PhyVGGT: Feed-Forward Articulated Object Reconstruction from Monocular Videos

Qi Shao, Zhihao Cao, Shuhao Zhai, Anh Nguyen, Hong Liang, Baoru Huang, and Liang Li

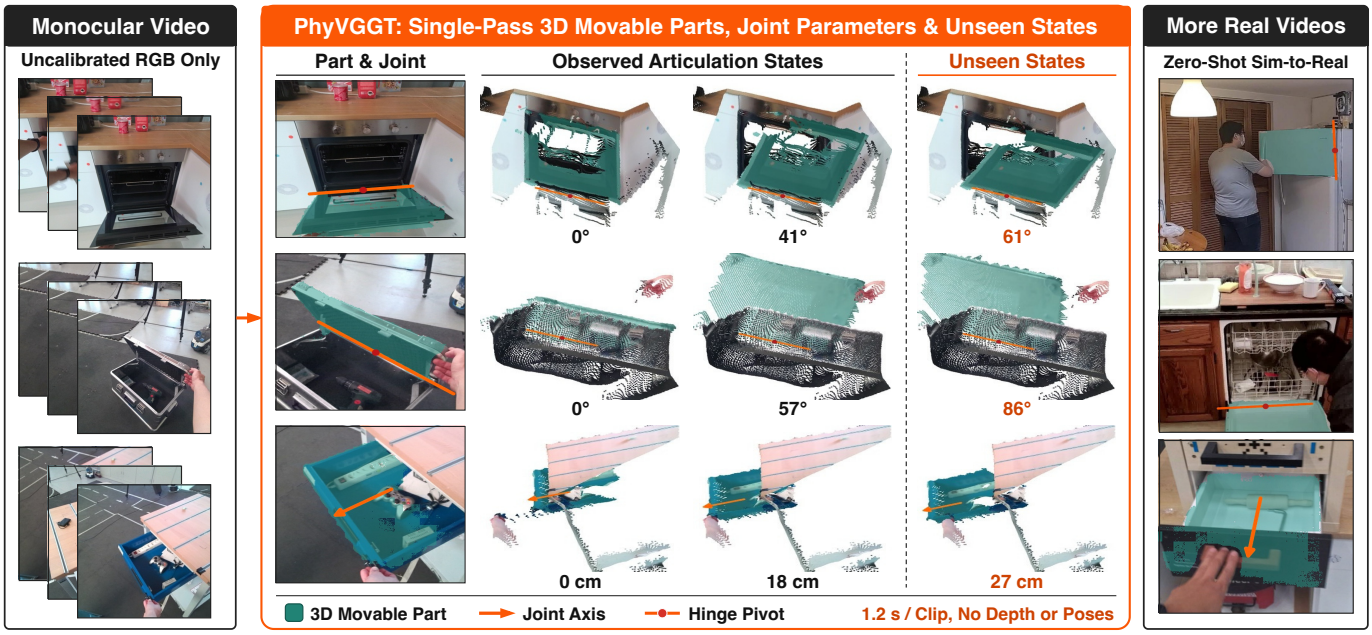


Fig. 1. **Overview of single-pass 3D articulated object reconstruction from uncalibrated monocular RGB videos.** Given casual RGB frames (*left*), PhyVGGT segments 3D movable parts (teal), estimates joint axes (orange) and hinge pivots (red) across observed articulation states, and drives each part to unseen states in 1.2 s per clip (*centre*), generalising zero-shot from synthetic training to diverse real-world scenes (*right*).

***Abstract*—Reconstructing an articulated object with its 3D movable parts, joint parameters and unseen articulation states from an uncalibrated monocular video is essential for robot manipulation and interactive digital twins. Existing per-scene optimisation methods require minutes to hours and external depth or camera poses, whereas feed-forward 3D foundation models reconstruct dense point maps in one pass without segmenting movable parts or inferring joints. Even on these monocular point maps, regressing or fitting continuous joints directly fails on casual videos, since cross-frame attention blurs moving part boundaries in deep layers, underestimates motion along the viewing ray and leaves continuous axes ill-conditioned under small opening angles. To address these challenges, we propose PhyVGGT, a single-pass framework on a frozen 3D foundation model that anchors moving-part matching and closed-form joint estimation to the undistorted geometry of the stationary base. First, our Part and Motion Decoder fuses intermediate per-frame tokens with global tokens and references part queries to a static-base query to segment 3D parts, align point maps on the static base and predict learned Bingham priors. Second, our Kinematic Inference Layer fits an orthogonal Manhattan frame to static surface normals, selects the joint axis in closed form under a Bingham posterior combining pairwise motion scatter with the learned priors, and solves a linear system for the hinge pivot to drive unseen states. Trained solely on synthetic data, PhyVGGT achieves state-of-the-art monocular performance across five benchmarks, outperforming FAMOS and monocular ArtiPoint, surpassing posed RGB-D ArtiPoint in median error, and running in 6.8 s per 48-frame video against hours of per-scene optimisation in Articulat3D. Project page: https://phyvggt.github.io.**

***Index Terms*—Articulated object reconstruction, joint estimation, 3D foundation models, monocular video, feed-forward reconstruction.**

I. Introduction

Robotic manipulation and interactive simulation require reconstructing an articulated object in 3D, including its geometry, movable parts, joint types, axes, pivots and joint states. Although feed-forward 3D foundation models [1–3] recover dense point maps and cameras from uncalibrated images in one pass and token query heads segment static spatial relations [4], neither models how object parts move. Conversely, per-scene optimisation methods [5–8] fit explicit motion models to multi-state scans or videos, yet require minutes to hours per object alongside sensor depth, camera poses or known kinematic trees. While recent feed-forward articulation predictors [9–11] avoid per-scene fitting, they regress unconstrained joint vectors or fit continuous displacements directly in 3D, where monocular depth errors corrupt the recovered axis and pivot.

Attaching a prediction head or motion fitter directly to a frozen 3D foundation model fails on casual videos because

alternating intra- and cross-frame attention [2] treats the scene as static. In deep global-attention layers, moving part edges blur into the stationary body and camera drift mixes with part motion, while cross-frame context biases moving-part depth towards other frames and underestimates displacement along the viewing ray. When the canonical rest pose is also unknown, Fisher information for continuous axis and pivot estimation vanishes quadratically under small opening angles [12], leaving unconstrained regression and motion fitting ill-conditioned.

Unlike moving parts, the stationary base is unaffected by cross-frame depth bias and remains geometrically accurate across frames. In manufactured cabinets, appliances and furniture, flat faces meet at right angles, so revolute and prismatic axes align with the three orthogonal directions of the static base's Manhattan frame [13] while hinge pivots lie along the boundary between the part and the base. Benefiting from this undistorted anchor, we redesign both feature decoding and joint inference around the stationary base. Referencing part queries to the static base cancels camera drift during dense matching, and restricting candidate axes to the static Manhattan frame turns ill-conditioned continuous regression into closed-form three-way selection that tolerates up to 45° of tilt towards the viewing ray.

Based on this design, we propose PhyVGGT, a single-pass framework on a frozen visual geometry model [2] that reconstructs 3D movable parts, joint parameters and unseen articulation states from an uncalibrated monocular video, as shown in Figure 1. First, our Part and Motion Decoder introduces multi-layer per-frame and global token fusion together with an explicit static-base query, recovering sharp 3D part masks, correspondences aligned to the static base and learned Bingham priors. Second, our Kinematic Inference Layer fits an orthogonal Manhattan frame to static surface normals via fourth-order alignment, selects the joint axis in closed form under a Bingham posterior [14] combining pairwise motion scatter with the learned priors, and solves a positive-definite linear system for the hinge pivot to drive unseen states.

Our contributions are as follows:

- We introduce PhyVGGT, a single-pass framework on a frozen 3D model that anchors part matching and closed-form joint estimation to the stationary base to reconstruct 3D parts, joints and unseen states from a monocular video.
- We propose a Part and Motion Decoder that fuses multi-layer per-frame and global tokens and references part queries to a static-base query to recover sharp 3D part masks and correspondences aligned on the static base.
- We design a Kinematic Inference Layer that restricts joint estimation to the Manhattan frame of the static base, removing depth bias along the viewing ray through closed-form Bingham axis selection and linear hinge pivot estimation.

II. Related Work

A. Articulated Object Reconstruction

Early learning methods predict part coordinates and joint parameters from depth scans [15, 16], whereas per-scene optimisation fits neural radiance fields or 3D Gaussians to discrete articulation states [5, 6] or continuous videos [7, 8]. Articulation in Prime [17] shortens per-object fitting to roughly ten minutes, yet still requires external depth maps, camera poses and part masks. Casual and egocentric pipelines likewise depend on RGB-D input [18]: ArtiPoint [19] fits joints to tracked 3D interaction points under known odometry, MoMa-SG [20] constructs articulated scene graphs from RGB-D sequences, and PAWS [21] lifts 2D hand trajectories with sensor depth and poses before querying a vision-language model over surface candidates. Avoiding per-scene optimisation and depth sensors, single-image predictors infer openable parts from appearance priors [22–25], ArtSplat [9] and FAMOS [10] predict joints from multi-state point clouds, and sim2art [11] estimates part masks and continuous joint vectors from monocular videos. Yet predicting unconstrained vectors in $\mathbb{R}^3$ or fitting continuous axes directly to monocular point maps propagates depth errors along the viewing ray into the joint estimate, whereas PhyVGGT anchors part matching and closed-form Bingham inference to the Manhattan frame of the static base in one RGB pass.

B. Feed-Forward 3D Foundation Models

DUSt3R [1] and MASt3R [26] predict dense 3D point maps from uncalibrated image pairs in a shared coordinate frame, while VGGT [2] and π^3 [3] recover point maps, camera poses and point tracks across an entire video in one forward pass. Recent extensions adapt visual geometry transformers to causal streaming and dynamic 4D scenes [27–30], and such reconstruction priors also support monocular dense mapping across multiple agents [31, 32]. RelationVGGT [4] attaches a lightweight query decoder to a frozen VGGT backbone for static 3D spatial relation segmentation. Although these foundation models reconstruct accurate stationary 3D surfaces in a single pass, their alternating intra- and cross-frame attention blurs moving part boundaries in deep layers and biases moving-part depth towards context frames, underestimating motion along the viewing ray. PhyVGGT overcomes both issues by fusing intermediate per-frame and global tokens relative to a static-base query and anchoring joint inference to undistorted static surface normals.

