arXiv:2606.15908v2 [cs.CV] 18 Jun 2026

High-Fidelity 4D Hand-Object Capture via Multi-View Spatiotemporal Tracking and Physics-Aware Gaussians

Bo Peng[1,2*], Xu Chen[1], Yi Gu[1,3], Hidenobu Matsuki[1], Mingsong Dou[1], Jingjing Shen[1], Deying Kong[1], Juyong Zhang[2], and Zhengyang Shen[1*]

[1] Google XR
[2] University of Science and Technology of China (USTC)
[3] The Hong Kong University of Science and Technology (Guangzhou)
[*]Equal contribution

Abstract. The growing demand for high-fidelity 4D hand-object interaction (HOI) data in embodied AI and spatial computing is currently bottlenecked by the reliance on pre-scanned object templates and physical markers. While recent methods have demonstrated promising results in reconstructing 4D hand-object interaction from videos, they are highly sensitive to initial estimates of hand and object poses. Yet, estimating these poses from images is challenging, in particular under severe occlusion which is inherent in hand-object interaction scenarios. We propose a novel system for the robust and accurate reconstruction of hands and objects from synchronized and calibrated multi-view videos without requiring any templates or markers. Our system consists of two main components with key innovations: (1) a multi-view feed-forward transformer model that aggregates cross-view geometry and temporal cues to provide a reliable, metric-consistent initialization for both poses and dense object geometry, and (2) a hand-object physics-aware Gaussian-based optimization framework to refine the initial estimates, integrating tetrahedral constraints, collision refinement, and appearance decomposition to produce physically plausible and visually accurate reconstruction. Validated on public benchmarks and an extensive internal dataset, our pipeline achieves highly robust, artifact-free reconstruction, providing an efficient foundation for automated 4D asset generation. Our project page are available at `https://hostpg.github.io/`.

Keywords: 4D Reconstruction · Hand-Object Interaction · Gaussian Splatting · Spatiotemporal Prior · Digital Assets

1 Introduction

The rapid evolution of embodied AI and immersive spatial computing has catalyzed a surging demand for high-fidelity 4D hand-object interaction (HOI) data [3, 11, 16–18, 20, 21, 48, 61]. Conventionally, capturing such data requires known object templates and physical markers [16] or manually annotated keypoints [3] to track the object and hand poses. This strict dependency prevents

these methods from generalizing to diverse, unknown objects in the wild, severely bottlenecking the data production throughput of existing capture pipelines.

Several recent methods have attempted to reconstruct 4D interacting hands and objects from video alone [9, 15, 17, 41], bypassing markers and templates by optimizing 3D Gaussian or NeRF representations. While promising, these methods are highly sensitive to initialization. An inaccurate estimate of initial hand or object poses often leads the optimization toward poor local minima or total divergence. Despite rapid progress [8, 49, 56, 65], estimating hand and/or object poses from images remains challenging due to inherent depth ambiguity and severe dynamic occlusions. Consequently, existing video-based capture methods often lack robustness or must impose restrictive assumptions, such as requiring objects to remain static [41] or move at low velocities [9].

To bridge this gap, we propose a novel system capable of reconstructing hands and unknown objects during natural interactions from multi-view video captured by a calibrated, 360-degree synchronized camera rig (comprising 21 cameras at 30 fps) with significantly improved robustness and accuracy. Our framework adopts a two-stage design with key innovations at each phase. We first employ a novel multi-view, spatiotemporal feed-forward model for robust initialization, followed by a physics-aware Gaussian-based optimization framework that refines geometry and appearance.

Our first innovation addresses the initialization bottleneck via the hand object spatiotemporal transformer (HOST). Existing hand or object pose estimators [8, 49, 56, 65] are primarily designed for monocular input and rely heavily on learned priors, which often fail to resolve the severe depth ambiguities and occlusions inherent in complex interactions. While recent foundational 3D models (e.g., VGGT [52]) effectively leverage multi-view cues, they are generally tailored for reconstruction of static scenes. HOST extends these foundational models to the task of multi-view tracking of dynamic interacting hands and objects. By introducing a factorized spatiotemporal attention mechanism, HOST efficiently aggregates information across viewpoints and frames, producing accurate and temporally coherent object and hand pose estimates. In addition, even if the object is partially occluded in some frames, HOST is able to reconstruct a complete and dense point cloud by using the geometric cues in all frames and viewpoints. HOST effectively resolves depth ambiguities where monocular trackers fail, ensuring metric 3D consistency and providing a reliable geometric foundation for subsequent refinement.

Our second innovation is a hand-object physics-aware Gaussian (HOPG) optimization framework, which elevates this initialization into high-fidelity 4D assets. We model the two hands and the unknown object with independent 2D Gaussian [25] clouds, which are initialized with the tracked pose parameters as well as the reconstructed dense point cloud from HOST. We optimize the 2D Gaussians along with the object and hand poses by minimizing the photometric loss. To prevent implausible distortions or volume collapse of the hand during close interactions [9, 40], we anchor the hand Gaussian on a tetrahedral mesh proxy and regularize the optimization with a volumetric As-Rigid-As-Possible

(ARAP) energy [47]. Furthermore, we incorporate a dedicated physics-aware collision resolution step to ensure no penetration. To reconstruct high-fidelity appearance of hand and object, we further introduce a temporal batching strategy and a decoupled model [45] that disentangles view-independent albedo and view-dependent specular. This decomposition prevents environmental lighting from being "baked" into the geometry, recovering the intrinsic surface colors.

Leveraging our framework, we processed a large-scale dataset of 77 complex sequences involving 32 unique objects with complex geometries. Achieving an automated success rate exceeding 93%, we demonstrate our system's exceptional robustness for automated 4D asset generation. We extensively evaluate both our isolated modules (HOST and HOPG) and the full pipeline across a public benchmark [37], the VEPHand dataset [1], and our custom dataset. The results demonstrate the effectiveness of our framework across three core tasks: it establishes superior robust pose tracking, excels in dynamic novel view synthesis, and recovers highly accurate dense geometry.

In summary, our key contributions are:

- A system that robustly and accurately reconstructs hand object interaction from synchronized multi-view videos (e.g., using a 21-camera 360-degree rig) without requiring pre-scanned templates or markers, automating the pipeline for efficient data acquisition.
- A spatiotemporal transformer model that explicitly leverages cross-view and temporal cues, providing a reliable geometric initialization even under severe occlusion.
- A physics-aware Gaussian optimization framework integrating tetrahedral constraints, collision refinement, and appearance decomposition to produce physically plausible and visually accurate 4D reconstructions.

2 Related Works

Neural Hand and HOI Rendering. Reconstructing high-fidelity dynamic hand-object interactions (HOI) is a long-standing challenge. To address this, foundational work first focused on bare-hand neural rendering. Early neural implicit methods, such as HandAvatar [6], achieved high fidelity by disentangling hand geometry and texture but suffered from slow rendering speeds. While subsequent works like LiveHand [38] overcame this to achieve real-time performance via mesh-guided sampling, 3DGS has recently emerged as a fundamentally more efficient alternative. GauHuman [24] models articulated humans by warping 3D Gaussians via Linear Blend Skinning, and HandSCS [13] explicitly handles severe hand articulations by introducing a Structural Coordinate Space to bridge sparse skeletal joints and dense Gaussians.

Building upon these bare-hand reconstruction foundations [13, 19, 27], a substantial body of work [59, 64] tackles HOI through per-scene optimization to recover precise surfaces and contact mechanics. Recent methods achieve this using articulated 3D Gaussians (MANUS [41], DyTact [9]), compositional implicit fields (HOLD [15]), or generative diffusion priors for unobserved regions [34, 53,

68]. However, these optimization frameworks suffer from two critical limitations: they inevitably collapse when their required tracking initializations drift under severe occlusions, and they frequently "bake" environmental illumination into the geometry. To bypass the fragility of traditional tracking bottlenecks, recent advances like EasyHOI [36] leverage large vision-language models to directly estimate hand-object geometry from monocular in-the-wild inputs before executing optimization.

Hand-Object Pose Tracking. To provide the tracking priors required by the aforementioned optimization pipelines, an emerging line of inference-based methods trains feed-forward models to directly regress hand and object states from images. For instance, HORT [8] employs a coarse-to-fine transformer architecture to efficiently reconstruct dense 3D point clouds of hand-held objects and their hand-relative poses. For complex hand motions, Dyn-HaMR [65] recovers global 4D trajectories from monocular videos by combining hierarchical hand tracking with generative motion priors and SLAM. To capture hands across alternative viewpoints, multi-view frameworks have also been explored; notably, POEM [62] introduces a point-embedded transformer for single-frame multi-view hand reconstruction. Additionally, foundation models like SAM 3D [5] have generalized single-image 3D perception, demonstrating remarkable capabilities in predicting object geometry, texture, and scene layout directly from in-the-wild 2D images.

However, monocular methods remain inherently susceptible to severe depth ambiguities, while single-frame multi-view fusion methods often fail to maintain temporal consistency. In contrast, our HOST module explicitly aggregates continuous cross-view and temporal cues to resolve these ambiguities, establishing a metrically reliable and temporally coherent initialization for the downstream optimization.

