HGGT: Robust and Flexible 3D Hand Mesh Reconstruction from Uncalibrated Images

Yumeng Liu[1], Xiao-Xiao Long[2], Marc Habermann[3], Xuanze Yang[1]
Cheng Lin[4], Yuan Liu[5], Yuexin Ma[6], and Ligang Liu[1*]

[1] University of Science and Technology of China
[2] Nanjing University
[3] Max-Planck-Institut für Informatik
[4] Macau University of Science and Technology
[5] Hong Kong University of Science and Technology
[6] ShanghaiTech University

Project Page: https://lym29.github.io/HGGT/

arXiv:2603.23997v3 [cs.CV] 5 Sep 2026

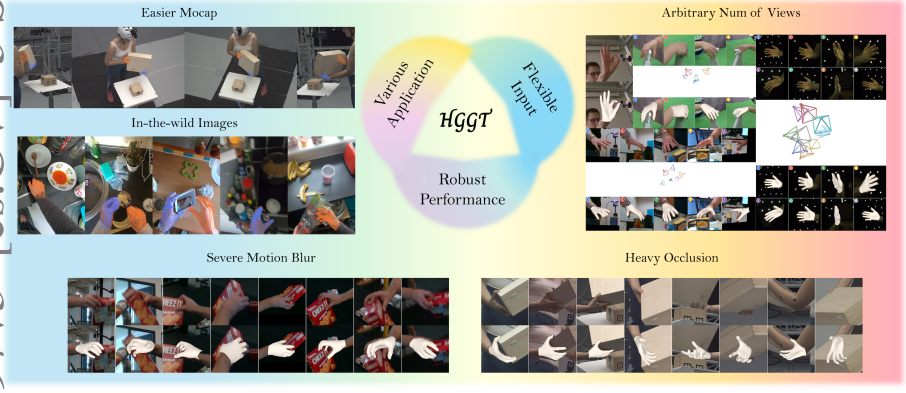


Fig. 1: We introduce **H**and **G**eometry **G**rounding **T**ransformer (**HGGT**), a scalable and generalized solution for 3D hand mesh recovery. Our method unifies diverse data sources to achieve robust performance across varying camera viewpoints and environments.

Abstract. Recovering high-fidelity 3D hand geometry from images is a critical task in computer vision, holding significant value for domains such as robotics, animation and VR/AR. Crucially, scalable applications demand both accuracy and deployment flexibility, requiring the ability to leverage massive amounts of unstructured image data from the internet or enable deployment on consumer-grade RGB cameras without complex calibration. However, current methods face a dilemma. While single-view approaches are easy to deploy, they suffer from depth ambiguity and occlusion. Conversely, multi-view systems resolve these uncertainties but typically demand fixed, calibrated setups, limiting their real-world utility. To bridge this gap, we draw inspiration from 3D foundation models that learn explicit geometry directly from visual data. By reformulating hand reconstruction from arbitrary views as a visual-geometry grounded task, we propose a feed-forward architecture that, for the first time in

* Corresponding author.

literature, jointly infers 3D hand meshes and camera poses from uncalibrated views. Extensive evaluations show that our approach outperforms state-of-the-art benchmarks and demonstrates strong generalization to uncalibrated, in-the-wild scenarios.

1 Introduction

The hand serves as our primary interface for interacting with the world. The detection and reconstruction of the hand has, thus, long been a pivotal task in computer vision, as it is key to a wide range of applications, such as analyzing human behavior [31], enhancing VR realism [13,29], and empowering robots with human-level dexterity [39].

The advent of Vision Transformers (ViT) has recently revolutionized 3D human recovery. HaMeR [28] demonstrated that scaling up ViT backbones can significantly outperform traditional CNN-based regressors for monocular hand reconstruction. However, as a single-view method, it remains vulnerable to depth ambiguity and occlusion. To overcome these monocular limitations, POEM [44] introduced a Point-Embedded Transformer that leverages multi-view geometry to fuse features onto a set of 3D basis points. However, POEM requires accurate camera calibration to perform feature aggregation, which limits its flexibility in in-the-wild, uncalibrated environments.

In parallel, the broader 3D vision community has witnessed a paradigm shift towards feed-forward 3D reconstruction, exemplified by DUSt3R [38], Mast3R [21], and VGGT [36]. These foundation models demonstrate that scene attributes like camera extrinsics, depth, and point tracks can be learned directly from large-scale data via a unified transformer, obviating the need for traditional optimization. Inspired by this, concurrent works such as Human3R [5] and UniSH [22] have attempted to leverage these priors for human pose estimation. However, they typically treat the foundation model as a frozen module and focus solely on aligning human poses with the reconstructed 3D scene. Nevertheless, directly applying such scene-level models to human reconstruction or hand pose estimation in a feedforward manners remains relatively unexplored despite its potential benefits. We found that directly applying these methods to multi-view hand images performs poorly since these images exhibit minimal visual overlap between views. Therefore, these models often struggle to infer accurate camera poses, necessitating a more integrated approach that jointly reasons about hand geometry and camera pose.

To bridge this gap, we introduce a feedforward framework that empowers the VGGT backbone to tackle the challenge of 3D hand reconstruction from uncalibrated multi-view images. Specifically, we append a transformer decoder where learnable hand and camera tokens act as task-specific queries to retrieve geometric and view-dependent cues from the holistic image context via cross-attention. These decoded features are then routed into specialized heads to predict hand pose and camera parameters, respectively. This design effectively disentangles the representation of hand geometry from camera viewpoints, ensuring robust estimation.

To maximize the performance of our proposed framework, extensive data diversity is essential. Specifically, the model must learn to handle arbitrary camera setups and complex backgrounds, which requires training on a rich mix of scenarios. We achieve this by integrating three distinct types of data: First, we incorporate massive **in-the-wild monocular images** to expose the model to diverse environments and lighting conditions, thereby significantly enhancing its robustness and generalization in unconstrained scenarios. Second, **real-world multi-view datasets** provide high quality 3D annotation, ensuring accurate supervision for geometric alignment. Finally, to break free from the fixed camera rigs typical for studio collections, we introduce a **large-scale synthetic dataset** with randomized viewpoints; this is critical for driving the model to learn intrinsic geometric invariance rather than memorizing specific capture setups.

Extensive experiments on public benchmarks demonstrate the effectiveness of our proposed framework. Our model outperforms state-of-the-art methods in both 3D mesh recovery and 2D keypoint detection.

In summary, our main contributions are as follows:

- We propose the **first feed-forward framework** for calibration-free multi-view hand geometry recovery. Our unified transformer architecture jointly estimates camera poses and hand meshes, eliminating the need for pre-calibration or iterative optimization.
- We introduce a new **synthetic dataset** and a **mixed-data training strategy** that effectively leverages real monocular data, real multi-view data and synthetic multi-view data. The diversity of data resource significantly enhances the model's generalization capabilities across different domains.
- We conduct **extensive benchmarking** on open datasets, demonstrating that our method exhibits superior robustness in uncalibrated settings, outperforming existing solutions.

2 Related Work

Multi-view 3D Reconstruction. Reconstructing 3D geometry from multiple images, commonly known as multi-view stereo (MVS), is a fundamental problem in computer vision. MVS techniques aim to recover detailed 3D shapes by leveraging consistent features and geometric cues observed from multiple views. Early traditional methods rely on handcrafted photometric consistency and propagation strategies [10, 33]. The advent of deep learning introduced learning-based MVS, which integrates differentiable geometric modules like voxel-based cost volumes [14, 45] or point-based aggregations [3]. These methods typically utilize camera parameters to perform explicit homography warping for feature alignment across views. Recently, a paradigm shift has occurred, pioneered by DUSt3R [38] and MASt3R [21], which directly estimate aligned dense point clouds from a pair of views. A series of follow-up works [35, 37, 42, 47] have further exemplified this paradigm. VGGT [36] advances this by introducing a scalable framework capable of processing an arbitrary number of views jointly. It

employs a novel Alternating Attention mechanism to capture dense multi-view correlations and directly infer camera parameters and 3D geometry in a single feed-forward pass. Leveraging this robust global geometric understanding, we adopt the VGGT backbone to drive our hand reconstruction. However, naively applying VGGT to uncalibrated in-the-wild images results in erroneous camera estimation and distorted point clouds, we will give a more detailed discussion in Section 4.

Hand Reconstruction. Recovering 3D hand pose and shape from images remains a fundamental challenge in computer vision. Recovering 3D hand pose and shape from images remains a fundamental challenge in computer vision. A significant body of work [27, 40, 48, 49] focuses on the monocular setting, [17, 19, 28, 30] directly regressing the parameters of the differentiable MANO model [32] to constrain the solution space. Another line is non-parametric approach that directly regresses the vertices of the MANO mesh [6,12,20]. To effectively extract features for accurate estimation from a single image, researchers have explored various architectures, ranging from early Graph Convolutional Networks (GCNs) [4, 12] to alternative representations such as voxels [24] or UV maps [2]. Despite these architectural advancements, monocular methods remain fundamentally limited by scale-depth ambiguity. To address this, POEM [44] extended the Transformer paradigm to multi-view settings, modeling the hand mesh as a point set within a multi-view stereo framework. However, this approach heavily relies on pre-calibrated camera parameters to fuse visual features across different views. In contrast, we tackle the more challenging uncalibrated setting. We demonstrate the effectiveness of a VGGT-based architecture in this context. By directly regressing MANO parameters and 2D keypoints, our approach validates that such a concise, direct paradigm yields robust performance, effectively resolving hand reconstruction problem without relying on complex intermediate representations or explicit camera calibration.

3 Method

We present **HGGT**, a feed-forward framework designed to estimate 3D hand pose and shape from uncalibrated multi-view images. We begin by formalizing the problem setting in Section 3.1. Following this, we describe the data preparation process in Section 3.2. The core architecture and design choices of our model are detailed in Section 3.3 with an illustration of our pipeline in Figure 2. Finally, we discuss the optimization objectives in Section 3.4 and the training strategy in Section 3.5.

