm and 10mm thresholds (F@5mm, F@10mm). (3) *Interaction Quality*: We compute the Hand-relative Chamfer Distance (CD_h), also in cm^2, by aligning the object to the hand's root frame to measure relative spatial consistency. (4) *Robustness*: Crucially, we report the *Success Rate* (SR). A sequence is considered a failure if the method encounters initialization breakdown (e.g., SfM collapse) or suffers from catastrophic tracking drift.

4.1 Comparisons

Quantitative Evaluation. As presented in Table 1, AGILE establishes a new state-of-the-art benchmark for monocular HOI reconstruction. We consistently outperform both optimization-based



Fig. 4. **Qualitative Results of Agentic Generation.** We visualize the intermediate stages of our pipeline across diverse object categories. Despite severe hand occlusion in the input keyframes, our VLM-guided approach successfully synthesizes consistent multi-view images and reconstructs high-fidelity 3D meshes. Notably, the *texture refinement* step significantly enhances surface details and sharpness compared to the initial raw generation.

Table 1. **Quantitative comparison on DexYCB and HO3D-v3 datasets.** The **best** results are highlighted in bold, and the second best are underlined. Baseline metrics (†) are averaged over the *successful subset* only, creating a survivor bias.

	DexYCB Dataset						HO3D-v3 Dataset					
Method	MPJPE (mm)↓	CD (cm^2)↓	F@5 (%)↑	F@10 (%)↑	CD_h (cm^2)↓	SR (%)↑	MPJPE (mm)↓	CD (cm^2)↓	F@5 (%)↑	F@10 (%)↑	CD_h (cm^2)↓	SR (%)↑
HOLD† [Fan et al. 2024b]	30.86	19.30	33.20	54.94	170.9	45.0	22.09	1.11	81.75	92.42	18.66	**100.0**
MagicHOI† [Wang et al. 2025a]	21.20	2.05	45.67	67.14	661.90	25.0	7.38	0.90	76.74	91.59	21.81	83.3
Ours	**19.06**	**0.52**	**83.21**	**95.43**	**94.60**	**100.0**	**3.92**	**0.27**	**86.63**	**97.77**	**15.81**	**100.0**

(HOLD) and generative-based (MagicHOI) baselines across all metrics. In terms of geometric fidelity, AGILE achieves a Chamfer Distance (CD) of **0.52** cm^2 on DexYCB, reducing the reconstruction error by nearly **75%** compared to MagicHOI (2.05 cm^2). This substantial gap highlights the advantage of our agentic generation pipeline over standard SfM-based initialization, which often yields noisy or incomplete geometry.

Interaction Stability and Drift. A critical differentiator lies in the interaction quality (CD_h). Prior methods suffer from severe *interaction drift* due to scale ambiguity and lack of physical constraints. MagicHOI, for instance, exhibits a drastically high CD_h of 661.90 cm^2, indicating that the object frequently loses contact or drifts far from the hand. In contrast, AGILE maintains a significantly lower CD_h of 94.60 cm^2. This confirms that our *Contact-Aware Optimization* effectively anchors the object to the hand, ensuring the reconstruction remains physically cohesive and valid for simulation.

Robustness and Survivor Bias. As shown in the Success Rate (SR) column, baselines exhibit high failure rates on DexYCB: MagicHOI fails on 75% of the sequences, and HOLD fails on 55%. It is important to note that the metrics reported for these baselines are biased, as they are averaged only over the *easiest* subset of sequences where initialization succeeded (Survivor Bias). Conversely, AGILE achieves a 100% Success Rate, evaluating on the entire dataset including the most challenging occlusion cases. Remarkably, even while averaging over these harder samples, AGILE still achieves a lower global pose error (MPJPE: 19.06 mm) than MagicHOI (21.20 mm) computed on its easy subset. This demonstrates that AGILE is not only more robust but also fundamentally more accurate.

Generation Quality. Figure 4 visualizes intermediate outputs of our agentic generation pipeline across diverse object categories. Despite severe hand occlusion in the input keyframes, the VLM-guided synthesis yields consistent multi-view images, and the texture refinement step substantially enhances surface details over the raw generation.