III. Method

Given an uncalibrated monocular RGB video $\{\mathbf{I}_t\}_{t=1}^{T}$, we aim to reconstruct an articulated object consisting of a stationary base b and up to K movable rigid parts in a single forward pass. Each movable part k connects to the base via a one-degree-of-freedom joint defined by its kinematic type $y_k \in \{\text{rev}, \text{pris}\}$, unsigned unit axis $\mathbf{a}_k \in \mathbb{RP}^2$, hinge pivot $\mathbf{o}_k \in \mathbb{R}^3$ and per-frame joint states $\{q_{kt}\}_{t=1}^{T}$. Together, these parameters specify a unit twist $\boldsymbol{\xi}_k$, equal to $(\mathbf{a}_k, \mathbf{o}_k \times \mathbf{a}_k)$ for a revolute joint and $(\mathbf{0}, \mathbf{a}_k)$ for a prismatic joint, which transforms any homogeneous 3D point $\mathbf{x}$ on part k from frame 1 to frame t:

$$\mathbf{x}_t = \exp\big(q_{kt}\,\hat{\boldsymbol{\xi}}_k\big)\,\mathbf{x}, \qquad q_{k1} = 0, \tag{1}$$

where $\exp(q_{kt}\hat{\boldsymbol{\xi}}_k) \in SE(3)$ is generated by $\hat{\boldsymbol{\xi}}_k \in \mathfrak{se}(3)$. From RGB frames alone, a frozen VGGT [2] backbone predicts world-frame point maps $\{\mathbf{X}_t\}_{t=1}^{T}$, tracking features and multi-layer tokens. Benefiting from the undistorted stationary surfaces

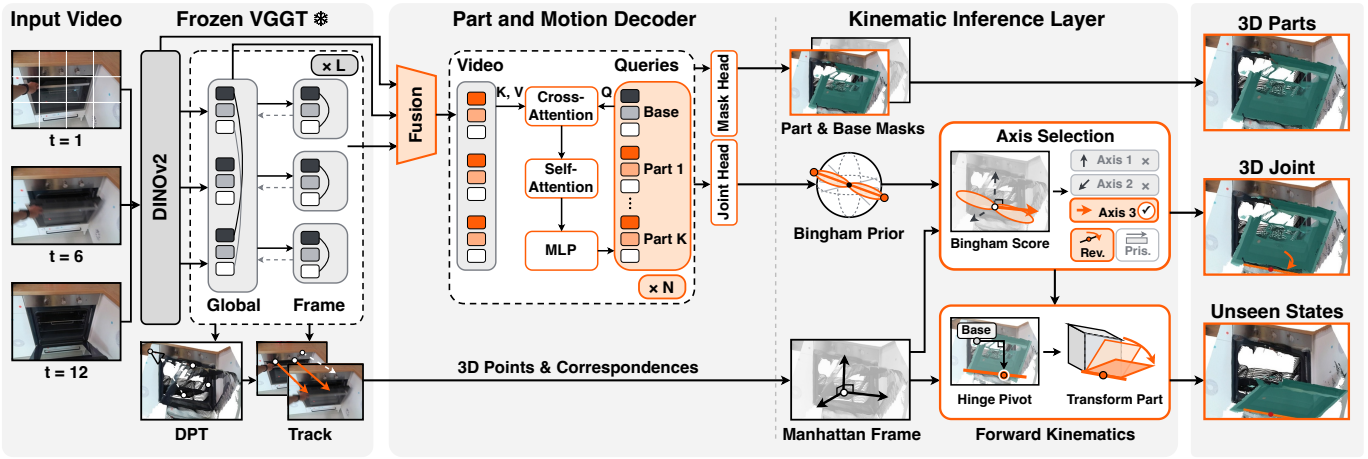


Fig. 2. **Architecture overview of PhyVGGT.** Given an uncalibrated monocular video, a frozen VGGT backbone extracts patch, per-frame and global tokens alongside 3D point maps (*DPT*) and tracking features (*Track*). First, our trainable Part and Motion Decoder fuses these streams and updates learnable base and part queries across N attention blocks to output part masks (*Mask Head*), correspondences and Bingham priors (*Joint Head*). Second, our Kinematic Inference Layer fits a *Manhattan Frame* to static normals, scores its three axes via a three-way softmax (*Axis Selection*) and solves for joint type, *Hinge Pivot* and *Unseen States* via *Forward Kinematics*.

in $\{\mathbf{X}_t\}_{t=1}^{T}$, PhyVGGT anchors two sequential modules to the static base as shown in Figure 2: our Part and Motion Decoder in Sec. III-A segments 3D parts and recovers rigid part motion relative to the static base, and our Kinematic Inference Layer in Sec. III-B selects joint axes on the Manhattan frame of the static base and solves for pivots and states.

A. Part and Motion Decoder

In VGGT's alternating aggregator, per-frame self-attention preserves sharp intra-frame part edges, whereas deep cross-frame global attention captures inter-frame motion while smoothing moving boundaries into the static body. To retain both cues, our Part and Motion Decoder extracts per-frame and global tokens from multiple VGGT layers and fuses them with DINOv2 [33] patch tokens $\mathbf{D}_t$ and VGGT tracking features to form per-frame features $\mathbf{U}_t^{\mathrm{f}}$ and global features $\mathbf{U}_t^{\mathrm{g}}$. To decode parts relative to the static body, we maintain $K+2$ learnable queries $\mathbf{Q} = [\mathbf{q}_{\mathrm{bg}}, \mathbf{q}_{\mathrm{b}}, \mathbf{q}_1, \ldots, \mathbf{q}_K]$ for background, the static base and K movable parts, replicated per frame as $\mathbf{Q}_t$. In each of the N residual decoder blocks shown in the middle of Figure 2, $\mathbf{Q}_t + \mathbf{Q}$ first cross-attends to keys and values from $\mathbf{U}_t^{\mathrm{f}}$ within frame t to localise sharp part boundaries, followed by a feed-forward layer. Clip queries $\mathbf{Q}$ then add the temporal mean of $\{\mathbf{Q}_t\}$, cross-attend to keys and values from $\{\mathbf{U}_t^{\mathrm{g}}\}_{t=1}^{T}$ across frames, separate part instances via self-attention, and update each part query $\mathbf{q}_k$ through an MLP on $[\mathbf{q}_k, \mathbf{q}_{\mathrm{b}}, \mathbf{q}_k - \mathbf{q}_{\mathrm{b}}]$ before adding $\mathbf{Q}$ back to $\mathbf{Q}_t$.

From the decoded features, upsampled patch tokens are combined with VGGT point-map, tracking and image features into pixel embeddings $\mathbf{U}_t(u) \in \mathbb{R}^d$ and coarse mask logits $\mathbf{c}_t(u)$. Projecting $\mathbf{Q}_t$ to $\tilde{\mathbf{Q}}_t \in \mathbb{R}^{(K+2)\times d}$ yields pixel mask probabilities:

$$\mathbf{P}_t(u) = \mathrm{softmax}\big(\mathbf{c}_t(u) + d^{-1/2}\tilde{\mathbf{Q}}_t\,\mathbf{U}_t(u)\big). \tag{2}$$

In parallel, a matching head maps tracking features, query-conditioned tokens and patch tokens $\mathbf{D}_t$ to a unit pixel descriptor $\mathbf{d}_t(u)$. With temperature $\tau > 0$, weighting descriptor similarity by $\mathbf{P}_t(u)\cdot\mathbf{P}_r(v)$, the probability that pixel u in frame t and pixel v in reference frame r belong to the same slot, restricts correspondences within each rigid part:

$$a_t^r(u,v) = \frac{\exp(\tau^{-1}\mathbf{d}_t(u)^\top\mathbf{d}_r(v))\,\mathbf{P}_t(u)\cdot\mathbf{P}_r(v)}{\sum_m \exp(\tau^{-1}\mathbf{d}_t(u)^\top\mathbf{d}_r(m))\,\mathbf{P}_t(u)\cdot\mathbf{P}_r(m)}, \tag{3}$$

which maps pixel u to its matched 3D point $\mathbf{C}_t^r(u) = \sum_v a_t^r(u,v)\mathbf{X}_r(v)$ in frame r. Using static-base correspondences $\mathbf{C}_t^1(u) \leftrightarrow \mathbf{X}_t(u)$, iteratively reweighted orthogonal Procrustes alignment first estimates a rigid transformation per frame to align $\{\mathbf{X}_t\}_{t=1}^{T}$ on the static base and cancel camera drift. On these aligned point maps, applying the same Procrustes solver to each part's correspondences between frames r and t recovers its pairwise rotation $\mathbf{R}_{kt}^r \in SO(3)$, rotation vector $\mathbf{z}_{kt}^{r,\mathrm{rev}} = \frac{1}{2}(\mathbf{R}_{kt}^r - \mathbf{R}_{kt}^{r\top})^\vee$ and centroid translation $\mathbf{z}_{kt}^{r,\mathrm{pri}} = \bar{\mathbf{x}}_{kt} - \bar{\mathbf{c}}_{kt}^r$ between matched centroids $\bar{\mathbf{c}}_{kt}^r$ and $\bar{\mathbf{x}}_{kt}$, along with total inlier weight ω_{kt}^r. Finally, the joint head attends across features along the boundary between the part and the base alongside pairwise motion summaries to predict joint-type logits, a boundary pivot prior $\hat{\mathbf{o}}_k \in \mathbb{R}^3$, and Cholesky factors $\mathbf{L}_k$ forming positive-semidefinite Bingham prior matrices $\mathbf{M}_k = \mathbf{L}_k\mathbf{L}_k^\top$ for $\mathbf{M}_k \in \{\mathbf{M}_k^{\mathrm{rev}}, \mathbf{M}_k^{\mathrm{pri}}\}$.