Feed-Forward Geometric Estimation. The success of large-scale pre-training has shifted 3D geometric estimation toward feed-forward regression [23, 55]. Pioneered by DUSt3R [32, 54, 60], foundational works demonstrate that dense 3D point maps can be directly regressed from unposed images. Building upon this, VGGT [52] and its successors [46, 58] introduced large networks [14] for joint camera and scene geometry prediction. To adapt these spatial priors for dynamic 4D environments without incurring prohibitive computational overhead, recent works like 4D-VGGT [51] and MonST3R [66] introduce spatiotemporal decoupling to model structural consistency and motion continuity separately. Concurrently, drawing inspiration from these visual-geometry grounded architectures, HGGT [35] jointly infers 3D hand meshes and camera poses from uncalibrated arbitrary views. Unlike these foundational or single-entity models, which primarily output generic scene point clouds or focus exclusively on bare hands, HOST is explicitly designed to parse complex hand-object interactions. By introducing a spatiotemporal attention mechanism, we efficiently aggregate multi-view and temporal cues to jointly regress hand and object states and entity-specific dense geometry. This effectively translates powerful generic 3D priors into a special-

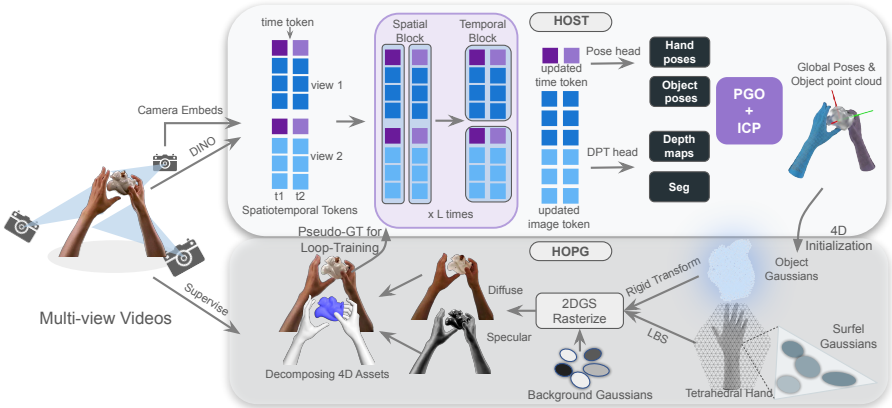

Fig. 1: Overview of our framework. Our pipeline operates in two stages. First, the Hand Object Spatiotemporal Transformer (HOST) processes multi-view videos to robustly regress parametric hand/object poses, dense object point clouds, and segmentation masks, yielding a metric 3D initialization. Second, the Hand Object Physics-aware Gaussian (HOPG) module leverages this initialization to optimize a hybrid 2D Gaussian representation. By enforcing structural constraints and explicitly decoupling diffuse and specular appearance, HOPG effectively prevents illumination bake-in, reconstructing high-fidelity 4D assets.

ized articulated tracker, providing a reliable 4D initialization for our downstream HOPG pipeline.

3 Method

Given synchronized multi-view RGB videos and calibrated camera parameters, our framework first employs the Hand Object Spatiotemporal Transformer (Sec. 3.1) to establish a robust multi-view spatiotemporal prior, followed by a Hand Object Physics-aware Gaussian (HOPG) refiner (Sec. 3.2). The full pipeline is illustrated in Fig. 1.

3.1 HOST: Hand-Object Spatiotemporal Transformer

The first stage of our framework establishes a robust geometric prior to guide the downstream volumetric refinement. Given a sequence of F frames across V synchronized static views, HOST regresses temporally coherent 4D attributes and dense geometric features. Formally, HOST defines a mapping $\mathcal{F}$ as follows:

$$\left(\mathbf{D}_{1:F,1:V}, \mathbf{\Theta}^{h}_{1:F}, \mathbf{\Theta}^{o}_{1:F}\right) = \mathcal{F}\left(\mathbf{I}_{1:F,1:V}, t_{1:F}, \mathbf{E}_{1:V}, \mathbf{K}_{1:V}\right), \tag{1}$$

where $\mathbf{D}_{1:F,1:V}$ denote the dense predictions (i.e., semantic segmentation masks, depth maps, and point maps) across all F frames and V viewpoints; $\mathbf{\Theta}^{h}_{1:F}$ and $\mathbf{\Theta}^{o}_{1:F}$ refer to the full sequences of hand and object parameters, respectively;

$\mathbf{I}_{1:F,1:V}$ denotes the input multi-view video sequence; $\mathbf{E}_{1:V}$ and $\mathbf{K}_{1:V}$ denote the camera extrinsic and intrinsic parameters for all views, and $t_{1:F}$ represent the relative frame identifiers.

Spatiotemporal Feature Backbone. HOST builds upon the VGGT [52] transformer backbone but extends it to a spatial and temporal structure. To encode spatial and temporal information, each image $\mathbf{I}_{f,v}$ is patchified into tokens $\mathbf{T}_{\text{image}}$ using a frozen DINOv2 [39] backbone, and augmented with view-consistent, ray-based camera embeddings [28]. Simultaneously, relative frame identifiers t_f are mapped to view-shared time tokens $\mathbf{T}_{\text{time}}$, which are then concatenated with $\mathbf{T}_{\text{image}}$ to form the final input sequence $(\mathbf{T}_{\text{time}}, \mathbf{T}_{\text{image}})$ for the factorized attention layers.

Processing massive 4D sequences (V views, F frames, N patches) via monolithic global attention, such as the frame-global loop in VGGT, incurs a prohibitive computational complexity of $\mathcal{O}((VFN)^2)$. To circumvent this bottleneck while preserving rich 4D context, we propose the spatiotemporal mechanism. By retaining foundational intra-frame attention but explicitly factorizing the intractable global attention into alternating Spatial and Temporal Blocks (Fig. 1), the feature update cycle for the k-th group is formulated as:

$$\mathcal{B}_{\text{spatial}} : \mathbf{T}'_k = \mathcal{A}_{\text{view}}\left(\mathcal{A}_{\text{frame}}(\mathbf{T}_k)\right), \tag{2}$$

$$\mathcal{B}_{\text{temporal}} : \mathbf{T}_{k+1} = \mathcal{A}_{\text{temporal}}\left(\mathcal{A}_{\text{frame}}(\mathbf{T}'_k)\right), \tag{3}$$

Here, $\mathcal{A}_{\text{frame}}$ performs local intra-frame self-attention (costing $\mathcal{O}(VFN^2)$). The Spatial Block ($\mathcal{B}_{\text{spatial}}$) aggregates stereo cues via cross-view attention $\mathcal{A}_{\text{view}}$ for each frame independently, bounding its cost to $\mathcal{O}(F(VN)^2)$. Conversely, the Temporal Block ($\mathcal{B}_{\text{temporal}}$) aggregates motion cues via temporal attention $\mathcal{A}_{\text{temp}}$ for each view independently, costing $\mathcal{O}(V(FN)^2)$. This factorized design successfully aggregates global 4D context while drastically reducing the quadratic computational overhead. In practice, we train the framework using a chunking strategy, e.g., a typical chunk we use in training is 5 frames with 6 views.

Dense Geometry Heads. We employ a Dense Prediction Transformer (DPT) architecture [42] to decode the features into a set of per frame and per view dense maps $\mathbf{D}$: multi-class probability map for the left hand, right hand, and the interacting object; depth map and point map. This joint supervision encourages the network to disentangle the tightly interacting entities implicitly, providing a clean, entity-specific geometric foundation.

Global Kinematics Heads. To estimate the global 4D kinematic state, we pass the final stage time tokens $\hat{\mathbf{T}}_{\text{time}}$, which encode the overall spatiotemporal dynamics, to dedicated MLP heads that regress the per-frame kinematic states for hand and object parameters. The hand parameters $\mathbf{\Theta}^h = \{\alpha, \beta\}$ here refer to the identity α and poses β w.r.t. a given parametric hand model [44].

For the rigid object transform, we adopt a hybrid state representation comprising the absolute centroid translation $\mathbf{t}_f^{\text{abs}}$ in world coordinates and the relative rotation $\Delta\mathbf{R}_{1\to f}$ (with respect to the first frame of the input chunk). This asymmetric design explicitly addresses the distinct observability of translation and rotation. Because the canonical orientation of an unseen object is inherently

ambiguous without a predefined template, predicting relative rotation is strictly necessary. While it may seem intuitive to also predict relative translations, integrating frame-to-frame offsets inevitably accumulates trajectory drift over time. Instead, since the object's absolute 3D geometric center remains continuously observable from spatiotemporal cues, we directly regress its absolute translation. This firmly anchors the spatial trajectory, eliminating long-term positional drift and confining any unavoidable error accumulation strictly to the rotational domain.

Training Losses. We follow the VGGT [52] for dense prediction (segmentation, depth and point map) training and apply geodesic distance [22, 26] on rotation and L2 losses on translation and registered meshes from forward kinematics.