3.1 Problem Definition

Our goal is to estimate the 3D hand mesh and camera parameters from an arbitrary number of RGB images without prior calibration. Formally, given a set of S images $\{\mathbf{I}_s\}_{s=1}^{S}$ captured from unknown viewpoints, HGGT jointly predicts:

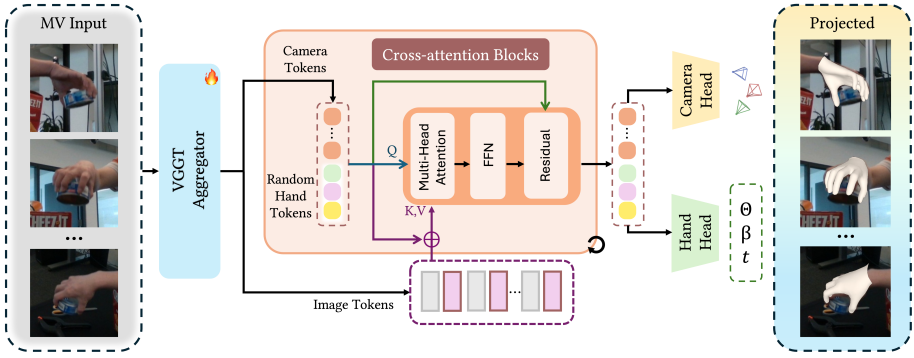


Fig. 2: The pipeline of HGGT. Given uncalibrated multi-view images, we first employ a VGGT Aggregator to extract image tokens and initial camera tokens. These are processed alongside learnable random hand tokens via a series of Cross-attention Blocks. Finally, two parallel heads predict the camera parameters and the canonical MANO parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$, which can be re-projected onto the input views for verification.

(i) the MANO hand parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$ defined in the first camera coordinate system; and *(ii)* view-specific camera pose encodings $\{\mathbf{p}_s\}_{s=1}^{S}$. Here, $\boldsymbol{\theta} \in \mathbb{R}^{48}$ represents the articulated hand pose in axis-angle format, where the first three dimensions correspond to the global orientation; $\boldsymbol{\beta} \in \mathbb{R}^{10}$ denotes the shape parameters; and $\mathbf{t} \in \mathbb{R}^3$ represents the global translation. To this end, we employ distinct prediction heads atop a shared feature backbone, enabling effective multi-task learning and facilitating mutual supervision across tasks.

3.2 Dataset Preparation

We construct a hybrid training dataset by amalgamating three distinct data sources to ensure coverage of both semantic diversity and geometric precision. First, we incorporate large-scale in-the-wild monocular image collections [28] to encompass a broad spectrum of illumination conditions and background complexities. Second, we utilize established real-world multi-view datasets [44] for their high-fidelity 3D annotations; however, acknowledging that these are predominantly captured via fixed rigs with limited extrinsic variation, we further augment our training set with a synthetic component.

Specifically, to introduce randomized camera settings absent in real-world captures, we leverage the pre-computed grasp interactions from GraspXL [46] (based on Objaverse objects [7]). We subsequently render these interactions with photorealistic skin textures using the asserts provided by Dart [11] to minimize the synthetic-to-real domain gap, as visualized in Fig. 3. This synthetic subset significantly expands our data scale, comprising 85,683 image sets for training and 21,242 for validation, with each set containing 10 rendered views.



Fig. 3: Samples from our synthetic dataset. It contains diverse photorealistic hand-object interactions.

3.3 Model Design

An illustration of our pipeline is shown in Figure 2. The pipeline originates with the VGGT aggregator, which encodes multi-view inputs to simultaneously extract image features and initialize camera tokens. These representations are then fed into a **Unified Cross-attention Refinement Module** alongside a set of learnable random hand tokens, where both the camera and hand tokens are iteratively refined. This mechanism allows the model to effectively aggregate visual cues of hand across different views. Finally, the updated tokens are decoded by task-specific heads: a **Camera Head** that predicts the extrinsic parameters for each view, and a **Hand Head** that regresses the MANO parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$ to recover the 3D hand geometry.

Multi-View Feature Aggregation. Our backbone utilizes the VGGT [36] alternating-attention architecture to extract multi-view features. It processes image patches and a learnable *camera token* through interleaved frame-level and global attention layers. We resume the pretrained weight from VGGT and unfreeze all the layer of its aggregator to train on our task.

Unified Cross-Attention Refinement. Following the feature extraction stage, we introduce a set of learnable latent embeddings, referred to as **hand tokens**, to represent the global hand geometry. These tokens are concatenated with the camera tokens (initialized by the VGGT Aggregator) and processed by 4 stacked cross-attention blocks. In this mechanism, the task-specific tokens (i.e., both hand and camera tokens) function as Queries ($\mathbf{Q}$), while the encoded multi-view image features serve as Keys ($\mathbf{K}$) and Values ($\mathbf{V}$). This design allows the hand tokens to explicitly query and aggregate relevant geometric cues from all uncalibrated views, iteratively refining their representations alongside the camera tokens using cross-view information.

Parametric Hand Head. To recover the dense 3D mesh, we decode the refined hand tokens into the parameters of the MANO model [32]. Specifically, the

updated hand tokens are fed into a lightweight Multi-Layer Perceptron (MLP) module named **Hand Head**, which directly regresses the pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and translation $\mathbf{t}$ parameters. These predicted parameters are subsequently passed through the differentiable MANO layer to generate the final 3D hand vertices $\mathcal{V} \in \mathbb{R}^{778\times3}$ and joints $\mathcal{J} \in \mathbb{R}^{21\times3}$.

Camera Pose Estimation Head The camera head estimates the extrinsic and intrinsic parameters for each view relative to the first frame. Specifically, it predicts a 9-dimensional encoding $\mathbf{p}_s = [\mathbf{T}_s, \mathbf{q}_s, \mathbf{f}_s] \in \mathbb{R}^9$ for each view s, comprising the translation $\mathbf{T}_s \in \mathbb{R}^3$, a unit quaternion rotation $\mathbf{q}_s \in \mathbb{R}^4$, and Field of View (FoV) $\mathbf{f}_s \in \mathbb{R}^2$. This head shares the same architecture as the camera head in VGGT [36].

3.4 Loss Function

Our training objective is designed to decouple the optimization of intrinsic hand properties from the geometric imaging process. The total loss $\mathcal{L}_{\text{total}}$ is composed of three distinct components:

Hand Loss. This component focuses exclusively on the reconstruction of the hand mesh in canonical 3D space, independent of the camera viewpoint. We supervise the MANO parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$ and the resulting root-relative 3D joint coordinates J_{3D}. However, since all of single-view datasets lack ground-truth MANO annotations and some of them may only contain 2D annotation lacking 3D joints supervision [28], we apply supervision only when available. Consequently, the loss is formulated as:

$$\mathcal{L}_{\text{hand}} = \mathcal{I}_{\text{mano}} \left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2 \right) + \mathcal{I}_{\text{joints}} \|J_{3D} - \hat{J}_{3D}\|_2^2, \tag{1}$$

where $\mathcal{I}_{\text{mano}}$ and $\mathcal{I}_{\text{joints}}$ are an indicator function that equals 1 if ground-truth MANO parameters or 3D joints are provided, and 0 otherwise (e.g., for single-view samples).

Camera Loss. To ensure geometric consistency between the reconstructed 3D hand and the 2D image plane, we introduce a camera alignment loss. Since relative camera pose estimation relies on multi-view cues, this loss is activated exclusively for samples with multiple views ($S > 1$). Specifically, we enforce explicit constraints on the camera extrinsics as follows:

$$\mathcal{L}_{\text{cam}} = \frac{1}{S} \sum_{s=1}^{S} (\|\mathbf{T}_s - \hat{\mathbf{T}}_s\|_2^2 + \|R(\mathbf{q}_s) \ominus R(\hat{\mathbf{q}}_s)\|_2^2 + ||\mathbf{f}_s - \hat{\mathbf{f}}_s||_2^2), \tag{2}$$

where $\hat{\cdot}$ denotes the predicted parameters, $\mathcal{R}(\cdot)$ is the rotation matrix derived from the quaternion, and $\ominus$ denotes the geodesic distance on $SO(3)$.

Projection Consistency Loss. To bridge the gap between the predicted 3D structure and the observed 2D images, we employ a projection consistency loss. This is particularly crucial for single-view samples where ground-truth 3D annotations are unavailable; in such cases, 2D reprojection serves as the sole supervision to align the predicted hand and camera. The loss comprises two key components:

1. **Reprojection Alignment:** We project the *predicted* 3D joints onto the image plane using the *predicted* camera intrinsics and extrinsics. By minimizing the distance between these projected points and the ground-truth 2D keypoints, we simultaneously constrain the 3D hand geometry and the camera pose:

$$\mathcal{L}_{\text{reproj}} = \frac{1}{S \cdot J} \sum_{i=1}^{S} \| \Pi(J_{3D}^{\text{pred}}, K_{\text{pred}}, [R|T]_{\text{pred}}) - J_{2D}^{\text{GT}} \|_2^2. \tag{3}$$

   This term ensures that the reconstructed 3D hand, when viewed through the predicted camera, visually matches the evidence in the input image.
2. **Negative Depth Penalty:** To prevent degenerate solutions where points are mathematically projected but physically located behind the camera, we impose a negative depth penalty. Given the predicted depth $z_{i,j}$ for the j-th keypoint in the i-th view, this penalty is defined as:

$$\mathcal{L}_{\text{neg}} = \frac{1}{S \cdot J} \sum_{i=1}^{S} \sum_{j=1}^{J} \left(\max(0, -z_{i,j})\right)^2, \tag{4}$$

   where S is the number of views and J is the number of hand keypoints.

Finally, the total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \sum_{l=1}^{L} \gamma^{L-l} (\lambda_{\text{hand}} \mathcal{L}_{\text{hand}}^{(l)} + \lambda_{\text{cam}} \mathcal{L}_{\text{cam}}^{(l)}) + \lambda_{\text{proj}} \mathcal{L}\text{proj}^{(L)}, \tag{5}$$

where L denotes the total number of cross-attention blocks, and the superscript (l) indicates the predictions from the l-th block. We apply intermediate supervision to the hand and camera parameters with exponentially increasing stage weights controlled by $\gamma = 0.6$. The 2D projection loss $\mathcal{L}_{\text{proj}}$, however, is applied exclusively to the final output to ensure terminal geometric consistency.