B. Kinematic Inference Layer

Given the pairwise motion vectors $\mathbf{z}_{kt}^r \in \{\mathbf{z}_{kt}^{r,\mathrm{rev}}, \mathbf{z}_{kt}^{r,\mathrm{pri}}\}$ and inlier weights ω_{kt}^r from the decoder, summing outer products over all frame pairs cancels single-frame reference noise and forms the rotational and translational motion scatter matrices:

$$\mathbf{S}_k = \sum_{r,t} \omega_{kt}^r\,\mathbf{z}_{kt}^r\mathbf{z}_{kt}^{r\top}. \tag{4}$$

Under noise-free 3D geometry, $\mathbf{S}_k$ has rank one along $\mathbf{a}_k$. On monocular point maps, cross-frame attention scales motion along the viewing ray by $\gamma \in (0,1)$ and tilts the leading eigenvector of $\mathbf{S}_k$ by up to $\tan\theta \le (1-\gamma)/(2\sqrt{\gamma})$ regardless of frame count, while small opening angles under an unknown

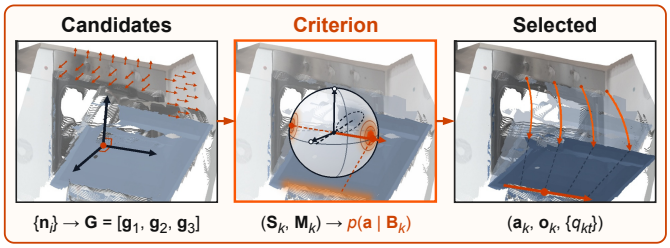


Fig. 3. **Mechanism of the Kinematic Inference Layer on real point maps.** Dense static-base normals $\{\mathbf{n}_i\}$ form an orthogonal Manhattan frame $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3]$ (*Candidates*). Combining pairwise motion scatter $\mathbf{S}_k$ with the learned Bingham prior $\mathbf{M}_k$ defines an antipodal Bingham posterior $p(\mathbf{a} \mid \mathbf{B}_k)$ (*Criterion*), selecting $\mathbf{a}_k$ to solve for the hinge pivot $\mathbf{o}_k$ and states $\{q_{kt}\}$ (*Selected*).

rest pose leave continuous axis estimation ill-conditioned, as derived in the supplementary material. Instead of estimating an unconstrained direction on $\mathbb{RP}^2$, selecting $\mathbf{a}_k$ from three orthogonal candidate axes in Figure 3 recovers the exact direction whenever this tilt stays below $45°$, which holds for any $\gamma > 3 - 2\sqrt{2} \approx 0.172$.

To obtain these orthogonal candidates without depth bias from moving parts, we compute finite-difference unit normals $\{\mathbf{n}_i\}_{i\in\mathcal{O}}$ on the point-map grid of the stationary object regions. In manufactured cabinets, appliances and furniture, planar faces meet at right angles [13, 34], so $\{\mathbf{n}_i\}$ cluster around an orthogonal Manhattan frame $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3] \in SO(3)$. Whereas point-cloud PCA rotates arbitrarily when two object dimensions are similar, maximising fourth-order normal alignment [35] locks directly onto the perpendicular face normals on the left of Figure 3:

$$\mathbf{G} = \arg\max_{\mathbf{G}\in SO(3)} \sum_{i\in\mathcal{O}} \sum_{j=1}^{3} \left(\mathbf{n}_i^\top \mathbf{g}_j\right)^4, \tag{5}$$

which we solve from second-moment eigenvectors by projected gradient ascent on $SO(3)$ via polar decomposition.

Since $\pm\mathbf{a}_k$ define the same physical joint line, we combine the motion scatter $\mathbf{S}_k$ and the learned prior $\mathbf{M}_k \in \{\mathbf{M}_k^{\text{rev}}, \mathbf{M}_k^{\text{pri}}\}$ under an antipodally symmetric Bingham distribution [14] $p(\mathbf{a} \mid \mathbf{B}_k) \propto \exp(\mathbf{a}^\top \mathbf{B}_k \mathbf{a})$ on $\mathbb{RP}^2$. Here $\alpha_k > 0$ normalises $\mathbf{S}_k$ by total inlier weight and point variance, multiplied by the part's squared radius of gyration for revolute scatter so both joint types are dimensionless, while $\kappa_k > 0$ weights $\mathbf{M}_k$ and increases when translational motion aligns with the viewing ray:

$$\mathbf{B}_k = \alpha_k \mathbf{S}_k + \kappa_k \mathbf{M}_k. \tag{6}$$

Although normalising $p(\mathbf{a} \mid \mathbf{B}_k)$ over $\mathbb{S}^2$ requires confluent hypergeometric functions, restricting $\mathbf{a}_k$ to the three columns of $\mathbf{G}$ with inverse temperature $\beta > 0$ reduces exact posterior inference to the closed-form three-way softmax in the middle of Figure 3:

$$w_{kj} = \frac{\exp\left(\beta\, \mathbf{g}_j^\top \mathbf{B}_k\, \mathbf{g}_j\right)}{\sum_{m=1}^{3} \exp\left(\beta\, \mathbf{g}_m^\top \mathbf{B}_k\, \mathbf{g}_m\right)}, \qquad j \in \{1, 2, 3\}. \tag{7}$$

Orienting each candidate by $s_{kj} = \operatorname{sign}(\mathbf{g}_j^\top \hat{\mathbf{a}}_k)$ using the leading eigenvector $\hat{\mathbf{a}}_k$ of $\mathbf{M}_k$, we select $j^* = \arg\max_j w_{kj}$ with straight-through gradients during training:

$$\mathbf{a}_k = s_{kj^*}\, \mathbf{g}_{j^*}. \tag{8}$$

Given $\mathbf{a}_k$, we infer the joint type y_k by combining the joint head's type logits with the reduction in Procrustes residual from pure translation to rigid rotation, penalising lever arms larger than the object radius. For a revolute joint with pairwise translation residual $\mathbf{t}_{kt}^r = \bar{\mathbf{x}}_{kt} - \mathbf{R}_{kt}^r \bar{\mathbf{c}}_{kt}^r$, normalised pair weight $\bar{\omega}_{kt}^r$ and projected rotation operator $\mathbf{A}_{kt}^r = (\mathbf{I} - \mathbf{R}_{kt}^r)(\mathbf{I} - \mathbf{a}_k\mathbf{a}_k^\top)$, we estimate the hinge pivot $\mathbf{o}_k = \arg\min_{\mathbf{o}} E_k(\mathbf{o})$ by minimising:

$$E_k(\mathbf{o}) = \sum_{r,t} \bar{\omega}_{kt}^r \left\| \mathbf{A}_{kt}^r \mathbf{o} - \mathbf{t}_{kt}^r \right\|^2 + \lambda_{\text{o}} \|\mathbf{o} - \hat{\mathbf{o}}_k\|^2. \tag{9}$$

Setting $\nabla_{\mathbf{o}} E_k(\mathbf{o}) = \mathbf{0}$ yields a 3×3 positive-definite linear system that solves for the rotation centre in the plane orthogonal to $\mathbf{a}_k$ from rigid motion while fixing $\mathbf{a}_k^\top \mathbf{o}_k = \mathbf{a}_k^\top \hat{\mathbf{o}}_k$ along the axis to the boundary prior $\hat{\mathbf{o}}_k$ on the right of Figure 3. Projecting frame-1-to-t rotation angles or centroid translations onto $\mathbf{a}_k$ recovers $\{q_{kt}\}_{t=1}^T$, and substituting $(\mathbf{a}_k, \mathbf{o}_k, q_{kt})$ into Eq. (1) transforms the 3D part points to unseen states in one pass.

C. Training Objective

Keeping the VGGT backbone frozen, we train PhyVGGT in two stages on synthetic clips, first optimising the segmentation and correspondence trunk and then training the joint priors and calibration parameters of the Kinematic Inference Layer. Predicted part slots are matched to ground-truth parts via Hungarian assignment on pixel mask overlap. After $\mathrm{Sim}(3)$ scene alignment, the first stage sums cross-entropy and Dice losses on $\mathbf{P}_t$, first-frame InfoNCE on $\mathbf{d}_t(u)$ and 3D canonical correspondence losses on $\mathbf{C}_t^1(u)$, while the second stage sums binary cross-entropy on joint-type logits, Bingham quadratic and antipodal cosine losses on joint axes, smooth-L_1 losses on per-frame screw motion vectors and motion span, and a point-to-line distance loss on $\mathbf{o}_k$. All training clips are rendered from PartNet-Mobility [36] with moving cameras, textured indoor rooms, articulated hands grasping handles and hinged doors converted into sliding doors, supervising 3D motion-based joint estimation without real-world videos.

IV. Experiments

A. Experimental Setup

Datasets. We evaluate zero-shot synthetic-to-real transfer across three real-world video benchmarks alongside in-domain and cross-category generalisation on two synthetic video benchmarks. On real-world interactions, Arti4D [19] captures egocentric handheld RGB-D videos of drawers, doors and appliances across 329 held-out test clips in three unseen scenes, D3D-HOI [37] provides 170 third-person static-camera test clips annotated with 3D joint axes, and RBO [38] contributes 357 third-person RGB-D videos across 14 objects with motion-capture ground truth. On synthetic PartNet-Mobility [36] videos, we train on 6,264 clips from ten categories and evaluate on held-out test splits of 290 joints from the ten trained categories

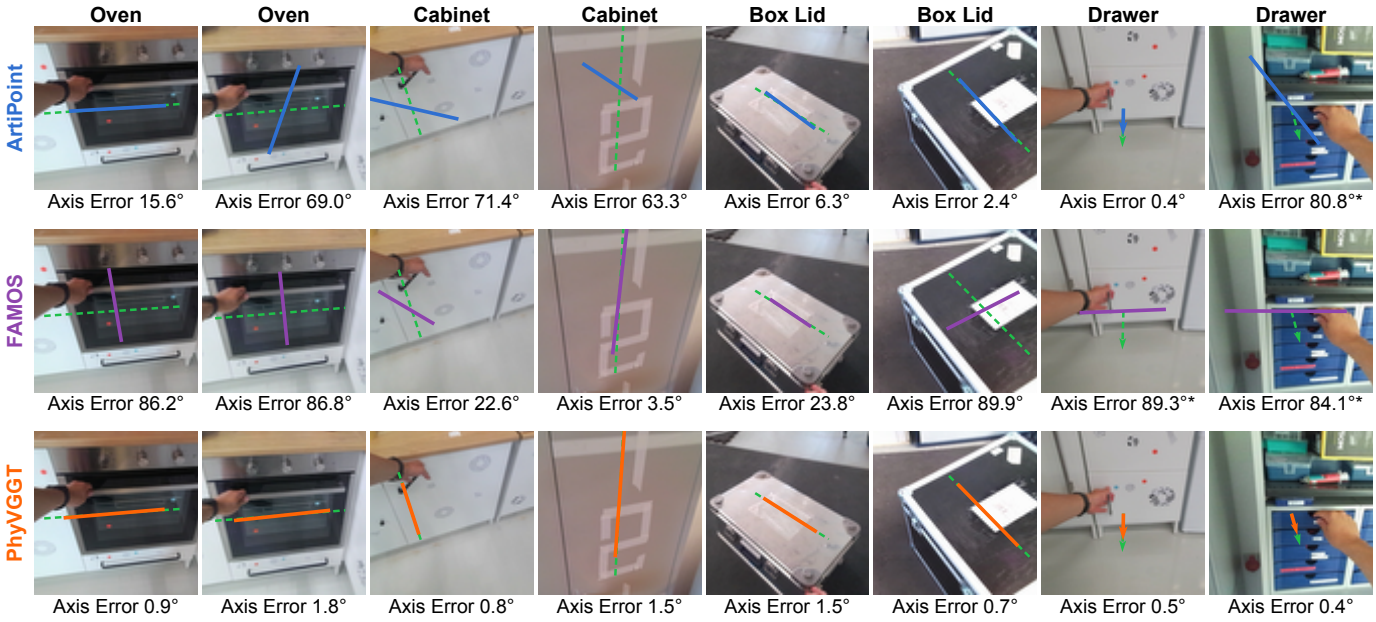


Fig. 4. **Qualitative comparison of 3D joint axis estimation on Arti4D test videos.** Blue shows ArtiPoint with RGB-D input and ground-truth poses, purple shows FAMOS on PhyVGGT point maps, orange shows RGB-only PhyVGGT, and green dashed lines denote ground truth. Arrows indicate prismatic joints while lines indicate revolute joints, with * marking wrong joint types.

