arXiv:2603.25791v1 [cs.CV] 26 Mar 2026

ArtHOI: Taming Foundation Models for Monocular 4D Reconstruction of Hand-Articulated-Object Interactions

Zikai Wang[1] Zhilu Zhang[1,*] Yiqing Wang[2] Hui Li[1] Wangmeng Zuo[1]
[1]Harbin Institute of Technology [2]Shanghai Jiao Tong University

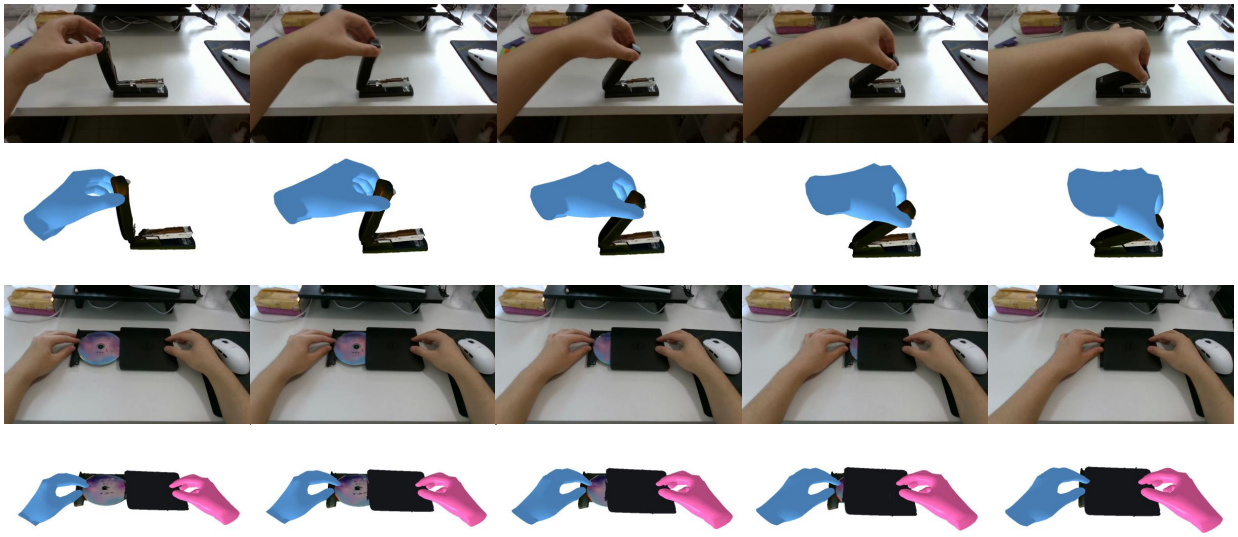

Figure 1. Given a monocular RGB video sequence of hands interacting with an unknown articulated object, our method, ArtHOI, reconstructs 4D human-object interactions (HOI) without any pre-defined object templates or multi-view scan initialization. Here we show two examples of input videos and the reconstructed HOI results.

Abstract

Existing hand-object interactions (HOI) methods are largely limited to rigid objects, while 4D reconstruction methods of articulated objects generally require pre-scanning the object or even multi-view videos. It remains an unexplored but significant challenge to reconstruct 4D human-articulated-object interactions from a single monocular RGB video. Fortunately, recent advancements in foundation models present a new opportunity to address this highly ill-posed problem. To this end, we introduce ArtHOI, an optimization-based framework that integrates and refines priors from multiple foundation models. Our key contribution is a suite of novel methodologies designed to resolve the inherent inaccuracies and physical unreality of these priors. In particular, we introduce an Adaptive Sampling Refinement (ASR) method to optimize object's metric scale and pose for grounding its normalized mesh in world space. Furthermore, we propose a Multimodal Large Language Model (MLLM) guided hand-object alignment method, utilizing contact reasoning information as constraints of hand-object mesh composition optimization. To facilitate a comprehensive evaluation, we also contribute two new datasets, ArtHOI-RGBD and ArtHOI-Wild. Extensive experiments validate the robustness and effectiveness of our ArtHOI across diverse objects and interactions. Project: https://arthoi-reconstruction.github.io.

* Corresponding author. Email: cszlzhang@outlook.com

1. Introduction

Hand-Object Interactions (HOI) reconstruction [7, 8, 12, 14, 16, 40, 56, 60, 69] aims at obtaining a physically plausible 3D representation of hands, objects, and their interplay from visual observations. It plays a crucial role in various applications, including human behavior analysis [25], robotic manipulation [24, 42, 75], and augmented reality [55].

Early works usually require predefined object templates [3, 4, 7, 14, 15, 19] or category-specific knowledge [5, 31, 68], which limited their applicability to unconstrained, wild scenarios. While recent template-free and category-independent methods [1, 12, 40, 58, 60] have demonstrated improved generalization, they largely operate under the assumption of rigid objects. Furthermore, we also note that significant progress [24, 27, 32, 33, 38, 50, 57, 64, 71, 72] has been made in 4D articulated object reconstruction through optimization-based [24, 32, 38, 50, 71] and learning-based [22, 44] techniques, but these methods typically rely on pre-scanning objects (for canonical shape) [24, 50, 64] or even multi-view videos [39, 72]. Consequently, in uncontrolled environments where articulated objects (*e.g.*, scissors, eyeglasses, and laptops) are manipulated naturally, HOI reconstruction from monocular videos remains an unexplored challenge.

It is an inherently ill-posed task due to limited visual cues and frequent occlusions, making the design of an effective and robust method non-trivial. In contrast, humans can effortlessly perceive such complex interactions, a capability that stems from accumulated knowledge and experience. Drawing inspiration from this human faculty, we argue that a promising solution lies in leveraging the rich priors of various foundation models. Specifically, these models can provide critical geometric, motion, and semantic information. For instance, image-to-3D [18, 26, 29, 41, 66] can recover 3D shape of an articulated object, and pose estimation [45, 61] can compute its 6D transformation relative to the camera. Furthermore, depth estimation [6, 51] and tracking [23, 65] can offer metric geometry and motion cues, respectively. For the hand, specialized models [49, 52] can reconstruct its 3D mesh. Multimodal Large Language Models (MLLMs) [10, 59] can infer the interaction state between the hand and the object.

Nevertheless, a naive integration of these foundation models is prone to failure, as their individual predictions sometimes contain inaccuracies and some are not inherently grounded in the physical reality. In particular, image-to-3D models typically generate geometry in a normalized, object-centric coordinate system, lacking the metric scale required to determine the object's true pose in world space. Furthermore, even if the 4D representation of the object is accurately reconstructed, simply composing it with a hand mesh often leads to physically implausible results, such as interpenetration or disjointed contact, due to spatial misalignments between the two.

To address these issues, we propose ArtHOI, a novel framework for reconstructing 4D hand-articulated-object interactions from a monocular video, which optimizes the inconsistency and mismatch problems while collaboratively leveraging priors of foundation models. In particular, firstly, we propose an Adaptive Sampling Refinement (ASR) method to estimate the metric scale and 6-DoF pose of the canonical articulated object. It is used to recover 3D mesh in world space from the generated normalized one and prepare the object motion reconstruction. Secondly, for hand-object mesh composition, we elaborate the prompts for MLLM to infer frame-wise contact states and fingers. The contact information is then used as optimization constraints to jointly refine the object scale and hand pose, improving their spatial alignment.

Specifically, the ArtHOI pipeline mainly comprises four stages: data preprocessing, canonical object mesh reconstruction, part-wise object motion reconstruction, and hand-object alignment. First, the preprocessing stage leverages foundation vision models to extract hand and object masks, metric depths, camera parameters, *etc*. A video inpainting model is applied to restore the object regions occluded by the hand. Second, we deploy an image-to-3D model to generate a normalized 3D mesh from the inpainted object. This mesh is then scaled and oriented in world space using our proposed ASR method. Third, we initialize coarse motion trajectories for each object part using a dense tracking model. These trajectories, along with part visibilities, are then used to solve for the per-part $\mathrm{SE}(3)$ transformations over time. Finally, hand reconstruction is performed, and hand-object interaction is refined via our MLLM-guided alignment method.

To facilitate a more comprehensive evaluation, we supplement the existing RSRD [24] dataset with two new benchmarks: ArtHOI-RGBD, comprising RGBD videos captured with a RealSense camera, and ArtHOI-Wild, consisting of challenging videos collected from the internet. Experiments demonstrate our ArtHOI effectively reconstructs physically plausible 4D HOI across diverse objects and interactions. Notably, our method achieves superior performance even when compared to RSRD [24] that relies on pre-scanned object geometry as input.

Our contributions are summarized as follows:

- We introduce ArtHOI, an optimization-based framework that reconstructs 4D hand-articulated-object interactions from monocular videos via integrating and refining priors from multiple foundation models.
- We propose an Adaptive Sampling Refinement (ASR) method to optimize object's metric scale and pose, which serves object mesh reconstruction in the world space.
- We propose an MLLM-guided hand-object alignment method that performs contact reasoning for constrainting hand-object mesh composition.
- We conduct extensive experiments on existing and newly introduced challenging datasets, which demonstrated the superior robustness and effectiveness of our method across diverse objects and interactions.