3.5 Training Strategy

A major challenge in 3D hand reconstruction is the scarcity of large-scale datasets with diverse viewpoints and 3D annotations. While multi-view datasets provide essential geometric cues, they are often limited in scale. Conversely, single-view datasets are abundant but typically contain only 2D keypoint annotations. To maximize our model's performance and generalization capability, it is crucial to harness the complementary strengths of these data sources.

However, integrating these datasets in a unified training framework is non-trivial due to structural heterogeneity. First, the **variable input cardinality** prevents standard uniform batching, where single-view samples ($S = 1$) coexist with multi-view samples ($S \in [2, 24]$). Second, this leads to **computational imbalance**, as resource consumption scales linearly with view counts, rendering static batching strategies inefficient. Finally, the **training objectives differ**: single-view data assumes a fixed canonical camera, whereas multi-view data necessitates additional camera and projection losses to enforce geometric consistency.

To address these issues, we employ a specialized training strategy via gradient accumulation. By accumulating gradients over alternating single view and multi-view steps, we simulate a joint batch that balances these heterogeneous objectives within a single weight update. Following the efficient protocol of VGGT [36], we enforce a constant image budget N_{img} to maximize hardware utilization, dynamically adjusting the batch size to $B = \lfloor N_{\text{img}}/S \rfloor$ for each step. By accumulating gradients over alternating single view and multi-view steps, we simulate a joint batch that optimizes both objectives simultaneously.

4 Experiments

4.1 Implementation Details

We implement our framework using PyTorch, training a model with approximately 1.4 billion parameters on a cluster of 8 NVIDIA A100 GPUs. The optimization utilizes AdamW [23] with a weight decay of 0.05. To ensure stable convergence given the memory constraints, we set the micro-batch size to 32 images per GPU and perform gradient accumulation over 4 steps before each weight update. The learning rate follows a composite schedule: it linearly warms up from 1×10^{-8} to 1×10^{-6} during the first 5% of the 80,000 total iterations, followed by a cosine decay to 1×10^{-8}, with the entire training process taking approximately 5 days. We employ an alternating training strategy between single-view and multi-view data. For single-view data, we fix the number of views to $N = 1$. For multi-view data, the maximum number of sampled frames depends on the available camera views provided by each dataset. Specifically, we dynamically sample $N \in [2, 10]$ frames per sequence for our synthetic dataset, $N \in [2, 5]$ for HO3D, $N \in [2, 8]$ for DexYCB, ARCTIC, and InterHand, and $N \in [2, 4]$ for OakInk. Input images are processed with a patch size of 14 and resized to a maximum dimension of 518 pixels, with aspect ratios randomized between 0.5 and 1.0. Finally, to enhance robustness, we apply photometric augmentations same as VGGT [36], including color jittering, Gaussian blur, and grayscale conversion. During evaluation, we use the original image order to evaluate our method.

4.2 Training Data Details

To ensure robust generalization and accurate 3D hand reconstruction across diverse scenarios, we train our model on a comprehensive mixture of single-view

and multi-view datasets, encompassing both real-world captures and synthetic renderings.

Single-View Datasets For single-view supervision, we adopt the extensive data collection curated by HaMeR [28]. This compilation consolidates multiple datasets providing 2D or 3D hand annotations, specifically FreiHAND [51], HO3D [15], MTC [40], RHD [50], InterHand2.6M [25], H2O3D [15], DEX YCB [1], COCO WholeBody [18], Halpe [9], and MPII NZSL [34]. This yields approximately 2.7M training examples. While the majority of these images are captured in controlled environments, around 5% of the training examples (from COCO WholeBody, Halpe, and MPII NZSL) consist of in-the-wild images, which significantly enhances the model's robustness to diverse backgrounds and lighting conditions.

Multi-View Real Datasets For real-world multi-view data, we utilize the standardized collection introduced by POEM [44]. This integrates observations and 3D annotations from multiple calibrated cameras across several prominent hand-object interaction datasets, including DexYCB [1], HO3D [15], OakInk [43], ARCTIC [8], and InterHand2.6M [25]. In total, this compilation provides approximately 438K multi-view training frames (comprising roughly 25K, 9K, 58K, 135K, and 210K frames from the respective datasets).

Multi-View Synthetic Datasets A major limitation of existing real-world multi-view datasets is that they are typically captured under fixed camera configurations, leading to a severe lack of viewpoint diversity. To address this issue, we synthesize a large-scale multi-view dataset that provides highly diverse camera viewpoints. Specifically, we render hand-object interaction sequences utilizing the rich kinematic and geometric data from GraspXL [46]. To ensure high visual realism and appearance diversity, we apply a wide variety of high-quality object materials sourced from the Objaverse [7] and DART [11] datasets during the rendering process. This synthetic subset significantly expands our data scale, comprising 85,683 multi-view frames for training and 21,242 for validation. With each frame containing 10 rendered views, this yields a substantial total of 856,830 synthetic training images. In conclusion, this synthetic data serves as a valuable complement to existing real-world datasets, providing diverse camera viewpoints, realistic rendering results, and rich annotations (including hand pose, as well as object and hand masks and depth maps).

4.3 Evaluation Metric

We report the Mean Per Joint Position Error (**MPJPE**) and Mean Per Vertex Position Error (**MPVPE**) in millimeters. In a our setting, the absolute depth of the wrist is inherently unobservable due to scale ambiguity. While our predicted camera layout is structurally consistent with the ground truth, as shown in the

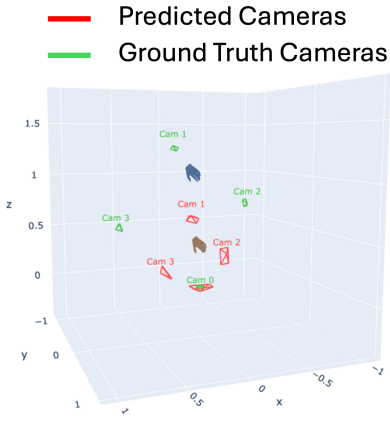


Fig. 4: Camera Layout Visualization.

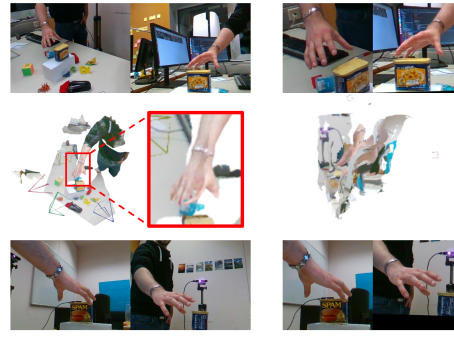


Fig. 5: Failure cases of off-the-shelf VGGT on HO3D datasets.

example of Figure 4, aligning the global coordinate system requires known wrist depth, which is mathematically equivalent to the Root-Relative (RR) evaluation setting. Therefore, to ensure a fair and valid comparison, we strictly report results under root-relative alignment and Procrustes analysis (PA). Additionally, we evaluate the Area Under the Curve (AUC) of the Percentage of Correct Keypoints (PCK) to assess prediction robustness.

4.4 Limitations of Off-the-shelf VGGT.

We observe that directly feeding multi-view hand images into the off-the-shelf VGGT results in severe misalignment, regardless of the input strategy. As shown in Figure 5, when using cropped images (the standard protocol for hand reconstruction), the minimal visual overlap between sparse views causes the geometric estimation to collapse entirely. However, simply reverting to full-frame images does not resolve the issue. As shown in the left of Figure 5, although the background context helps establish a rough global coordinate system, the specific point cloud of the hand remains poorly aligned. Our quantitative analysis in Table 1 also confirms that the direct use of off-the-shelf VGGT leads to sub-optimal results for current state-of-the-art methods. This is because VGGT's camera estimation is dominated by static, high-texture background features, failing to capture the fine-grained geometric consistency required for the relatively small and weakly textured hand region.

4.5 Main Results

We present the quantitative comparison with the state-of-the-art method POEM in Table 1. It is important to note that POEM relies on ground-truth (GT) camera parameters as input, while our method aims for reconstruction from uncalibrated images. To enable a fair evaluation under practical settings, we additionally evaluate POEM using predicted cameras from two sources: (1) the original VGGT estimation, and (2) our proposed camera prediction module.

Table 1: The "Cali-Free" column denotes calibration dependency. Rows in gray (marked with ✗) represent methods utilizing ground-truth camera parameters, serving as an *upper-bound reference*. Rows in black (marked with ✓) represent calibration-free approaches. Within the calibration-free setting, we highlight the **best** results in bold and the second-best with an underline.