Inference. To process arbitrarily long videos under memory constraints, we split sequences into overlapping chunks using a sliding window with a stride of one. For hands, simply averaging the overlapping predictions provides sufficient temporal smoothing. For rigid objects, however, we prevent long-term trajectory drift by aggregating chunk-wise predictions in two steps:

We first employ *Pose Graph Optimization (PGO)* [12, 31] to combine per-window pose estimates into a continuous trajectory. Specifically, we optimize absolute object pose estimate $\mathbf{\Theta}^o = (\mathbf{R}, \mathbf{t})$ with the predicted relative rotation edges $\mathcal{E}$ and absolute translations:

$$\underset{\{\mathbf{R}_f, \mathbf{t}_f\}}{\arg\min} \sum_{(i,j)\in\mathcal{E}} \left\| \mathbf{R}_i^T \mathbf{R}_j - \Delta\mathbf{R}_{i\to j} \right\|_F^2 + \lambda_{\text{abs}} \sum_f \left\| \mathbf{t}_f - \mathbf{t}_f^{\text{abs}} \right\|^2 . \tag{4}$$

To eliminate residual drift, we then unproject the masked depth maps into 3D point clouds and align them via frame-to-model ICP [2, 69] with the PGO trajectory as a robust initialization. In this way, we are able to infer accurate object pose trajectories even in long videos. Fusing these aligned observations yields the canonical object point cloud $\mathbf{V}_{\text{obj}}^{\text{cano}}$.

In summary, this inference strategy aggregates per-window predictions into globally consistent hand parameters $\mathbf{\Theta}^h$, absolute object poses $\mathbf{\Theta}^o$, and an optional canonical point cloud of the object $\mathbf{V}_{\text{obj}}^{\text{cano}}$. These outputs are then used to initialize the subsequent HOPG refinement.

3.2 HOPG: Hand-Object Physics-aware Gaussian Optimization

In this stage, we jointly optimize three distinct 2D Gaussian fields, hand, object, and background, leveraging the pose and shape estimates from HOST as an initialization.

Volumetric Hand. We utilize a volumetric deformation model inspired by VEPHand [1] as our geometric proxy. To faithfully model anatomical thickness and to resolve collision during close interactions, we embed the hand surface within a canonical tetrahedral mesh. However, rather than optimizing per-frame volumetric variables, we predict these intrinsic tetrahedral deformations dynamically via a small MLP network.

Specifically, let $\mathbf{v}_c$ denote the canonical surface vertices and $\boldsymbol{\rho}_c$ represent the vertices of the underlying canonical tetrahedral mesh. Unlike standard parametric models that apply Linear Blend Skinning (LBS) directly to the base surface, our volumetric model first applies non-rigid, pose-dependent deformations in the tetrahedral domain:

$$\text{(Standard Surface)} \quad \mathbf{v}_p = \mathrm{LBS}(\mathbf{v}_c, \alpha, \beta), \tag{5}$$

$$\text{(Our Volumetric)} \quad \boldsymbol{\rho}'_c = \boldsymbol{\rho}_c + \mathrm{MLP}(\boldsymbol{\rho}_c, \beta), \tag{6}$$

$$\mathbf{v}'_c = \mathcal{M}_{\mathrm{tet}\to\mathrm{surf}}(\boldsymbol{\rho}'_c), \tag{7}$$

$$\mathbf{v}_p = \mathrm{LBS}(\mathbf{v}'_c, \alpha, \beta), \tag{8}$$

where α and β represent the identity shape and kinematic pose parameters, respectively. In our formulation, a lightweight MLP predicts pose-dependent offsets for the tetrahedral control points to yield the deformed volume $\boldsymbol{\rho}'_c$. We then compute the updated canonical surface vertices $\mathbf{v}'_c$ via a predefined barycentric mapping operator $\mathcal{M}_{\mathrm{tet}\to\mathrm{surf}}$. Geometrically, each surface vertex is initialized inside a specific tetrahedron. This operator calculates the deformed surface position simply by interpolating the displaced tetrahedral vertices $\boldsymbol{\rho}'_c$ using fixed barycentric weights. Consequently, the surface acts as a continuous skin strictly bound to the underlying volumetric flesh. Finally, standard Linear Blend Skinning (LBS) articulates these volume-aware canonical surface vertices into the final posed space $\mathbf{v}_p$.

Following PICA [40], we anchor 2D Gaussians to this deformed mesh using barycentric interpolation, ensuring the radiance field strictly inherits the kinematic consistency of the hand model.

Rigid Object. The object is represented by a set of canonical 2D Gaussians, seamlessly initialized from the fused canonical point cloud $\mathbf{V}_{\mathrm{obj}}^{\mathrm{cano}}$ provided by HOST. During refinement, we jointly optimize these canonical Gaussian attributes and the per-frame rigid transformations.

Background. We model the static environment using a separate set of 2D Gaussians initialized on a large enclosing sphere.

Appearance Model. Standard Gaussian Splatting [29] bakes view-dependent lighting into the spherical harmonics (SH) of each primitive. In highly dynamic hand-object interactions, this entanglement between extreme rotations and illumination causes severe flickering and texture baking. To reconstruct physically and temporally consistent assets, we explicitly factorize intrinsic albedo, diffuse shading, and specularity. Following Saito et al. [45], the dynamic diffuse color $\mathbf{C}_{\mathrm{diff}}$ of a Gaussian is formulated by convolving environmental light SH coefficients $\mathbf{L}$ with dynamically rotated canonical transfer coefficients $\mathbf{d}(\mathbf{R})$, modulated by a learnable view-independent albedo $\mathbf{a}$:

$$\mathbf{C}_{\mathrm{diff}} = \mathbf{a} \odot \sum_{i=1}^{(n+1)^2} \mathbf{L}_i \mathbf{d}_i(\mathbf{R}), \tag{9}$$

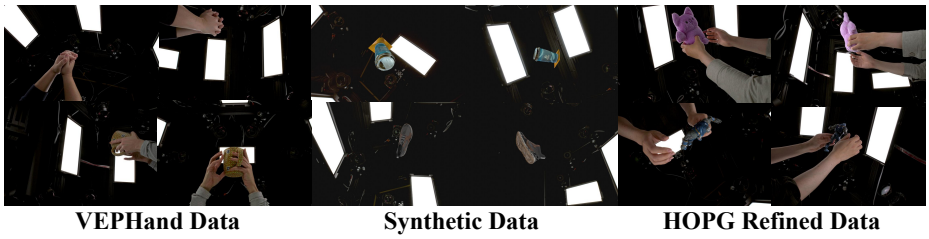

Fig. 2: Overview of our composite training dataset. Left: Multi-view captures from the VEPHand [1] dataset. **Middle:** Synthetic data rendered in Blender using diverse textured objects from Texverse [67] Dataset. **Right:** Custom captures refined via our HOPG module.

where n is the SH degree, and $\mathbf{R}$ defines the Gaussian's rotation from canonical to world space. Explicitly applying $\mathbf{R}$ ensures the transfer coefficients remain aligned with the global environment light, effectively decoupling illumination from the primitive's motion. Furthermore, to prevent view-dependent highlights from polluting this intrinsic albedo, we model specularity using a dedicated lightweight MLP [50]. The network processes the Integrated Directional Encoding (IDE) of the reflection vector $\mathbf{r}$ and the Gaussian's learnable surface roughness ρ to output a specular residual. The final composited color is then evaluated as $\mathbf{C}_{\text{total}} = \mathbf{C}_{\text{diff}} + s \cdot \text{MLP}_{\text{spec}}(\text{IDE}(\mathbf{r}, \rho))$, where s is a learnable, primitive-specific specular intensity.

Losses. The data terms are driven by standard photometric losses (L1, D-SSIM, LPIPS). For physical fidelity, we impose an As-Rigid-As-Possible (ARAP) energy [47] on the underlying tetrahedral mesh for volume preservation on deformed canonical tetrahedra $\rho'_{\mathbf{c}}$ and a ray-tracing based collision loss [1] on posed surface mesh $\mathbf{v}_p$ to prevent both inter and self penetration.

3.3 Dataset Roadmap

To ensure kinematic diversity and geometric generalizability, HOST is trained on a comprehensive composite dataset. The foundation comprises 1,600 multi-view sequences from VEPHand Dataset [1], featuring high-quality scans and hand-object registrations across 17 3D-printed CAD models. To overcome the limited object diversity of this set, we synthesize a supplementary dataset in Blender, rendering 500 high-resolution textured objects [67] simulated with randomized 6D trajectories. Finally, to capture real-world complexity, we collect a custom dataset of 47 everyday objects augmented with miniature ArUco markers to reliably acquire initial relative poses. These captures are then processed by our HOPG refiner. Once converged, HOPG acts as a robust pseudo-label generator, yielding refined ground-truth equivalents (cleaner masks, accurate depth, and collision-free poses). These high-fidelity annotations are then fed back to fine-tune HOST, progressively closing the domain gap without manual intervention.

4 Experiment

To comprehensively evaluate our framework's capabilities in photorealistic novel view synthesis (NVS), physically plausible geometry recovery, and industrial-level scalability, we conduct extensive experiments on a composite evaluation suite comprising both public benchmarks and diverse in-house captures.