2. Related Works

2.1. Hand-Object Interaction Reconstruction

Reconstructing hand-object interaction (HOI) from monocular RGB images or video [1, 4, 5, 7–9, 12, 19, 21, 40, 47, 56, 58, 60, 62, 69] is intrinsically difficult due to severe occlusions and depth ambiguities [1, 12, 60]. Early solutions addressed this by assuming known object templates [3, 4, 7, 14, 15, 19] or pretraining on small-scale 3D object datasets [5, 56, 68]. More recent, model-free approaches exploit priors from large reconstruction or foundation models: some employ pretrained large reconstruction models (LRMs) to obtain an initial object shape [40, 62], while others use novel-view synthesis [60] to recover geometry under sparse view inputs. Nonetheless, many of these methods are restricted to image inputs, rigid-object assumptions, or static contact states [1, 19, 40, 62] during optimization; consequently they do not handle dynamic interactions or complex articulated objects well. Importantly, rich real-world priors can serve not only for shape initialization but also for articulated motion analysis and dynamic contact reasoning. By fully exploiting such priors from multiple foundation models [2, 6, 23, 26, 54, 61], our work advances 4D reconstruction of dynamic hand-articulated-object interactions from casual monocular videos.

2.2. 4D Reconstruction of Articulated Object

Reconstructing real-world articulated objects from limited input remains a challenging problem. Earlier methods typically require 3D point-cloud inputs [22, 37, 46] or multi-view observations [20, 72, 74]; constrained by these requirements, they usually rely on synthesized [13, 34, 43] or laboratory-captured datasets [11, 24] and thus do not generalize well to in-the-wild data. Recent work has begun to reconstruct articulated objects from monocular RGB video captured in the wild [24, 27, 38, 50, 57, 63, 64], achieving promising results by combining flexible 3D representations with rich priors from foundation models such as DINOv2 [48], SAM [54], and dense tracking models [23, 65, 73]. However, most of these approaches assume an initial pre-scanned sequence (object observed from surrounding viewpoints) [24, 38, 50] or depend on predefined part libraries [27, 67]. This initialization provides full-view coverage and a static geometry prior but is impractical for casual capture. Moreover, existing methods typically model only the articulated object and ignore the interacting hand present in real manipulation videos. While effective in controlled settings, these limitations hinder applicability to natural interaction scenarios. By leveraging and coordinating multiple foundation-model priors, our approach relaxes these restrictions, enables joint reconstruction of hands and articulated objects from casually captured monocular interaction videos.

3. Method

Our ArtHOI framework mainly consists of four stages. In Sec. 3.1, we employ a set of foundation models to preprocess the input video and extract multi-dimensional priors. Sec. 3.2 constructs a canonical representation of the articulated object, including its mesh, metric scale, and 6-DoF global pose. In Sec. 3.3, we estimate part-wise $\mathrm{SE}(3)$ motion trajectories from dense tracking priors via an occlusion-aware optimization. Finally, Sec. 3.4 integrates a hand reconstruction model to recover 4D hand mesh, and employs MLLM-guided HOI alignment optimization that resolves spatial mismatches between the reconstructed hands and the object. The pipeline of ArtHOI can be seen in Fig. 2.

3.1. Data Preprocessing

Given a monocular video $\mathcal{V} = \{\mathbf{I}_i\}_{i=1}^N$ of N RGB frames, we first apply several foundation vision models to extract informative priors. Object masks $\{\mathbf{M}_i\}_{i=1}^N$ and human masks are obtained using a video segmentation model [54]. Metric depth maps $\{\mathbf{D}_i\}_{i=1}^N$ and camera intrinsics $\mathbf{K}$ of the input video are estimated with a monocular depth estimator [6]. To mitigate hand-object occlusions, we apply a video inpainting model [30] to remove the human from the input video, producing an inpainted video $\mathcal{V}' = \{\mathbf{I}'_i\}_{i=1}^N$ containing only the object. The inpainted video is further processed with the same preprocessing pipeline to extract object-only masks $\{\mathbf{M}'_i\}_{i=1}^N$ and depth maps $\{\mathbf{D}'_i\}_{i=1}^N$.

We then leverage priors from a large image-to-3D reconstruction model, HunYuan3D [26], to recover the complete geometry of the articulated object. Specifically, let the inpainted canonical frame be denoted by $\mathbf{I}'_c$, we extract the object image from $\mathbf{I}'_c$ using its mask $\mathbf{M}'^o_c$, and feed the cropped object image into HunYuan3D to obtain its 3D mesh.

3.2. Metric Pose and Scale Optimization of Object

Here we align the normalized mesh produced by HunYuan3D with other priors (including the estimated metric depth $\mathbf{D}'^o_c$ and object mask $\mathbf{M}'^o_c$) to obtain a metric canonical mesh in world space. It is achieved by optimizing metric scale s^o_c and 6-DoF pose $\mathbf{T}^o_c$ of the object.

A natural option is to directly apply a state-of-the-art 6-DoF pose estimator, *e.g.*, FoundationPose [61], on the inpainted frame $\mathbf{I}'_c$ with $\mathbf{D}'^o_c$ and $\mathbf{M}'^o_c$. However, while FoundationPose performs well when given accurate metric depth and a metric-scaled ground-truth mesh, its performance degrades notably in our setting due to the inconsistencies between the generated mesh and inaccurate depth, leading to poor or unstable predictions.

To reconcile these heterogeneous priors, we introduce an **Adaptive Sampling Refinement (ASR)** method. ASR first computes a coarse scale estimate for the normalized mesh by using back-projected metric depth, then iteratively samples candidate scales from an adaptive range around initial

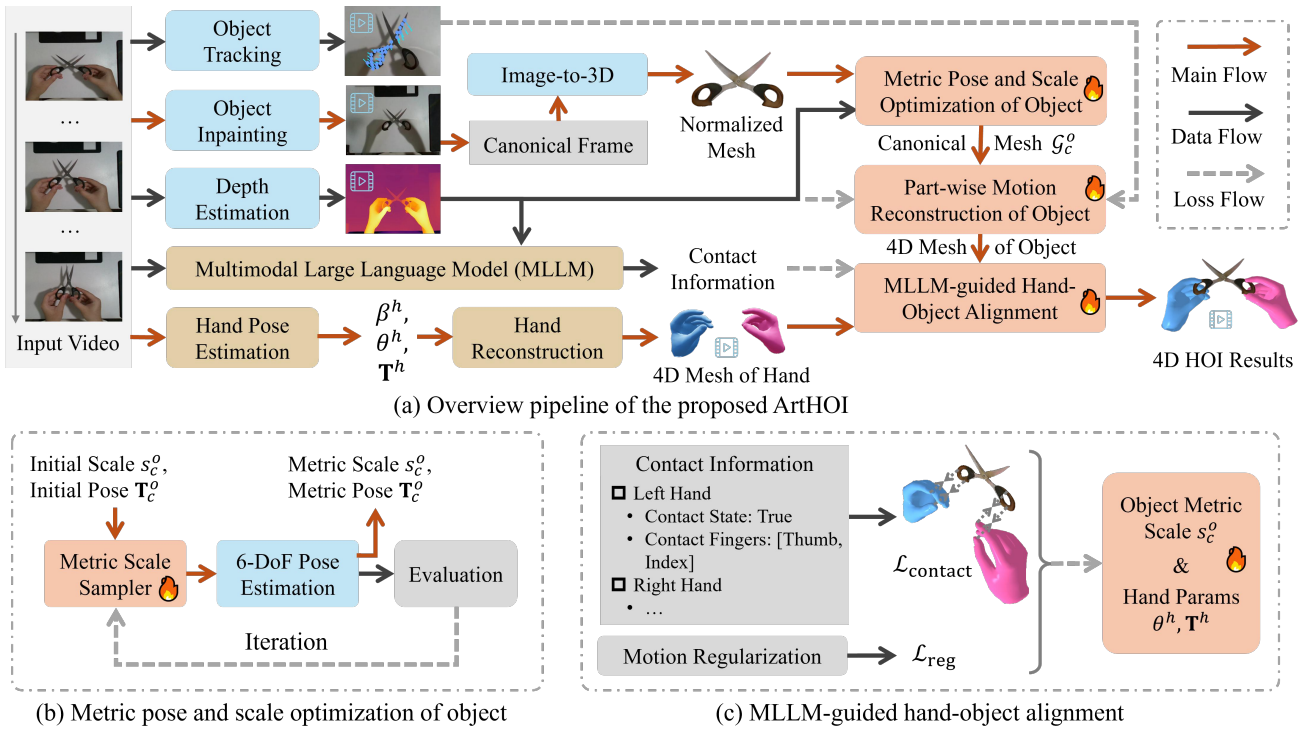

Figure 2. Pipeline of our ArtHOI. ArtHOI is an optimization-based framework (see subfigure (a)) that integrates and refines priors from multiple foundation models for monocular 4D reconstruction of human-articulated-object interactions. In particular, the proposed object's metric scale and pose optimization (see subfigure (b)) recovers 3D mesh in world space from a normalized one, while MLLM-guided hand-object alignment method (see subfigure (c)) promotes physically plausible hand-object mesh composition.

estimate. For each sampled candidate scale, ASR queries FoundationPose to produce pose hypothesis, and evaluates each hypothesis by rendering the posed mesh and matching the rendered silhouette against the preprocessed object mask. The sampling range is adaptively adjusted based on recent refinement progress: if no improvement is observed in recent iterations, the sampling range is expanded; otherwise it is kept unchanged. The algorithm selects the final scale and pose with the best rendered feedback. By searching metric scales and validating pose hypotheses, ASR robustly coordinates the normalized mesh, noisy depth, and pose predictions to yield a reliable metric scale s_c^o and pose $\mathbf{T}_c^o$. The detailed procedure is given in Algorithm 1.

