Follow My Hold: Hand-Object Interaction Reconstruction through Geometric Guidance

Ayce Idil Aytekin[1*] Helge Rhodin[1,2] Rishabh Dabral[1] Christian Theobalt[1]
[1] Max Planck Institute for Informatics and Saarland University
[2] Bielefeld University

Figure 1. **In-the-wild results.** From a single RGB frame from in-the-wild (top row), our method reconstructs detailed 3D hand–object interactions (bottom row). It does so by guiding a latent diffusion model with multi-modal cues derived from the input image and several foundation models.

Abstract

We propose a novel diffusion-based framework for reconstructing 3D geometry of hand-held objects from monocular RGB images by leveraging hand-object interaction as geometric guidance. Our method conditions a latent diffusion model on an inpainted object appearance and uses inference-time guidance to optimize the object reconstruction, while simultaneously ensuring plausible hand-object interactions. Unlike prior methods that rely on extensive post-processing or produce low-quality reconstructions, our approach directly generates high-quality object geometry during the diffusion process by introducing guidance with an optimization-in-the-loop design. Specifically, we guide the diffusion model by applying supervision to the velocity field while simultaneously optimizing the transformations of both the hand and the object being reconstructed. This optimization is driven by multi-modal geometric cues, including normal and depth alignment, silhouette consistency, and 2D keypoint reprojection. We further incorporate signed distance field supervision and enforce contact and non-intersection constraints to ensure physical plausibility of hand-object interaction. Our method yields accurate, robust and coherent reconstructions under occlusion while generalizing well to in-the-wild scenarios. The project page is at https://aidilayce.github.io/FollowMyHold-page/.

1. Introduction

Our hands are how we shape the world; by picking, pushing, holding, slicing, or sculpting. From watering flowers to assembling furniture, hand-object interactions are ubiquitous. However, from a 3D reconstruction perspective, these interactions are inherently ambiguous: hands occlude objects and shapes overlap. This makes an accurate 3D reconstruction of hand-object interaction (HOI) difficult as ambiguity is even greater when working with a single image, where depth and contact cues are limited. Tackling this practical setting is crucial to enable robust and scalable 3D understanding in AR/VR, embodied AI, and robotics.

It is noteworthy that hand and object reconstruction have been long-standing problems [12, 21, 24, 25, 31, 45, 46]

*Corresponding author: aaytekin@mpi-inf.mpg.de

with several challenges arising out of the ill-posedness of the task. Unlike clothed bodies or textured objects, hands are of rather homogeneous colors, making it challenging to reconstruct them from a single image. Likewise, reconstructing objects in 3D, despite the impressive progress in the recent years [29, 40, 43, 50, 56], remains a challenging task since objects come in such diverse shapes that building a strong prior is difficult. The reconstruction problem becomes significantly more complex when we combine the two tasks into one: reconstructing the 3D objects when they are undergoing hand-held interaction.

Needless to say, one cannot solve the HOI reconstruction problem without reasonably modeling the occlusion caused by the hands. In this line, existing works have attempted several approaches, such as template-based optimization [20–23, 54], training on 3D hand-object data [24, 27, 53], data-driven normal prediction [18] or 6DoF prediction [4, 45]. However, such models either require explicit geometric priors in the form of a template or are better suited for reconstruction with multiple views/frames. Moreover, methods that perform a direct regression of the partial point-clouds [48] lack the ability to complete the unobserved portions of the captured object.

Recently, methods like EasyHOI [32] and Gen3DSR [13] have attempted to leverage the large-scale 2D and 3D foundational models to propose pipelines that combine the implicit priors in different foundation models into one. Despite impressive out-of-domain generalizability, they tend to be brittle, and failure of a single component during optimization often results in a catastrophic failure of the reconstruction. Instead of directly adopting a generative model's output as the final reconstruction, we propose to embed 2D foundational model supervision directly into the 3D generative sampling process to improve robustness.

With this core idea, we introduce **FollowMyHold**, which *guides* a pretrained, flow-based image-to-3D generator *during* inference with geometric supervision. Concretely, we (i) extract complementary cues with 2D/3D foundation models: interaction masks [33] and hand detection [38], inpainted object appearance [28] with Gemini prompts [44], an initial hand mesh [37], an initial coarse HOI mesh obtained from Hunyuan3D-2 [56], and a partial HOI point cloud [48]; (ii) register all outputs into a shared image-aligned frame; and (iii) steer a rectified-flow 3D generator (Hunyuan3D-2 [56]) with an optimization-in-the-loop design for sampling. Our staged optimization design (hand, object, then joint refinement) applies pixel-aligned 2D losses (normal, depth, silhouette, keypoints) together with 3D interaction constraints (intersection & proximity), yielding physically plausible, globally consistent HOI reconstructions.

We evaluate on well-established HOI image-to-3D reconstruction benchmarks like OakInk [52], Arctic [17] and DexYCB [5]. FollowMyHold sets the state-of-the-art among generative methods, surpassing EasyHOI in object reconstruction accuracy and achieving almost two-fold higher reconstruction rate, with strong robustness and in-the-wild generalization. Extensive evaluation validates the accuracy and robustness of our method. The code will be publicly available.

2. Related Work

2.1. Reconstructing Hand-Object Interactions

Recovering the 3D structure of interacting hands and objects from a single image is challenging due to occlusions and limited 3D supervision. Many existing methods simplify the problem by assuming access to 3D object templates [3, 4, 7, 12, 19–23, 54], estimating only the hand and object poses from video sequences [6, 36] or multi-view inputs [39]. Template-free approaches learn from 3D interaction datasets [9, 10] but often struggle due to limited object diversity in their training set. Recent work [32] uses foundation models to reconstruct hands and objects separately but fixes object geometry early, making the pipeline brittle to initial errors. In contrast, our method jointly optimizes object shape, pose, and hand pose in a category-agnostic manner, leveraging foundation model priors while guiding the model via geometric signals.