(Syn-10) and 510 joints across 38 categories including 28 unseen categories (Syn-38), as well as 48-frame reconstruction across all 20 scenes and 48 joints of Articulat3D-Sim [8]. Component ablations and backbone transfer are evaluated on four disjoint development splits: Arti4D-dev on scene *din080* with 85 clips, D3D-HOI-dev with 86 clips, Syn-10-val with 313 joints and Syn-38-val with 534 joints.

Metrics. Following [7, 19], we measure axis error as the acute angle on $\mathbb{RP}^2$ and revolute pivot error as 3D line-to-line distance, reporting median errors and joint accuracy within $5°$ and $10°$ requiring both correct joint type and axis.

Baselines. Among uncalibrated monocular RGB methods, we evaluate three complementary baseline groups: same-backbone ablations on frozen VGGT via *Direct regression* with a continuous MLP joint head trained on the same features, masks and synthetic clips as PhyVGGT and *Motion estimate* taking the leading eigenvector of $\mathbf{S}_k$ without Bingham selection on the Manhattan frame, appearance and point-cloud regressors via single-image OPDMulti [22] testing appearance priors without multi-frame 3D geometry and feed-forward FAMOS [10] testing multi-state continuous point-cloud regression, and continuous trajectory fitting via monocular ArtiPoint [19] on π^3 [3] and VGGT [2]. Privileged baselines marked with * include posed RGB-D ArtiPoint [19] with sensor depth and ground-truth poses, and per-scene 3D Gaussian optimisation Articulat3D [8] with ground-truth depth, poses, masks and kinematic trees.

Implementation. At test time, PhyVGGT processes $T = 12$ uniform 504×504 RGB frames per clip, and $T = 48$ on Articulat3D-Sim, using $K = 8$ part queries, $N = 3$ decoder blocks, 256-D queries, 64-D pixel and correspondence embeddings and intermediate VGGT layers $\{4, 11, 17, 23\}$. Reweighted Procrustes with a Cauchy loss runs three iterations for static-base alignment and six for part motion, while Eq. (9) solves a 3×3 linear system with learned regulariser $\lambda_o \geq 0.02$. Manhattan frame optimisation runs 48 projected gradient steps on a stride-8 normal grid, pooling background normals below 500 foreground normals or 50% inliers and falling back to $\hat{\mathbf{a}}_k$ on untrusted frames deviating over $38°$. We train on frozen VGGT-1B with AdamW using 8-frame synthetic clips of batch size 4 for 4,500 first-stage steps at learning rate 2×10^{-4} and 2,000 second-stage steps at 3×10^{-4}. On real videos, person pixels are masked [39] prior to matching, and on clips where a single part is operated the active slot is chosen automatically by motion amplitude weighted by root mask area. All runtimes are measured on a dedicated NVIDIA A100 GPU.

TABLE I
COMPARISON ON REAL-WORLD ARTI4D AND D3D-HOI. JOINT ACCURACY AND MEDIAN AXIS ERROR IN DEGREES. *SENSOR DEPTH AND POSES.

	Arti4D (329)			D3D-HOI (170)		
Method	$\leq 5°$	$\leq 10°$	Med.	$\leq 5°$	$\leq 10°$	Med.
ArtiPoint, RGB-D*	0.422	0.614	6.30	–	–	–
Direct regression	0.006	0.020	57.07	0.022	0.063	37.15
OPDMulti	0.146	0.292	64.15	0.147	0.418	11.55
ArtiPoint, π^3	0.198	0.389	15.90	–	–	–
FAMOS	0.241	0.369	23.24	0.108	0.145	68.93
Motion estimate	0.090	0.280	21.92	0.186	0.369	15.93
PhyVGGT (Ours)	**0.493**	**0.604**	**4.79**	**0.629**	**0.724**	**2.77**

B. Benchmark Comparisons

Table I evaluates zero-shot transfer from synthetic training to handheld egocentric videos on Arti4D and third-person videos on D3D-HOI. Benefiting from the undistorted geometry of

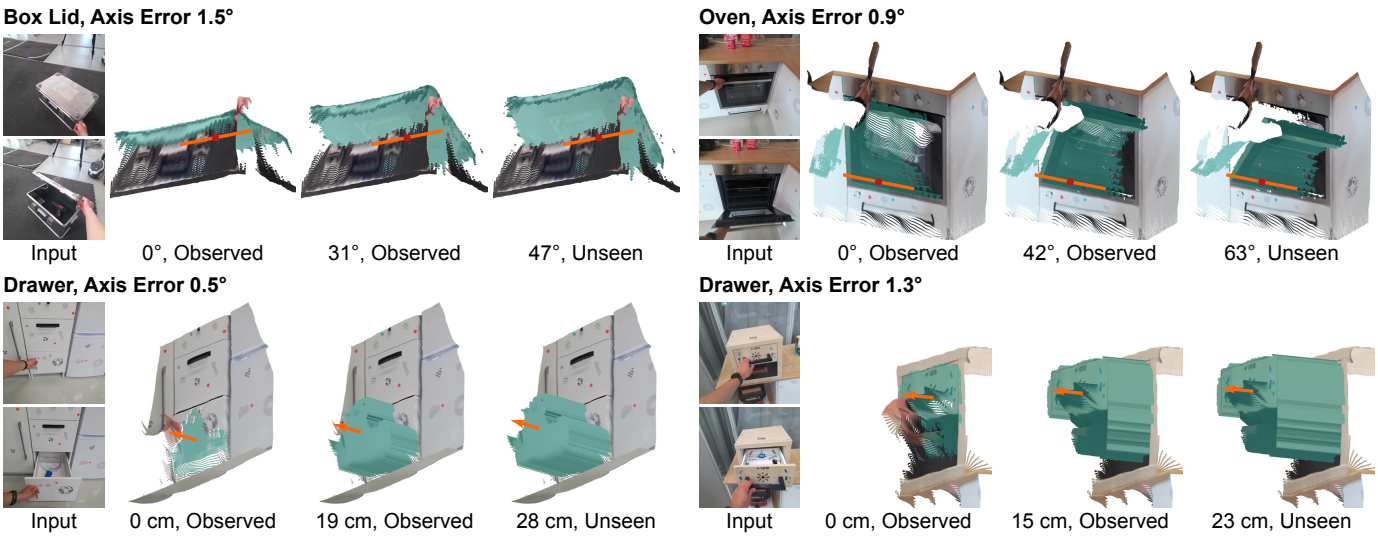


Fig. 5. **Qualitative 3D part reconstructions and unseen articulation states on Arti4D test videos.** For each scene, we show two input RGB frames alongside PhyVGGT's 3D point map near the rest state, in which the segmented **movable part (teal)** is driven by the predicted **joint axis (orange)** and **hinge pivot (red)** to the start and end of the observed motion and then to an unseen articulation state via forward kinematics. Drawers show the volume swept by the part along the predicted axis, with axis angular errors listed above.

stationary object faces, we redesigned axis estimation in our Kinematic Inference Layer around the object's Manhattan frame instead of camera coordinates, avoiding overfitting to synthetic camera viewpoints under steep handheld pitches and lifting 10° accuracy to 60.4% and 72.4% against 2.0% and 6.3% for Direct regression and 29.2% and 41.8% for single-image OPDMulti. By introducing closed-form Bingham selection that fuses pairwise motion scatter with learned Bingham priors, the Kinematic Inference Layer also removes depth underestimation along the viewing ray and ambiguity under small opening angles, outperforming multi-state FAMOS at 60.4% and 72.4% against 36.9% and 14.5% and raising prismatic drawer accuracy to 48.2% and 58.3% against 14.7% and 11.1% for unconstrained Motion estimate. On Arti4D, referencing part queries to the static-base query and aligning static-base correspondences in our Part and Motion Decoder cancels handheld camera drift, improving accuracy to 60.4% against 38.9% for monocular ArtiPoint on π^3. Even without the sensor depth and ground-truth camera poses used by RGB-D ArtiPoint, fitting the Manhattan frame to flat static cabinet normals avoids depth sensor noise on small handles, raising 5° accuracy to 49.3% against 42.2% and reducing median error to 4.79° against 6.30° in Figures 4 and 5.

On tabletop interactions in Table II, hands frequently occlude moving parts while perturbing the object pose. By separating static-base masks from moving parts and gating correspondences within each slot, our Part and Motion Decoder isolates unoccluded base regions for reliable fourth-order normal alignment, enabling PhyVGGT to reach 24.1% overall on RBO and 31.1% on stationary-base clips against 18.5% for Motion estimate, 11.9% for FAMOS and 9.8% for monocular ArtiPoint. Even compared with privileged RGB-D ArtiPoint using sensor depth and motion-capture poses, anchoring revolute inference to planar base faces and boundary priors in our Kinematic Inference Layer raises revolute accuracy to 27.5% against 25.3% and halves median axis error to 23.3° against 51.3°.

TABLE II
COMPARISON ON REAL-WORLD RBO. JOINT ACCURACY, MEDIAN AXIS ERROR IN DEGREES, AND REVOLUTE/PRISMATIC ACCURACY AT 10°. *SENSOR DEPTH AND POSES.

Method	$\leq 5^\circ$	$\leq 10^\circ$	Med.	Rev.	Pris.
ArtiPoint, RGB-D*	0.148	0.246	51.3	0.253	0.229
Direct regression	0.015	0.046	48.7	0.041	0.059
OPDMulti	0.050	0.129	44.7	0.126	0.135
ArtiPoint, π^3	0.034	0.098	64.2	0.119	0.042
FAMOS	0.045	0.119	68.9	0.153	0.024
Motion estimate	0.063	0.185	29.2	0.206	0.128
PhyVGGT (Ours)	**0.134**	**0.241**	**23.3**	**0.275**	**0.149**

TABLE III
COMPARISON ON SYNTHETIC PARTNET-MOBILITY. JOINT ACCURACY AND MEDIAN AXIS ERROR ON 10 TRAINED CATEGORIES IN SYN-10 AND 38 CATEGORIES WITH 28 UNSEEN IN SYN-38. ‡PRETRAINED ON OPENABLE FURNITURE.