3.3. Part-wise Motion Reconstruction

To effectively exploit both spatial and temporal cues while handling part-wise occlusions, we leverage dense tracking priors [23, 65] to obtain coarse part motions and then optimize per-part SE(3) transformations over time.

Concretely, denote part masks of i-th frame as $\{\mathbf{M}_i'^{p_k}\}_{k=1}^K$, we first partition the canonical object mesh $\mathcal{G}_c^o$ into parts by applying PartField [36] to group vertices and using these masks for partition. We run CoTracker [23] on the inpainted video $\mathcal{V}'$ to produce temporally coherent point tracks together and per-point visibilities. For the k-th part, we sample Q query pixels inside its mask $\mathbf{M}'^{p_k}$ and track sampled queries using CoTrackerV3 [23], which outputs a 2D point trajectory together with a per-frame visibility indicator. Then we lift them to 3D using the depth map $\mathbf{D}_i'$, yielding the 3D track and visibility pair $(\mathbf{z}_{i,q}^k, v_{i,q}^k)$, where $v_{i,q}^k \in \{0,1\}$. Therein, outlier tracks are removed by a lightweight post-processing operation.

We then optimize per-part SE(3) transformations across frames, denoted $\{\mathbf{T}_i^{p_k}\}_{i=1}^N$, by enforcing consistency with 3D tracking priors under visibility constraints. For the k-th part in i-th frame, let $\mathbb{S}$ be a set of sampled reference frames, the tracking loss is

$$\mathcal{L}_{\text{track}} = \sum_{j\in\mathbb{S}} \sum_{q\in\mathbb{W}_{i,j}^k} \left\| \mathbf{z}_{j,q}^k - (\mathbf{T}_i^{p_k})^{-1}\mathbf{T}_j^{p_k}\,\mathbf{z}_{i,q}^k \right\|, \quad (1)$$

where $\mathbb{W}_{i,j}^k = \{q \mid v_{i,q}^k = 1 \wedge v_{j,q}^k = 1\}$ is the set of tracks visible in both frames i and j. To regularize the temporal motion dynamics, we further apply a smoothness constraint:

$$\mathcal{L}_{\text{smooth}} = \sum_{i=2}^{N-1} \left\| \Delta^2 \mathbf{T}_i^{p_k} \right\|. \quad (2)$$

where Δ^2 denotes the discrete second-order difference operator applied along the temporal dimension, *i.e.*, $\Delta^2\mathbf{T}_i^{p_k} = \mathbf{T}_{i+1}^{p_k} - 2\mathbf{T}_i^{p_k} + \mathbf{T}_{i-1}^{p_k}$.

Algorithm 1 Adaptive Sampling Refinement (ASR)

Require: Normalized object mesh $\mathcal{G}^o$; RGB $\mathbf{I}'_c$, depth $\mathbf{D}'_c$ and mask $\mathbf{M}'^o_c$ of canonical frame; camera intrinsics $\mathbf{K}$; number of iterations J; initial sampling range δ

Ensure: Metric scale s^o_c and pose $\mathbf{T}^o_c$ of canonical object, scaled canonical object mesh $\mathcal{G}^o_c$

$s^o_{\text{coarse}} \leftarrow \textsc{CoarseScaleEstimation}(\mathcal{G}^o, \mathbf{D}'_c, \mathbf{M}'^o_c)$
$(\mathcal{L}_{\text{best}}, j_{\text{best}}) \leftarrow (-\infty, 0)$
for $j = 1$ **to** J **do**
  if $j_{\text{best}} < \frac{j}{2}$ **then**
    $\delta \leftarrow 2\delta$ ▷ Adaptively expand the range
  end if
  $\hat{s^o_c} \leftarrow s^o_{\text{coarse}} \cdot \textsc{RandomSample}(-\delta, \delta)$
  $\hat{\mathcal{G}}^o_c \leftarrow \textsc{Scale}(\mathcal{G}^o, \hat{s^o_c})$
  $\hat{\mathbf{T}}^o_c \leftarrow \textsc{FoundationPose}(\hat{\mathcal{G}}^o, \mathbf{I}'_c, \mathbf{D}'_c, \mathbf{M}'^o_c, \mathbf{K})$
  $\hat{\mathbf{M}}^o_c \leftarrow \textsc{RenderSilhouette}(\hat{\mathbf{T}}^o_c \cdot \hat{\mathcal{G}}^o, \mathbf{K})$
  $\mathcal{L}_{\text{iou}} \leftarrow \textsc{iou}(\hat{\mathbf{M}}^o_c, \mathbf{M}'^o_c)$
  if $\mathcal{L}_{\text{iou}} > \mathcal{L}_{\text{best}}$ **then**
    $s^o_c \leftarrow \hat{s^o_c}$, $\mathbf{T}^o_c \leftarrow \hat{\mathbf{T}}^o_c$, $\mathcal{L}_{\text{best}} \leftarrow \mathcal{L}_{\text{iou}}$, $j_{\text{best}} \leftarrow j$
  end if
end for
$\mathcal{G}^o_c \leftarrow \textsc{Scale}(\mathcal{G}^o, s^o_c)$
return $s^o_c, \mathbf{T}^o_c, \mathcal{G}^o_c$

Finally, the overall objective for part-wise motion optimization is formulated as

$$\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{track}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}. \tag{3}$$

3.4. MLLM-guided Articulated HOI Alignment

We employ the off-the-shelf hand pose estimator WiLoR [52] to reconstruct MANO-based 4D hands, parameterized by articulated hand joint poses $\{\theta^h_i\}^N_{i=1} \in \mathbb{R}^{N\times 45}$, hand shape $\beta^h \in \mathbb{R}^{10}$ and global transformation $\{\mathbf{T}^h_i\}^N_{i=1}$. To handle missing or unreliable predictions due to occlusions, we apply spherical linear interpolation (SLERP) on hand pose and global transformation to temporally smooth and fill in the hand poses and transformations.

Separated reconstruction of 4D articulated objects and hands often produces spatio-temporal misalignments due to inconsistencies among different priors, motivating a joint optimization for articulated HOI. To enable dynamic interaction reasoning, we leverage Multimodal Large Language Models (MLLMs) [10, 59] to infer contact information, including the binary contact state and contacting fingers for each frame, leveraging their rich real-world priors and multimodal reasoning capabilities. However, naively querying MLLMs for contact estimation is insufficient: diverse camera viewpoints often lead to left–right hand confusion, while limited RGB cues make it difficult to distinguish true physical contact from mere proximity.

To mitigate these issues, we design a structured prompting strategy. First, we ask the MLLM to determine the camera perspective (egocentric vs. exocentric) of the video and incorporate this information into subsequent contact queries. Next, we infer frame-wise contact information—including hand laterality, binary contact state, and contacting fingers—by iteratively querying each frame with the constructed prompt. To provide richer contextual cues, we concatenate k neighboring RGB frames along with their colorized depth maps to form a large image prompt. This pipeline yields more reliable frame-wise estimates for subsequent optimization. We denote the set of frames where the hand is in contact with the object as $\mathbb{C}$, and the set of contacting fingers in the i-th frame as $\mathbb{Y}_i$.

We leverage the retrieved contact information as frame-wise constraints to guide 4D hand-object interaction alignment. Our optimization follows a two-stage procedure. Given that WiLoR [52] provides reliable metric scale priors of hand, while estimated depth may remain ambiguous, the first stage optimizes only object scale s^o_c to align with the hand. In the second stage, we fix the optimized object scale and jointly refine the hand pose parameters θ^h_i and global transformations $\mathbf{T}^h_i$ to further enhance the spatial consistency between the interacting hand and object.