2.2. Monocular 3D Object Reconstruction

Recovering 3D shape from a single image remains one of the most fundamental problems in 3D vision. From early CNN-based models [29, 40] to retrieval-based systems [43] and volumetric methods [34], a wide array of strategies have been proposed. Recent trends shift toward generative reconstruction, with Large Reconstruction Models (LRMs) [26, 42, 47, 49, 50, 56] enabling high-fidelity geometry from sparse input. However, these models are typically trained under assumptions of object-centric, unobstructed inputs. In practice, reconstructions degrade sharply in the presence of occlusion, such as during human-object interaction. Our approach adapts large-scale object reconstruction models to this more complex setting by integrating additional cues derived from hand-object contact and visibility, guiding them toward more plausible completions even under partial observation.

2.3. Hand Mesh Recovery

Estimating hand pose and shape from images has seen major progress with the introduction of parametric models like MANO, allowing dense predictions from monocular RGB [14, 37, 38]. Recent approaches either directly regress MANO parameters [1, 2] or optimize them via image-level constraints [55, 57]. While these techniques perform well in isolation, they often fail to account for physical plausibil-

ity in the presence of interacting objects. In our method, we employ HaMeR [37], a state-of-the-art hand mesh recovery method, and use this mesh as a signal to guide the object reconstruction process.

3. Preliminaries: Rectified Flow and Notation

We review rectified flow model, which is the backbone of our 3D generator (Hunyuan3D-2), to clarify how inference-time guidance operates.

Rectified flow $\hat{\mathbf{x}}_1$ formulation. Rectified flow models are a class of generative models trained with the flow-matching objective [16], framing generation as solving an ordinary differential equation (ODE). Instead of predicting noise (as in DDPMs, which approximate the score function), the model learns a *velocity field* that moves a noisy sample toward a clean target over time. Given a noisy sample $\mathbf{x}_t$ at time t and conditioning $\mathbf{c}$, the model predicts $\mathbf{v}_\theta(x_t, t, \mathbf{c})$, and the update is

$$\mathbf{x}_{t+\Delta t} \;=\; \mathbf{x}_t \;+\; (\sigma_{t+\Delta t} - \sigma_t)\, \mathbf{v}_\theta(x_t, t, \mathbf{c}), \qquad (1)$$

where $\sigma_t \in [0, 1]$ encodes the flow time at step $t \in [0, 1]$. At inference, we *steer* generation without changing the model weights by modifying $\mathbf{v}_\theta(x_t, t, \mathbf{c})$ using gradients of a task objective G (Sec. 4), e.g., geometry-consistency losses, to nudge the trajectory toward lower G. To evaluate G we often need the estimate of the clean sample $\hat{\mathbf{x}}_1$ (following Hunyuan3D-2's $\mathbf{x}_1$-target formulation), as it is the current approximation of the underlying target. This is recovered from the flow $\mathbf{v}_\theta(x_t, t, \mathbf{c})$ using Eq. 4.54 in [30],

$$\hat{\mathbf{x}}_1 \;=\; \mathbf{x}_t \;+\; (1 - \sigma_t)\, \mathbf{v}_\theta(x_t, t, \mathbf{c}). \qquad (2)$$

Derivation details are in the supplemental material.

Notation. We denote meshes by $\mathcal{M}$, with superscripts indicating their canonical spaces: $\mathcal{M}^{\mathrm{U}}$ for **Hu**nyuan, $\mathcal{M}^{\mathrm{H}}$ for **H**aMeR, and $\mathcal{M}^{\mathrm{I}}$ for **i**mage-aligned space. Transformations between spaces are represented by a similarity transform $\mathcal{T}$, parameterized by scale $s \in \mathbb{R}$, rotation $R \in \mathrm{SO}(3)$ and the translation $\mathbf{t} \in \mathbb{R}^3$. For example, a mesh in HaMeR's output space could be transformed to image-aligned space via $\mathcal{M}^{\mathrm{I}} = \mathcal{T}_{\mathrm{H}\to\mathrm{I}} \cdot \mathcal{M}^{\mathrm{H}}$.

4. Method

Reconstructing a physically plausible 3D hand-object interaction from a single RGB image is a fundamentally ill-posed task. Occlusion, entangled geometry, and limited depth cues make it difficult to recover accurate hand and object shapes, and consequently their spatial arrangement. Recent foundation models offer strong priors for hands (e.g., HaMeR), 3D geometry (e.g., Hunyuan3D-2), and pixel-aligned partial geometry with depth cues (e.g., MoGe-2). Each model presents a set of complementary strengths and weaknesses, and our goal is to propose a method that effectively navigates these conflicting properties. However, the methods are incompatible, e.g., they operate in their own canonical coordinate space, with mismatching assumptions about scale, orientation, and alignment.

Simply connecting the outputs of these models, as attempted by recent methods like EasyHOI [32] and Gen3DSR [13], does not work well as the misalignment of canonical spaces prevents direct comparison or joint optimization of model outputs. Our method directly tackles this coordination challenge by explicitly aligning model outputs into a shared, *image-aligned* reference frame. This alignment unlocks a crucial capability: we can now compare the rendered normals, disparity, and silhouette maps from our 3D predictions directly against 2D supervision maps rendered from MoGe-2 [48], a state-of-the-art method for point-cloud prediction. This enables pixel-accurate, differentiable supervision throughout the optimization process.

At a high level, our method (1) uses foundation models to segment and inpaint the image and produce initial 3D predictions, including a hand mesh, a partial point cloud of HOI, and a coarse HOI mesh (Sec. 4.1.1 and Sec. 4.1.2), (2) aligns all outputs to a shared coordinate frame via a two-stage transformation chain (Sec. 4.1.3), and (3) renders from this unified frame and applies gradient-based guidance during diffusion using both 2D supervision (e.g., normals, depth, silhouette) and 3D interaction constraints (e.g., intersection and proximity losses) (Sec. 4.2.1). See Fig. 2 for an overview of our method.