4.1 Experimental Setup

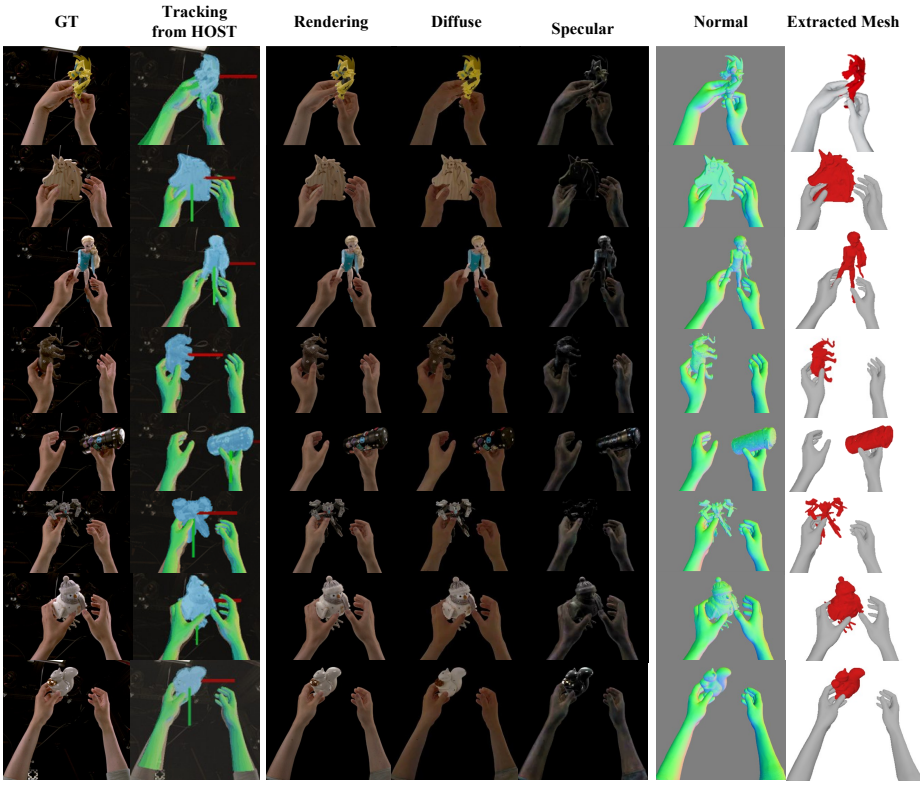

Fig. 3: Automated, Scalable High-Fidelity Capture across Diverse Objects. Given calibrated multi-view images (GT), HOST provides robust tracking initializations. HOPG then achieves photorealistic rendering by decomposing appearance into diffuse and specular components. This explicit lighting decoupling prevents artifacts from baking into the geometry, yielding clean normal maps and high-fidelity 3D meshes.

Implementation Details HOST is built on a ViT-Large backbone with frozen DINOv2 [39] patch embeddings and initialized with pre-trained VGGT [52] weights. We optimize the network using AdamW in bfloat16 precision. To stabilize early training, we introduce a 2,000-step geometry warmup where we only optimize the backbone and dense geometry heads. The model is first pre-trained for

200k steps on 64 TPU v5 accelerators ($lr = 8.0 \times 10^{-5}$) on VEPHand data, and subsequently fine-tuned for 20k steps on 8 NVIDIA H100 GPUs ($lr = 1.0 \times 10^{-5}$) on synthetic and our HOPG refined data. For the HOPG refinement, we use a batch size of 5 frames to explicitly enforce the decoupled appearance model. A typical sequence of 21 views and 600 frames takes about 30k iterations (roughly 5 hours) to converge on a single NVIDIA A100 GPU.

Scalability Validation. To validate scalability, we deploy our pipeline on a newly captured multi-view dataset of diverse objects. As qualitatively demonstrated in Fig. 3, our framework operates entirely automatically, successfully reconstructing 72 out of 77 interaction sequences across 32 unique objects, thereby achieving an exceptional ∼93.5% success rate. Furthermore, we also showcase the high-resolution 3D object meshes extracted via TSDF fusion [10] in Fig. 4. The few failures stem from challenging object properties, such as geometrically thin structures (pens) and highly symmetric, textureless surfaces (tapes), which induce severe pose ambiguities during HOST initialization.

Evaluation Datasets. We comprehensively evaluate our framework across three distinct domains. First, a representative 10-sequence subset from our custom captures serves as the primary testbed for overall pipeline comparisons. Second, we benchmark our isolated HOPG module on InterHand2.6M [37]. Finally, to rigorously evaluate both our feed-forward initialization and the final dense geometry reconstruction, we utilize the VEPHand Dataset [1], which provides high-quality per-frame volumetric pseudo-ground truth across 7 complex sequences.

4.2 Novel View Synthesis

In this section, we first evaluate novel view synthesis capability of HOPG on InterHand2.6M and then assess the full pipeline on our Custom Dataset.

HOPG on InterHand2.6M. We follow the HandSCS [13] protocol (10 training, 15 held-out views) across their exact 9 selected sequences. As shown in Table 1 (left), our method is highly competitive. Standard NeRf [7,38] or 3DGS baselines [13] artificially inflate photometric scores by relying on semi-transparent volume accumulation and baked-in illumination. In contrast, our formulation trades unconstrained photometric memorization for physical accuracy by enforcing strict, opaque 2D surfaces and explicitly disentangling albedo from specular components. We also provide qualitative rendering results in Fig. 6a.

Full pipeline on Custom Data. To evaluate the complete framework, taking only raw multi-view videos and camera parameters as input, we compare against representative dynamic scene methods: 4DGS [57] and Deformable 3DGS [63]. To focus strictly on the interaction, we utilize SAM2 [30,43] to isolate the foreground hand and object. As reported in Table 1 (right), our explicitly constrained representation outperforms these baselines by a massive margin (>4.5 dB in PSNR). Qualitatively (Fig. 5), general dynamic methods lack structural priors, causing their optimization to degrade into severe motion blur and catastrophic geometric collapse under highly non-linear kinematics. Conversely, guided by the accurate metric priors from HOST, our framework seamlessly tracks these motions while maintaining sharp visual boundaries.

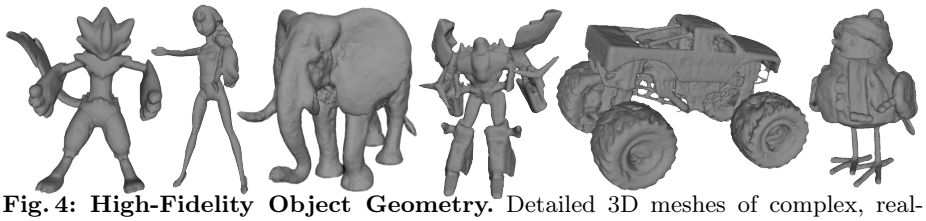

Fig. 4: High-Fidelity Object Geometry. Detailed 3D meshes of complex, real-world objects extracted via TSDF fusion.

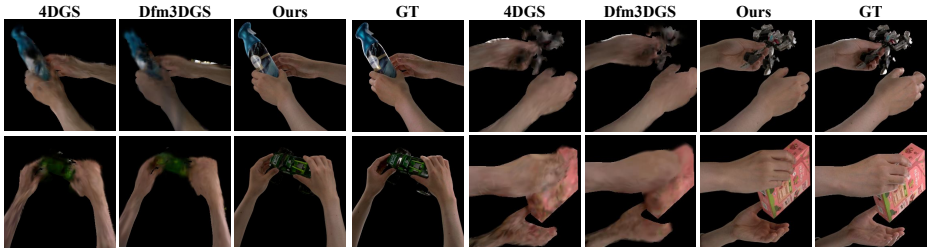

Fig. 5: Qualitative comparison on our Custom-Captured Dataset. Our method successfully tracks difficult motions and recovers complex object geometries, whereas prior-free baselines suffer from severe motion blur and geometric collapse.

4.3 Geometry Evaluation

To verify that our framework recovers 3D geometry rather than merely overfitting photometric projections, we evaluate it on the VEPHand Dataset with per-frame GT scan. We report the one-way Surface Accuracy (point-to-surface distance from the predicted vertices to the ground-truth scans, which include the forearm). Importantly, as our goal is to capture the complete interaction, all reported metrics in Table 2 are computed on the combined hand-object meshes. Specifically, we compare our HOST initialization against state-of-the-art feedforward trackers, and assess the final dense geometry refined by our full pipeline (Table 2).

We first evaluate feedforward approaches by comparing our HOST results against state-of-the-art pose estimators under a sparse 6-view setup. We compare against HORT [8], a composite baseline comprising DynHaMR [65] for hands and MV-SAM3D [33] for objects, and the multi-view hand reconstruction framework POEM [62] combined with MV-SAM3D. To establish an optimally competitive multi-view baseline for these methods, we provide them with significant informational advantages. For the inherently monocular hand trackers (HORT and DynHaMR), we elevate their predictions into metric space using ground-truth camera parameters, aggregating their root-relative articulations via a robust 3D spatial median across all 6 views followed by a rigid registration step. For the object reconstruction baseline, we utilize MV-SAM3D, an open-source multi-view extension. To ensure a highly competitive evaluation, we explicitly provide it with ground-truth camera parameters and depth maps. This allows the model to reliably determine cross-view visibility, enabling it to effectively handle occlusions.