datasets	#	methods	Cali-Free	Root Relative				Procrustes Alignment			
				MPVPE$_\downarrow$	AUC$_{V20\uparrow}$	MPJPE$_\downarrow$	AUC$_{J20\uparrow}$	MPVPE$_\downarrow$	AUC$_{V20\uparrow}$	MPJPE$_\downarrow$	AUC$_{J20\uparrow}$
DexYCB-Mv (8 views)	1	POEM$_{v1}$	✗	7.21	0.65	7.30	0.62	4.00	0.80	3.93	0.78
	2	POEM$_{v2}$-large	✗	7.50	0.64	7.54	0.63	3.84	0.81	3.68	0.81
	3	VGGT+POEM	✓	39.08	0.10	39.99	0.10	11.06	0.54	10.20	0.53
	4	Ours+POEM	✓	19.69	0.29	20.18	0.27	**4.38**	0.78	**4.19**	**0.79**
	5	Ours	✓	**13.29**	**0.41**	**13.79**	**0.39**	4.47	**0.78**	4.57	0.77
OakInk-Mv (4 views)	6	POEM$_{v1}$	✗	7.63	0.64	7.46	0.63	4.21	0.79	4.00	0.78
	7	POEM$_{v2}$-large	✗	8.47	0.59	8.47	0.58	4.49	0.78	4.24	0.77
	8	VGGT+POEM	✓	47.29	0.12	48.66	0.12	13.18	0.48	12.08	0.47
	9	Ours+POEM	✓	26.58	0.22	27.34	0.21	**5.49**	**0.73**	**5.14**	**0.72**
	10	Ours	✓	**20.13**	**0.37**	**20.59**	**0.37**	6.14	0.70	6.12	0.70
InterHand-Mv (8 views)	11	POEM$_{v2}$-large	✗	9.04	0.62	8.82	0.60	5.23	0.76	4.72	0.70
	12	VGGT+POEM	✓	49.92	0.07	50.57	0.08	20.16	0.27	18.49	0.23
	13	Ours+POEM	✓	**19.18**	**0.29**	**19.47**	**0.27**	**5.98**	**0.73**	**5.37**	**0.69**
	14	Ours	✓	23.22	0.20	27.46	0.20	6.05	0.71	13.81	0.59
ARCTIC-Mv (8 views)	15	POEM$_{v2}$-large	✗	7.13	0.64	6.80	0.65	4.43	0.78	3.91	0.78
	16	VGGT+POEM	✓	20.19	0.25	20.67	0.24	8.21	0.60	7.29	0.62
	17	Ours+POEM	✓	19.67	**0.25**	**19.93**	**0.25**	**5.35**	**0.73**	**4.75**	**0.74**
	18	Ours	✓	**19.48**	0.21	20.04	0.19	6.57	0.67	6.61	0.67
					AUC$_{V50\uparrow}$		AUC$_{J50\uparrow}$		AUC$_{V50\uparrow}$		AUC$_{J50\uparrow}$
HO3D-Mv (5 views)	19	POEM$_{v1}$	✗	21.45	0.58	21.94	0.55	9.97	0.80	9.60	0.78
	20	POEM$_{v2}$-large	✗	10.58	0.79	10.58	0.79	4.59	0.91	4.13	0.91
	21	VGGT+POEM	✓	56.07	0.23	57.53	0.20	21.05	0.59	20.29	0.47
	22	Ours+POEM	✓	28.48	0.48	29.37	0.46	6.05	0.88	5.57	0.88
	23	Ours	✓	**9.98**	**0.80**	**10.03**	**0.80**	**3.30**	**0.93**	**3.33**	**0.93**

Comparison with GT-based Methods. Despite operating without ground-truth camera information, our method achieves performance comparable to POEM (with GT inputs). Notably, on the HO3D dataset, our approach even surpasses the GT-aided POEM, demonstrating the robustness of our learned geometry.

Comparison under Uncalibrated Settings. When POEM is restricted to using cameras predicted by off-the-shelf VGGT, our method significantly outperforms it. As shown in Table 1, equipping POEM with our estimated cameras yields better results than the VGGT baseline. Despite this improvement, our method still surpasses POEM across the majority of benchmarks, with only a marginal deficit observed on the OakInk dataset. This highlights the robustness of our pose estimation framework, demonstrating its ability to maintain high accuracy across diverse geometric scenarios despite the challenges of uncalibrated inputs.

Table 2: Ablation Studies. We analyze the impact of (a) freezing different stages of the VGGT backbone and (b) incorporating our rendered synthetic dataset. The default setting is highlighted in gray.

Experiment	Setting	Trainable Params	Root Relative				Procrustes Alignment			
			MPVPE↓	AUC_{V50} ↑	MPJPE↓	AUC_{J50} ↑	MPVPE↓	AUC_{V50} ↑	MPJPE↓	AUC_{J50} ↑
(a) VGGT Freezing Strategy (on Real+Syn Data)										
Freeze Layers	24	0.46B	25.87	0.51	26.41	0.50	17.82	0.65	18.19	0.64
	20	0.56B	18.16	0.65	17.60	0.66	13.72	0.73	14.06	0.72
	10	0.81B	12.57	0.75	12.88	0.74	4.32	0.92	4.39	0.92
	None	1.37B	**9.98**	**0.80**	**10.03**	**0.80**	**3.30**	**0.93**	**3.33**	**0.93**
(b) Training Data Composition										
Data Source	Real Only	-	15.30	0.61	15.33	0.61	5.97	0.89	6.04	0.88
	Real + Synthetic	-	**9.98**	**0.80**	**10.03**	**0.80**	**3.30**	**0.93**	**3.33**	**0.93**

4.6 Ablation Study

We conduct ablation studies to validate the effectiveness of our design choices, focusing on the freezing strategy of the VGGT backbone and the contribution of synthetic training data. We evaluate different experiment setting on HO3D dataset, and all results are reported in Table 2.

Impact of Backbone Freezing. We first investigate how different freezing strategies affect the model performance. As shown in the top section of Table 2, fixing the backbone parameters leads to suboptimal performance. Specifically, the freezing none setting significantly outperforms the freeze all (24 layers) setting, reducing MPJPE by a large margin. This suggests that although the pre-trained VGGT features are robust, the domain gap between the pre-training task and 3D hand reconstruction is non-negligible. A full trainable VGGT backbone allow the model to adapt its feature representation to capture the geometric details required for accurate hand mesh recovery.

Effectiveness of Synthetic Data. We further analyze the impact of incorporating our rendered synthetic dataset into the training pipeline. Comparing the *Real Only* baseline with our final *Real + Synthetic* setting in Table 2, we observe a consistent performance boost across all metrics (e.g., MPVPE decreases significantly). We attribute this improvement to two factors: (1) The synthetic data provides perfect ground-truth annotations, mitigating the noise inherent in real-world labels; (2) The diverse poses and viewpoints in the synthetic set act as strong regularization, preventing the model from overfitting to the limited distribution of real-world data.

4.7 Qualitative Comparison

Figure 6 visualizes the reconstruction results of our method compared to existing baselines on the InterHand2.6M [25], OakInk [43], and HO3D [15, 16] datasets. For more comprehensive visualizations across additional datasets, please refer to the Supplementary Material. As shown in Figure 6, the baseline using VGGT-predicted cameras (row a) often suffers from significant misalignment due to

inaccurate camera estimation. In contrast, our full method (row d) achieves reconstruction quality comparable to the baseline using Ground Truth cameras (row b). This demonstrates that our approach effectively recovers accurate camera parameters and hand geometry, producing tight alignment and physically plausible poses without relying on ground truth intrinsics.

To evaluate the robustness of our framework, we test it under extreme conditions as shown in Figure 7. Despite the challenges posed by severe motion blur (top row), where high-frequency edge details are degraded by rapid hand movements, our method successfully recovers the global hand pose and maintains structural integrity. Furthermore, our approach demonstrates remarkable stability under heavy occlusion (bottom row). In scenarios where the hand is heavily obstructed by large objects our model effectively leverages the sparse visual cues combined with learned priors.

4.8 Limitation and Future Work

Similar to state-of-the-art methods like HaMeR [28] and POEM [44], our framework currently relies on off-the-shelf 2D detectors (e.g., ViTPose [41]) for hand localization and cropping in unconstrained settings. Furthermore, as the estimated camera parameters are defined within a normalized coordinate space, recovering the absolute metric depth of the wrist from uncalibrated images remains an open challenge. Addressing metric scale recovery from uncalibrated views represents a promising direction for future research. Beyond isolated hand reconstruction, our large-scale synthetic data offers rich geometric annotations that are often scarce in real-world hand-object interaction datasets. We hope that both the proposed HGGT framework and this dataset can contribute meaningfully to future efforts in broader tasks, such as robotic grasping and human-object interaction reconstruction.

5 Conclusion

In this work, we introduced the first feed-forward framework for calibration-free multi-view hand reconstruction. Supported by our mixed-data training strategy and our newly created large-scale synthetic dataset of rendered multi-view hand-object data, our model demonstrates superior robustness and generalization, outperforming state-of-the-art methods on uncalibrated scenarios.

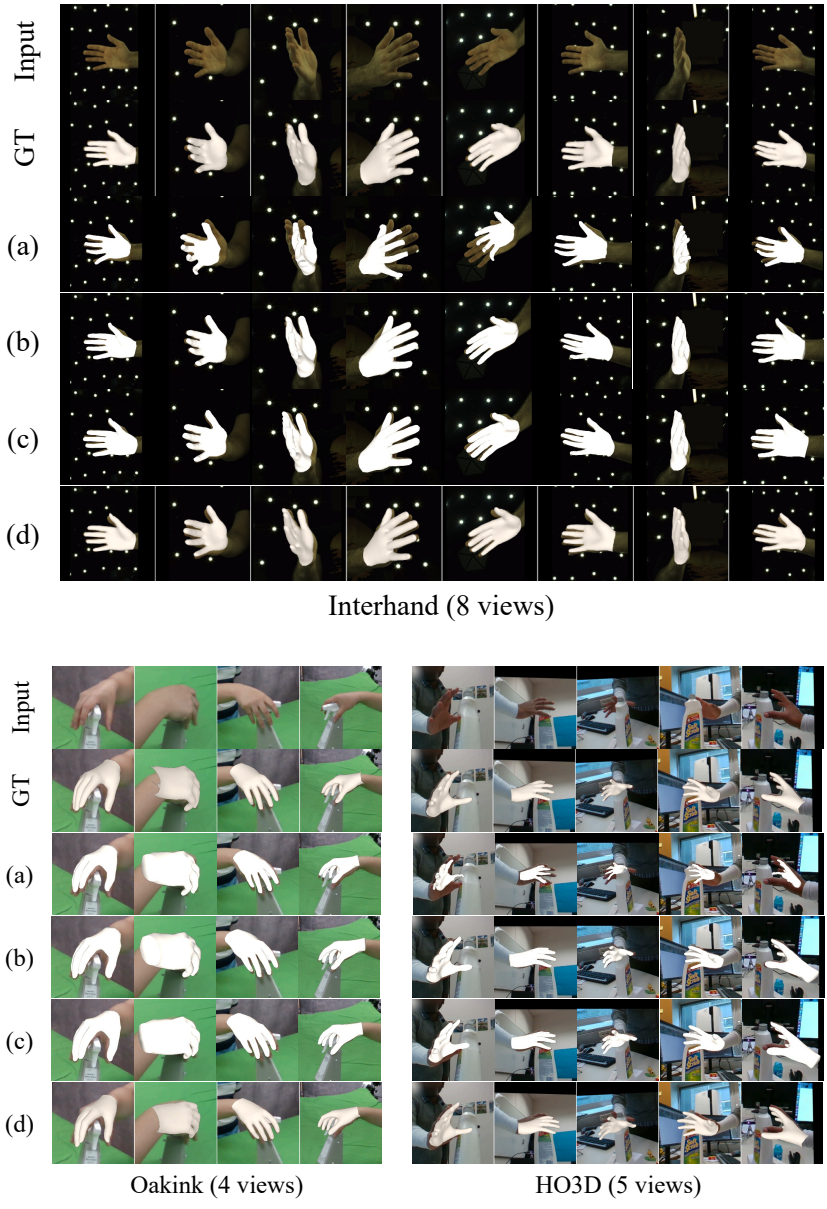


Fig. 6: Qualitative comparison on InterHand2.6M, OakInk and HO3D. We compare our method against baselines on open benchmarks. Rows correspond to: **Input** RGB, Ground Truth, **(a)** Cameras Predicted by VGGT + POEM_{v2}-large, **(b)** POEM_{v2}-large, **(c)** Cameras Predicted by Ours + POEM_{v2}-large, and **(d)** Ours (Full).