Benchmark	Method	$\leq 5^\circ$	$\leq 10^\circ$	Med.
Syn-10 (290)	Direct regression	0.061	0.198	20.60
	OPDMulti‡	0.028	0.072	75.03
	FAMOS	0.426	0.529	7.08
	Motion estimate	0.366	0.543	7.96
	PhyVGGT (Ours)	**0.762**	**0.778**	**1.53**
Syn-38 (510)	Direct regression	0.042	0.128	33.40
	OPDMulti‡	–	–	–
	FAMOS	0.314	0.398	32.82
	Motion estimate	0.259	0.410	15.76
	PhyVGGT (Ours)	**0.610**	**0.631**	**2.27**

On synthetic PartNet-Mobility in Table III, combining pairwise motion scatter with joint-type and Bingham priors in our Kinematic Inference Layer separates sliding doors from hinged doors by 3D motion instead of static appearance while

TABLE IV
COMPARISON ON 48-FRAME ARTICULAT3D-SIM. *PER-SCENE OPTIMISATION WITH GT DEPTH, POSES, MASKS AND KINEMATIC TREES.

Method	Trained Med.	All joints (48) $\leq 5°$	All joints (48) $\leq 10°$	Unseen $\leq 10°$	Time
Articulat3D*	2.63	0.479	0.521	0.316	2.6–11.9 h
Direct regression	21.67	0.035	0.139	0.000	6.8 s
Motion estimate	11.79	0.215	0.347	0.158	6.8 s
PhyVGGT (Ours)	**2.20**	**0.431**	**0.486**	**0.246**	**6.8 s**

TABLE V
3D PART, SURFACE AND HINGE PIVOT RECONSTRUCTION. UPPER ROWS REPORT 3D MOVABLE / PART IOU AND SURFACE ERROR, AND LOWER ROWS REPORT HELD-OUT HINGE MEDIAN ERROR IN CM AND ACCURACY WITHIN 10 CM.

3D Part & Surface Geometry	Syn-10-val	Syn-38-val	Pt. err.
Final layer only ($l = 23$)	0.877 / 0.663	0.590 / 0.542	0.032
w/o Tracking features	0.888 / 0.694	0.608 / 0.564	0.032
PhyVGGT (Ours)	**0.891 / 0.701**	**0.617 / 0.576**	**0.032**
3D Hinge Pivot (Med. cm / $\leq$10 cm)	Syn-10 (290)	Syn-38 (510)	RBO (357)
Direct regression	6.49 / 0.64	5.70 / 0.68	9.34 / 0.53
FAMOS [10]	4.35 / 0.69	4.96 / 0.67	11.14 / 0.46
PhyVGGT (Ours)	**3.16 / 0.83**	**3.36 / 0.82**	**8.76 / 0.55**

TABLE VI
ADDITIVE ABLATION STUDY ON THE FOUR DEVELOPMENT SPLITS. JOINT ACCURACY AT $10°$ AS EACH MODULE AND CUE IS INTRODUCED.

Configuration	Arti4D-dev	D3D-dev	Syn-10-val	Syn-38-val
Direct regression	0.063	0.058	0.188	0.125
+ Part and Motion Decoder	0.259	0.198	0.390	0.303
+ Motion scatter $\mathbf{S}_k$ only	0.482	0.628	0.497	0.386
+ Bingham prior $\mathbf{M}_k$ on $\mathbf{G}$ only	0.522	0.620	0.778	0.601
+ **Full Kinematic Inference Layer (Ours)**	**0.725**	**0.744**	**0.803**	**0.636**

TABLE VII
CONTROLLED VGGT-1B COMPARISON AND ZERO-SHOT BACKBONE TRANSFER. UPPER ROWS SHARE VGGT-1B POINT MAPS AND MASKS AT $10°$, AND LOWER ROWS (†) SWAP UNSEEN BACKBONES ZERO-SHOT.

Method	Backbone	Arti4D-dev	D3D-dev	Syn-10-val
Direct regression	VGGT-1B	0.063	0.058	0.188
ArtiPoint [19]	VGGT-1B	0.271	–	–
FAMOS [10]	VGGT-1B	0.376	0.116	0.502
Motion estimate	VGGT-1B	0.482	0.628	0.497
PhyVGGT (Ours)	VGGT-1B	**0.725**	**0.744**	**0.803**
PhyVGGT (Ours)	StreamVGGT† [29]	0.753	0.783	0.743
PhyVGGT (Ours)	STream3R† [30]	0.769	0.748	0.801
PhyVGGT (Ours)	$\pi^{3\dagger}$ [3]	0.765	0.775	0.802

removing tilt along the viewing ray, achieving 77.8% accuracy and $1.53°$ median error on Syn-10 against 54.3% for Motion estimate, 52.9% for FAMOS and 7.2% for appearance-based OPDMulti. Since 92.2% of ground-truth axes across all 38 categories in Syn-38 align within $10°$ of the static Manhattan frame $\mathbf{G}$, anchoring selection to $\mathbf{G}$ transfers directly to 28 unseen object categories at 63.1% accuracy and $2.27°$ median error against 41.0% for Motion estimate and 39.8% for FAMOS, reaching 83.3% even on tilted Laptop lids whose hinges align with the orthogonal keyboard rear edge. On 48-frame Articulat3D-Sim in Table IV, accumulating all-pairs motion scatter across longer sequences raises single-pass accuracy to 48.6% against 34.7% for Motion estimate and 13.9% for Direct regression. Although privileged Articulat3D [8] optimises 3D Gaussians per scene for 2.6–11.9 h using ground-truth depth, poses, masks and kinematic trees, PhyVGGT operating from uncalibrated RGB alone in 6.8 s achieves a lower median error of $2.20°$ against $2.63°$ on seen categories at over $1{,}300\times$ the speed. In Table V, introducing multi-layer per-frame token fusion and tracking features in our Part and Motion Decoder preserves sharp intra-frame boundaries while strengthening cross-frame matching, improving 3D movable and part IoU to 0.891/0.701 on Syn-10-val against 0.877/0.663 from the final layer alone and cutting 3D match error to 0.91 cm against 1.28 cm without tracking features. Building on the selected Manhattan axis, our positive-definite linear solver in Eq. (9) prevents axis tilt from shifting the lever arm, lowering held-out hinge pivot error to 3.16 cm, 3.36 cm and 8.76 cm on Syn-10, Syn-38 and RBO against 4.35 cm, 4.96 cm and 11.14 cm for FAMOS.

Qualitatively on real videos, Figure 5 illustrates four Arti4D test scenes with axis errors between $0.5°$ and $1.5°$. The recovered joints rotate the lid and oven door about their hinges, slide both drawers outwards and extrapolate cleanly to 1.5 times the observed motion range.

C. Ablation Study

Table VI examines how each design in PhyVGGT builds performance across the four development benchmarks. Introducing our Part and Motion Decoder over Direct regression fuses intermediate per-frame tokens with global tokens and references part queries to the static base, recovering sharp part masks and aligned correspondences to lift accuracy to 25.9% on Arti4D-dev and 39.0% on Syn-10-val against 6.3% and 18.8%. Within this decoder, fusing intermediate per-frame layers with the final block preserves intra-frame part edges, raising accuracy to 80.3% against 76.1% on Syn-10-val and 72.5% against 69.8% on Arti4D-dev over reading layer $l = 23$ alone alongside the 3D IoU gain in Table V. On top of the aligned correspondences, using pairwise motion scatter $\mathbf{S}_k$ alone captures active displacement to reach 48.2% and 62.8% on Arti4D-dev and D3D-HOI-dev yet remains tilted towards the viewing ray, whereas using the Bingham prior $\mathbf{M}_k$ on the Manhattan frame $\mathbf{G}$ alone locks onto undistorted orthogonal normals to reach 77.8% on Syn-10-val yet lacks motion cues to tell which face normal moves on casual videos. Unifying both cues in our full Kinematic Inference Layer combines undistorted Manhattan candidates with observed motion evidence, raising accuracy across all four splits to 72.5%, 74.4%, 80.3% and 63.6% and cutting real-world median error to $3.6°$ and $2.9°$ against $31.4°$ and $30.7°$ from the learned prior alone.

Inside the Kinematic Inference Layer, formulating Eq. (8) as closed-form discrete selection over $\mathbf{G}$ prevents depth noise along the viewing ray from blending orthogonal Bingham modes, improving accuracy to 72.5%, 74.4% and 80.3% against 70.6%, 46.5% and 55.6% for continuous posterior expectation on Arti4D-dev, D3D-HOI-dev and Syn-10-val. Extracting $\mathbf{G}$ via

fourth-order normal alignment in Eq. (5) is equally important, locking onto perpendicular object faces even under oblique handheld views and symmetric cabinet dimensions where camera-axis candidates drop to 4.3% and 43.0% and PCA candidates drop to 6.3% and 8.5%. Consequently, on shared VGGT-1B point maps and masks in Table VII, our Kinematic Inference Layer outperforms ArtiPoint at 72.5% against 27.1%, FAMOS at 72.5%, 74.4% and 80.3% against 37.6%, 11.6% and 50.2%, and Motion estimate at 48.2%, 62.8% and 49.7%.

Measuring frozen point maps directly without any trained head confirms that sliding motion compresses to $\gamma \in [0.60, 0.69]$ along the viewing ray across all four backbones, including 0.69 on VGGT against 0.99 across the ray, tilting $\mathbf{S}_k$ by at most 14.5° within our 45° bound. Under small opening angles where continuous Fisher curvature weakens, learned Bingham priors further disambiguate the three Manhattan axes, lifting accuracy to 75.0% on 10°–20° synthetic rotations and 58.6% on small-motion D3D-HOI-dev clips against 41.7% and 21.8% for Motion estimate in the supplementary material. Since both modules rely only on static surface normals and relative token correspondences instead of camera calibration, plugging PhyVGGT directly into unseen frozen backbones in Table VII transfers zero-shot in 1.2 s per clip, even raising real-world accuracy to 75.3%–76.9% on Arti4D-dev and 74.8%–78.3% on D3D-HOI-dev across StreamVGGT [29], STream3R [30] and π^3 [3].