Table 1: Quantitative NVS results. Comparison against specialized hand avatars on InterHand2.6M (left, 10-view train / 15-view test) and general dynamic baselines on our Custom-Captured Dataset (right, 17-view train / 4-view test).

InterHand2.6M			
Method	PSNR ↑	SSIM ↑	LPIPS ↓
LiveHand [38]	31.06	**0.977**	0.0384
HandAvatar [7]	32.14	0.962	0.0437
GauHuman [24]	31.73	0.957	0.0462
HandSCS [13]	**35.51**	0.974	**0.0292**
Ours (HOPG)	35.07	0.970	0.0348

Custom Dataset			
Method	PSNR ↑	SSIM ↑	LPIPS ↓
4DGS [57]	26.34	0.943	0.0814
Dfm3DGS [63]	26.32	0.947	0.0809
Ours (Full)	**30.86**	**0.968**	**0.0370**

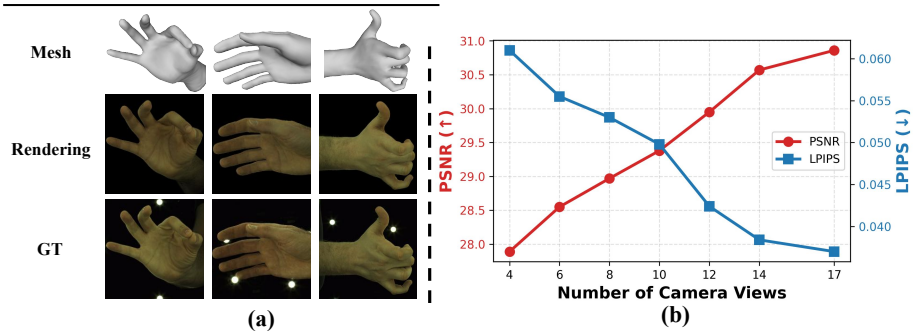

Fig. 6: InterHand2.6M Validation and View Robustness. (a) Our HOPG module achieves high-fidelity rendering and accurate geometry on bare-hand benchmarks. **(b)** End-to-end reconstruction remains highly stable across varying numbers of input cameras, maintaining strong accuracy even under extreme sparsity (4 views).

Despite arming these baselines with optimal 3D lifting and pseudo-ground-truth depth, our HOST module outperforms them end-to-end (achieving 3.07 mm vs. 5.81 mm for the monocular-lifted baseline, and 5.30 mm for the multi-view POEM baseline). This performance gap stems from the fact that while POEM leverages multi-view cues, it relies on static frame-by-frame fusion. In contrast, our factorized spatiotemporal attention effectively leverages continuous 4D context to maintain accurate and stable trajectories, whereas static multi-view trackers can still struggle with drift and temporal inconsistency during complex interactions. Building upon this robust initialization, our HOPG module successfully refines the dense geometry, further reducing the surface error to 2.35 mm.

4.4 Ablation Studies

We investigate the impact of our core design components on our Custom-Captured Dataset. A comprehensive analysis of our physical constraints and collision optimization is presented in Fig. 7.

Physics-Aware Geometry Constraint. We evaluate the necessity of our tetrahedral representation (Tet + 3D ARAP) compared to a standard surface-based approach (Surf + 2D ARAP). Since our pipeline operates entirely without

Table 2: Geometry Evaluation. One-way Surface Accuracy (mm, ↓) evaluated on combined hand-object meshes of the VEPHand Dataset. Our HOST module provides a superior metric initialization compared to feed-forward baselines, including multi-view alternatives, while our full pipeline further refines the geometry to achieve highly competitive dense accuracy. Processing time denotes the average time per frame measured on a single NVIDIA A100 GPU.

Method	Accuracy (mm) ↓	Processing Time ↓
HORT [8]	7.22	2s
DynHaMR [65] + MV-SAM3D [33]	5.81	1min
POEM [62] + MV-SAM3D [33]	5.30	1min
HOST (Ours)	3.07	2s
Full Pipeline (Ours)	**2.35**	0.5min

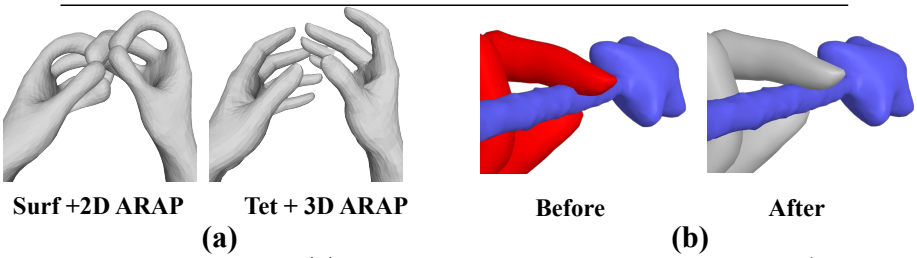

Fig. 7: Ablation studies. (a) Geometry: Our tetrahedral formulation (Tet + 3D ARAP) prevents the severe geometric flattening and volume loss seen in surface-only baselines. **(b) Collision:** Our offline collision optimization successfully resolves non-physical inter-penetrations.

ground-truth segmentation masks, optimizing exclusively via photometric rendering loss exposes the critical weakness of 2D ARAP: its inability to constrain internal volume, as illustrated in Fig. 7(a). In contrast, computing the ARAP energy over the 3D tetrahedral domain explicitly locks the internal structure. This rigorously preserves anatomical thickness, successfully preventing both texture-driven flattening and background-induced geometric collapse.

Collision Optimization. While photometric losses ensure visual alignment, they cannot constrain invisible regions. As shown in Fig. 7(b), our offline physics-based collision penalty is essential to resolve non-physical inter-penetrations, yielding plausible contact mechanics without sacrificing rendering quality.

Appearance Decoupling. As shown in Fig. 3, our physics-based model successfully factorizes the rendering into diffuse and specular components. Explicitly modeling view-dependent highlights prevents lighting artifacts from erroneously baking into the geometry. While standard 3DGS frequently warps surfaces or creates unnatural bumps to fake specular glints, our strict disentanglement eliminates these structural hallucinations. This yields pristine, noise-free normal maps, providing the geometric and photometric purity essential for extracting high-quality, relightable 4D digital assets.

Number of Camera Views. We also evaluate the system's performance under varying numbers of input camera views. In this ablation, the view restriction is strictly applied end-to-end. As shown in Fig. 6(b), reconstruction quality scales

predictably with the camera count. Notably, even in a sparse 4-view setup, the strong spatiotemporal priors from HOST prevent the optimization from collapsing, maintaining a PSNR of nearly 27.89 dB. This confirms that our framework remains highly robust under severe view scarcity.

5 Limitations

While our framework achieves high-fidelity 4D reconstruction, it operates under certain limitations. First, our system relies on synchronized and calibrated multi-view camera setups, which constrains its immediate deployment to unconstrained daily environments. Second, our pipeline currently assumes rigid objects, which prevents it from handling highly deformable items during interactions. Lastly, although our spatiotemporal prior is highly robust, systematic failure cases can still emerge under challenging scenarios: we observed tracking instabilities on geometrically thin or tiny objects (e.g., coins), and extreme motion blur or prolonged cross-hand occlusions could potentially lead to trajectory drift. Future work will integrate non-rigid deformation fields and diversify the training corpus to transition towards uncalibrated, in-the-wild captures.

6 Conclusion

We present a unified framework that bridges accurate kinematic tracking and photorealistic novel view synthesis for 4D hand-object interactions. By coupling a Hand Object Spatiotemporal Transformer (HOST) with a Physics-aware Gaussian (HOPG) module, our method robustly resolves the severe geometric flattening and volume loss common in unconstrained rendering. Furthermore, explicitly decoupling diffuse and specular reflections yields clean, noise-free normal maps, protecting the underlying 3D assets from baked-in lighting artifacts. By automating the workflow from multi-view video to high-fidelity digital assets, this framework offers an efficient solution for large-scale 4D asset creation.