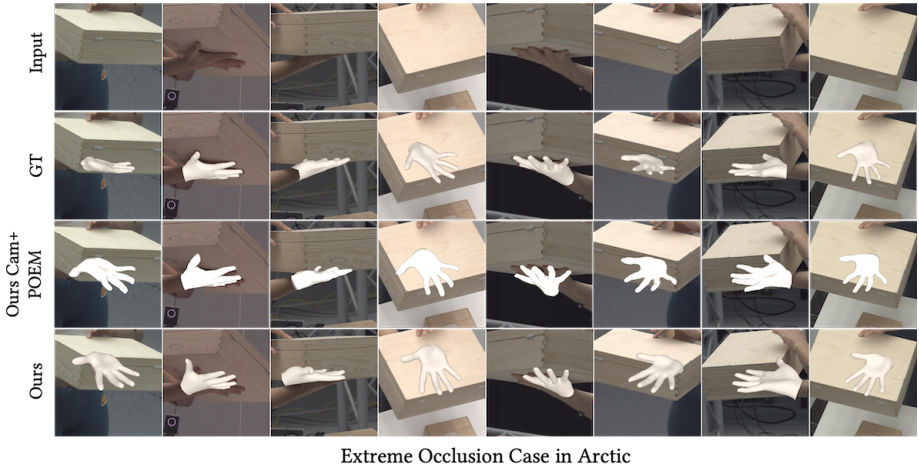


Fig. 7: Robustness in Extreme Scenarios. Our method maintains reasonable reconstruction quality even under challenging conditions, such as severe motion blur (top) and heavy object occlusion where the right hand is visible in only 2 out of 8 views (bottom).

References

1. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9044–9053 (2021)
2. Chen, P., Chen, Y., Yang, D., Wu, F., Li, Q., Xia, Q., Tan, Y.: I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12929–12938 (2021)
3. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: ICCV (2019)
4. Chen, X., Liu, Y., Ma, C., Chang, J., Wang, H., Chen, T., Guo, X., Wan, P., Zheng, W.: Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In: CVPR (2021)
5. Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., Pons-Moll, G.: Human3r: Everyone everywhere all at once. arXiv preprint arXiv:2510.06219 (2025)
6. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: European Conference on Computer Vision. pp. 769–787. Springer (2020)
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
8. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
9. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE transactions on pattern analysis and machine intelligence **45**(6), 7157–7173 (2022)
10. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE international conference on computer vision. pp. 873–881 (2015)
11. Gao, D., Xiu, Y., Li, K., Yang, L., Wang, F., Zhang, P., Zhang, B., Lu, C., Tan, P.: DART: Articulated Hand Model with Diverse Accessories and Rich Textures. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
12. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3d hand shape and pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10833–10842 (2019)
13. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
14. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2020)

15. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3196–3206 (2020)
16. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: CVPR (2022)
17. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11807–11816 (2019)
18. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: European Conference on Computer Vision. pp. 196–214. Springer (2020)
19. Kong, D., Zhang, L., Chen, L., Ma, H., Yan, X., Sun, S., Liu, X., Han, K., Xie, X.: Identity-aware hand mesh estimation and personalization from rgb images. In: European Conference on Computer Vision. pp. 536–553. Springer (2022)
20. Kulon, D., Guler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4990–5000 (2020)
21. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r (2024)
22. Li, M., Li, P., Zhang, Z., Lu, J., Zhao, C., Xue, W., Liu, Q., Peng, S., Zhang, W., Luo, W., et al.: Unish: Unifying scene and human reconstruction in a feed-forward pass. arXiv preprint arXiv:2601.01222 (2026)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: European Conference on Computer Vision. pp. 752–768. Springer (2020)
25. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: European Conference on Computer Vision (ECCV) (2020)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
27. Panteleris, P., Oikonomidis, I., Argyros, A.: Using a single rgb frame for real time 3d hand pose estimation in the wild. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 436–445. IEEE (2018)
28. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2024)
29. Pei, S., Chen, A., Lee, J., Zhang, Y.: Hand interfaces: Using hands to imitate objects in ar/vr for expressive interactions. In: Proceedings of the 2022 CHI conference on human factors in computing systems. pp. 1–16 (2022)
30. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12242–12254 (2025)
31. Qi, J., Ma, L., Cui, Z., Yu, Y.: Computer vision-based hand gesture recognition for human-robot interaction: a review. Complex & Intelligent Systems **10**(1), 1581–1606 (2024)

32. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
33. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 501–518. Springer (2016)
34. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1145–1153 (2017)
35. Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z.: Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. arXiv preprint arXiv:2412.06974 (2024)
36. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
37. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state (2025)
38. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
39. Wang, Y., Ye, J., Xiao, C., Zhong, Y., Tao, H., Yu, H., Liu, Y., Yu, J., Ma, Y.: Dexh2r: A benchmark for dynamic dexterous grasping in human-to-robot handover. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12702–12712 (2025)
40. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10974 (2019)
41. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)
42. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. arXiv preprint arXiv:2501.13928 (2025)
43. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: Oakink: A large-scale knowledge repository for understanding hand-object interaction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20953–20962 (2022)
44. Yang, L., Zhong, L., Zhu, P., Zhan, X., Kong, J., Xu, J., Lu, C.: Multi-view hand reconstruction with a point-embedded transformer (2024), https://arxiv.org/abs/2408.10581
45. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)
46. Zhang, H., Christen, S., Fan, Z., Hilliges, O., Song, J.: GraspXL: Generating grasping motions for diverse objects at scale. In: European Conference on Computer Vision (ECCV) (2024)
47. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views (2025), https://arxiv.org/abs/2502.12138
48. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular rgb image. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2354–2364 (2019)

49. Zhou, Y., Habermann, M., Xu, W., Habibie, I., Theobalt, C., Xu, F.: 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. pp. 5346–5355 (2020)
50. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE international conference on computer vision. pp. 4903–4911 (2017)
51. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 813–822 (2019)

HGGT: Robust and Flexible 3D Hand Mesh Reconstruction from Uncalibrated Images

Supplementary Material

A Implementation Details

To facilitate future research and ensure the reproducibility of our work, we are committed to open-sourcing our complete pipeline. We will publicly release:

- The complete training and evaluation code of our model.
- The large-scale multi-view synthetic dataset introduced in this work.
- The data generation pipeline used to synthesize the dataset.

A.1 Network Architecture.

Our multi-view 3D hand reconstruction framework builds upon the Visual Geometry Grounded Transformer (VGGT) architecture. The backbone employs a DINOv2 ViT-L/14 [26] encoder for patch embedding, followed by a VGGT aggregator with 24 transformer blocks. The aggregator operates with an embedding dimension of 1024, 16 attention heads, and an MLP expansion ratio of 4.0. It utilizes alternating attention between frame-level and global-level tokens with rotary position embeddings (RoPE frequency=100). Each view maintains 4 register tokens alongside camera and patch tokens. For hand pose estimation, we introduce a unified cross-attention refinement module, which takes hand tokens, camera tokens and concatenated image features from the aggregator and processes them through 4 cross-attention blocks, each with 16 attention heads and an MLP ratio of 4.

A.2 Training Setup.

In addition to the optimization settings detailed in the main text, we employ mixed-precision training with BF16 automatic mixed precision (AMP) to improve memory efficiency. Furthermore, we apply gradient clipping with an L2 norm threshold of 1.0 across all modules to ensure training stability.

A.3 Loss Weights.

Our multi-task loss function combines several objectives with carefully tuned weights. For hand pose estimation, we use a hand loss weight of 1.0, with subweights of 0.1 for pose parameters, 0.05 for shape coefficients, 5.0 for 3D keypoints, 1.0 for single-view 2D keypoints, and 10.0 for multi-view 2D keypoint consistency. The camera loss receives a weight of 5.0. Crucially, we employ intermediate supervision across the 4 cross-attention blocks. The hand and camera losses are computed at the output of each block and aggregated using exponentially increasing stage weights to emphasize the later refinement stages. In

contrast, the 2D projection loss is applied exclusively to the final block's output to ensure terminal geometric consistency without disrupting the intermediate feature learning.

B Additional Ablation Studies

In this section, we provide further ablation studies to validate the specific design choices in our network architecture and training strategy. The quantitative results are summarized in Table 3.

Table 3: Additional ablation studies on the intermediate supervision strategy, projection loss placement, and the number of cross-attention blocks. PA-MPVPE and PA-MPJPE are reported in millimeters (mm).