V. Conclusion

We proposed PhyVGGT, a single-pass framework on a frozen 3D model that reconstructs 3D movable parts, joints and unseen states from an uncalibrated video by anchoring part matching and closed-form joint estimation to the stationary base. Our Part and Motion Decoder fuses multi-layer tokens relative to a static-base query for accurate 3D correspondences, while our Kinematic Inference Layer combines Bingham axis selection on the Manhattan frame with linear pivot estimation to remove depth bias along the viewing ray. Trained solely on synthetic data, PhyVGGT achieves state-of-the-art monocular performance across five benchmarks, surpasses posed RGB-D ArtiPoint in median error, runs in 6.8 s per 48-frame video against hours of per-scene optimisation in Articulat3D, and transfers zero-shot across four 3D backbones. Although PhyVGGT already recovers articulated geometry and motion from a casual video, passive RGB frames do not reveal stationary or occluded parts or physical parameters such as mass, friction, springs and latches, which we will recover through interactive manipulation and force sensing in future work.

References

[1] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUSt3R: Geometric 3D Vision Made Easy," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[2] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotný, "VGGT: Visual Geometry Grounded Transformer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 5294–5306.

[3] Y. Wang, J. Zhou, H. Zhu, W. Chang, Y. Zhou, Z. Li, J. Chen, J. Pang, C. Shen, and T. He, "π^3: Permutation-Equivariant Visual Geometry Learning," in *International Conference on Learning Representations*, 2026.

[4] M. Kim, J. Choe, J. Lee, Y.-C. F. Wang, and S. J. Kim, "RelationVGGT: Visual Geometry Transformers for 3D Spatial Relation Segmentation," in *Advances in Neural Information Processing Systems*, 2026.

[5] J. Liu, A. Mahdavi-Amiri, and M. Savva, "PARIS: Part-level Reconstruction and Motion Analysis for Articulated Objects," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 352–363.

[6] Y. Liu, B. Jia, R. Lu, J. Ni, S.-C. Zhu, and S. Huang, "ArtGS: Building Interactable Replicas of Complex Articulated Objects via Gaussian Splatting," in *International Conference on Learning Representations*, 2025.

[7] Y. Liu, B. Jia, R. Lu, C. Gan, H. Chen, J. Ni, S.-C. Zhu, and S. Huang, "VideoArtGS: Building Digital Twins of Articulated Objects from Monocular Video," *arXiv preprint arXiv:2509.17647*, 2025.

[8] L. Guo, H. Zhao, X. Zhao, R. Fu, L. Zhuang, S. Huang, Z. Li, and H. Zou, "Articulat3D: Reconstructing Articulated Digital Twins From Monocular Videos with Geometric and Motion Constraints," in *European Conference on Computer Vision*, 2026, pp. 58–76.

[9] I. Lee, Y. Kim, E. Sohn, J. Lee, J. You, J. Lee, and J.-H. Kim, "ArtSplat: Feed-Forward Articulated 3D Gaussian Splatting from Sparse Multi-State Uncalibrated Views," *arXiv preprint arXiv:2605.24304*, 2026.

[10] K. Qu, T. Sun, M. Viola, L. Zhu, Z. Zhou, S. D. Sarkar, K. Schindler, and I. Armeni, "FAMOS: Feed-Forward 3D Articulation Modeling from Sparse Observations," *arXiv preprint arXiv:2609.20817*, 2026.

[11] A. Artykov, T. Ravaud, C. Sautier, and V. Lepetit, "sim2art: Accurate Articulated Object Modeling from a Single Video using Synthetic Training Data Only," *arXiv preprint arXiv:2512.07698*, 2025.

[12] H. J. Woltring, R. Huiskes, A. de Lange, and F. E. Veldpaus, "Finite Centroid and Helical Axis Estimation from Noisy Landmark Measurements in the Study of Human Joint Kinematics," *Journal of Biomechanics*, vol. 18, no. 5, pp. 379–389, 1985.

[13] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher, III, "A Mixture of Manhattan Frames: Beyond the Manhattan World," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3770–3777.

[14] C. Bingham, "An Antipodally Symmetric Distribution on the Sphere," *The Annals of Statistics*, vol. 2, no. 6, pp. 1201–1225, 1974.

[15] X. Li, H. Wang, L. Yi, L. Guibas, A. L. Abbott, and S. Song, "Category-Level Articulated Object Pose Estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3703–3712.

[16] A. Jain, R. Lioutikov, C. Chuck, and S. Niekum, "ScrewNet: Category-Independent Articulation Model Estimation From Depth Images Using Screw Theory," in *IEEE*

International Conference on Robotics and Automation, 2021, pp. 13 670–13 677.

[17] A. Artykov, T. Ravaud, N. Violante-Grezzi, and V. Lepetit, "Articulation in Prime: Primitive-Based Articulated Object Understanding from a Single Casual Video," *arXiv preprint arXiv:2605.18645*, 2026.

[18] W. Peng, J. Lv, C. Lu, and M. Savva, "iTACO: Interactable Digital Twins of Articulated Objects from Casually Captured RGBD Videos," in *International Conference on 3D Vision*, 2026, pp. 520–531.

[19] A. Werby, M. Büchner, A. Röfer, C. Huang, W. Burgard, and A. Valada, "Articulated Object Estimation in the Wild," in *Conference on Robot Learning*, 2025, pp. 3828–3849.

[20] M. Büchner, A. Röfer, T. Engelbracht, T. Welschehold, Z. Bauer, H. Blum, M. Pollefeys, and A. Valada, "Articulated 3D Scene Graphs for Open-World Mobile Manipulation," in *Conference on Robot Learning*, 2026.

[21] Y. Wang, Y. Miao, W. Zhao, W. Yang, Z. Wang, J. Pajarinen, L. Van Gool, D. P. Paudel, J. Kannala, X. Wang, and A. Solin, "PAWS: Perception of Articulation in the Wild at Scale from Egocentric Videos," *arXiv preprint arXiv:2603.25539*, 2026.

[22] X. Sun, H. Jiang, M. Savva, and A. X. Chang, "OPD-Multi: Openable Part Detection for Multiple Objects," in *International Conference on 3D Vision*, 2024, pp. 169–178.

[23] H. Li, H. Xie, J. Xu, B. Wen, F. Hong, and Z. Liu, "MonoArt: Progressive Structural Reasoning for Monocular Articulated 3D Reconstruction," in *European Conference on Computer Vision*, 2026, pp. 79–96.

[24] Z. Wu, Y. Xin, C. Hou, M. Chen, Y. Lyu, J. Zhang, and S. Zhang, "URDF-Anything+: End-to-End Generation for Simulation-Ready Articulated Assets," *arXiv preprint arXiv:2603.14010*, 2026.

[25] L. Le, J. Xie, W. Liang, H.-J. Wang, Y. Yang, Y. J. Ma, K. Vedder, A. Krishna, D. Jayaraman, and E. Eaton, "Articulate-Anything: Automatic Modeling of Articulated Objects via a Vision-Language Foundation Model," in *International Conference on Learning Representations*, 2025.

[26] V. Leroy, Y. Cabon, and J. Revaud, "Grounding Image Matching in 3D with MASt3R," in *European Conference on Computer Vision*, 2024, pp. 71–91.

[27] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.-H. Yang, "MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion," in *International Conference on Learning Representations*, 2025.

[28] H. Feng, J. Zhang, Q. Wang, Y. Ye, P. Yu, M. J. Black, T. Darrell, and A. Kanazawa, "St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World," in *IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8503–8513.

[29] D. Zhuo, W. Zheng, J. Guo, Y. Wu, J. Zhou, and J. Lu, "Streaming Visual Geometry Transformer," in *International Conference on Learning Representations*, 2026.

[30] Y. Lan, Y. Luo, F. Hong, S. Zhou, H. Chen, Z. Lyu, S. Yang, B. Dai, C. C. Loy, and X. Pan, "STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer," in *International Conference on Learning Representations*, 2026.

[31] Z. Cao, Q. Shao, S. Zhai, F. Tian, A. Nguyen, H. Wang, and B. Huang, "CoMo3R-SLAM: Collaborative monocular dense SLAM with learned 3D reconstruction priors for outdoor multi-agent systems," *arXiv preprint arXiv:2605.30488*, 2026.

[32] Z. Cao, Q. Shao, S. Zhai, J. Zhang, A. Nguyen, H. Wang, and B. Huang, "MAGS-SLAM: Monocular multi-agent Gaussian splatting SLAM for geometrically and photometrically consistent reconstruction," *IEEE Transactions on Automation Science and Engineering*, 2026.

[33] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. G. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," *Transactions on Machine Learning Research*, 2024.

[34] Z. Mandi, Y. Weng, D. Bauer, and S. Song, "Real2Code: Reconstruct Articulated Objects via Code Generation," in *International Conference on Learning Representations*, 2025.

[35] Y. Zhai, Z. Yang, Z. Liao, J. Wright, and Y. Ma, "Complete Dictionary Learning via ℓ^4-Norm Maximization over the Orthogonal Group," *Journal of Machine Learning Research*, vol. 21, no. 165, pp. 1–68, 2020.

[36] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. Guibas, and H. Su, "SAPIEN: A SimulAted Part-Based Interactive ENvironment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 094–11 104.

[37] X. Xu, H. Joo, G. Mori, and M. Savva, "D3D-HOI: Dynamic 3D Human-Object Interactions from Videos," *arXiv preprint arXiv:2108.08420*, 2021.

[38] R. Martín-Martín, C. Eppner, and O. Brock, "The RBO Dataset of Articulated Objects and Interactions," *The International Journal of Robotics Research*, vol. 38, no. 9, pp. 1013–1019, 2019.

[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

Supplementary Material

S1. Mathematical Derivations and Proofs

A. Fisher Information of Continuous Joint Estimation

Proposition. Consider a movable rigid part observed across T frames whose measured per-frame rotation and centroid carry isotropic Gaussian errors of variance σ_ω^2 and σ_e^2, while both the joint states $\{q_t\}$ and the canonical rest pose $(\mathbf{R}_0, \mathbf{c}_0)$ remain unknown. In the tangent space at the ground-truth joint, the axis tilt and the hinge pivot decouple and satisfy the Fisher information matrices:

$$\mathbf{I}_\mathbf{a} = \frac{T(1-\bar{R}^2)}{\sigma_\omega^2}\mathbf{I} + \frac{r^2}{\sigma_e^2}\mathbf{K}, \quad \mathbf{I}_\mathbf{o} = \frac{\mathbf{K}}{\sigma_e^2} + \frac{\mathbf{J}\mathbf{K}\mathbf{J}^\top}{\sigma_e^2 + r^2\sigma_\omega^2}, \tag{S1}$$

where $\bar{R} = \left\| \frac{1}{T}\sum_{t=1}^{T}(\cos q_t, \sin q_t) \right\|$ is the circular mean resultant length of the states, $\mathbf{K}$ is the 2×2 scatter of $(\cos q_t, \sin q_t)$ about its mean, r is the perpendicular distance from the part centroid to the hinge axis and $\mathbf{J}$ is a 90° rotation in the plane orthogonal to $\mathbf{a}$. Both $\mathbf{I}_\mathbf{a}$ and $\mathbf{I}_\mathbf{o}$ vanish quadratically as the observed motion range $Q \to 0$.