Qualitative Evaluation. Qualitative comparisons in Figure 5 corroborate the quantitative findings. Baselines often produce fragmented geometry or suffer from "floating object" artifacts where the object drifts away from the interaction zone. Leveraging generative priors, AGILE reconstructs watertight, high-fidelity meshes. Moreover, the temporal consistency of our results eliminates the jitter observed in prior arts, faithfully reflecting the input video.

Bimanual Interactions Evaluation. AGILE seamlessly extends to bimanual settings by jointly optimizing both hands with independent contact constraints. As Table 2 demonstrates, our method establishes a new state-of-the-art, consistently outperforming both HOLD and BIGS [On et al. 2025] across hand pose accuracy, object geometry, and interaction quality metrics. This confirms AGILE's inherent robustness and its capability to generalize to complex multi-hand manipulations without requiring architectural modifications.

4.2 Ablation Study

We conduct a detailed ablation study on the HO3D dataset to dissect the contribution of each component in our framework. We evaluate the impact on three key dimensions: Hand Pose Accuracy (MPJPE), Object Geometric Fidelity (CD), and Interaction Alignment (CD_h). The quantitative results are summarized in Table 3.

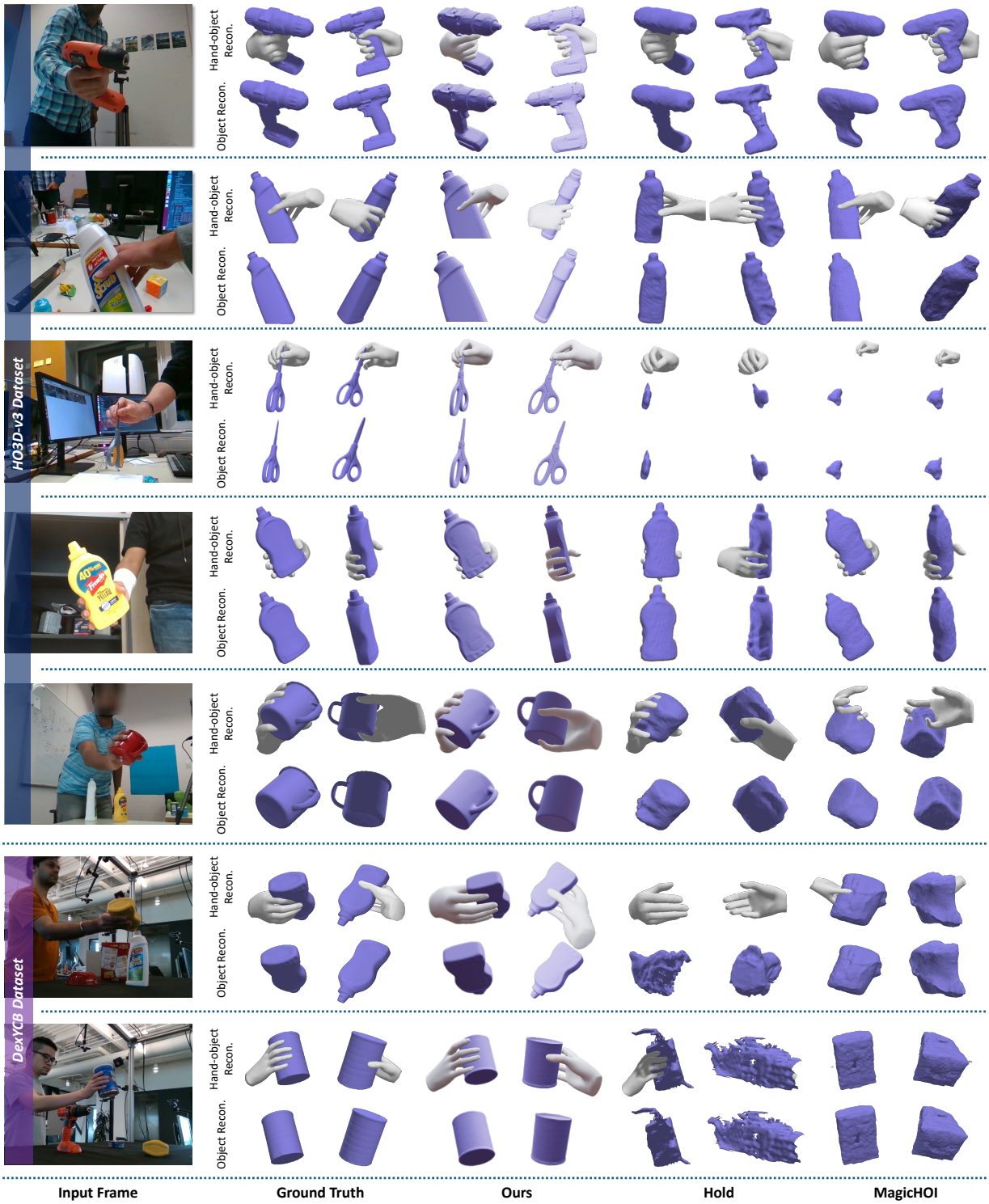


Fig. 5. **Qualitative Comparison**. We compare our reconstructed hands and objects with baseline methods on the HO3D-v3 and DexYCB dataset, showing camera views as well as side views of the object-only and hand-object interaction results.

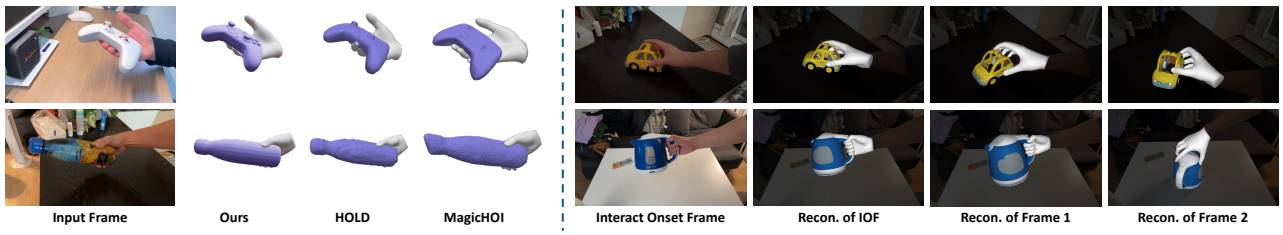