Configuration	**PA-MPVPE** ↓	**PA-MPJPE** ↓	**PA-AUC**$_{V50}$ ↑	**PA-AUC**$_{J50}$ ↑
Intermediate Supervision Strategy				
No Intermediate Sup.	3.52	3.53	0.92	0.92
Constant Weights	3.38	3.37	0.93	0.92
Exponential (Ours)	3.30	3.33	0.93	0.93
Number of Cross-Attention Blocks				
2 Blocks	4.51	4.55	0.87	0.86
3 Blocks	3.67	3.70	0.90	0.90
4 Blocks (Ours)	3.30	3.33	0.93	0.93
Placement of 2D Projection Loss				
All Blocks	3.47	3.45	0.91	0.91
Final Block Only (Ours)	3.30	3.33	0.93	0.93

B.1 Intermediate Supervision Strategy

As described in the main text, we apply intermediate supervision across the 4 cross-attention blocks using exponentially increasing stage weights. To validate this design, we compare our approach with two baselines: (1) **No Intermediate Supervision**, where losses are only applied to the final output, and (2) **Constant Weights**, where all intermediate blocks receive equal loss weights (*i.e.*, 1.0). As shown in Table 3, the exponentially increasing strategy yields the best performance. Constant weights tend to over-constrain the early layers, while our strategy effectively guides the progressive refinement without disrupting early feature extraction.

B.2 Number of Cross-Attention Blocks

We also investigate the impact of the number of cross-attention blocks in our refinement module. We test configurations with 2, 3, and 4 blocks. As reported

in Table 3, increasing the number of blocks brings improvements in hand reconstruction accuracy, demonstrating the scalability of our method.

B.3 Placement of 2D Projection Loss

A key detail in our training setup is applying the 2D projection loss exclusively to the final block's output. We ablate this by comparing it against a variant where the 2D projection loss is applied to all intermediate blocks. The results in Table 3 demonstrate that applying the projection loss to all layers degrades the performance. This confirms our hypothesis that enforcing strict 2D geometric consistency in early layers disrupts the intermediate 3D feature learning, whereas applying it only at the terminal stage ensures final alignment without feature distortion.

C More Qualitative Visualizations

We provide additional qualitative results in this section to further demonstrate the effectiveness and robustness of our proposed method. These extensive visualizations showcase our model's capability in accurately predicting camera poses and reconstructing complex hand-object interactions across diverse and challenging scenarios.

C.1 Open Benchmark.

In this subsection, we present comprehensive visual comparisons on several widely used open benchmarks. Specifically, Figure 8 illustrates the qualitative results on the ARCTIC and DexYCB datasets, respectively. In these challenging scenarios, our full pipeline consistently yields better alignment and more accurate 3D reconstructions compared to the baselines.

Furthermore, Figure 11, Figure 9, Figure 10, Figure 12 and Figure 13 provide additional qualitative visualizations on open dataset. These extra examples further demonstrate our model's robustness when handling diverse scenarios.

C.2 In-the-wild Video Inference

To further demonstrate the practical applicability of our method in real-world scenarios, we evaluate it on continuous in-the-wild video sequences. As shown in Figure 14, the videos are captured using a casual setup consisting of only two smartphones placed at random positions.

Given the input videos, our model processes them frame-by-frame to predict the 3D hand reconstructions. To highlight the robustness of our approach, we compare our results with POEM_{v2}-large. Since POEM_{v2}-large strictly requires calibrated cameras, we supply it with the camera parameters predicted by our method to enable the comparison. The qualitative comparisons across three different demo sequences are illustrated in Figure 15, Figure 16, and Figure 17. Please refer to the video in our project page for dynamic visualization results.

Notably, the inference of our HGGT is highly efficient, requiring only 0.23 seconds per frame. However, our current pipeline relies on an external hand detector to obtain initial bounding boxes, which introduces an additional processing time of approximately 1.2 seconds per frame. We acknowledge this dependency as a limitation of the current system and plan to address it in future work by integrating a more efficient, end-to-end detection module to further accelerate the overall pipeline.

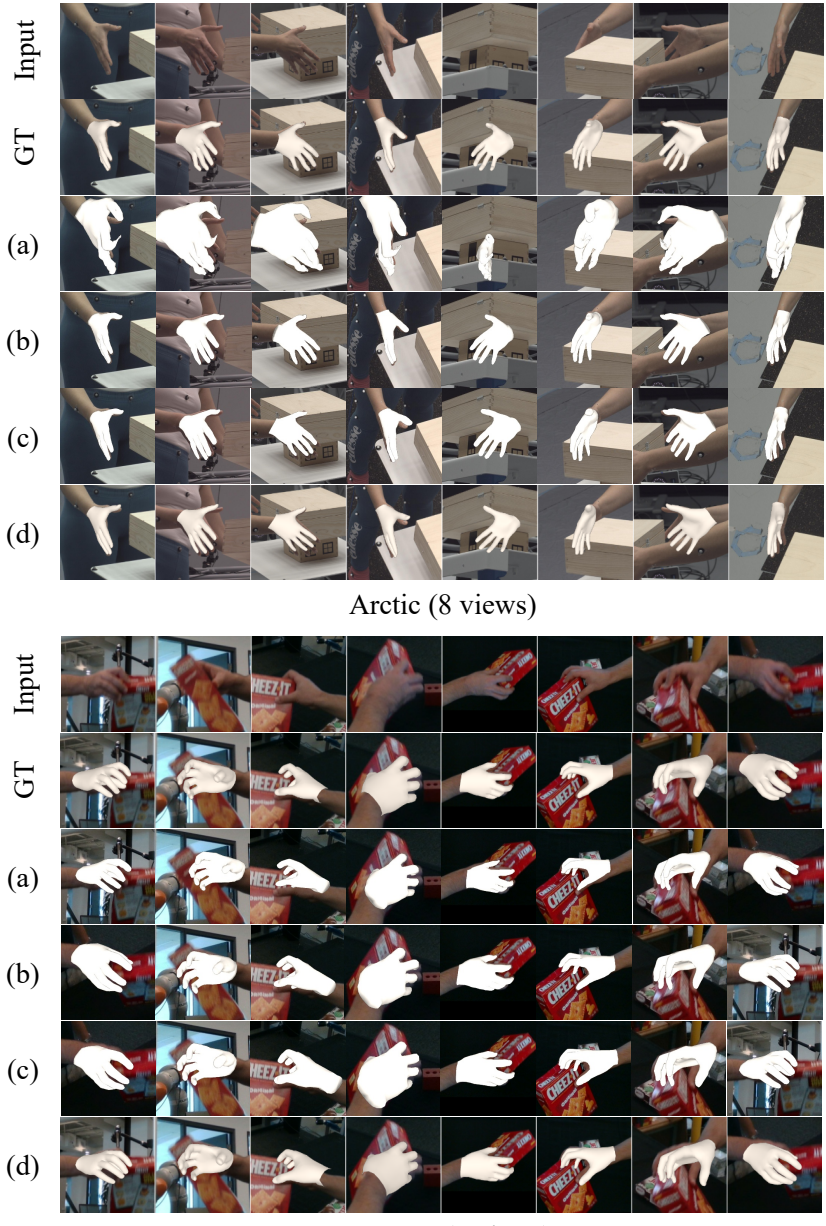


Fig. 8: Qualitative comparison on the DexYCB dataset. We compare our method against baselines on DexYCB dataset. Rows correspond to: Input, Ground Truth, **(a)** Cameras Predicted by VGGT + POEM$_{v2}$-large, **(b)** POEM$_{v2}$-large, **(c)** Cameras Predicted by Ours + POEM$_{v2}$-large, and **(d)** Ours (Full).

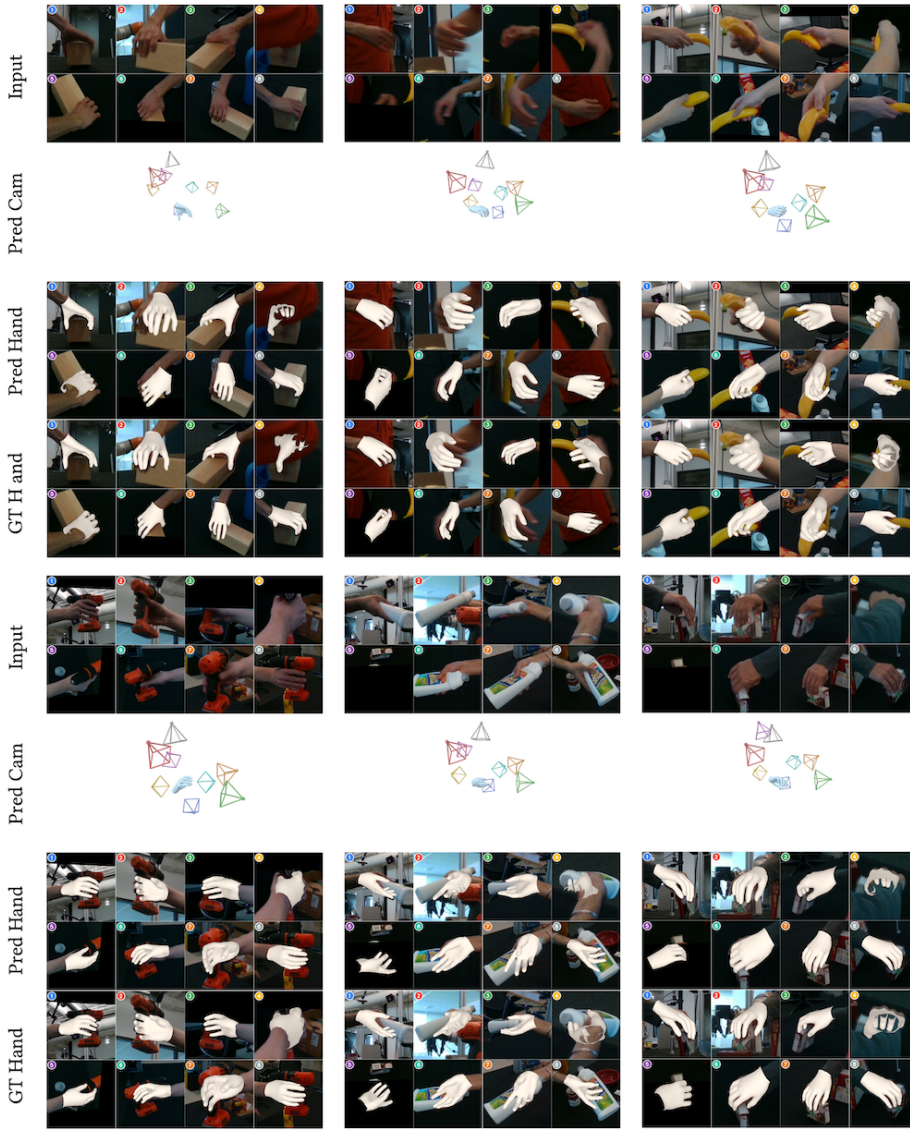


Fig. 9: Additional qualitative results on the DexYCB [1] dataset.