Proof. Align $\mathbf{a} = \hat{\mathbf{z}}$ and identify the plane orthogonal to $\mathbf{a}$ with $\mathbb{C}$ so that $\mathbf{a} \times \mathbf{v}$ maps to iv, placing the centroid at $\rho = r \in \mathbb{R}$ at the mean state. In frame t, $\mathbf{R}_t = \exp(q_t[\mathbf{a}]_\times)\mathbf{R}_0$ and $\mathbf{c}_t = \mathbf{o} + \exp(q_t[\mathbf{a}]_\times)(\mathbf{c}_0 - \mathbf{o})$ with $u_t = e^{iq_t}$. Perturbing the axis by a tangent vector $\boldsymbol{\delta} \perp \mathbf{a}$ written as $d \in \mathbb{C}$, the rest rotation by $\boldsymbol{\phi}$, the rest centroid by $w + w_z\mathbf{a}$, the pivot by $n \in \mathbb{C}$ and the states by ε_t yields four independent residual blocks: normal rotation $y_t = id + u_t(\phi - id)$, axial rotation $\psi_t = \varepsilon_t + \phi_z$, normal centroid $z_t = (1-u_t)n + u_t w + i\varepsilon_t u_t r$ and axial centroid $h_t = w_z + \mathrm{Re}(\bar{d}r) - \mathrm{Re}(u_t\bar{d}r)$. Since $\{d, \phi_\perp, w_z\}$ enter only (y_t, h_t) whereas $\{n, w, \varepsilon_t, \phi_z\}$ enter only (z_t, ψ_t), the joint Fisher matrix is block diagonal. Taking the Schur complement of $\sum_t |id + u_t g|^2/\sigma_\omega^2$ over $g = \phi - id$ yields $|d|^2 T(1-\bar{R}^2)/\sigma_\omega^2$, while profiling w_z out of $\sum_t h_t^2/\sigma_e^2$ yields $(r^2/\sigma_e^2)\mathbf{d}^\top\mathbf{K}\mathbf{d}$, establishing $\mathbf{I}_\mathbf{a}$. Rotating z_t by $\bar{u}_t$ preserves isotropic noise and gives $\bar{u}_t z_t = \bar{u}_t n + (w - n) + i\varepsilon_t r$. Eliminating ε_t against ψ_t inflates the imaginary-axis variance to $\sigma_e^2 + r^2\sigma_\omega^2$, and profiling out $w - n$ yields $\mathbf{I}_\mathbf{o}$. For a prismatic joint, profiling the rest position out of $\mathbf{y}_t^\perp = \mathbf{p}^\perp + q_t\boldsymbol{\delta} + \mathbf{e}_t^\perp$ yields $\sum_t (q_t - \bar{q})^2/\sigma_e^2$ per normal coordinate, which likewise vanishes quadratically with the slide range. □

B. Viewing-Ray Tilt Bound and Manhattan Selection Tolerance

Decompose a unit joint axis as $\mathbf{a} = \cos\varphi\,\mathbf{r} + \sin\varphi\,\mathbf{a}_\perp$ relative to the camera viewing ray $\mathbf{r}$. When inter-frame displacement along $\mathbf{r}$ is scaled by a factor $\gamma \in (0, 1)$, the observed motion vector aligns with $\gamma\cos\varphi\,\mathbf{r} + \sin\varphi\,\mathbf{a}_\perp$, tilting away from $\mathbf{a}$ by an angle θ satisfying:

$$\tan\theta = \frac{(1-\gamma)\tan\varphi}{\gamma + \tan^2\varphi} \le \frac{1-\gamma}{2\sqrt{\gamma}}, \tag{S2}$$

where equality holds at $\tan\varphi = \sqrt{\gamma}$. This systematic tilt is independent of frame count T and remains strictly below 45° whenever $\gamma > 3 - 2\sqrt{2} \approx 0.172$.

An unsigned axis on $\mathbb{RP}^2$ corresponds to a hemisphere of area 2π, on which a cap of angular radius ϵ_a covers area $\approx \pi\epsilon_a^2$, requiring $B_{\mathrm{loc}} = \log_2(2/\epsilon_a^2)$ bits to distinguish continuous hypotheses versus $B_{\mathrm{sel}} = \log_2|\mathcal{G}|$ bits over a finite candidate set $\mathcal{G}$. Under an antipodally symmetric Bingham density $p(\mathbf{a}) = \frac{1}{{}_1F_1(1/2,3/2,\mathbf{B}_k)}\exp(\mathbf{a}^\top\mathbf{B}_k\mathbf{a})$ on $\mathbb{RP}^2$ [14], restricting the support to uniform point masses on the Manhattan frame $\mathcal{G} = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$ replaces the confluent hypergeometric normaliser ${}_1F_1$ with the three-term sum $\sum_{m=1}^{3}\exp(\beta\,\mathbf{g}_m^\top\mathbf{B}_k\mathbf{g}_m)$, yielding Eq. (7) of the main paper in closed form. Whenever a rank-one motion scatter $\mathbf{S}_k \propto \tilde{\mathbf{a}}\tilde{\mathbf{a}}^\top$ tilts by $\theta < 45^\circ$ from the true Manhattan axis $\mathbf{g}_{j^*}$, its quadratic projection satisfies $\cos^2\theta > 1/2 > \sin^2\theta \ge (\mathbf{g}_m^\top\tilde{\mathbf{a}})^2$ for $m \ne j^*$, guaranteeing MAP selection of $\mathbf{g}_{j^*}$ without zero-mean or independent-frame noise assumptions.

C. All-Pairs Scatter Identity and Camera-Drift Cancellation

Let part k have correspondence weight n_t and centroid $\mathbf{d}_t$ in frame t, with total weight $n = \sum_t n_t$, weighted mean $\bar{\mathbf{d}}$ and centred covariance $\mathbf{C} = \sum_t n_t(\mathbf{d}_t - \bar{\mathbf{d}})(\mathbf{d}_t - \bar{\mathbf{d}})^\top$. Expanding the reference-frame scatter $\mathbf{S}^r = \sum_t n_t(\mathbf{d}_t - \mathbf{d}_r)(\mathbf{d}_t - \mathbf{d}_r)^\top$ yields $\mathbf{S}^r = \mathbf{C} + n(\mathbf{d}_r - \bar{\mathbf{d}})(\mathbf{d}_r - \bar{\mathbf{d}})^\top$, showing that a single reference frame injects a rank-one bias along its own noise realisation with full sequence weight n. Summing $n_r\mathbf{S}^r$ across all reference frames cancels this single-frame bias to give $2n\mathbf{C}$, whose expectation under isotropic frame noise $\sigma^2\mathbf{I}$ preserves $\mathbf{a}$ as its exact leading eigenvector.

Suppose observed point maps obey $\tilde{\mathbf{X}}_t(u) = \mathbf{X}_t(u) + \mathbf{F}_t(\mathbf{X}_t(u)) + \boldsymbol{\xi}_{kt}$, where $\mathbf{F}_t$ is an L-Lipschitz camera-induced deformation shared across the scene and $\|\boldsymbol{\xi}_{kt}\| \le \epsilon$ is a part-specific depth error. Without cross-frame structural constraints, continuous fitting cannot separate prismatic axes within $2\arcsin(2\epsilon'/Q)$ or parallel hinges within $\epsilon'/\sin(Q/4)$ for $\epsilon' = \epsilon/(1+L)$. Since the static base experiences $\mathbf{F}_t$ without articulation, aligning static-base correspondences cancels the shared camera drift $\mathbf{F}_t$, leaving only the residual $\boldsymbol{\xi}_{kt}$ within the 45° tolerance of our Kinematic Inference Layer.

S2. Architecture, Training and Evaluation Protocol

PhyVGGT builds directly on official frozen VGGT-1B [2] weights, including its built-in DINOv2-L [33] patch stem. From VGGT's 24-block alternating-attention aggregator, the Part and Motion Decoder reads intermediate tokens at layers $\{4, 11, 17, 23\}$, applies separate layer normalisation and two-layer MLP projections to the concatenated per-frame self-attention pyramid and global cross-frame pyramid, concatenates them with layer-normalised patch tokens $\mathbf{D}_t$, and refines them through alternating transformer blocks initialised from VGGT, using three frame-global pairs for part masks and two for correspondences. Projecting the refined per-frame stream plus the per-frame pyramid yields $\mathbf{U}_t^{\mathrm{f}} \in \mathbb{R}^{P\times 256}$, whereas projecting the refined global stream concatenated with the global pyramid and pooled features from VGGT's tracking head produces $\mathbf{U}_t^{\mathrm{g}} \in \mathbb{R}^{P\times 256}$. The query bank holds $K + 2 = 10$ learnable 256-D queries $[\mathbf{q}_{\mathrm{bg}}, \mathbf{q}_{\mathrm{b}}, \mathbf{q}_1, \ldots, \mathbf{q}_8]$, updated across $N = 3$

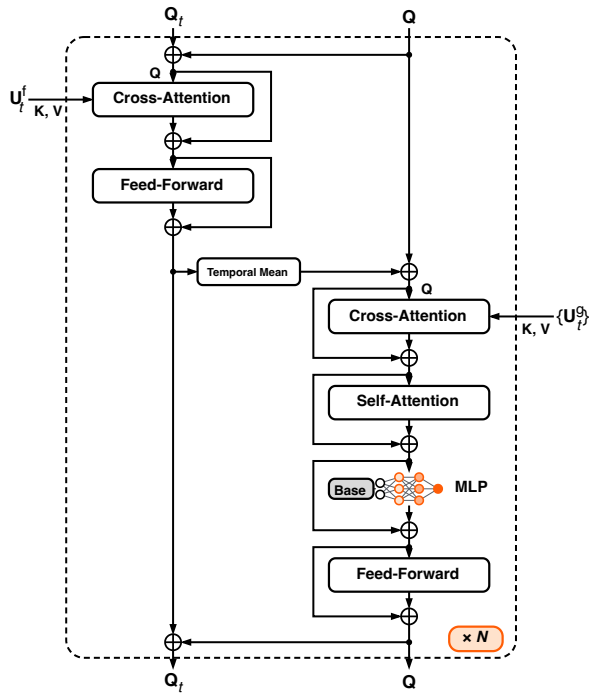


Fig. S1. **One block of the Part and Motion Decoder.** Learnable per-frame tokens $\mathbf{Q}_t$ and clip tokens $\mathbf{Q}$ serve as attention queries, while fused features $\mathbf{U}_t^{\mathrm{f}}$ and $\{\mathbf{U}_t^{\mathrm{g}}\}$ provide keys and values. Every sub-layer adds its output back residually via $\oplus$, and the MLP updates only part queries from $[\mathbf{q}_k, \mathbf{q}_{\mathrm{b}}, \mathbf{q}_k - \mathbf{q}_{\mathrm{b}}]$.

eight-head decoder blocks followed by patch upsampling and query-pixel dot products in Eq. (2) to output 504×504 mask probabilities $\mathbf{P}_t$ and 64-D unit-normalised correspondence descriptors $\mathbf{d}_t(u)$. Figure S1 details the order of attention, residual and feed-forward layers inside one decoder block.