References

1. Anonymous: View-efficient photometric hand performance capture at scale (2026), under review
2. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. Spie (1992)
3. 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)
4. Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G.: Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. IEEE Transactions on Visualization and Computer Graphics **31**(9), 6100–6111 (2024)

5. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025)
6. Chen, X., Wang, B., Shum, H.Y.: Hand avatar: Free-pose hand animation and rendering from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
7. Chen, X., Wang, B., Shum, H.Y.: Hand avatar: Free-pose hand animation and rendering from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8683–8693 (2023)
8. Chen, Z., Potamias, R.A., Chen, S., Schmid, C.: Hort: Monocular hand-held objects reconstruction with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6046–6057 (2025)
9. Cong, X., Xing, A., Pokhariya, C., Fu, R., Sridhar, S.: Dytact: Capturing dynamic contacts in hand-object manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7191–7203 (2025)
10. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312 (1996)
11. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018)
12. Dellaert, F.: Factor graphs and gtsam: A hands-on introduction. Georgia Institute of Technology, Tech. Rep **2**(4) (2012)
13. Dong, Y., Wang, W., Wang, Q., Yang, J., Liu, H., Zhu, X., Slabaugh, G., Yuan, S.: Handscs: Structural coordinate space for animatable hand gaussian splatting. arXiv preprint arXiv:2503.14736 (2025)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Fan, Z., Parelli, M., Kadoglou, M.E., Chen, X., Kocabas, M., Black, M.J., Hilliges, O.: 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. pp. 494–504 (2024)
16. 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 of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12943–12954 (2023)
17. Fu, R., Zhang, D., Jiang, A., Fu, W., Funk, A., Ritchie, D., Sridhar, S.: Gigahands: A massive annotated dataset of bimanual hand activities. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17461–17474 (2025)
18. 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)
19. Guo, Z., Zhou, W., Wang, M., Li, L., Li, H.: Handnerf++: Modeling animatable interacting hands with neural radiance fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
20. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)

21. 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)
22. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. International journal of computer vision **103**(3), 267–305 (2013)
23. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
24. Hu, S., Hu, T., Liu, Z.: Gauhuman: Articulated gaussian splatting from monocular human videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20418–20431 (2024)
25. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
26. Huynh, D.Q.: Metrics for 3d rotations: Comparison and analysis. Journal of Mathematical Imaging and Vision **35**(2), 155–164 (2009)
27. Ivashechkin, M., Mendez, O., Bowden, R.: Handocc: Nerf-based hand rendering with occupancy networks. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2025)
28. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
31. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: g 2 o: A general framework for graph optimization. In: 2011 IEEE international conference on robotics and automation. pp. 3607–3613. IEEE (2011)
32. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024)
33. Li, B.: Mv-sam3d: Adaptive multi-view 3d reconstruction. `https://github.com/devinli123/MV-SAM3D` (2025), accessed: March 2026
34. Liu, X., Ren, P., Qi, Q., Sun, H., Zhuang, Z., Wang, J., Liao, J., Wang, J.: Generalizable hand-object modeling from monocular rgb images via 3d gaussians. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
35. Liu, Y., Long, X.X., Habermann, M., Yang, X., Lin, C., Liu, Y., Ma, Y., Wang, W., Liu, L.: Hggt: Robust and flexible 3d hand mesh reconstruction from uncalibrated images. arXiv preprint arXiv:2603.23997 (2026)
36. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7037–7047 (2025)
37. 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. pp. 548–564. Springer (2020)
38. Mundra, A., Wang, J., Habermann, M., Theobalt, C., Elgharib, M., et al.: Livehand: Real-time and photorealistic neural hand rendering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 18035–18045 (2023)

39. 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)
40. Peng, B., Tao, Y., Zhan, H., Guo, Y., Zhang, J.: Pica: Physics-integrated clothed avatar. IEEE Transactions on Visualization and Computer Graphics (2025)
41. Pokhariya, C., Shah, I.N., Xing, A., Li, Z., Chen, K., Sharma, A., Sridhar, S.: Manus: Markerless grasp capture using articulated 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2197–2208 (2024)
42. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
43. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
44. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
45. Saito, S., Schwartz, G., Simon, T., Li, J., Nam, G.: Relightable gaussian codec avatars. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 130–141 (2024)
46. Shen, Y., Zhang, Z., Qu, Y., Zheng, X., Ji, J., Zhang, S., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
47. Sorkine, O., Alexa, M., et al.: As-rigid-as-possible surface modeling. In: Symposium on Geometry processing. vol. 4, pp. 109–116 (2007)
48. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: European conference on computer vision. pp. 581–600. Springer (2020)
49. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025), `https://arxiv.org/abs/2511.16624`
50. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(11), 9426–9437 (2024)
51. Wang, H., Zhou, H., Liu, H., Yan, L.: 4d-vggt: A general foundation model with spatiotemporal awareness for dynamic scene geometry estimation. arXiv preprint arXiv:2511.18416 (2025)
52. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
53. Wang, S., He, H., Parelli, M., Gebhardt, C., Fan, Z., Song, J.: 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. pp. 5957–5968 (2025)
54. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)

55. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
56. Wang, Y., Xu, H., Heng, P.A., Fu, C.W.: Unihope: A unified approach for hand-only and hand-object pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12231–12241 (2025)
57. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
58. Wu, X., Bai, Y., Li, M., Wu, X., Zhao, X., Lai, Z., Liu, W., Wang, X.: 4dlangvggt: 4d language-visual geometry grounded transformer. arXiv preprint arXiv:2512.05060 (2025)
59. Xie, X., Wen, B., Chang, Y., Rabeti, H., Li, J., Yuan, Y., Pons-Moll, G., Birchfield, S.: Cari4d: Category agnostic 4d reconstruction of human-object interaction. arXiv preprint arXiv:2512.11988 (2025)
60. 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. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21924–21935 (2025)
61. 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)
62. Yang, L., Zhong, L., Zhu, P., Zhan, X., Kong, J., Xu, J., Lu, C.: Multi-view hand reconstruction with a point-embedded transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
63. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
64. Ye, Y., Gupta, A., Tulsiani, S.: What's in your hands? 3d reconstruction of generic objects in hands. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3895–3905 (2022)
65. Yu, Z., Zafeiriou, S., Birdal, T.: Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27716–27726 (2025)
66. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. arXiv preprint arXiv:2410.03825 (2024)
67. Zhang, Y., Zhang, L., Ma, R., Cao, N.: Texverse: A universe of 3d objects with high-resolution textures. arXiv preprint arXiv:2508.10868 (2025)
68. Zhenyuan, L., Guo, Y., Li, X., Bickel, B., Zhang, R.: Bigs: Bidirectional primitives for relightable 3d gaussian splatting. In: 2025 International Conference on 3D Vision (3DV). pp. 1022–1031. IEEE (2025)
69. Zhou, Q.Y., Park, J., Koltun, V.: Fast global registration. In: European conference on computer vision. pp. 766–782. Springer (2016)

Supplementary Material

This supplementary document provides additional details omitted from the main text due to space constraints. It is organized into two primary sections: Section A presents extended ablation studies. Section B provides comprehensive implementation and training details, including the full formulation of our HOST, multi-task loss functions, and hyperparameter schedules for both the HOST and HOPG modules. Additionally, a supplementary demo video is provided to showcase our results. The video specifically contrasts the robust initialization provided by the HOST model with the final high-fidelity 4D reconstructions.

A Additional Ablation Studies

A.1 HOST: Network Architecture and Training Data

To isolate the impact of our architectural designs and training data, we conduct an ablation study on the raw network predictions, bypassing downstream refinements such as Pose Graph Optimization (PGO) and ICP. We evaluate performance using direct regression metrics: Hand Vertex Reconstruction Error (MSE between vertices derived from predicted and ground-truth MANO parameters), Object Rotation Error (Geodesic), and Object Translation Error (MSE). This evaluation is performed on a reserved validation set of 200 non-synthetic samples (5 frames $\times$ 6 views), with results summarized in Table 3.

Table 3: Ablation Study on HOST Architecture and Data. Performance of core modules and data components using direct regression metrics on raw network outputs. All MSE metrics are scaled by 10^3 for readability.

Model Configuration	**Hand Vertices** $\downarrow$ (MSE $\times 10^3$)	**Obj. Rot.** $\downarrow$ (Geodesic)	**Obj. Trans.** $\downarrow$ (MSE $\times 10^3$)
(a) Naive VGGT Attn. (Frame-Global)	6.65	0.0928	55.10
(b) w/o Camera Embedding	7.12	0.0933	64.30
(c) w/o Time Token	7.36	0.1011	74.10
(d) w/o Dense Geometry Heads	6.86	0.0939	65.20
(e) w/o Synthetic Data Training	7.28	0.2849	87.80
(f) w/o HOPG Pseudo-GT	8.06	0.1135	78.50
HOST (Full Model)	**6.64**	**0.0901**	**54.50**

Architectural Designs. Replacing our factorized spatiotemporal attention with the naive frame-global attention from VGGT (Row a) yields comparable regression accuracy but significantly increases computational complexity and memory footprint, limiting efficient multi-view processing. The inclusion of camera embeddings (Row b) and time tokens (Row c) is essential for the network to resolve

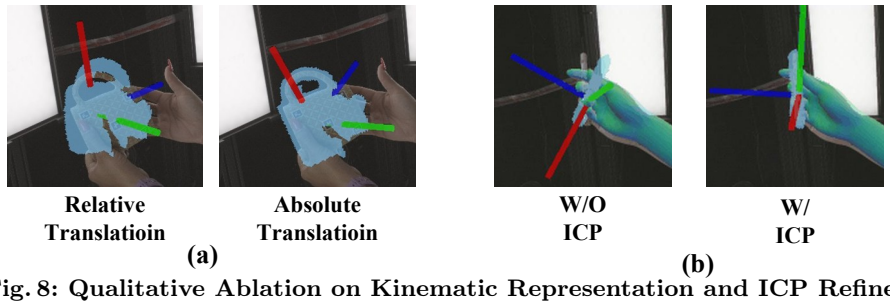

Fig. 8: Qualitative Ablation on Kinematic Representation and ICP Refinement. (Left) Compared to relative offsets, our absolute centroid regression provides a global anchor that eliminates cumulative translational drift. (Right) PGO aggregates short-window predictions but is susceptible to rotational drift over long sequences. Our ICP stage uses predicted geometric priors as absolute anchors to correct these offsets and ensure long-term pose consistency.

spatiotemporal cues. Specifically, when time tokens are omitted, the model predicts hand and object states by averaging visual image tokens, which leads to a decrease in both localization accuracy and temporal consistency. Furthermore, although removing the dense geometry heads (row d) only results in a slight decrease in the raw regression metrics, these heads are necessary for providing the 3D spatial priors (e.g. depth and masks) needed for the Pose Graph Optimization and ICP refinement stages.