4.1. Initialization with Foundational Models

4.1.1 2D Signal Extraction

Given as input a single RGB image $I_{\text{full}} \in \mathbb{R}^{\hat{H}\times\hat{W}\times 3}$, we first localize the interaction region using foundation models. Specifically, LangSAM [33] is used to extract the hand (M_h) and object (M_o) masks, while WiLoR's hand detector [38] detects hand bounding boxes. Together they define a crop yielding $I_{\text{crop}} \in \mathbb{R}^{H\times W\times 3}$. We then use a diffusion-based inpainter (FLUX.1 Kontext[dev] [28] guided by a Gemini [44] text prompt) to remove the hand and complete the occluded object appearance, producing the inpainted object image I_{obj}. We additionally mask I_{crop} with the union of M_h and M_o to obtain I_{hoi}, which serves as input to MoGe-2 and Hunyuan3D-2.

4.1.2 3D Geometry Initialization

We precompute 3D cues to guide diffusion-based hand-object reconstruction. To begin with, we employ a state-of-the-art hand-reconstruction model, HaMeR [37], that provides an initial hand mesh $\mathcal{M}_h^H$ and its 2D keypoints $\hat{K}_{2D}$ from I_{crop}. Then, MoGe-2 [48] is used to estimate the partial point cloud $P_{\text{hoi}} \in \mathbb{R}^{N_m\times 3}$ and camera ϕ from I_{hoi}. The

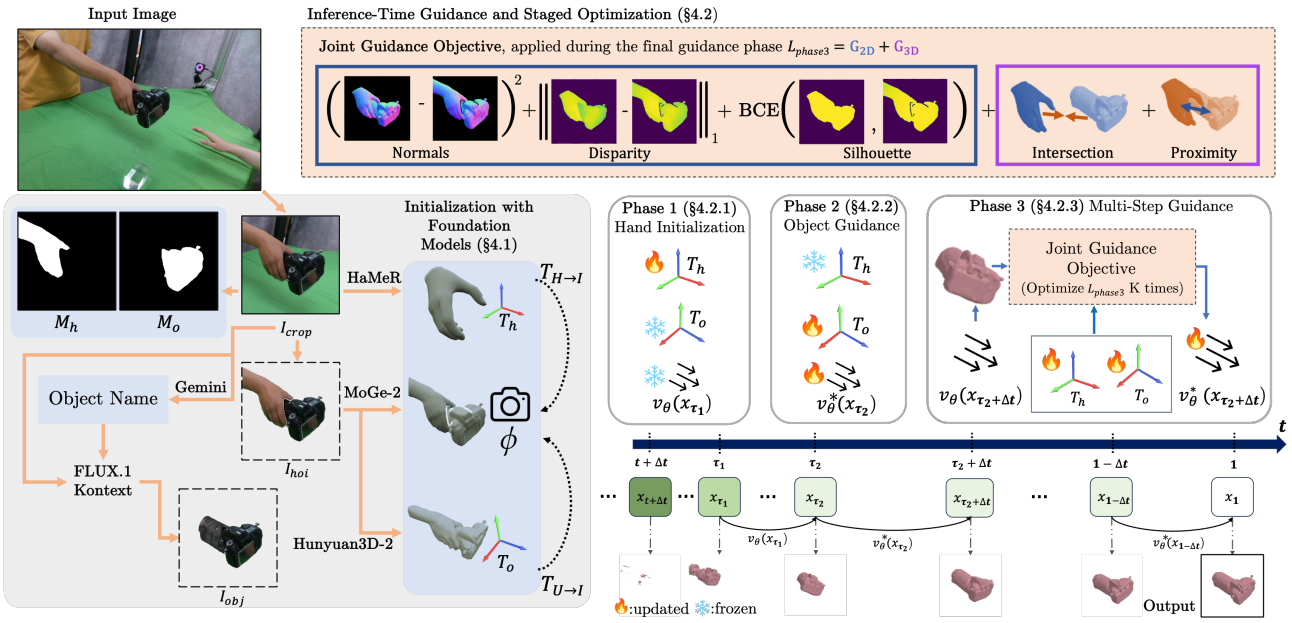

Figure 2. **Overview of FollowMyHold**. Given a single RGB frame, we (1) isolate the interaction region and derive binary hand/object masks with LangSAM and WiLoR's hand detector; (2) inpaint the occluded object appearance using FLUX.1 Kontext + Gemini (§4.1). Next, we obtain three complementary 3D cues: a HaMeR hand mesh, a MoGe-2 partial point cloud (with camera pose ϕ), and a coarse Hunyuan3D-2 HOI mesh. A two-step ICP registers all cues into a common image-aligned frame. Finally, we perform inference-time guidance with a staged optimization (§4.2): Phase 1 optimizes the hand transform T_h; Phase 2 optimizes the object transform T_o and guides the velocity field; Phase 3 jointly refines (T_h, T_o) while guiding with pixel-aligned 2D losses (G_{2D}: normals, disparity, silhouette) and 3D constraints (G_{3D}: intersection, proximity). The right bottom row shows progressive object refinement over diffusion steps.

partial point cloud P_{hoi} output by MoGe-2 only consists of the points corresponding to the unoccluded portions of the input object image I_{hoi}, as shown in Fig. 2. We then render the normal map, disparity map, and silhouette from P_{hoi} using ϕ to serve as the supervision maps for the geometric guidance objective during the sampling process.