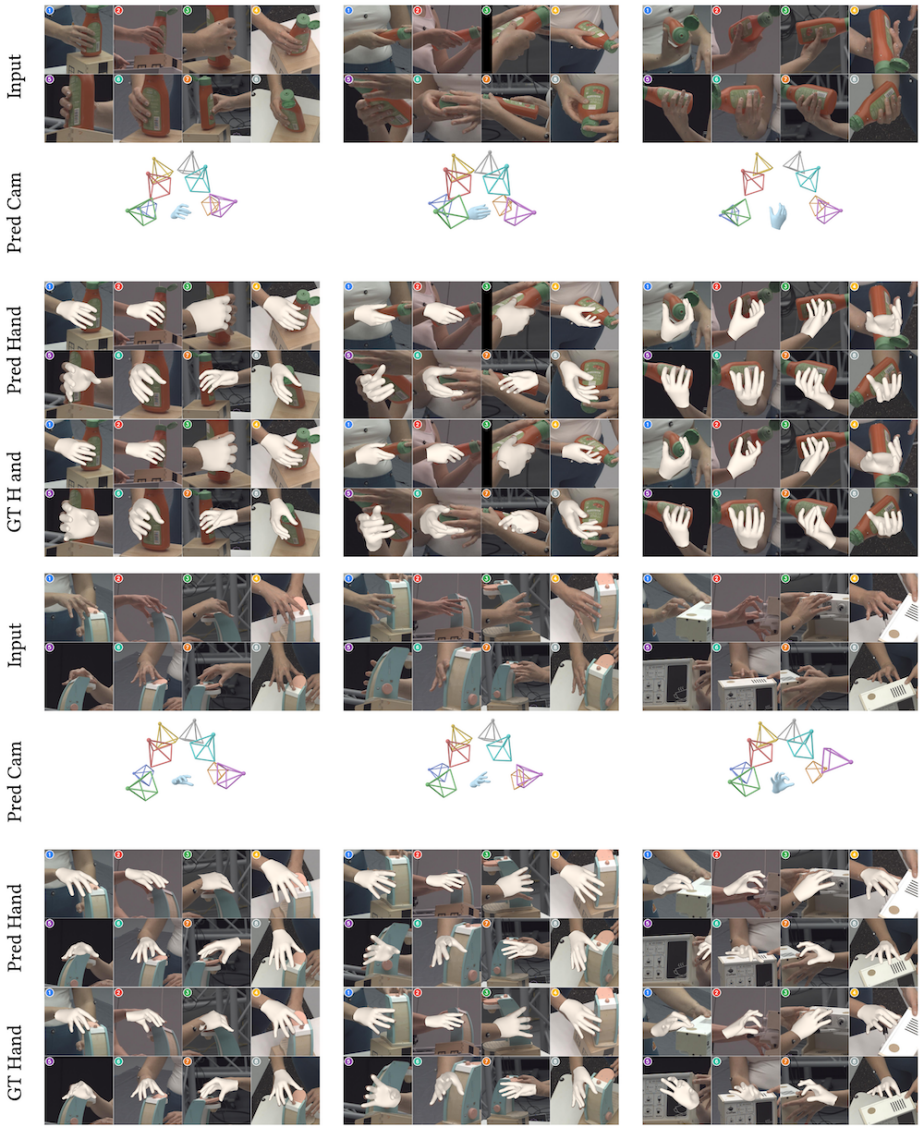


Fig. 10: Additional qualitative results on the Arctic [8] dataset.

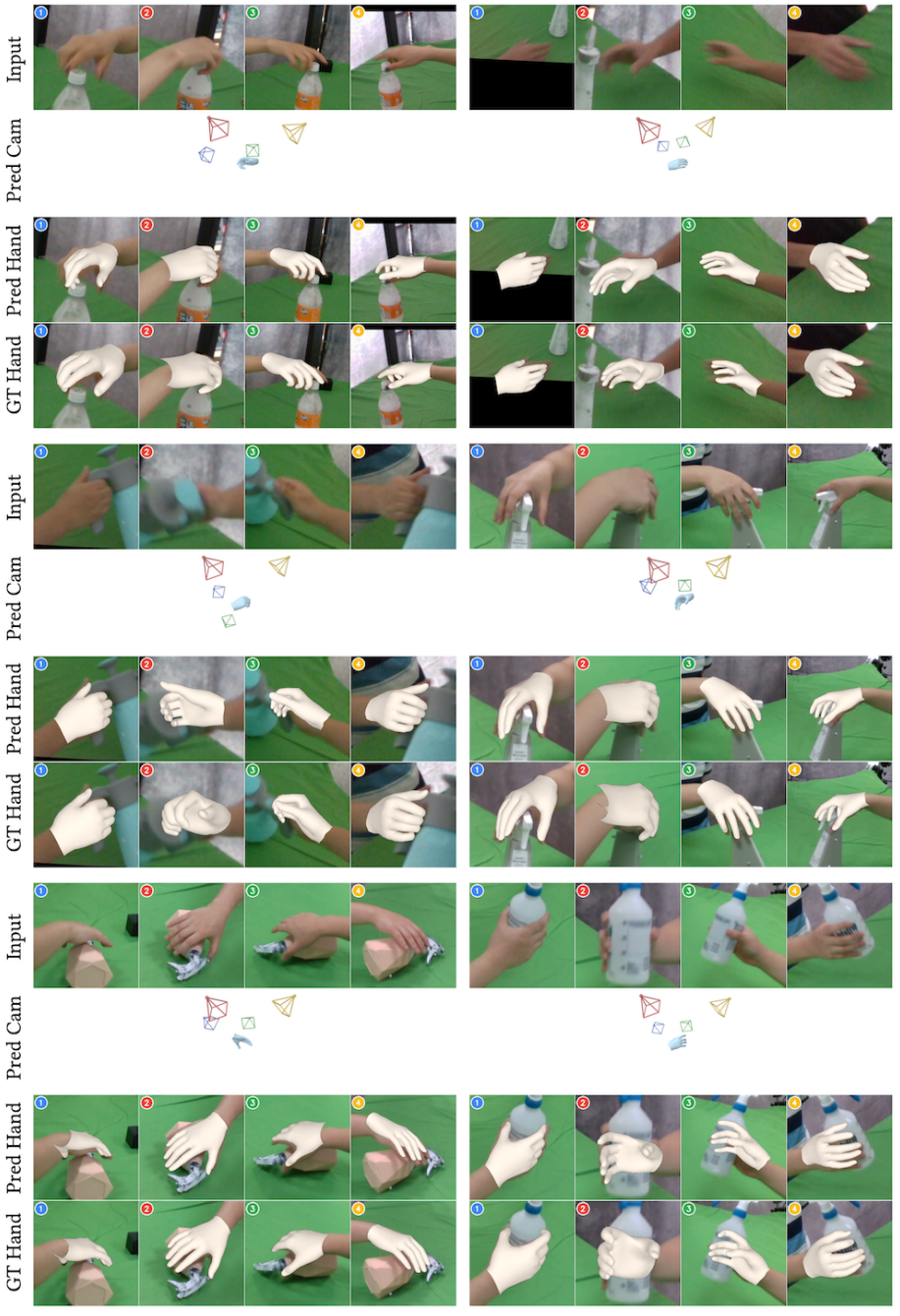


Fig. 11: Additional qualitative results on the OakInk [43] dataset.



Fig. 12: Additional qualitative results on the HO3D [15] dataset.

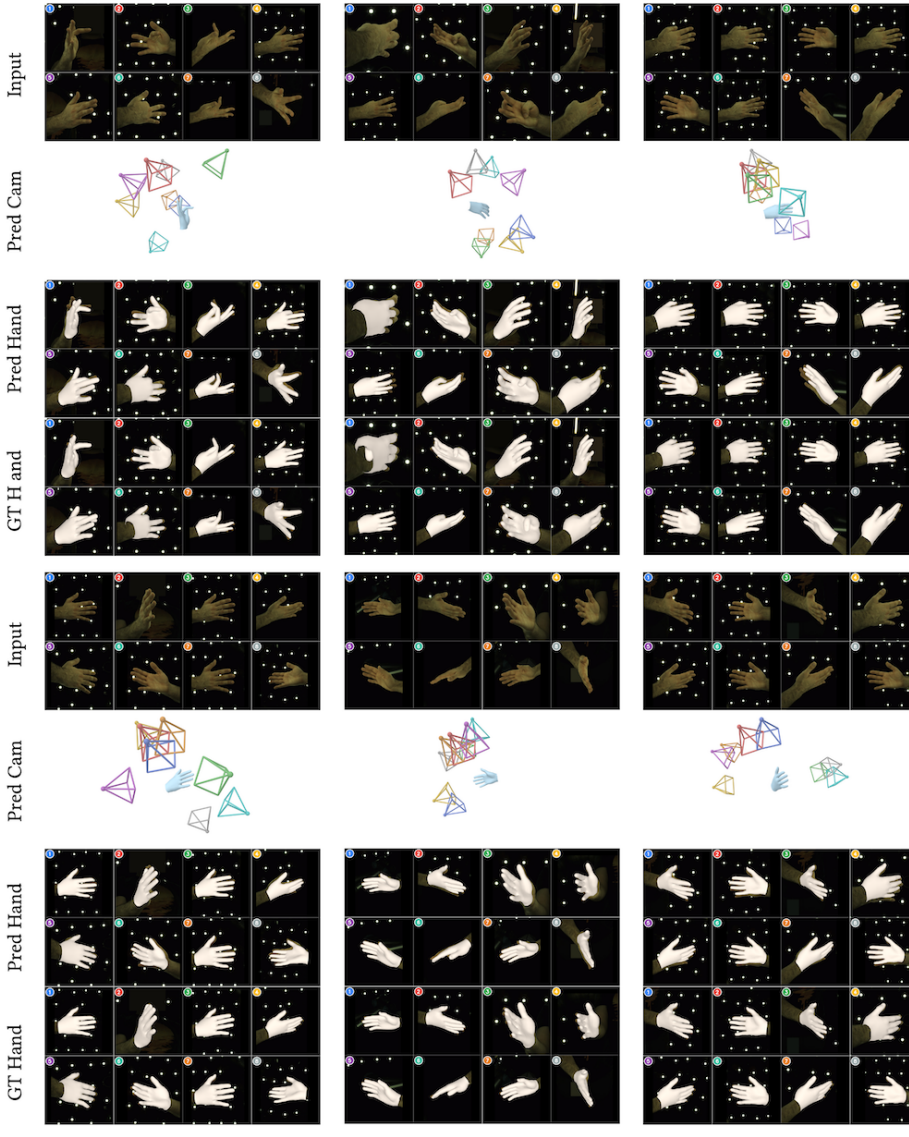


Fig. 13: Additional qualitative results on the InterHand2.6M [25] dataset.