Let $\mathbb{T}_i$ denote the set of MANO fingertip vertices corresponding to $\mathbb{Y}_i$. The contact loss $\mathcal{L}_{\text{contact}}$ minimizes the distance from each fingertip to the closest point from object mesh $\mathcal{G}^o_i$. It can be written as

$$\mathcal{L}_{\text{contact}} = \sum_{i\in\mathbb{C}} \sum_{\mathbf{v}_t\in\mathbb{T}_i} \min_{\mathbf{v}_o\in\mathcal{G}^o_i} \left\|\mathbf{v}_o - \mathbf{v}_t\right\|_2. \tag{4}$$

To further regularize the optimization, we introduce a motion regularization term $\mathcal{L}_{\text{reg}}$ over hand parameters θ^h_i and global transforms $\mathbf{T}^h_i$. This term combines an acceleration prior on $\mathbf{T}^h_i$ and an ℓ_1 penalty between the optimized pose with the initial pose $\theta^{h,init}_i$, *i.e.*,

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{acc}} \left\|\Delta^2\mathbf{T}^h\right\|_2 + \lambda_\theta \sum_{i=1}^{N} \left\|\theta^h_i - \theta^{h,init}_i\right\|_1. \tag{5}$$

Finally, the overall HOI alignment loss can be written as

$$\mathcal{L}_{\text{hoi}} = \mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{reg}}. \tag{6}$$

4. Experiments

4.1. Datasets

We capture five demonstration sequences of common articulated objects using an Intel RealSense stereo camera at 1280×720 and 30 FPS with accurate metric depth; we denote this collection as **ArtHOI-RGBD**. In addition, we collect eight in-the-wild clips from internet sources and smartphone recordings, denoted **ArtHOI-Wild**. Experiments are

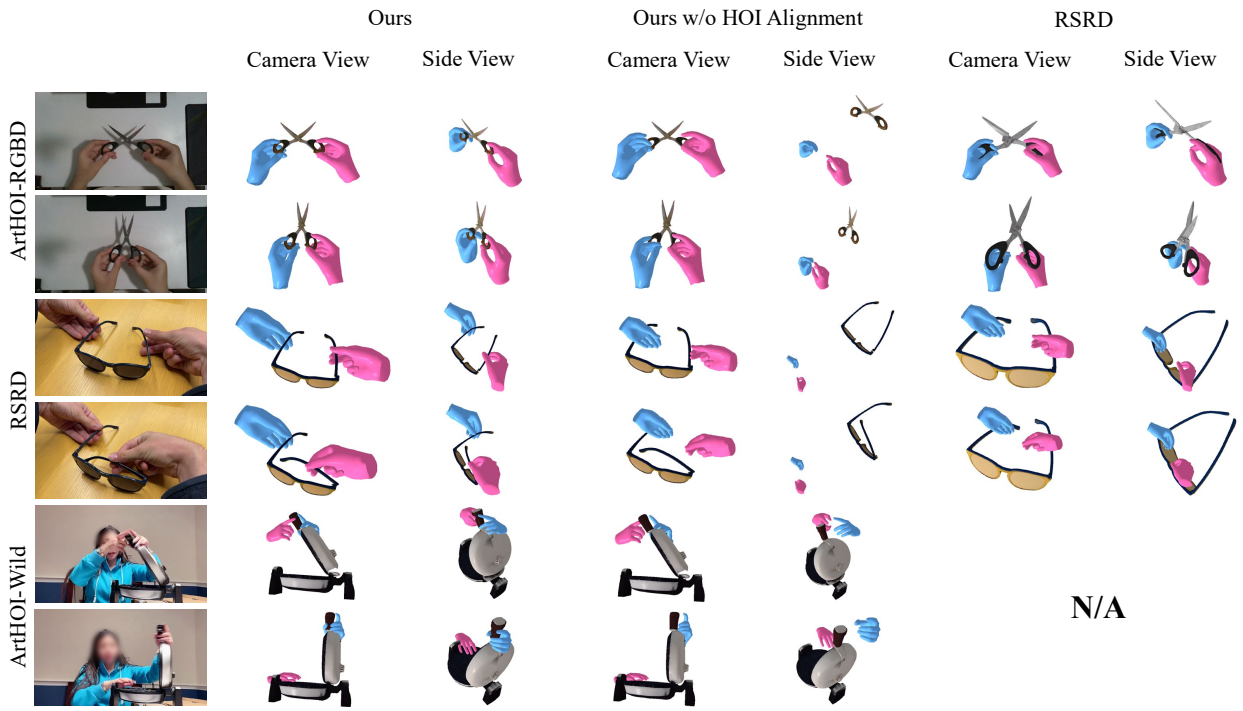

Figure 3. This gallery showcases the results of our hand-articulated-object reconstruction on three data sources: ArtHOI-RGBD, RSRD and ArtHOI-Wild.(more results in the supp.). The first column shows sampled input frames. We present the camera view and a side view to display the reconstructed HOI meshes. Hand reconstructions for RSRD are produced using the same WiLoR model as ours for a fair comparison. Note that RSRD is unable to process the video from ArtHOI-Wild, as it requires an object surrounding scan that is unavailable for internet videos.

performed on these two collections, and we additionally evaluate on nine videos from the RSRD dataset, as well as a three-object subset of ARCTIC [11], covering diverse objects and manipulation scenarios.

Because the ground-truth depth in ArtHOI-RGBD provides only partial surface observations, we develop a 3D annotation tool (built on Viser [70]) to label part-wise object motions across frames for all five videos and four RSRD videos under the help of depth maps as geometric guidance. To obtain complete object geometry, we additionally capture a surrounding scan for each object to reconstruct full ground-truth meshes (used by RSRD). We also annotate hand-object contact states for all used videos.

4.2. Implementation Details

Our system can be implemented on an NVIDIA A6000 GPU, with a total computation time of $\sim$1 hour for a monocular video input with 100 frames under 960×540 resolution. We use Video-Depth-Anything [6] for depth estimation with UnidepthV2 [51] for metric scaling and camera parameter recovery. We adopt Segment-Anything 2 [54] for mask segmentation. DiffuEraser [30] is used for inpainting. The canonical meshes of articulated objects are generated using HunYuan3D [26] from inpainted canonical frames.

In ASR, we run 20 iterations with an initial sampling range $\delta = 0.03$. Part motion reconstruction uses 500 iterations per frame with Adam optimizer and a linearly decayed learning rate from 0.02 to 0.002. The loss weights are set to $\lambda_{\text{match}} = 1.0$ and $\lambda_{\text{smooth}} = 0.01$.

For articulated HOI alignment, we employ Qwen-VL-Max [2] for MLLM-based contact reasoning, followed by 800 optimization steps over all frames with Adam. The learning rate decreases from 10^{-3} to 10^{-4} and the loss weights are set to $\lambda_{\text{contact}} = 1$, $\lambda_{\text{accel}} = 1$, and $\lambda_{\theta} = 50.0$.

4.3. Evaluation Settings

As no existing method reconstructs hand-articulated-object interactions from monocular RGB video without pre-scanned or template object templates, we compare against RSRD [24], a recent 4D articulated HOI reconstruction approach that requires pre-scanned sequences of the object, and EasyHOI[40], a monocular image HOI reconstruction method by apply it frame-by-frame.

For evaluating 4D reconstruction of articulated objects, we report the Chamfer distance (CD) and the Maximum Symmetry-Aware Surface Distance (MSSD) [17] and F-score at 5mm and 10mm thresholds. For evaluating hand-object alignment, we adopt the Collision-Contact (Co^2)

Table 1. 4D reconstruction accuracy of articulated object on monocular RGB videos from ArtHOI-RGBD dataset. Lower CD/MSSD and higher F-scores indicate better performance.

Objects	Method	CD (mm) ↓	MSSD (mm) ↓	F10 ↑	F5 ↑
Headphone	EasyHOI [40]	209.295±107.36	291.02±105.03	1.26	0.59
	RSRD [24]	14.708±0.18	41.06±3.65	41.67	20.91
	Ours	**8.124**±0.44	**30.43**±2.20	**69.68**	**42.19**
Scissor	EasyHOI [40]	170.946±110.29	210.08±114.00	4.14	2.24
	RSRD [24]	13.841±5.89	31.53±5.20	37.98	17.55
	Ours	**4.256**±1.02	**15.14**±3.14	**92.57**	**65.00**
Candy Box	EasyHOI [40]	50.693±55.08	82.36±67.44	25.96	12.94
	RSRD [24]	7.768±4.88	31.40±25.10	83.11	55.25
	Ours	**4.104**±1.33	**17.67**±5.90	**92.55**	**71.63**
CD Drive	EasyHOI [40]	648.704±391.84	827.14±400.53	0.08	0.04
	RSRD [24]	282.330±192.30	348.59±233.49	10.90	6.92
	Ours	**3.334**±0.20	**9.71**±1.89	**96.01**	**78.75**
Stapler	EasyHOI [40]	116.813±96.20	198.25±97.52	8.58	5.28
	RSRD [24]	288.704±144.44	363.92±106.98	0.80	0.34
	Ours	**4.487**±0.97	**20.15**±5.81	**91.63**	**67.94**

Table 2. 4D reconstruction accuracy of articulated object on monocular RGB videos from RSRD [24] dataset. Lower CD/MSSD and higher F-scores indicate better performance.