At the core of our framework lies Hunyuan3D-2 [56], a large-scale latent diffusion model for 3D shape generation. The diffusion transformer $\mathcal{E}_{\text{hoi}}$ produces the HOI latent $\mathbf{x}_{\text{hoi}} = \mathcal{E}_{\text{hoi}}(\mathbf{x}_t; t, \mathbf{c}_{\text{hoi}})$, where $\mathbf{c}_{\text{hoi}}$ are DINOv2 [35] features of I_{crop}. The generated latent $\mathbf{x}_{\text{hoi}}$ is decoded into a signed distance field (SDF) by Hunyuan3D-2's decoder $\mathcal{D}$ over a query grid $\mathbf{q} \in [-1, 1]^3$, and meshed via FlexiCubes [41]: $\text{SDF}_{\text{hoi}} = \mathcal{D}(\mathbf{x}_{\text{hoi}}, q)$ and $\mathcal{M}^U_{\text{hoi}} = \text{FlexiCubes}(\text{SDF}_{\text{hoi}})$. Although trained on single objects, Hunyuan3D-2 preserves a coarse hand–object arrangement on HOI inputs, as shown in Fig. 2 in the Initialization with Foundation Models part.

In our work, we propose to leverage this coarse alignment as a valuable cue for transforming and aligning meshes across canonical spaces.

4.1.3 Transforming Across Canonical Spaces with ICP

While critical to our pipeline, the outputs from HaMeR, Hunyuan3D-2, and MoGe-2 lie in different coordinate spaces. To enable joint reasoning and rendering, we estimate rigid transformations between these coordinate spaces using Iterative Closest Point (ICP) alignment.

We first align the coarse HOI mesh to the partial point cloud (yielding $T_{U\to I}$), then align the hand mesh to the coarse HOI mesh (resulting in $T_{H\to U}$) to compose them and obtain $T_{H\to I}$; this two-step ICP avoids the pitfalls of directly aligning the hand to an incomplete point cloud. These transformations bring all the geometric entities into a unified coordinate frame (image frame of MoGe-2 partial point cloud) for consistent modeling and inference-time guidance. Note that we initialize ICP with a similarity transformation that aligns the centroids and global scales of the source and the target.

4.2. Inference-Time Guidance and Staged Optimization

We combine inference-time guidance with staged optimization of transformation parameters. We first explain our guidance strategy in rectified flow and then the role of each optimization phase.

4.2.1 Inference-time Guidance

Our base image-to-3D generative model, Hunyuan3D-2, uses a rectified flow model (Sec. 3) designed for efficient

image-to-shape generation. We modify this model by integrating gradient-based supervision directly into the flow trajectory at inference time, enabling end-to-end refinement of object geometry and pose under loss-specific constraints. We structure our reconstruction process into three phases, each addressing specific challenges in HOI recovery.

Before introducing our geometric guidance, we let the denoising process reach time step $t \geq \tau_1$, at which the denoised latents reach a coarse but sufficient quality. For early diffusion steps $t < \tau_1$, we apply standard rectified flow updates in Eq. 1 without guidance.

Algorithm 1 Inference-time Multi-step Optimization

Require: $\mathcal{R}_S$: Supervision maps
Require: $\mathbf{w_{opt}}$: geometric optimization guidance strengths
Require: DifferentiableRenderer
$x_T \sim \mathcal{N}(0, I)$ ▷ Sample noisy latent vector
for all t from 0 to 1 **do**
  $\mathbf{v}_t \leftarrow \mathcal{E}(x_t, t, c_{\text{obj}})$
  if $t > \tau_2$ **then** ▷ Geometric guidance in Phase 3
    $\mathbf{Z}_t = [\mathbf{v}_t, T_{o,t}, T_{h,t}]$ ▷ Define optimization variables
    for $k = 1$ to K **do** ▷ Multiple gradient descent steps
      $\hat{\mathbf{x}}_1^k = \mathbf{x}_t + (1 - \sigma_t) \cdot \mathbf{v}_t^k$
      $\mathcal{M}_o^U = \text{FlexiCubes}(\mathcal{D}(\hat{\mathbf{x}}_1^k, \mathbf{q}))$
      $\mathcal{M}_o^I = T_{o,t}^k T_{U\to I} \mathcal{M}_o^U$, $\mathcal{M}_h^I = T_{h,t}^k T_{H\to I} \mathcal{M}_h^H$
      $\mathcal{R}_t^k \leftarrow \text{DifferentiableRenderer}(\mathcal{M}_o^I \cup \mathcal{M}_h^I)$
      ▷ Compute 2D and 3D geometric losses
      $G_{2D} \leftarrow \mathcal{L}_{2D}(\mathcal{R}_t^k, \mathcal{R}_S)$, $G_{3D} \leftarrow \mathcal{L}_{3D}(\mathcal{M}_o^I, \mathcal{M}_h^I)$
      ▷ Gradient-based updates
      $\mathbf{Z}_t^k \leftarrow \mathbf{Z}_t^k - \mathbf{w_{opt}} \nabla_{\mathbf{Z}_t^k}(G_{2D}(\mathbf{Z}_t^k) + G_{3D}(\mathbf{Z}_t^k))$
    end for
    $\mathbf{v}_t^* = \mathbf{v}_t^K$ ▷ Update $\mathbf{v}$ with the steered velocity field
  end if
  $\mathbf{x}_t \leftarrow \mathbf{x}_t + \mathbf{v}_t^*$ ▷ Rectified Flow step
end for
▷ Obtain final hand and object meshes
$\mathcal{M}_o^I = T_{o,1} T_{U\to I} \text{FlexiCubes}(\mathcal{D}(\mathbf{x}_1, q))$
$\mathcal{M}_h^I = T_{h,1} T_{H\to I} \mathcal{M}_{h,1}^I$

Algorithm 1 describes our full inference-time multi-step optimization used in Phase 3, where both the object and hand transformations T_o and T_h are jointly optimized and the velocity field $\mathbf{v}_\theta(x_t, t, \mathbf{c}_{obj})$ is guided using both 2D and 3D objectives. We denote the flow field $\mathbf{v}_\theta(x_t, t, \mathbf{c})$ as $\mathbf{v}_t$ for simplicity. $\mathcal{R}_s$ represents supervision maps rendered from P_{hoi}: normal, disparity, and silhouette maps. Unlike traditional inference-time guidance, we follow an optimization-in-the-loop strategy, jointly optimizing the transformations (T_h and T_o) along with the flow field estimated by the model. In particular, for each diffusion time step $t \in [\tau_2, 1]$, we perform K inner gradient descent iterations to jointly optimize the transformations and steer $\mathbf{v}_\theta(x_t, t, \mathbf{c})$ using 2D and 3D geometric objectives G_{2D} and G_{3D}.