Training Data Distribution. Rows (e) and (f) evaluate the impact of our training data. Excluding the synthetic object dataset (Row e) reduces the network's ability to generalize for 6D rotation estimation on unseen objects. Additionally, fine-tuning with HOPG-refined pseudo-ground truth (Row f) is necessary to bridge the domain gap.

A.2 HOST: Translation Representation

As discussed in the main paper, we use an asymmetric kinematic representation of relative rotation combined with absolute centroid translation. We validate this design by comparing it to a fully relative representation. As illustrated in Fig. 8 (left), relative translation inevitably leads to cumulative error and severe positional drift over long sequences. In contrast, regressing the absolute geometric center provides a global anchor for the trajectory. This ensures that translation errors remain bounded and consistent, regardless of sequence duration.

A.3 HOST: ICP Correction

Although our PGO module successfully combines the predicted relative rotations of the short-window transformer into a continuous trajectory, the resulting global poses remain susceptible to cumulative rotational drift over long sequences. To address this issue, we introduce an ICP-based refinement stage that uses the

dense geometry priors (depth and masks) predicted by our dense heads. Aligning the object point cloud to these per-frame spatial anchors provides an absolute geometric reference through ICP correction, which rectifies the accumulated drift. As illustrated in Fig. 8 (right), this step ensures long-term pose consistency.

B Method and Implementation Details

In this section, we provide comprehensive architectural details, loss formulations, and training hyperparameters for both the Hand Object Spatiotemporal Transformer (HOST) and the Hand Object Physics-aware Gaussian (HOPG) refinement module.

B.1 Hand Object Spatiotemporal Transformer (HOST)

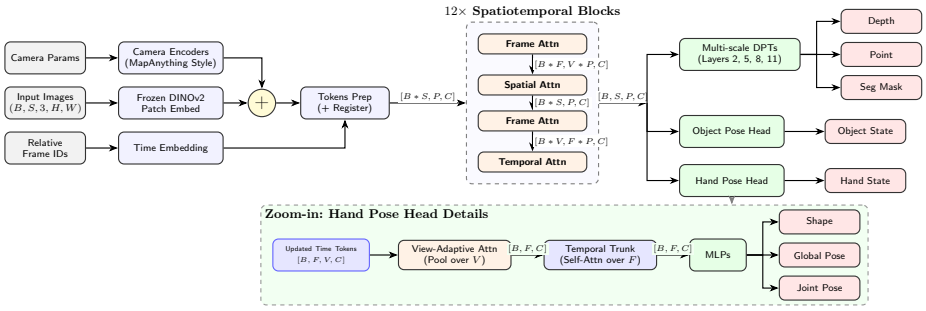

Fig. 9: Detailed Architecture of HOST. The network extracts features using a frozen DINOv2 backbone integrated with camera and time embeddings. Given an input sequence of F frames from V views ($S = F \times V$ images total), 12 spatiotemporal blocks process the tokens by alternating between spatial (cross-view) and temporal (cross-frame) attention. These shared feature representations are then fed into task-specific heads: a dense geometry decoder for depth, point, and mask prediction, and parametric heads for hand and object pose regression. In the pose heads (bottom), updated time tokens from all V views are adaptively pooled into canonical per-frame features, which are then processed by a temporal trunk to ensure motion continuity prior to kinematic parameter regression.

Patch and Camera Embeddings. We process 308×308 images using a frozen DINOv2 [39] (ViT-L/14) patch embedding layer. To inject geometric and temporal awareness, we follow MapAnything [28] to augment each patch token with ray-based camera embeddings, which encode camera geometry via 4-layer MLPs. For temporal information, we map relative frame IDs through a learnable time embedding layer to obtain frame-specific time tokens. These time tokens are

concatenated with the image tokens and 4 register tokens [52] before being fed into the spatiotemporal aggregator.

Spatiotemporal Attention. The core aggregator consists of $L = 12$ Transformer blocks. As detailed in the main paper, we decompose the full global attention into a sequential factorized scheme (Frame $\rightarrow$ Spatial $\rightarrow$ Frame $\rightarrow$ Temporal) to maintain computational tractability. In our implementation, each sub-attention layer retains the same hidden dimension $C = 1024$ as the VGGT backbone [52]. To effectively leverage pre-trained geometric priors, we initialize the aggregator by mapping weights from VGGT: the frame attention weights are directly transferred to our frame attention layers, while the weights from VGGT's global attention layers are sequentially mapped to initialize our spatial and temporal attention blocks.

Task-Specific Heads. For dense predictions including depth, masks, and point maps, we employ the DPT-based decoders [42] following VGGT [52], which leverage multi-scale features from intermediate layers $\{2, 5, 8, 11\}$.

For kinematic states, we first extract the updated time tokens $\hat{\mathbf{T}}^{f,v}_{\text{time}}$ from all views. To resolve occlusions and compute robust frame-level features $\mathbf{z}_f$, we introduce a view-adaptive attention mechanism applied separately to each entity. For the hand or object, a dedicated MLP maps the corresponding time tokens to softmax-normalized visibility weights $\alpha_{f,v}$ across all V views:

$$\alpha_{f,v} = \frac{\exp(\text{MLP}_{\text{attn}}(\hat{\mathbf{T}}^{f,v}_{\text{time}}))}{\sum_{k=1}^{V} \exp(\text{MLP}_{\text{attn}}(\hat{\mathbf{T}}^{f,k}_{\text{time}}))} \tag{10}$$

The resulting aggregated feature $\mathbf{z}_f = \sum_{v=1}^{V} \alpha_{f,v} \hat{\mathbf{T}}^{f,v}_{\text{time}}$ is then processed by a separate temporal trunk—consisting of self-attention blocks over the temporal dimension F—to ensure motion smoothness and continuity for each entity. Finally, dedicated MLP decoders regress these temporally smoothed features into kinematic parameters. Specifically, the object head predicts absolute 3D translations and relative 6D rotations, while the hand head regresses 10D shape parameters, 9D global poses, and 63D local joint Euler angles [44].

Loss Formulation The HOST is trained end-to-end using a multi-task loss:

$$\mathcal{L}_{total} = \lambda_{depth}\mathcal{L}_{depth} + \lambda_{point}\mathcal{L}_{point} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{obj}\mathcal{L}_{obj} + \lambda_{hand}\mathcal{L}_{hand} \tag{11}$$

During the joint training stage, we set $\lambda_{depth} = 0.5$, $\lambda_{point} = 0.5$, $\lambda_{mask} = 1.0$, $\lambda_{obj} = 1.0$, and $\lambda_{hand} = 1.0$.

Dense Losses. We supervise segmentation masks with a standard binary cross-entropy loss. For depth and 3D point maps, we adopt the same loss formulation as VGGT [52].

Kinematic Losses. We employ the geodesic loss for all 3D rotations:

$$\mathcal{L}_{geo}(\mathbf{R}_{pred}, \mathbf{R}_{gt}) = \arccos\left(\frac{\text{tr}(\mathbf{R}_{pred}\mathbf{R}_{gt}^T) - 1}{2}\right) \tag{12}$$

The object pose is supervised through global translation and rotation errors:

$$\mathcal{L}_{obj} = \lambda_{trans}\|\mathbf{t}_{pred} - \mathbf{t}_{gt}\|_2 + \lambda_{rot}\mathcal{L}_{geo}(\mathbf{R}_{pred}, \mathbf{R}_{gt}) \tag{13}$$

To accurately recover the articulated hand, we jointly supervise the global pose, local joint rotations, shape parameters, and the resulting mesh vertices.

$$\begin{aligned}\mathcal{L}_{hand} = \lambda_{trans}\|\mathbf{t}_{pred} - \mathbf{t}_{gt}\|_2 + \lambda_{rot}\mathcal{L}_{geo}(\mathbf{R}_{pred}, \mathbf{R}_{gt}) \\ + \lambda_{joint}\mathcal{L}_{geo}(\mathbf{R}_{pred}^{joint}, \mathbf{R}_{gt}^{joint}) + \lambda_{shape}\|\alpha_{pred} - \alpha_{gt}\|_2 \\ + \lambda_{vert}\|\mathbf{V}_{pred} - \mathbf{V}_{gt}\|_2\end{aligned} \tag{14}$$

For object and hand loss, we set the loss weights to $\lambda_{trans} = 5.0$ and $\lambda_{vert} = 5.0$, while all other coefficients are set to 1.0.