Objects	Method	CD (mm) ↓	MSSD (mm) ↓	F10 ↑	F5 ↑
Scissor	EasyHOI [40]	205.567±214.40	270.99±232.23	6.04	3.07
	RSRD [24]	68.564±20.94	109.31±19.91	2.09	1.20
	Ours	**5.447**±3.34	**13.12**±6.38	**80.95**	**61.81**
LED Light	EasyHOI [40]	52.435±61.43	100.18±72.43	16.43	7.64
	RSRD [24]	**10.144**±1.31	**32.80**±5.72	**63.21**	**46.21**
	Ours	10.836±7.50	35.30±21.84	60.95	31.81
Bear	EasyHOI [40]	23.785±2.85	77.29±14.88	27.45	12.89
	RSRD [24]	**8.739**±1.03	**28.70**±3.63	**65.63**	**32.73**
	Ours	12.374±2.34	30.31±5.76	45.48	22.86
Sunglasses	EasyHOI [40]	123.385±56.65	304.35±84.76	3.13	1.61
	RSRD [24]	31.985±8.46	164.60±63.75	28.44	15.66
	Ours	**9.956**±4.38	**41.38**±20.84	**65.14**	**44.39**

Table 3. Comparison on a subset of ARCTIC [11]. 'Cont.Acc' denotes binary contact accuracy and 'Fing.Acc' denotes main contacting finger (thumb, index, middle) accuracy of MLLM reasoning results.

	Method	CD(mm) ↓	F10 ↑	MSSD(mm) ↓	Cont.Acc% ↑	Fing.Acc% ↑
Mixer	EasyHOI [40]	226.0	0.05	326.1	N/A	N/A
	RSRD [24]	Failed to reconstruct articulated object				
	Ours	**12.1**	**0.55**	**41.4**	**82.5**	**76.2**
Box	EasyHOI [40]	207.9	0.03	356.4	N/A	N/A
	RSRD [24]	Failed to reconstruct articulated object				
	Ours	**14.0**	**0.44**	**51.2**	**76.6**	**62.3**
Scissors	EasyHOI [40]	436.4	0.01	497.1	N/A	N/A
	RSRD [24]	Failed to reconstruct articulated object				
	Ours	**58.6**	**0.10**	**185.2**	**57.9**	**54.4**

score from Open3DHOI [62] to evaluate 3D interaction quality, computing both contact and collision scores on annotated contact frames, and only the collision score on in-contact frames.

4.4. Quantitative Results

We evaluate our method on three aspects: the accuracy of articulated object reconstruction, the quality of overall hand-object interaction (HOI) alignment and the accuracy of MLLM-driven contact reasoning results.

Articulated Object Reconstruction Quality. We evaluate articulated object 4D reconstruction on annotated sequences from ArtHOI-RGBD, RSRD and ARCTIC [11]. For a fair comparison, 3D Gaussian part representation of RSRD is replaced with the corresponding mesh during evaluation. Tables 1, 2 and 3 shows that, on all five ArtHOI-RGBD sequences featuring challenging hand-part occlusions (e.g., Stapler) and part-part occlusions (e.g., CD Drive), our method achieves consistently lowest reconstruction errors. On the RSRD dataset, our results are comparable to RSRD despite not requiring any pre-scanning. In addition, our approach successfully handles ArtHOI-Wild and ARCTIC videos, whereas RSRD fails due to the absence of a surrounding scan.

Table 4. Comparison of Co^2 scores for unaligned and aligned articulated HOI reconstruction under different contact reasoning strategies. We evaluate four settings: (1) unaligned hand-object reconstruction, (2) RSRD with WiLoR [52] hands, (3) our alignment using a mask-intersection contact heuristic (w/o MLLM), and (4) our full alignment with MLLM-based contact reasoning (w/ MLLM). Lower is better. RSRD fails on ArtHOI-Wild due to missing object-scanning inputs.

	ArtHOI-RGBD	RSRD [24]	ArtHOI-Wild
No Alignment	0.972	0.517	0.514
RSRD [24] + WiLoR [52]	0.392	0.166	N/A
Ours w/o MLLM	0.046	0.035	0.059
Ours w/ MLLM	**0.029**	**0.022**	**0.039**

HOI Alignment Quality. We assess the final HOI alignment using the collision-contact (Co^2) score. Table 4 and Fig. 3 compare unaligned outputs, RSRD (with WiLoR [52] hand estimates) and our MLLM-guided alignment. Our optimization, guided by MLLM-derived contact cues, produces the lowest Co^2 scores and visually plausible, well-aligned 4D reconstructions, outperforming competing strategies that lack scale-aware or temporally consistent contact constraints.

MLLM Contact Reasoning Accuracy. Table 6 reports contact accuracy and the false-positive (FP) rates. To account for temporal ambiguity at interaction boundaries, predictions within $\pm 1 - 3$ frames of the annotated contact window are counted as correct.The results show that our prompting scheme substantially reduces FP while improving accuracy, particularly on in-the-wild data.

4.5. Qualitative Results

Fig. 3 presents qualitative comparisons across all datasets. Our method robustly reconstructs articulated object geometry and motion, together with aligned interacting hands in both controlled and in-the-wild scenarios, demonstrating strong robustness and practical applicability in real-world settings. In-the-wild sequences often exhibit substantial occlusions between hand-part and part-part, making reconstruction particularly challenging. Even under such conditions, our framework maintains coherent geometry and motion across frames by leveraging consistent geometric,

Table 5. Comparison of canonical mesh pose and scale optimization. We compare with FoundationPose and Any6D [28]. Metrics include the IoU between rendered and ground-truth masks under the optimized pose, and the optimization success rate (SR%). A case is considered failed if subsequent part motion reconstruction or HOI alignment cannot proceed.

Method	ArtHOI-RGBD		RSRD [24]		ArtHOI-Wild	
	IoU↑	SR(%)↑	IoU↑	SR(%)↑	IoU↑	SR(%)↑
FoundationPose [61]	0.820	60%	0.706	78%	0.749	71%
Any6D [28]	0.876	60%	0.857	78%	0.683	57%
ASR (ours)	**0.905**	**100%**	**0.876**	**100%**	**0.882**	**100%**

Table 6. Ablation study on prompting strategies for MLLM contact reasoning, evaluated by accuracy and false positive rate (FP, %). "Temp." incorporates temporal context from neighboring frames. "Persp." indicates introducing camera-perspective cues; "MinFP" uses prompts designed to suppress false positives; and "Depth" augments image prompts with colorized depth. Results of ArtHOI-RGBD is excluded due to its near 100% accuracy.

Prompting Strategy				RSRD [24]		ArtHOI-Wild	
Temp.	Persp.	MinFP	Depth	Acc.↑	FP↓	Acc.↑	FP↓
				81.53	18.24	83.50	16.59
✓				82.75	17.08	82.62	17.02
✓		✓	✓	86.42	13.49	85.92	13.27
✓	✓		✓	86.27	13.66	86.21	13.79
✓	✓	✓		87.65	12.13	**87.52**	11.35
✓	✓	✓	✓	**88.58**	**11.20**	86.56	**9.81**

depth, and interaction cues extracted from diverse foundation models. In contrast, RSRD struggles with heavy occlusions and latent ambiguities and fails to produce precise part-motion trajectories. Importantly, our method generalizes robustly to in-the-wild videos, successfully recovering both part motion and hand alignment. In contrast, RSRD and other similar approaches require a pre-scanned object in a canonical state, which is infeasible for internet videos and often unattainable even for lab-captured interaction videos.

4.6. Ablation Study

Effect of Adaptive Sampling Refinement. We evaluate the effectiveness of Adaptive Sampling Refinement (ASR) by comparing it against directly applying FoundationPose [61] using only the coarse scale estimate. We further include Any6D [28], a model-free RGB-D method for scale and 6-DoF pose estimation, as it follows a conceptually similar strategy and can be adapted to our setting. For a fair comparison with Any6D, we use the same HunYuan3D mesh and match the number of scale samples used in ASR. Table 5 reports the 2D silhouette IoU and optimization success rate, where a failure is defined as any case in which subsequent part-motion reconstruction or HOI alignment cannot proceed. ASR achieves the highest IoU and success rates across all videos. In contrast, FoundationPose often fails due to inconsistencies between the generated mesh and noisy depth estimates, while Any6D struggles to recover a valid metric scale owing to its dependence on empirically tuned hyperparameters. Fig. 4 provides a qualitative comparison on an ArtHOI-Wild example.

(a) Input RGB (b) FoundationPose (c) Any6D (d) Ours

Figure 4. Qualitative comparison of metric scale and pose estimation on in-the-wild videos without ground-truth depth. Images are cropped and zoomed-in for better visualization.

Effect of MLLM-guided Hand-Object Interaction Alignment. We ablate the proposed MLLM-guided HOI alignment by comparing against three variants: (1) a baseline that removes the alignment module entirely, (2) RSRD hand-object reconstruction using WiLoR as the hand estimator, and (3) a simple heuristic that infers contact from hand-object mask intersection. As shown in Table 4, excluding MLLM-derived contact cues consistently degrades reconstruction accuracy. Qualitative results in Fig. 3 further highlight that, without scale and spatio-temporal optimization on hand and object parameters, the reconstructed 4D hand and articulated object suffer from severe spatial drift and scale inconsistency, revealing the necessity of MLLM-guided HOI alignment.