Across phases, 2D losses compare rendered maps against $\mathcal{R}_s$: normal alignment $L_{\text{norm},(\cdot)}$, L1 disparity $L_{\text{disp},(\cdot)}$, and binary cross-entropy silhouette $L_{\text{sil},(\cdot)}$, with $(\cdot) \in \{\text{h}, \text{o}, \text{hoi}\}$ for hand, object, and their union. Note that we treat the hand geometry as fixed during object reconstruction, as HaMeR provides reliable hand estimates even under occlusion. In contrast, the object is more underconstrained due to partial visibility and the wide range of plausible shapes. We use the hand as a stable geometric anchor to constrain object pose optimization. Overall, this staged strategy enables stable and physically plausible 3D interaction reconstruction from a single RGB image. The phases are explained in the following.

4.2.2 Phase 1: Hand-Only Optimization

Phase 1 only focuses on optimizing the hand to establish a spatially meaningful anchor before introducing guidance on the object. At diffusion step $t = \tau_1$, we optimize the hand transformation T_h, initialized with unit scale, identity rotation, and zero translation, while keeping the object transformation parameters T_o and the velocity field $\mathbf{v}_\theta(x_t, t, \mathbf{c}_{obj})$ fixed. We first transform the MANO hand mesh $\mathcal{M}_h^H$ into the image-aligned space as $\mathcal{M}_h^I = \mathcal{T}_{H\to I} \mathcal{M}_h^H$ and pose it with $\widetilde{H} = T_h \cdot \mathcal{M}_h^I$. Rendering $\widetilde{H}$ yields $L_{\text{norm,h}}, L_{\text{disp,h}}, L_{\text{sil,h}}$. We also add the ℓ_2 loss between the projected 3D hand keypoints and their corresponding 2D keypoints detected by HaMeR and regularize excessive translation using $L_{\text{reg,h}} = \left\|\mathbf{t}_h\right\|^2$. The Phase 1 objective L_{phase1} sums these terms (weights in supplemental material). Once complete, we obtain a well-posed hand mesh and proceed with the denoising process.

4.2.3 Phase 2: Object-Only Optimization

This phase addresses a key robustness challenge: if the object reconstructed at $t = \tau_1$ is heavily misaligned or poorly scaled, it can occlude or be occluded by the hand in the rendered views, degrading the effectiveness of 2D losses. To mitigate this, we coarsely align the object before enabling joint hand-object reasoning in Phase 3. At $t = \tau_2$, we optimize T_o and apply gradient-based guidance to $\mathbf{v}_\theta(x_t, t, \mathbf{c}_{obj})$, while keeping the hand transformation T_h fixed. The operations in this phase are similar to Algorithm 1 without the hand mesh. T_o is initialized as identity, and following Sec. 4.2.1, we extract the object's mesh, which is defined in the canonical space of Hunyuan3D-2. We then transform it to image-aligned space via $\mathcal{M}_o^I = \mathcal{T}_{U\to I} \mathcal{M}_o^I$ and pose it by $\widetilde{O} = T_o \cdot \mathcal{M}_o^I$. From $\widetilde{O}$, we render geometry maps and compute the supervision losses $L_{\text{norm,o}}, L_{\text{disp,o}}, L_{\text{sil,o}}$. We also compute a regularization loss $L_{\text{reg,o}}$ to penalize excessive translation, scale drift, and mesh noise. The Phase 2 loss L_{phase2} (sum of these terms; weights in supplemental material) serves as G_{2D} to steer $\mathbf{v}_\theta(x_t, t, \mathbf{c}_{obj})$ while updating T_o for k_2 steps. Once

complete, we obtain a coarsely posed object mesh and a supervision-steered velocity field $\mathbf{v}^*_\theta(x_t, t, \mathbf{c})$, and proceed by updating the object latent using Eq. 1.

4.2.4 Phase 3: Joint Optimization

In this final phase (at $t \in [\tau_2, 1]$), we jointly refine $\mathbf{v}_\theta(x_t, t, \mathbf{c}_{obj})$, T_o, and T_h. This phase introduces 3D geometric supervision alongside 2D objectives to ensure spatial plausibility and physical interaction.

2D geometric supervision: With $\widetilde{H}$ and $\widetilde{O}$ posed, we define $\widetilde{HOI} = \widetilde{H} \cup \widetilde{O}$ and compute $L_{\text{norm,hoi}}, L_{\text{disp,hoi}}, L_{\text{sil,hoi}}$.

3D geometric supervision: To ensure physically plausible hand-object interactions, we introduce intersection and proximity losses.

Intersection loss. We add an intersection penalty with the goal of introducing a guidance objective that nudges the sampling process towards a 3D latent that does not intersect with the hand. Specifically, we convert the posed hand and object meshes into signed distance fields (SDFs) on a shared volumetric grid and define the objective as

$$L_{\text{intersection}} = \frac{1}{K} \sum_{i=1}^{K} \mathbb{1}[\text{SDF}_h(f_i) < 0 \wedge \text{SDF}_o(f_i) < 0], \tag{3}$$

where $\mathbb{1}[\cdot]$ represents the indicator function. This loss penalizes regions where both SDFs are negative, indicating a volumetric overlap.