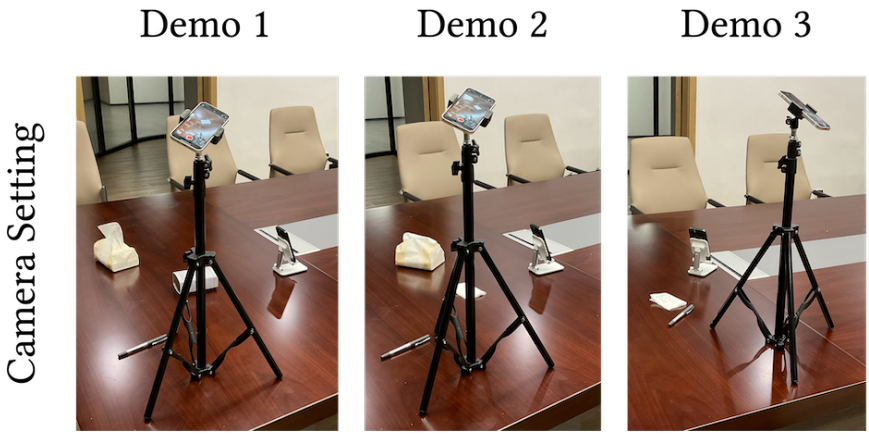


Fig. 14: Camera settings for in-the-wild video capture. The videos are recorded using two smartphones placed at random positions without prior calibration.

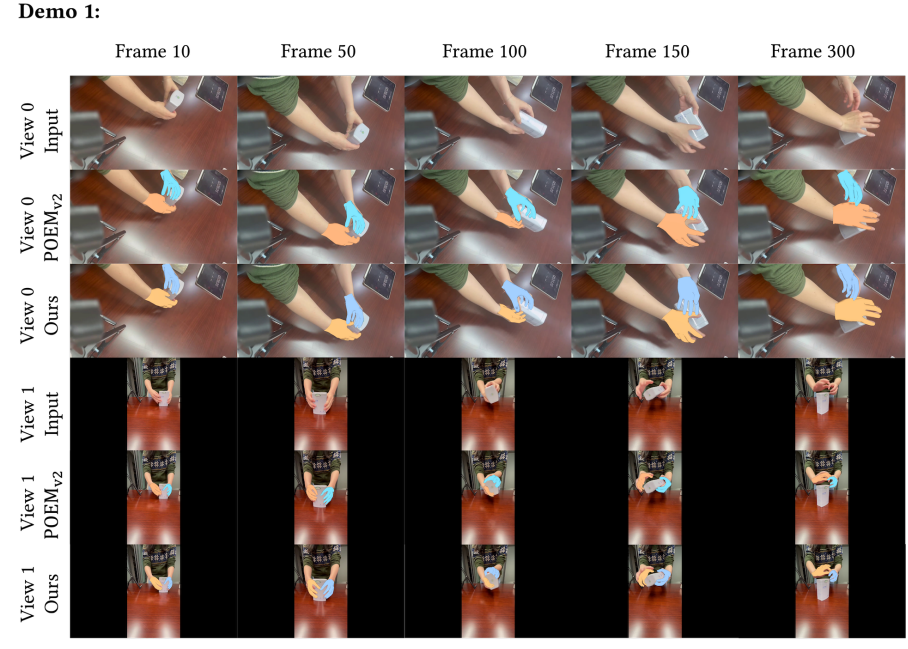


Fig. 15: Qualitative comparison with POEM$_{v2}$-large on in-the-wild video sequence (Demo 1).

Demo 2:

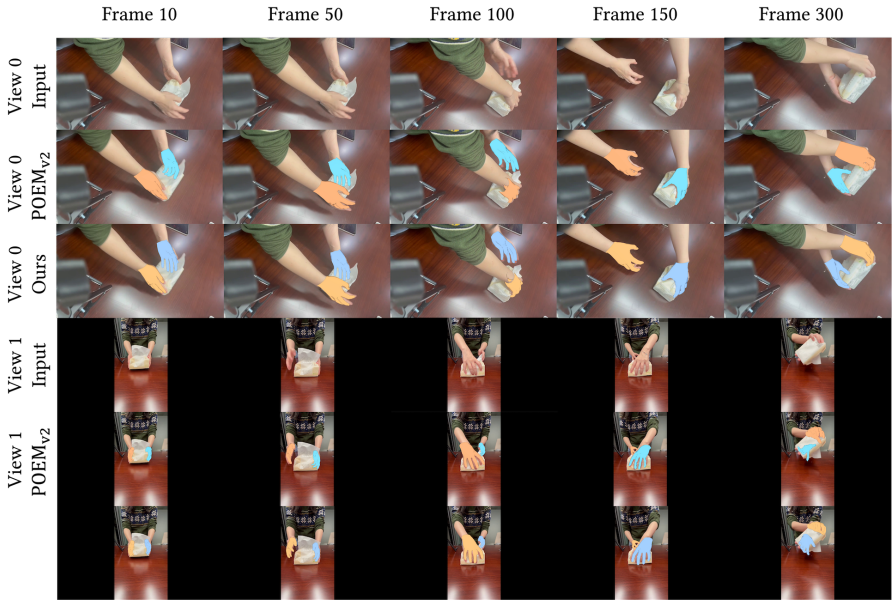


Fig. 16: Qualitative comparison with POEM$_{v2}$-large on in-the-wild video sequence (Demo 2).

Demo 3:

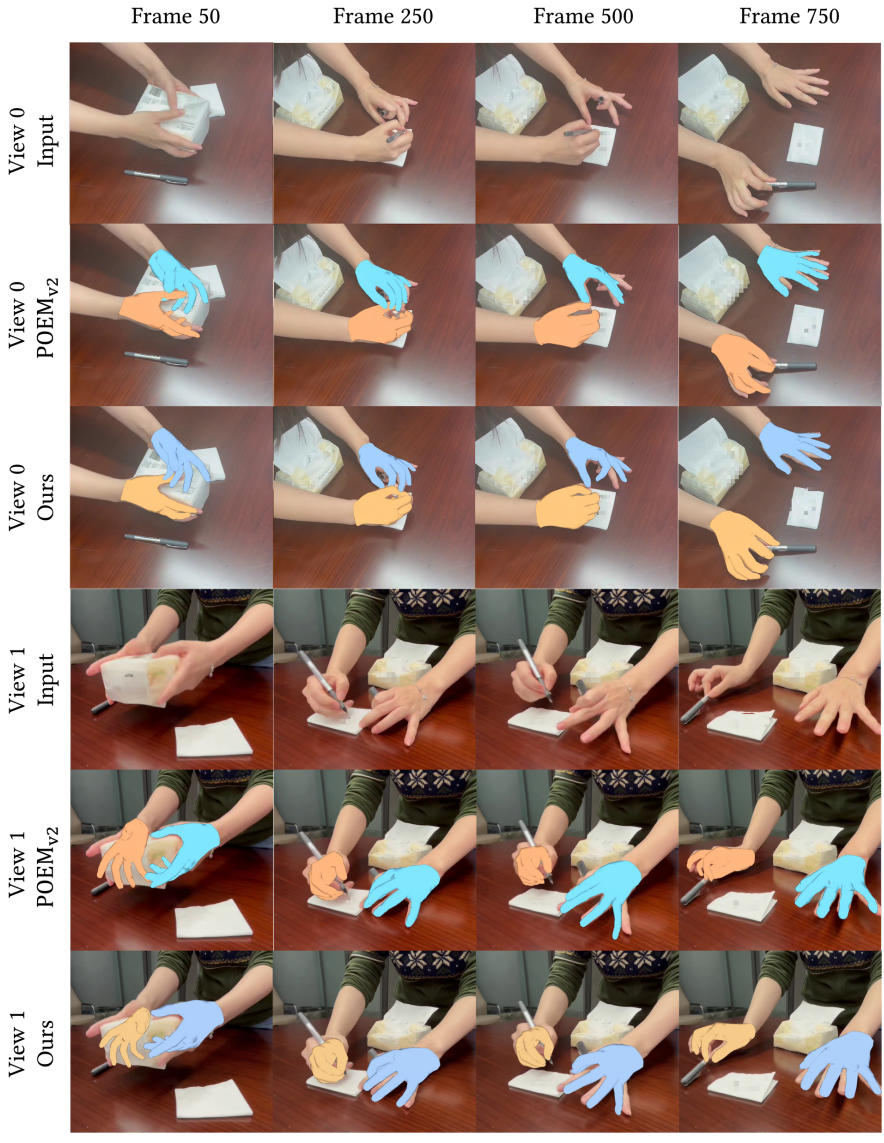


Fig. 17: Qualitative comparison with POEM$_{v2}$-large on in-the-wild video sequence (Demo 3).