Effect of Prompting Strategies in MLLM Reasoning. We ablate the effect of four prompting components: temporal context (Temp.), camera perspective cues (Persp.), false positive suppression (MinFP), and depth-augmented image prompts (Depth) by progressively enabling them and reporting accuracy and FP in Table 6. Incorporating temporal context provides modest gains, while adding perspective reasoning and MinFP prompts substantially reduces spurious contact predictions. Temporal and depth-augmented prompts further improve robustness on challenging in-the-wild videos where single-frame appearance cues are unreliable. The full combination of all components produces the best trade-off, achieving the highest accuracy and lowest FP across both RSRD and ArtHOI-Wild.

5. Conclusion

We presented a method for reconstructing 4D hand-object interactions with articulated objects from monocular videos. Our approach leverages rich priors from multiple foundation models and unifies them through optimization strategies that explicitly handle cross-prior inconsistencies and estimation noise. Extensive experiments on two datasets demonstrate that our model-free method outperforms prior approaches relying on pre-scanned articulated objects, and generalizes effectively to in-the-wild Internet videos, showcasing robust real-world applicability to articulated interactions.

Acknowledgement

This work was partially supported by the National Key RD Program of China under Grant No. 2022YFA1004100 and China Postdoctoral Science Foundation under Grant No. 2025M784371.

References

[1] Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance. *arXiv preprint arXiv:2508.18213*, 2025. 2, 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3, 6

[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 *European Conference on Computer Vision*, pages 361–378. Springer, 2020. 2, 3

[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, 3

[5] etc. Chao, Yu-Wei. Dexycb: A benchmark for capturing hand grasping of objects. In *Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 3

[6] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 2, 3, 6

[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. 1, 2, 3

[8] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-Aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022. 1

[9] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-Driven signed distance functions for 3D hand-object reconstruction. In *CVPR*, 2023. 3

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 5

[11] 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. 3, 6, 7

[12] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, 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. 1, 2, 3

[13] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023. 3

[14] 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. 1, 2, 3

[15] 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. 2, 3

[16] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1

[17] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024. 6

[18] 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

[19] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 3

[20] Xiaoxia Huang, Ian Walker, and Stan Birchfield. Occlusion-aware reconstruction and manipulation of 3d articulated objects. In *2012 IEEE international conference on robotics and automation*, pages 1365–1371. IEEE, 2012. 3

[21] Chaofan Huo, Ye Shi, and Jingya Wang. Monocular human-object reconstruction in the wild. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 5547–5555, New York, NY, USA, 2024. Association for Computing Machinery. 3

[22] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5616–5626, 2022. 2, 3

[23] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-

tracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025. 2, 3, 4

[24] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. In *8th Annual Conference on Robot Learning*, 2024. 1, 2, 3, 6, 7, 8

[25] Takashi Kikuchi and Shun Takeuchi. Self-supervised human-object interaction of complex scenes with context-aware mixing: Towards in-store consumer behavior analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 744–751, 2024. 1

[26] Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025. 2, 3, 6, 4

[27] Long Le, Jason Xie, William Liang, Hung-Ju Wang, Yue Yang, Yecheng Jason Ma, Kyle Vedder, Arjun Krishna, Dinesh Jayaraman, and Eric Eaton. Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. *arXiv preprint arXiv:2410.13882*, 2024. 2, 3

[28] Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6d: Model-free 6d pose estimation of novel objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11633–11643, 2025. 8

[29] Sixu Li, Chaojian Li, Wenbo Zhu, Boyang Yu, Yang Zhao, Cheng Wan, Haoran You, Huihong Shi, and Yingyan Lin. Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–13, 2023. 2

[30] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*, 2025. 3, 6

[31] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10175, 2020. 2

[32] Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 352–363, 2023. 2

[33] Jiayi Liu, Manolis Savva, and Ali Mahdavi-Amiri. Survey on modeling of human-made articulated objects, 2025. 2

[34] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 3

[35] Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9704–9715, 2025. 1

[36] Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9704–9715, 2025. 4

[37] Shaowei Liu, Saurabh Gupta, and Shenlong Wang. Building rearticulable models for arbitrary 3d objects from 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21138–21147, 2023. 3

[38] Yu Liu, Baoxiong Jia, Ruijie Lu, Chuyue Gan, Huayu Chen, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Videoartgs: Building digital twins of articulated objects from monocular video. *arXiv preprint arXiv:2509.17647*, 2025. 2, 3

[39] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Building interactable replicas of complex articulated objects via gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[40] 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. In *CVPR*, pages 7037–7047, 2025. 1, 2, 3, 6, 7

[41] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *CVPR*, 2024. 2

[42] Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026. 1

[43] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 3

[44] Sassan Mokhtar, Eugenio Chisari, Nick Heppert, and Abhinav Valada. Centerart: Joint shape reconstruction and 6-dof grasp estimation of articulated objects. In *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024. 2

[45] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[46] Neil Nie, Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Structure from action: Learning interactions for articulated object 3d structure discovery. *arXiv preprint arXiv:2207.08997*, 2022. 3

[47] Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek.

Bigs: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17437–17447, 2025. 3

[48] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3

[49] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 2

[50] Weikun Peng, Jun Lv, Cewu Lu, and Manolis Savva. iTACO: Interactable Digital Twins of Articulated Objects from Casually Captured RGBD Videos. In *3DV 2026*, 2025. 2, 3

[51] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6

[52] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025. 2, 5, 7, 4

[53] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1

[54] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 6

[55] Mike Salvato, Negin Heravi, Allison M Okamura, and Jeannette Bohg. Predicting hand-object interaction for improved haptic feedback in mixed reality. *IEEE Robotics and Automation Letters*, 7(2):3851–3857, 2022. 1

[56] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 1, 3

[57] Chaoyue Song, Jiacheng Wei, Chuan Sheng Foo, Guosheng Lin, and Fayao Liu. Reacto: Reconstructing articulated objects from a single video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5384–5395, 2024. 2, 3

[58] Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Jean-Sébastien Franco, and Grégory Rogez. Host3r: Keypoint-free hand-object 3d reconstruction from rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7204–7213, 2025. 2, 3

[59] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 5

[60] Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. 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*, pages 5957–5968, 2025. 1, 2, 3

[61] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 2, 3, 8

[62] Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and Yong-Lu Li. Reconstructing in-the-wild open-vocabulary human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17426–17436, 2025. 3, 7

[63] Abdelrhman Werby, Martin Büchner, Adrian Röfer, Chenguang Huang, Wolfram Burgard, and Abhinav Valada. Articulated object estimation in the wild. In *Conference on Robot Learning (CoRL)*, 2025. 3

[64] Mingxuan Wu, Huang Huang, Justin Kerr, Chung Min Kim, Anthony Zhang, Brent Yi, and Angjoo Kanazawa. Predict-optimize-distill: A self-improving cycle for 4d object understanding. *arXiv preprint arXiv:2504.17441*, 2025. 2, 3

[65] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Iurii Makarov, Bingyi Kang, Xin Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy. In *ICCV*, 2025. 2, 3, 4

[66] 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. 2

[67] Xianghao Xu, Yifan Ruan, Srinath Sridhar, and Daniel Ritchie. Unsupervised kinematic motion detection for part-segmented 3d shape collections. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[68] 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, 3

[69] Yu Yang, Zhilu Zhang, Xiang Zhang, Yihan Zeng, Hui Li, and Wangmeng Zuo. Physworld: From real videos to world models of deformable objects via physics-aware demonstration synthesis. *arXiv preprint arXiv:2510.21447*, 2025. 1, 3

[70] Brent Yi, Chung Min Kim, Justin Kerr, Gina Wu, Rebecca Feng, Anthony Zhang, Jonas Kulhanek, Hongsuk Choi, Yi Ma, Matthew Tancik, and Angjoo Kanazawa. Viser: Imperative, web-based 3d visualization in python, 2025. 6

[71] Qiaojun Yu, Xibin Yuan, Junting Chen, Dongzhe Zheng, Ce Hao, Yang You, Yixing Chen, Yao Mu, Liu Liu, Cewu Lu, et al. Artgs: 3d gaussian splatting for interactive visual-physical modeling and manipulation of articulated objects. *arXiv preprint arXiv:2507.02600*, 2025. 2

[72] Tianjiao Yu, Vedant Shah, Muntasir Wahed, Ying Shen, Kiet A Nguyen, and Ismini Lourentzou. Part^{2}gs: Part-aware modeling of articulated objects using 3d gaussian splatting. *arXiv preprint arXiv:2506.17212*, 2025. 2, 3

[73] Bowei Zhang, Lei Ke, Adam W Harley, and Katerina Fragkiadaki. Tapip3d: Tracking any point in persistent 3d geometry. *arXiv preprint arXiv:2504.14717*, 2025. 3

[74] Ge Zhang, Or Litany, Srinath Sridhar, and Leonidas Guibas. Strobenet: Category-level multiview reconstruction of articulated objects. *arXiv preprint arXiv:2105.08016*, 2021. 3

[75] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations. *arXiv preprint arXiv:2501.14208*, 2025. 1

ArtHOI: Taming Foundation Models for Monocular 4D Reconstruction of Hand-Articulated-Object Interactions

Supplementary Material

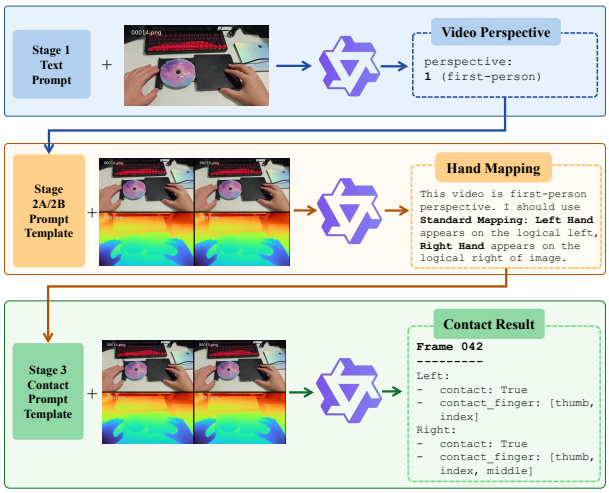

Figure A. Demonstration of our MLLM contact reasoning pipeline. For clarity, we merge 2 neighbouring frames, but in practice, it's typically set to 3. The top row shows RGB frames, the bottom row shows colorized depth maps. The MLLM analyzes visual and depth cues across frames to determine contact status and engaged fingers for each hand.