Proximity loss. While the intersection term ensures that the resulting SDF_o does not penetrate the hands, it does not guarantee that the object stays in proximity to the hand. Hence, we define the proximity loss using attraction terms with margin δ_{contact},

$$L_{\text{proximity}} = \frac{1}{|V_h^I|} \sum_{v_h \in V_h^I} \max(0, d_{\text{ho}}(v_h) - \delta_{\text{contact}}). \tag{4}$$

where $d_{\text{ho}}(v_h)$ represents one-sided distances from V_h^I to V_o^I. We compute one-sidedly from hand to object because the hand mesh typically has more reliable geometry. This avoids unstable gradients from noisy object predictions and encourages stable, plausible contact during optimization. Both losses provide complementary guidance: the intersection loss enforces separation, while the proximity loss promotes firm contact.

Final Phase 3 loss. The total loss L_{phase3} sums the HOI 2D terms, the intersection and proximity losses, and softly reuses L_{phase1} and L_{phase2} with lower weights to avoid overriding new signals. At each step ($\tau_2 < t \leq 1$) we optimize (T_h, T_o) and steer $\mathbf{v}_t$ for k_3 iterations using L_{phase3}, then update the latent via Eq. 1. At the end of this phase, we decode the final latent using the steered velocity field $\mathbf{v}^*_\theta(x_t, t, \mathbf{c})$, extract the object mesh, and pose both hand and object with the optimized T_h and T_o to obtain the final HOI reconstruction $\widetilde{HOI}$.

5. Experimental Results

We evaluate our approach on three publicly available datasets: OakInk [52], Arctic [17], DexYCB [5]. All datasets include human grasps annotated with 3D hand-object poses, shapes, and meshes. OakInk contains 100 rigid objects, Arctic 11 articulated objects, and DexYCB 20 YCB objects. We randomly sample 1000 images from each dataset as our test sets. Our model is not additionally trained as we use pretrained foundation models.

Comparison Baselines. We compare against the SOTA HOI methods in two groups: deterministic feed-forward (FF) and generative (Gen). For FF methods, following [32], we include IHOI [53], AlignSDF [8] and gSDF [10]. We also include HORT [11], which predicts point clouds; for fair comparison, we convert them to meshes using alpha shapes [15] as in their Supplementary Sec. B.1. Our primary focus, though, is on Gen models, where EasyHOI [32] is the current state-of-the-art leveraging 2D and 3D foundation models for hand–object reconstruction. However, we deviate from EasyHOI's evaluation scheme wherein new test inputs are sampled until a set of 500 successful generations are achieved. Instead, we randomly sample 1000 test images and evaluate all the baseline methods on these test images. Since AlignSDF, gSDF, and HORT were trained on DexYCB, we exclude them from DexYCB evaluation for fairness. Video-based methods are excluded since our approach is single-frame. All the results for all the methods are re-computed on our randomly sampled testset.

Performance Measures. We assess three main aspects of the hand-object interaction reconstruction task. Following [9], we report object accuracy via Chamfer Distance (CD) on 30K point samples and F-scores at 5mm/10mm (F5/F10). All object metrics (CD, F5, F10) capture both object shape and hand–object relative pose errors. For grasp plausibility, we measure hand–object Intersection Volume (I.V.) [32]. These accuracy metrics are computed only on the successfully reconstructed cases, excluding those where no output is provided, e.g., due to missing segmentation or failing optimization. To assess how robust a method is, we compute the Reconstruction Rate (R.R), *i.e.* the fraction of samples over the whole test set that the method produces an output for, regardless of quality. More details regarding evaluation are in the supplemental material.

Implementation Details. All experiments are conducted on an NVIDIA Tesla H100 NVL GPU. Our framework is implemented in PyTorch and integrated with the Hunyuan3D-2 pretrained diffusion backbone. We use the Adam optimizer to update the hand and object transforms and to steer the velocity field. As we optimize with respect to a single input view, the batch size is 1, and we use a total of

Table 1. Quantitative evaluation for HOI reconstruction. We do not evaluate AlignSDF, gSDF, and HORT on DexYCB as they are trained on it. F5 and F10 measure the F-scores of reconstructed object points within 5mm and 10mm of the ground-truth (GT) object, respectively. C.D. refers to the Chamfer Distance (in cm^2) between the reconstructed object and GT object. I.V. refers to Intersection Volume (in cm^3) between the hand mesh and the object mesh. R.R. refers to the Reconstruction Rate, defined as the fraction of objects for which a method produces an output, regardless of quality. Methods are also categorized as feed-forward (FF) or generative (Gen) approaches.

	Method	OakInk					Arctic					DexYCB				
		F5 ↑	F10 ↑	C.D. ↓	I.V. ↓	R.R. ↑	F5 ↑	F10 ↑	C.D. ↓	I.V. ↓	R.R. ↑	F5 ↑	F10 ↑	C.D. ↓	I.V. ↓	R.R. ↑
FF	IHOI	0.132	0.252	3.97	6.20	0.82	0.079	0.148	11.9	4.27	0.63	0.108	0.211	4.87	**6.41**	0.49
	AlignSDF	0.060	0.119	7.75	12.5	0.94	0.042	0.083	17.7	15.6	0.86	–	–	–	–	–
	gSDF	0.047	0.093	8.17	**0.62**	0.87	0.036	0.070	18.9	**1.23**	0.79	–	–	–	–	–
	HORT	**0.319**	**0.508**	2.22	16.6	**0.99**	0.101	0.190	11.0	30.8	0.88	–	–	–	–	–
Gen	EasyHOI	0.109	0.210	4.62	21.31	0.39	0.079	0.145	10.9	18.3	0.36	0.090	0.176	6.26	19.13	0.30
	Ours	0.179	0.322	**1.80**	5.96	0.87	**0.160**	**0.288**	**2.57**	5.08	**0.92**	**0.158**	**0.300**	**2.04**	7.02	**0.58**