Fig. 6. **Qualitative Evaluation on In-the-Wild Sequences.** *(Left)* Comparison against state-of-the-art baselines. While HOLD [Fan et al. 2024b] and MagicHOI [Wang et al. 2025a] suffer from geometric noise or over-smoothed artifacts due to unreliable initialization, AGILE recovers clean, high-fidelity meshes. *(Right)* Our reconstruction results across temporal sequences. Starting from the Interaction Onset Frame (IOF), our anchor-and-track strategy maintains robust alignment and physical plausibility throughout the dynamic interaction.



Fig. 7. **Qualitative Ablation Study.** Visual comparisons demonstrate that removing key components—such as agentic generation or interaction constraints—leads to severe geometric artifacts, texture degradation, and physical violations (e.g., interpenetration), validating the necessity of our full pipeline.



Fig. 8. **Real-to-Sim retargeting results**, demonstrating that AGILE enables stable kinematic transfer of reconstructed human hand–object interactions to a multi-fingered robotic hand without physics-based correction.

Table 2. **Quantitative comparison on the ARCTIC dataset (rigid-object subset). Best** results are in bold, second-best are underlined.

Method	MPJPE$_l$↓	MPJPE$_r$↓	CD$_o$↓	F@5↑	CD$_l$↓	CD$_r$↓
HOLD [Fan et al. 2024b]	27.1	24.7	2.07	37.1	105.9	123.5
BIGS [On et al. 2025]	34.1	36.1	1.36	56.4	46.1	31.3
Ours	**25.0**	**23.8**	**1.12**	**57.6**	**21.9**	**30.6**

The full model (a) achieves the best performance across all metrics, validating the synergy between high-fidelity asset generation and physics-aware tracking.