A. Implementation Details

A.1. Coarse Metric Scale Estimation of Object

We detail the coarse scale estimation introduced in Sec. 3.2. Given estimated metric depth maps, we first back-project them into 3D space using camera intrinsics $\mathbf{K}$ and the object mask. To suppress boundary noise, the mask is eroded prior to back-projection, followed by a Statistical Outlier Removal (SOR) filter to further clean the point cloud. We then compute the bounding boxes of both the normalized canonical object and the back-projected depth point cloud. The coarse metric scale s^o_{coarse} is obtained as the maximum ratio between their extents along the x- and y-axes. The z-axis (depth direction) is excluded because the back-projected point cloud only captures the visible object surface and is typically more noisy and unreliable in depth.

A.2. Object Part Segmentation

Sec. 3.3 describes the reconstruction of part-wise motion for articulated objects. Here, we provide additional details on the part partition process. We begin by applying PartField [35] to extract per-vertex feature fields, followed by agglomerative clustering to obtain vertex group labels.

Table A. Comparison of contact accuracy (Acc.) and false positive rate (FP) between our MLLM-based contact reasoning and a rule-based mask-intersection heuristic. While both methods perform similarly on the controlled RSRD dataset, the heuristic degrades notably on in-the-wild videos, whereas the MLLM remains robust. ArtHOI-RGBD is excluded due to its near-perfect accuracy.

	RSRD [24]		ArtHOI-Wild	
Contact Judge	Acc. ↑	FP ↓	Acc. ↑	FP ↓
Mask Intersection	0.86	0.14	0.76	0.23
MLLM	**0.89**	**0.11**	**0.87**	**0.10**

The object is then rendered in its canonical pose using PyTorch3D [53] to produce a 2D label map. Vertex groups are merged according to part masks, after which the mesh is finally split into individual parts.

A.3. MLLM Contact Reasoning

We adopt an image-text question-answer strategy to extract contact information for each frame of input video. The primary challenge of this task lies in suppressing false positives: in real-world videos, both humans and models often confuse near-contact with genuine physical contact, while clear separation is seldom misidentified as contact, making false negatives comparatively rare. To mitigate this, we augment RGB frames with colorized depth, incorporate neighboring-frame sampling to strengthen spatio-temporal cues, and explicitly instruct the MLLM to be cautious about false positives. Furthermore, because the input videos may be egocentric or exocentric, we identify video perspective beforehand to reduce hallucinations on hand laterality when reasoning about bimanual contact. Figures C, D, and E demonstrate the full prompt templates used in our pipeline.

Input and Output Format To provide richer contextual cues, we concatenate k neighboring frames ($k = 3$ in practice) along with their colorized depth maps into a single large image prompt, which the MLLM can jointly analyze for spatio-temporal consistency. The depth maps are visualized with a color gradient (blue for near, red for far), making depth discontinuities visually salient to the model. The output is a structured JSON containing: (i) frame count and which hands appeared in the video; (ii) for each frame, binary contact flags for left and right hands; (iii) lists of contacting fingers for each hand-frame pair, empty if no contact. This structured format enables downstream optimiza-

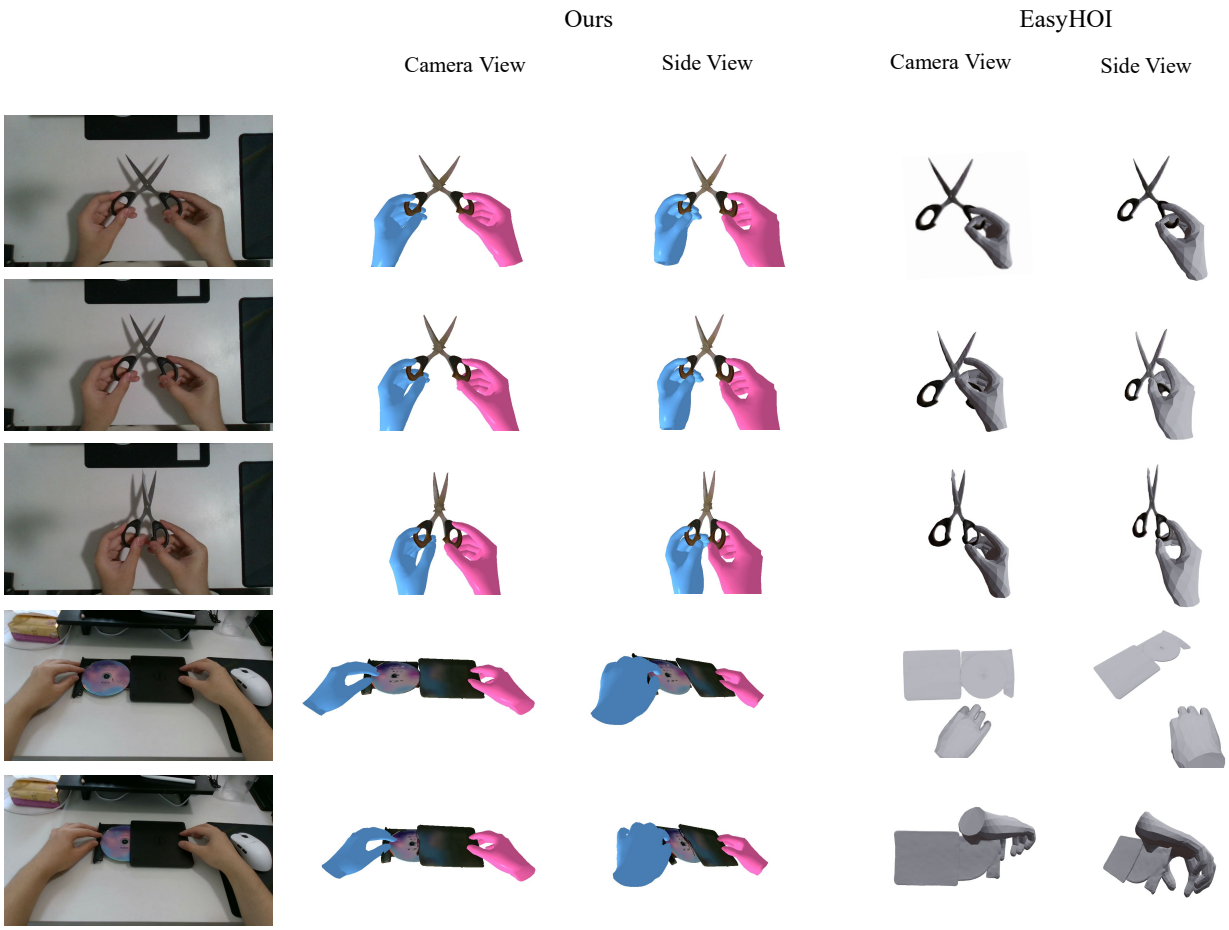

Figure B. Qualitative comparison between our method and EasyHOI [40] on ArtHOI-RGBD. EasyHOI often fails to recover articulated object scale and pose, and exhibits inconsistent hand-object alignment across frames.

tion to directly parse and apply contact constraints.

Three-Stage Prompting Strategy The MLLM contact reasoning pipeline consists of three carefully designed stages, as shown in Figures C, D, and E.

Stage 1: Perspective Detection. Video perspective (egocentric vs. exocentric) significantly affects hand laterality interpretation. In first-person perspective, a single visible hand is automatically from the operator's viewpoint, and spatial relationships are relatively straightforward. In third-person perspective, MLLM must infer the operator's orientation and account for mirror effects to correctly identify hands. By first explicitly determining the perspective, we reduce hallucinations on hand identity in subsequent reasoning stages.

Stage 2: Hand Mapping. After identifying perspective, hand mapping disambiguates left and right hands through perspective-specific heuristics. For first-person videos (Stage 2a), spatial positioning and thumb direction provide direct cues. For third-person videos (Stage 2b), the strategy shifts to analyzing relationship between camera and the operator's body. In this stage, the MLLM can map visible hands to left or right labels.