Training Strategy and Hyperparameters HOST is implemented in JAX/Flax and optimized using the AdamW optimizer in bfloat16 precision. Given that the backbone is initialized with robust pre-trained weights, we apply a learning rate multiplier of 5.0 to the newly added mask, hand, and object heads to harmonize their optimization with the deep backbone. The training pipeline is structured into two stages:

Stage 1: Large-scale Pre-training. We initialize our network using pre-trained weights from VGGT [52] to leverage robust geometric priors. Specifically, we load weights for the DINO backbone, attention aggregator, depth head, and point head directly. The model is trained on 64 TPU v5 accelerators for 200,000 steps with a total batch size of 64, using the large-scale VEPHand dataset [1] introduced in the main paper. The base learning rate is 8.0×10^{-5} with a 5,000-step linear warmup, followed by a piecewise constant decay by factors of $\{0.5, 0.2, 0.1\}$ at steps $\{60\text{k}, 80\text{k}, 120\text{k}\}$.

Stage 2: Fine-tuning. The network is fine-tuned for 200,000 steps on 8 NVIDIA H100 GPUs with a total batch size of 16, using a data mixture of the VEPHand dataset, synthetic object dataset [67] and pseudo-ground-truth sequences refined by HOPG. The base learning rate is reduced to 1.0×10^{-5} with a 5,000-step linear warmup to adapt to the target data distribution.

Geometry Warmup. At the beginning of both training stages, we enforce a geometry warmup for the initial 2,000 steps by setting $\lambda_{obj} = \lambda_{hand} = 0.0$. This allows the dense geometric features to stabilize before regressing kinematic poses.

B.2 Hand Object Physics-aware Gaussian (HOPG) Details

Mesh-Aligned 2D Gaussian Parameterization To reconstruct high-fidelity surfaces while preserving kinematic consistency, we anchor 2D Gaussians directly to the hand template meshes. Following the barycentric parameterization in PICA [40], we extend the formulation by employing a hybrid sampling strategy that anchors Gaussians to both triangular faces and mesh vertices. For each

vertex-anchored Gaussian, the normal vector $\mathbf{v}_0$ is derived from the vertex normal, while the tangent vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are determined via Gram-Schmidt orthogonalization against a reference axis. This hybrid parameterization ensures that the 2D Gaussians remain strictly coupled with the deforming mesh, preventing spatial scattering and ensuring surface integrity during articulated motion.

Pose-Conditioned Volumetric Deformation As described in the main paper, we predict the non-rigid displacements for the tetrahedral vertices $\boldsymbol{\rho}_c$ using a lightweight MLP consisting of 5 layers with a hidden dimension of 256. This network takes both the canonical coordinates $\boldsymbol{\rho}_c$ and the 63-dimensional axis-angle pose β as input to generate pose-dependent volumetric offsets. To ensure numerical stability, the final linear layer is initialized from a uniform distribution bounded by $[-10^{-5}, 10^{-5}]$. This initialization yields near-zero initial offsets, allowing the as-rigid-as-possible (ARAP) energy and Linear Blend Skinning (LBS) to provide a stable geometric starting point before the model learns fine-grained non-rigid deformations.

Normal-Depth Consistency To enhance surface smoothness and geometric accuracy, we employ the normal-depth consistency loss as proposed in 2DGS [25]. Specifically, we enforce an L1 alignment between the rendered normals and the pseudo-ground-truth normals derived from the spatial gradients of the depth map. Rather than using standard median depth, we utilize the depth Z_{pgsr} generated via Plane-based Gaussian Splatting (PGSR) [4], which provides more robust geometric constraints:

$$Z_{pgsr} = \frac{\sum \alpha_i d_i}{\sum \alpha_i |\mathbf{n}_i \cdot \mathbf{r}| + \epsilon} \tag{15}$$

As in the original 2DGS formulation, we apply a soft-blending mask to ignore pixels with grazing viewing angles, ensuring stable gradient propagation for the consistency loss.

ARAP and Offline Collision Optimization To prevent internal volume collapse during non-rigid deformation, we enforce an As-Rigid-As-Possible (ARAP) energy on the tetrahedral mesh. The energy is computed between the canonical tetrahedral vertices $\boldsymbol{\rho}_c$ and the deformed vertices $\boldsymbol{\rho}'_c$:

$$\mathcal{L}_{arap} = \sum_i \sum_{j \in \mathcal{N}(i)} w_{ij} \left\| (\boldsymbol{\rho}'_{c,i} - \boldsymbol{\rho}'_{c,j}) - \mathbf{R}_i (\boldsymbol{\rho}_{c,i} - \boldsymbol{\rho}_{c,j}) \right\|_2^2 \tag{16}$$

where $\mathbf{R}_i$ is the optimal rotation matrix estimated via Singular Value Decomposition (SVD) for the local neighborhood of the i-th tetrahedral cell.

Offline Collision Resolution. Since photometric supervision alone may not fully resolve physical penetrations, we incorporate a physics-aware refinement stage to enforce contact constraints. We categorize collisions into *self-collision*

(intra-hand) and *inter-collision* (hand-object). Specifically, for a set of query vertices $\{\mathbf{p}_q\}$, we first utilize a high-performance CUDA kernel to perform a point-in-mesh containment check, generating a binary mask $\mathcal{M}$ where $\mathcal{M}_q = 1$ for vertices situated inside the reference geometry. For each identified penetration, we compute its nearest reference surface point $\mathbf{p}_{ref}$ and minimize the distance to the surface projected along the query vertex's normal $\mathbf{n}_q$:

$$\mathcal{L}_{coll} = \sum_q \mathcal{M}_q \cdot ((\mathbf{p}_q - \mathbf{p}_{ref}) \cdot \mathbf{n}_q)^2 \tag{17}$$

Every 15,000 global training steps, we trigger a dedicated offline optimization phase. During this phase, we perform 1,000 iterations using the Adam optimizer to update the hand pose parameters $\Delta\beta$ exclusively, thereby effectively resolving any accumulated intersections.

Loss Functions and Optimization Details To explicitly decouple environmental lighting parameters from the intrinsic appearance of Gaussian primitives, the HOPG module is optimized using a batch size of $N_b = 5$. This joint optimization across multiple frames allows the model to differentiate between view-dependent lighting effects and the consistent underlying geometry and texture. For a typical 21-view, 600-frame sequence, the optimization runs for 30,000 iterations, taking approximately 5 hours on a single NVIDIA A100 GPU. The overall objective function is formulated as:

$$\begin{aligned}\mathcal{L}_{HOPG} = \lambda_{photo}\mathcal{L}_{L1} + \lambda_{dssim}\mathcal{L}_{D-SSIM} + \lambda_{lpips}\mathcal{L}_{LPIPS} \\ + \lambda_{arap}\mathcal{L}_{arap} + \lambda_{lap}\mathcal{L}_{lap} + \lambda_{norm}\mathcal{L}_{norm} + \lambda_{smooth}\mathcal{L}_{smooth}\end{aligned} \tag{18}$$

The primary photometric loss weights are set to $\lambda_{photo} = 1.0$, $\lambda_{dssim} = 0.2$, and $\lambda_{lpips} = 2.0$. For geometric regularization, we employ a Tet-ARAP weight of $\lambda_{arap} = 5000.0$ and a Laplacian smoothness weight of $\lambda_{lap} = 20000.0$ to preserve mesh structural integrity, alongside a PGSR normal consistency weight $\lambda_{norm} = 1.0$. Temporal smoothness penalties $\mathcal{L}_{smooth}$ are applied to both hand poses θ and object trajectories with $\lambda_{smooth} = 10.0$ to ensure kinematic continuity.

C Computational Efficiency and Scalability

As our system targets high-fidelity offline 4D asset reconstruction, we prioritize robustness and accuracy. Below, we detail the computational requirements of our two main stages.

HOST Inference: With a temporal window $F = 5$, HOST's cost comprises model inference (which scales roughly quadratically with the number of views V due to the attention mechanism) and view-agnostic post-processing (PGO+ICP, around 1 frame/s). Table 4 details the runtime and peak memory measurements on a single NVIDIA H100 GPU.

Table 4: HOST Inference Runtime and Peak Memory relative to the number of views (V) on a single H100 GPU.

Number of Views (V)	4	6	10	19
HOST Infer Time (s)	0.34	0.82	1.24	4.38
Peak VRAM (GB)	∼7.4	∼11.5	∼23.8	∼75.6

HOPG Optimization: For the HOPG refinement stage, the optimization takes approximately 5 hours per sequence on an NVIDIA A100 GPU. This computational cost is standard for high-fidelity, physics-aware offline asset reconstruction. Furthermore, to scale to large datasets, this sequence-level optimization is inherently decoupled and can be trivially distributed across compute clusters in parallel.