Impact of Agentic Generation Modules. We first analyze the components responsible for creating the 3D asset, which serves as the foundation for subsequent tracking.

Disabling the entire agentic loop (b) and lifting a 3D mesh directly from a single frame causes catastrophic geometric hallucinations in occluded regions. To further disentangle our agentic design from the innate capabilities of the base models, we ablate its two core components. Replacing the VLM critic with a max-mask heuristic (b_1) severely degrades accuracy, confirming that large masks often capture uninformative or self-occluded views rather than optimal geometric ones. Alternatively, retaining VLM selection but bypassing

Table 3. Detailed ablation study on the HO3D dataset. We systematically evaluate the impact of generative components and optimization objectives.

	Model Variant	Hand	Object Geometry		Interaction
		MPJPE ↓	CD ↓	F@5mm ↑	CD_h ↓
(a)	AGILE (Full)	**3.92**	**0.27**	**86.63**	**15.81**
Generative Components					
(b)	w/o Agentic Multi-view Gen.	4.02	2.85	30.12	40.12
(b_1)	Heuristic Selection	4.17	2.04	76.51	29.37
(b_2)	w/o Multi-View Synthesis	4.32	2.61	66.58	46.94
(c)	w/o Texture Refine.	4.16	0.30	86.02	20.18
Optimization Objectives					
(d)	w/o $\mathcal{L}_{\text{joint}}$	5.06	0.33	84.93	19.37
(e)	w/o $\mathcal{L}_{\text{mask}}$	3.97	0.40	81.14	23.99
(f)	w/o $\mathcal{L}_{\text{dino}}$	4.13	0.34	84.36	21.64
(g)	w/o $\mathcal{L}_{\text{interact}}$	4.17	0.35	86.10	54.40

multi-view synthesis (b_2) similarly fails, as single-view generation forces the model to hallucinate unseen topologies. These substantial degradations prove that our performance gains are driven by the agentic paradigm, not merely the underlying foundation models.

In variant (c), we skip the texture baking and refinement step. This ablation proves critical: high-quality texture is not merely for visual aesthetics but is a prerequisite for robust tracking. Since both our initialization (via FoundationPose) and our semantic consistency loss rely on discriminative surface features, noisy or blurred textures lead to initialization drift, causing a noticeable rise in interaction error (CD_h).

Impact of Optimization Objectives. Next, we evaluate the loss functions used during the decoupled optimization stages.

In variant (d), we drop $\mathcal{L}_{\text{joint}}$, leaving the hand pose to rely solely on the initial PnP estimation without further optimization. The significant degradation in MPJPE confirms that PnP initialization alone is insufficient for high-precision tracking. $\mathcal{L}_{\text{joint}}$ is indispensable for refining these initial estimates to resolve depth ambiguities and ensure temporally consistent hand articulation.

Variant (e) ablates $\mathcal{L}_{\text{mask}}$, which provides the strongest pixel-level geometric constraint for the rigid object. Although the object scale is pre-initialized, $\mathcal{L}_{\text{mask}}$ is critical for fine-tuning the scale at the anchor frame and guiding precise 6D pose optimization throughout the sequence. Without this constraint, the object pose fails to strictly adhere to visual boundaries, resulting in misalignment that negatively impacts both CD and F-scores.

Variant (f) ablates the DINO-based semantic feature loss $\mathcal{L}_{\text{dino}}$. In the absence of this consistency term, object tracking becomes prone to drift, particularly in textureless regions or under motion blur. The increase in CD_h confirms that semantic descriptors provide robust correspondence where pixel-level metrics alone are insufficient.

Finally, removing the interaction stability loss $\mathcal{L}_{\text{interact}}$ in variant (g) leads to severe physical violations. Without this constraint, the object tends to slide unnaturally or penetrate the hand mesh to minimize surface distance. Consequently, the fine-grained interaction quality degrades drastically (CD_h spikes to 54.40), highlighting the necessity of SDF-based priors for physically plausible reconstruction.