Stage 3: Frame-wise Contact Reasoning. Given correct hand identity, Stage 3 performs detailed frame-by-frame contact analysis. For each visible hand, the prompt guides the MLLM through a structured reasoning chain. The prompt emphasizes caution: uncertain cases should be marked as no-contact (`false`) to suppress false positives. This conservative bias aligns with our observation that false positive predictions in real-world contact cases are more often than false negatives.

B. Computational Performance

For a video sequence of 150 frames at a resolution of 960×540, preprocessing (mask segmentation, metric depth estimation, frame inpainting, hand estimation, and mesh re-

Stage 1 Perspective Detection Prompt (prompt_perspective)

You are given a set of images sampled from a video about a human manipulating an articulated object.
Please determine whether this video is from a **first-person perspective** or a **third-person perspective**.

If it is first-person perspective, output only the number 1.
If it is third-person perspective, output only the number 3.
Do not output any other text, explanation, or thoughts.

First-person: filmed from the operator's point of view; arms/hands extend from the bottom or sides of the frame; viewpoint aligned with the operator's head direction.
Third-person: filmed from an observer's point of view; shows the whole/most of the operator's body; viewpoint not aligned with the operator's head direction.

Judgment principles:
1. If only one hand appears, it must be first-person perspective.
2. If a human face appears, it must be third-person perspective.
3. In first-person perspective, the hand(s) usually occupy a large area of the image.

Figure C. Stage 1: Perspective Detection Prompt. This prompt determines whether the input video is from a first-person or third-person viewpoint, which is essential for correctly identifying hand laterality in subsequent stages.

Stage 2a Hand Mapping — First-Person

Step 1: Perspective Analysis & Hand Mapping (Reasoning Chain of Thought Example)
This video is **First-person** perspective. The camera mimics the operator's eyes. I need to determine carefully about hand side.
1. If there's two hands, then the hand on the left side of the image is the Left Hand, and the hand on the right side is the Right Hand. If only one hand appears, I need to determine carefully.
1. **Standard Mapping**: Usually, the Left Hand enters from the logical left, and the Right Hand from the logical right.
2. But I also need to examine the thumb direction to determine. Thumb of left hand is pointing right, and thumb of right hand is pointing left.

So I can check the thumb direction to determine the hand side, especially when only one hand appears.
3. **Apply Mapping**: Use these cues to strictly confirm the identity of any visible hands before proceeding.

Stage 2b Hand Mapping — Third-Person

Step 1: Perspective Analysis & Hand Mapping (Chain of Thought)
This video is **Third-person** perspective. You must determine the exact camera angle to identify hands correctly:
1. **Analyze Operator Orientation**: Look at the person's body/head in the RGB frames.

2. **Determine View Type & Arm Connectivity**:
-- **Frontal/Side-Front View**: Camera faces the person. → **Logic**: Mirror effect (Left Hand is on Right, Right Hand is on Left).
-- **Rear/Side-Rear View**: Camera looks at the person's back or side-back. Use **Arm Connectivity** to identify hands:
* **Right Side-Rear**: Camera observes from the operator's right-back. The **Right Arm** is visibly connected to the body on the right. → **Logic**: The hand connected to this visible right arm is the **Right Hand**. The other hand is the Left Hand.
* **Left Side-Rear**: Camera observes from the operator's left-back. The **Left Arm** is visibly connected to the body on the left. → **Logic**: The hand connected to this visible left arm is the **Left Hand**. The other hand is the Right Hand.

3. **Apply Mapping**: Based on the arm connections, strictly assign 'Left Hand' and 'Right Hand' labels before proceeding.

Figure D. Stage 2: Hand Mapping Prompt. This stage identifies and maps visible hands to left/right labels. Stage 2a handles first-person perspective videos using spatial positioning and thumb direction cues. Stage 2b handles third-person perspective videos by analyzing camera angle relative to the operator's body and arm connectivity patterns.

construction with HunYuan3D [26]) requires approximately 10 to 15 minutes. Optimizing the canonical object metric scale and pose takes less than 2 minutes. Part-wise motion recovery is the most time-consuming stage and takes roughly 30 minutes; during this stage, our pipeline could concurrently perform MLLM contact reasoning to obtain HOI contact information. Finally, aligning the separately reconstructed hands and the articulated object requires up to 5 minutes, yielding the final result. Overall, the full pipeline runtime is dominated by the coarse-to-fine part-wise motion reconstruction, which can be accelerated with a more optimized implementation.

For comparison, RSRD [24] reports similar overall runtime: about 40 minutes to reconstruct and segment the 3D part model from pre-scanned video, roughly 7 minutes for part-motion reconstruction and 4D hand estimation, yet it does not perform any hand-object joint optimization.

C. Additional Results

Qualitative Comparison with EasyHOI We compare our approach with EasyHOI [40], a monocular image HOI

Stage 3 Contact Reasoning Prompt (`prompt`, appended after Stage 2)

```
Step 2: Frame-by-Frame Contact Reasoning
The image contains K frames (horizontally merged) separated by black bars.
  -- Top row: RGB frames. -- Bottom row: Depth frames (blue=near, red=far).

For each frame, analyze the 'Left Hand' and 'Right Hand' (identified in Step 1) separately using the following Chain of
Thought:
  A. Visibility Check: Is the hand visible? If not, skip.


    (A.1) Left or Right hand may be occluded by the articulated object. I need to identify partial, occluded hand around
object, not missing them.
  B. Object contact estimation: Is the hand close enough to contact the articulated object (but not background) in the RGB
frame?
    -- If the hand is clearly distant from the object or merely contact, mark contact:false and skip to next.
    (B.1) I need to carefully identify the hand appearance, fully utilizing both RGB and depth.
    (B.2) I need to carefully determine if the hand is contacting the articulated object in a solid state, or it's in mere
contact (which I should output FALSE).


  C. Depth Map Verification (Critical Phase):
    -- Look at the bottom row (Depth map) corresponding to the hand's position.
    -- Does the hand's depth color seamlessly merge with the object's at the interaction point?
    -- Is there a sharp edge or color contrast separating them? If YES -> FALSE CONTACT.
    -- Mark contact: true only if depth values merge without discontinuity.
  D. Finger Analysis:
    -- If contact: true, identify specific fingers (thumb, index, middle) involved.
    -- If a finger is occluded or ambiguous, exclude it.


Step 3: Consistency & Final Decision
  -- Review your frame-by-frame findings.
  -- Ensure the contact status transitions logically (e.g., hand approaches -> touches -> leaves) and is fully supported by the
visual evidence in each frame independently.
  -- Combining the neighbouring frames, make sure judge merely contact frame as FALSE contact.
Step 4: Output Generation
IMPORTANT: If not 100% sure about contact ⇒ output false.


Do not simply output the same status for all frames; look for changes.
Please only output the JSON without any additional text or markdown format. output it just using text.
Output format (JSON only; r_/l_fingers: valid fingers only, empty list if no contact):


{ "frames_cnt": K,
  "appeared": ["left", "right"],  // list only hands that appeared
  "contacts": [

    { "frame": <frame_number noted on the corner of each frame, output the number using int, without .png or output string>,
      "r_contact": <bool>, "l_contact": <bool>,
      "r_fingers": ["thumb","index","middle"], // list valid fingers only, empty if no contact
      "l_fingers": [] },
    ...
  ]
}
```

Figure E. Stage 3: Frame-wise Contact Reasoning Prompt. This stage performs detailed analysis of each frame to determine contact state and identify engaged fingers. The critical depth map verification step (Phase C) distinguishes true physical contact from mere proximity using depth discontinuity analysis.

reconstruction method that also leverages foundation models. Since EasyHOI accepts only single images, we evaluate it frame by frame. For a fair comparison, we use the same foundation models as in our pipeline: WiLoR [52] for hand reconstruction and HunYuan3D [26] for object shape reconstruction.

Figure B shows EasyHOI results on ArtHOI-RGBD using single-frame input. EasyHOI generalizes poorly to articulated manipulation because it assumes a fixed object scale and 6-DoF pose, and instead optimizes camera parameters and object pose to fit each image. While this image-based paradigm can be efficient for isolated frames, it clearly fails to produce coherent results on videos.

Moreover, EasyHOI struggles to maintain consistent hand-object alignment across frames. It optimizes contact by considering the entire plausible hand interaction region, which is sufficient for rigid-object grasps, but without specifying contacting fingers, its performance degrades in articulated interactions. The frame-wise reconstruction paradigm also makes video processing computationally infeasible: reconstructing a 100 frame sequence requires roughly 3 hours or more. Finally, EasyHOI assumes a single-hand setting and cannot be easily extended to bimanual scenes without substantial code modifications.

Effect of MLLM Contact Reasoning We evaluate the effectiveness of our MLLM-based contact reasoning against a simple rule-based baseline that determines contact via mask intersection. As shown in Table A, while the mask-intersection heuristic shows slightly inferior performance on controlled lab datasets, its accuracy drops substantially on casually captured in-the-wild videos. In contrast, the MLLM leverages broader visual and semantic knowledge, enabling more reliable contact judgments under challenging real-world conditions.