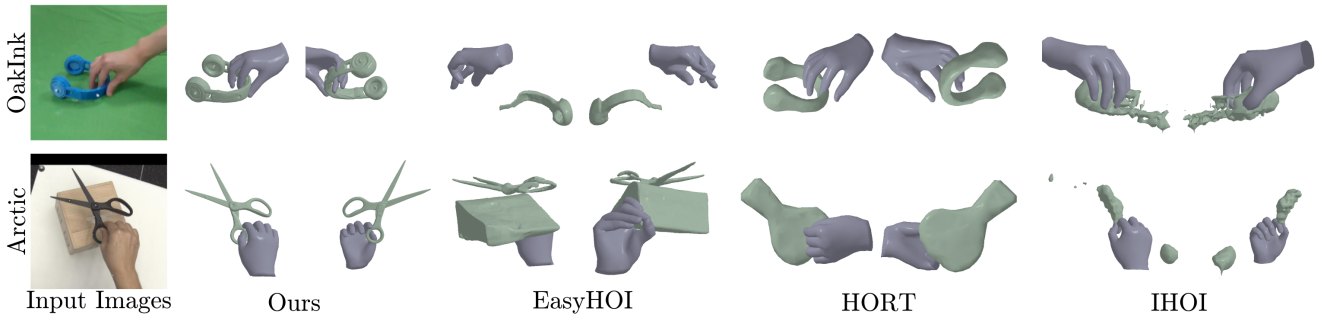

Figure 3. **Qualitative comparison of 3D HOI reconstructions on OakInk and Arctic datasets.** From left to right: input RGB frame; reconstructions produced by our method, by EasyHOI, by HORT, and by IHOI. We evaluate and visualize HORT results after converting the point clouds into meshes following the procedure mentioned in the HORT paper, Supplementary Section B.1. Our approach yields more accurate object geometry with more plausible hand-object interactions.

Figure 4. **Qualitative comparison of 3D HOI reconstructions on DexYCB dataset.** From left to right: input RGB frame; reconstructions produced by our method, by EasyHOI, and by IHOI. Since HORT is trained on DexYCB, we do not include it to the comparison. Our method yields more accurate geometry and plausible interactions.

20 diffusion inference steps, with hand-only optimization conducted at step 9 and gradient-based guidance applied from step 10 onward. The object guidance scale is fixed at 5.0. The remaining hyperparameters for each phase can be found in the supplemental material. Our method approximately takes 6.03 minutes per sample.

5.1. Comparison

We report the quantitative comparison results of our method against the baselines in Tab. 1. **Accuracy:** FollowMyHold achieves the lowest Chamfer Distance across all datasets, indicating superior *global* alignment of reconstructed objects. On OakInk, HORT achieves higher F-scores despite worse CD. This indicates that HORT reconstructions are coarser than us (higher CD) while managing to not fail catastrophically on the challenging cases (higher F5/F10) as also indicated by their high reconstruction rate, due to its feed-forward design. On Arctic, FollowMyHold generalizes better. HORT fails more frequently on thin, fine-grained, or larger tools of Arctic dataset, while FollowMyHold maintains coherent geometry and lower CD. Furthermore, our approach maintains a relatively low intersection volume, significantly outperforming EasyHOI and HORT. Although gSDF yields lower intersection volumes, it compromises significantly on object reconstruction accuracy and robustness. **Robustness:** FollowMyHold achieves successful reconstruction at a significantly higher rate than EasyHOI (87% vs 39% on OakInk), i.e. succeeds on more challenging cases while still being more accurate overall. In fact, our RR is comparable to feed-forward methods that rarely fail except under poor detections/segmentations. This demonstrates the strength of our optimization-in-the-loop guidance, combining accuracy with high success rates among generative methods like EasyHOI.

Figures 3 and 4 show that our method accurately reconstructs object geometries and plausible hand-object interactions, surpassing EasyHOI, HORT, and IHOI. EasyHOI struggles with scale consistency and object completeness,

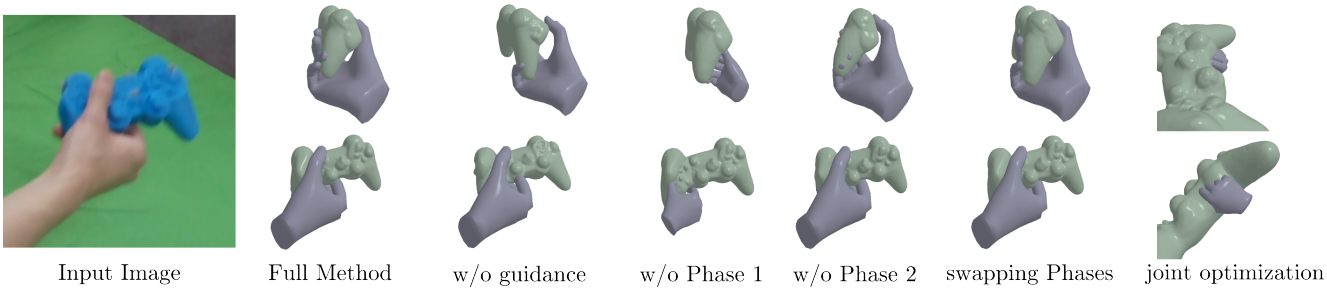

Figure 5. **Ablation of the phased optimization.** We visualize the ablation study on our staged optimization design. We compare HOI reconstruction quality when removing inference-time guidance, dropping Phase 1 or Phase 2, swapping their order, or performing joint optimization. The full method yields the most accurate grasp and object. The bottom row shows the camera viewpoint, while the top row provides an alternative view to highlight differences.

whereas IHOI, while better in grasp accuracy compared to EasyHOI, frequently fails in object reconstruction. HORT struggles on rare, fine-grained geometries (e.g., Fig. 3, second row, fourth column). Point-cloud meshing via alpha-shapes (HORT Supplementary Sec. B.1) can also break slender parts. Further qualitative results are provided in the supplemental material. We also demonstrate our results on in-the-wild images, collected from the internet and our daily lives in Fig. 1, which shows FollowMyHold's robustness and accuracy in real-world examples.