4.3 Application: Real-to-Sim Retargeting

To showcase the utility of our reconstructions for Embodied AI, we demonstrate a scalable Real-to-Sim retargeting pipeline. We import the reconstructed sequence into Isaac Gym [Makoviychuk et al. 2021] and map the human hand motion to a dexterous robotic hand (e.g., Shadow) via kinematic optimization [Qin et al. 2023], using the agentic-generated mesh as the manipulated asset.

As illustrated in Figure 8, AGILE effectively bridges the gap between video and simulation: (1) *Simulation-Ready Assets*: Our generated meshes are watertight and topologically clean, ready to be loaded into the simulator without manual cleanup. (2) *Faithful Kinematic Transfer*: The retargeted grasps closely follow the recovered hand-object configurations through purely kinematic mapping, providing a qualitative validation that AGILE recovers accurate relative poses and contact states. We note that this demonstration focuses on kinematic visualization rather than full physics-based rollout; closing the loop with contact dynamics and policy learning is left to future work. Nevertheless, by automating the path from monocular video to retargetable assets, our framework opens a route toward curating large-scale manipulation data from unconstrained videos for generalist policy learning.

5 Conclusion

In this work, we presented AGILE, a robust framework that addresses the persistent challenges of incomplete geometry and brittle initialization in monocular hand-object interaction reconstruction. By shifting the paradigm from *reconstruction-based* optimization to *agentic generation*, we leverage VLM-guided priors to synthesize high-fidelity, watertight object meshes even under severe hand occlusion. Complementing this, our *anchor-and-track* strategy eliminates the dependency on fragile SfM pipelines, enabling stable 6D pose estimation on challenging in-the-wild videos where prior methods routinely fail. Extensive experiments on DexYCB, HO3D-v3, ARCTIC, and in-the-wild sequences confirm that AGILE achieves state-of-the-art accuracy with a 100% success rate, producing simulation-ready digital twins that effectively bridge the gap between visual observation and physical simulation.

Limitations and Future Work. Our framework inherits limitations from the underlying foundation models for depth, segmentation, and multi-view generation, so extreme cases (e.g., transparent or highly reflective objects) may induce scale drift or local tracking errors. Notably, the static camera assumption applies only to the initial IOF detection and does not constrain core tracking, which remains valid under camera motion (see Supp.). Promising directions include extending agentic generation beyond rigid objects to articulated and deformable categories, and exploiting VLMs more broadly, both as a *reconstructibility pre-filter* screening viable segments and as a *multi-anchor detector* enabling bi-directional propagation against long-sequence drift.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576315).

References

Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 2019. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10843–10852.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9044–9053.

Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. 2002. The trimmed iterative closest point algorithm. In *2002 International Conference on Pattern Recognition*, Vol. 3. IEEE, 545–548.

Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. 2020. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5031–5041.

Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. 2024a. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *European Conference on Computer Vision*. Springer, 428–448.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. 2024b. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 494–504.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2020. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3196–3206.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400* (2023).

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9298–9309.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9970–9980.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. 2021. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. 2025. BIGS: Bimanual Category-agnostic Interaction Reconstruction from Monocular Videos via 3D Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17437–17447.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9826–9836.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2025. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12242–12254.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. 2023. AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System. In *Robotics: Science and Systems*.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

Javier Romero, Dimitrios Tzionas, and Michael J Black. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610* (2022).

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. arXiv:2508.10104 [cs.CV] https://arxiv.org/abs/2508.10104

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*. Springer, 1–18.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. 2025. SAM 3D: 3Dfy Anything in Images. (2025). arXiv:2511.16624 [cs.CV] https://arxiv.org/abs/2511.16624

Tencent Hunyuan3D Team. 2025. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. arXiv:2501.12202 [cs.CV]

Bugra Tekin, Federica Bogo, and Marc Pollefeys. 2019. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4511–4520.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 2025b. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5261–5271.

Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. 2025a. MagicHOI: Leveraging 3D Priors for Accurate Hand-object Reconstruction from Short Monocular Video Clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5957–5968.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. 2024. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17868–17879.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191* (2024).

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. 2021. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11097–11106.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. 2025. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 27716–27726.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. 2025. HaWoR: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1805–1815.

Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. 2019. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2354–2364.

Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 2020. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5346–5355.