5.2. Ablation Study

We analyze the effects of our staged design and influence of using guidance (Fig. 5 and Tab. 2). Our ablation set consists of 100 random samples from OakInk dataset. Our full method achieves 1.70cm^2 CD. Without guidance (generate object first, then optimize transforms), CD rises to 3.70cm^2, and confirms the importance of guidance for object refinement. Without Phase 1 (no hand-only optimization) CD degrades to 5.67cm^2, confirming the hand as a spatial anchor. Without Phase 2 (no object-only optimization) CD increases to 3.53cm^2 and indicates that object-before-joint refinement is necessary. Swapping the order of Phase 1 and Phase 2 results in 3.18cm^2, which demonstrates the importance of the phase order. When we perform hand-object joint optimization from the start, CD increases to 5.51cm^2 and highlights the importance of having phases of optimization. Additionally, to show that our main performance increase compared to EasyHOI comes from our method design rather than Hunyuan3D-2, we replace EasyHOI's mesh generator with Hunyuan3D-2, which only slightly improves CD (5.23 $\rightarrow$ 5.13cm^2). Our method still achieves a CD nearly 3$\times$ lower. Supplemental material includes the corresponding details and our ablation on loss terms.

6. Limitations and Future Work

Our inference-time guidance trades accuracy and robustness for compute: each step includes inner-loop gradient updates with backpropagation through the latent decoder, increasing runtime. Our method also assumes reliable upstream segmentation and inpainting, however artifacts at this stage can propagate into reconstruction. Another limitation lies in reconstructing thin objects where supervision signals are weaker. Examples are shown in supplemental material.

Future work includes developing more robust fusion of 2D and 3D signals together with learned interaction priors for uncertain regions, integrating our geometric objectives into diffusion training to remove the guidance inner loop, and extending the approach to video with temporal consistency (e.g., optical-flow alignment) for reconstructing stable frame-by-frame sequences for AR/VR and telepresence.

Table 2. Ablation study for each method phase on the ablation set.

Method	C.D. ↓	I.V. ↓
w/o inference-time guidance	3.70	5.83
w/o Phase 1	5.67	7.66
w/o Phase 2	3.53	4.82
swapping Phases	3.18	5.19
joint optimization	5.51	11.7
Full Method	**1.70**	**4.03**

7. Conclusion

We introduced **FollowMyHold**, a single-image HOI reconstruction method that guides a pretrained 3D rectified-flow model at inference using an optimization-in-the-loop design. We show that the proposed inference-time guidance from pixel-aligned 2D cues (normal and disparity alignment, silhouette consistency) together with 3D interaction constraints (intersection, proximity), applied within a staged optimization (hand $\rightarrow$ object $\rightarrow$ joint), produce physically plausible and globally consistent reconstructions. Interestingly, this strategy significantly improves the reconstruction robustness, in addition to the improved object reconstruction quality.

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1067–1076, 2019. 2

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019. 2

[3] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 2

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12417–12426, 2021. 2

[5] 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. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021. 2, 6, 8

[6] Jiayi Chen, Mi Yan, Jiazhao Zhang, Yinzhen Xu, Xiaolong Li, Yijia Weng, Li Yi, Shuran Song, and He Wang. Tracking and reconstructing hand object interactions from point cloud sequences in the wild. *arXiv preprint arXiv:2209.12009*, 2022. 2

[7] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30: 4008–4021, 2021. 2

[8] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *Computer Vision – ECCV 2022*, pages 231–248, Cham, 2022. Springer Nature Switzerland. 6

[9] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European conference on computer vision*, pages 231–248. Springer, 2022. 2, 6, 1

[10] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12890–12900, 2023. 2, 6

[11] Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. HORT: Monocular hand-held objects reconstruction with transformers. In *ICCV*, 2025. 6

[12] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 1, 2

[13] Andreea Dogaru, Mert Özer, and Bernhard Egger. Generalizable 3d scene reconstruction via divide and conquer from a single view. *arXiv preprint arXiv:2404.03421*, 2024. 2, 3

[14] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2024. 2

[15] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983. 6

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2, 6, 1, 7, 8

[18] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. 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*, pages 494–504, 2024. 2

[19] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2

[20] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 2

[21] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 1

[22] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 671–678. IEEE, 2010.

[23] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 2

[24] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid.

Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 1, 2

[25] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021. 1

[26] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[27] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2

[28] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 3

[29] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020. 2

[30] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024. 3, 1

[31] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14687–14697, 2021. 1

[32] Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. *arXiv preprint arXiv:2411.14280*, 2024. 2, 3, 6, 1

[33] Luca Medeiros. Language segment anything. https://github.com/luca-medeiros/lang-segment-anything, 2023. Accessed: 2025-05-23. 2, 3

[34] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[36] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos, 2022. 2

[37] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 2, 3

[38] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024. 2, 3

[39] Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15100–15111, 2023. 2

[40] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems*, pages 20154–20166. Curran Associates, Inc., 2020. 2

[41] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 4

[42] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2

[43] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3405–3414, 2019. 2

[44] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. 2, 3

[45] Bugra Tekin, Federica Bogo, and Marc Pollefeys. 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*, pages 4511–4520, 2019. 1, 2

[46] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1674, 2022. 1

[47] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 2

[48] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. 2, 3, 1

[49] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun

Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pages 57–74. Springer, 2024. 2

[50] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2

[51] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3, 4

[52] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 2, 6, 5, 7

[53] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3895–3905, 2022. 2, 6

[54] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 2

[55] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2354–2364, 2019. 2

[56] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 2, 4, 1

[57] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 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*, pages 5346–5355, 2020. 2