On a stride-8 grid of stationary object points, pooling static background normals whenever foreground normals fall below 500 or 50% inlier support, finite-difference cross products provide surface normals, from which we initialise $\mathbf{G} \in SO(3)$ [13] via second-moment eigenvectors and run 48 projected gradient ascent steps via polar decomposition on Eq. (5) [35]. Residual camera drift on $\{\mathbf{X}_t\}_{t=1}^{T}$ and pairwise rigid part motions $(\mathbf{R}_{kt}^{r}, \mathbf{z}_{kt}^{r})$ are estimated in vectorised closed form via Cauchy-reweighted orthogonal Procrustes alignment on confidence-weighted dense correspondences from Eq. (3) with $\tau = 0.05$, using three steps for static-base alignment and six steps for part motion to form the motion scatter $\mathbf{S}_k$ in Eq. (4). The joint head attends across part and base boundary tokens alongside pairwise motion summaries to output positive-semidefinite prior matrices $\mathbf{M}_k = \mathbf{L}_k \mathbf{L}_k^{\top}$ for $\mathbf{M}_k \in \{\mathbf{M}_k^{\mathrm{rev}}, \mathbf{M}_k^{\mathrm{pri}}\}$, type logits and a boundary pivot $\hat{\mathbf{o}}_k$. Combining $\mathbf{S}_k$ and $\mathbf{M}_k$ in Eq. (6) and evaluating the three-way softmax of Eq. (7) with $\beta = 6$ selects $\mathbf{a}_k$ via straight-through argmax in Eq. (8), falling back to $\hat{\mathbf{a}}_k$ on untrusted frames when $\hat{\mathbf{a}}_k$ deviates by over 38°. Revolute versus prismatic classification combines the network's type logits with Procrustes residual reduction and a lever-arm bound against object radius, while Eq. (9) solves a 3×3 linear system with learned regulariser $\lambda_{\mathrm{o}} \geq 0.02$ for the hinge pivot $\mathbf{o}_k$. On unsegmented multi-part real clips where a single part is operated

TABLE S1
ADDITIVE DESIGN PROGRESSION ON THE FOUR DEVELOPMENT BENCHMARKS. ENTRIES REPORT JOINT ACCURACY AT 10° AND MEDIAN AXIS ERROR IN DEGREES.

Variant	Arti4D-dev	D3D-dev	Syn-10-val	Syn-38-val
Direct regression	0.063 / 57.1	0.058 / 38.7	0.188 / 20.0	0.125 / 30.7
+ Part and Motion Decoder	0.259 / 31.5	0.198 / 36.4	0.390 / 16.1	0.303 / 26.8
+ Bingham prior on Manhattan frame	0.522 / 31.4	0.620 / 30.7	0.778 / 1.6	0.601 / 2.3
PhyVGGT (Ours)	**0.725 / 3.6**	**0.744 / 2.9**	**0.803 / 1.6**	**0.636 / 2.1**

per interaction window, person pixels are masked [39] and the active slot k^* is selected automatically by motion amplitude weighted by root mask area.

We train PhyVGGT in two stages with AdamW on 8-frame clips of batch size 4 at 504×504, running 4,500 first-stage steps at learning rate 2×10^{-4} followed by 2,000 second-stage steps at 3×10^{-4} across 6,264 synthetic clips rendered from ten PartNet-Mobility [36] categories: *StorageFurniture, Microwave, Refrigerator, Oven, Dishwasher, Door, Laptop, Table, Faucet* and *Box*. Evaluation uses $T = 12$ test frames per clip and $T = 48$ on Articulat3D-Sim, averaging models trained in this work over three random runs to reduce training variance. Half of the training clips composite articulated objects into textured indoor rooms with articulated human hands grasping the moving handle, and a subset procedurally converts hinged doors into sliding doors to supervise 3D motion-based joint classification.

Predicted parts are matched to ground-truth parts via Hungarian assignment on soft Dice overlap on PartNet-Mobility, via majority label on ground-truth part tracks across the 20 scenes and 48 active joints of Articulat3D-Sim [8] moving by at least 10° or 3 cm, and via the active-part score on the 329 clips of Arti4D [19], 170 clips of D3D-HOI [37] and 357 clips of RBO [38]. Predicted joints are aligned to ground truth via $\mathrm{Sim}(3)$ from the first and last frames on Arti4D and RBO, from static-base tracks on Articulat3D-Sim, and from scene points on PartNet-Mobility, while D3D-HOI is evaluated directly in camera coordinates since it annotates 3D axis directions without depth. FAMOS [10] receives ground-truth object masks on D3D-HOI and PartNet-Mobility, and PhyVGGT's predicted foreground masks on Arti4D and RBO.

S3. COMPREHENSIVE ABLATION AND DIAGNOSTIC STUDIES

Table S1 traces how each design choice builds accuracy when trained from scratch on frozen VGGT-1B. Introducing our Part and Motion Decoder over Direct regression fuses intermediate per-frame tokens with global tokens and references part queries to the static base, recovering sharp 3D part boundaries and aligned correspondences to lift accuracy to 25.9% on Arti4D-dev and 39.0% on Syn-10-val against 6.3% and 18.8%. Restricting candidate axes to the static base's Manhattan frame under learned Bingham priors removes tilt towards the viewing ray to reach 52.2%, 62.0%, 77.8% and 60.1%, and further incorporating pairwise motion scatter in the full Kinematic Inference Layer identifies which orthogonal axis moves on casual videos, raising accuracy to 72.5%, 74.4%,

TABLE S2
DETAILED INFERENCE, CANDIDATE-FRAME AND SAME-BACKBONE VGGT-1B COMPARISONS ON THE DEVELOPMENT SPLITS. ENTRIES REPORT JOINT ACCURACY AT 10°.

Configuration	Arti4D-dev	D3D-dev	Syn-10-val	Syn-38-val
PhyVGGT (Ours)	**0.725**	**0.744**	**0.803**	**0.636**
Kinematic Inference Layer Choices				
Bingham prior on Manhattan frame only	0.522	0.620	0.778	0.601
Motion scatter eigenvector only	0.482	0.628	0.497	0.386
Continuous Bingham expectation	0.706	0.465	0.556	–
Camera-frame orthogonal candidates	0.043	0.430	–	–
Point-cloud PCA candidates	0.063	0.085	–	–
Same-Backbone VGGT-1B Baselines				
Direct regression	0.063	0.058	0.188	0.125
ArtiPoint on frozen VGGT-1B	0.271	–	–	–
FAMOS on PhyVGGT point maps	0.376	0.116	0.502	–

TABLE S3
EFFECT OF MULTI-LAYER TOKEN FUSION ON 3D PART SEGMENTATION IOU, ALIGNED POINT-MAP ERROR AND JOINT ACCURACY. POINT-MAP ERROR IS NORMALISED BY MEAN CAMERA DISTANCE.

	Syn-10-val			Syn-38-val	
Configuration	$\leq 10^\circ$	Part IoU	Pt. err.	$\leq 10^\circ$	Part IoU
PhyVGGT (Ours)	**0.803**	**0.701**	0.032	**0.636**	**0.576**
Final layer only ($l = 23$)	0.761	0.663	0.032	0.627	0.542

80.3% and 63.6% with 3.6°, 2.9°, 1.6° and 2.1° median error in a single 1.16 s forward pass.

Tables S2 and S3 evaluate each design choice inside our Part and Motion Decoder and Kinematic Inference Layer. Fusing multi-layer per-frame and global tokens in the Part and Motion Decoder improves joint accuracy to 80.3% against 76.1% on Syn-10-val and 72.5% against 69.8% on Arti4D-dev compared with reading only VGGT's final aggregator block $l = 23$, while lifting 3D Part IoU to 0.701 against 0.663 on Syn-10-val and 0.576 against 0.542 on Syn-38-val by preserving sharp intra-frame part boundaries. Within the Kinematic Inference Layer, closed-form discrete Bingham selection outperforms continuous posterior expectation at 72.5%, 74.4% and 80.3% against 70.6%, 46.5% and 55.6% by preventing depth noise along the viewing ray from blending orthogonal Manhattan modes, while fourth-order normal alignment locks onto object faces where camera-axis candidates drop to 4.3% on Arti4D-dev and PCA candidates drop to 6.3% and 8.5%. Under a controlled backbone comparison on identical frozen VGGT-1B point maps, PhyVGGT surpasses ArtiPoint at 72.5% against 27.1% and FAMOS at 72.5%, 74.4% and 80.3% against 37.6%, 11.6% and 50.2% by replacing continuous fitting with Bingham selection on the Manhattan frame.

Table S5 measures inter-frame part displacement directly in the official frozen point maps of π^3 [3], VGGT [2], StreamVGGT [29] and STream3R [30] without any trained head. Every model underestimates sliding displacement along the viewing ray to $\gamma \in [0.60, 0.69]$ while preserving lateral motion across the ray at 0.95–1.05, and bidirectional cross-frame attention in VGGT and π^3 biases moving-part depth towards context frames with a coefficient of 0.17 to 0.22. Since $\gamma \geq 0.60$ induces a maximum tilt of $\arctan((1 - 0.60)/(2\sqrt{0.60})) = 14.5^\circ \ll 45^\circ$ by Eq. (S2), Figure S2 confirms that PhyVGGT consistently outperforms continuous Motion estimate across every motion-range bin, achieving 75.0% against 41.7% on 10°–20° synthetic rotations, 61.9% against 14.3% on 5%–10% synthetic translations, 57.5% against 24.1% on small-motion *din080* clips and 58.6% against 21.8% on small-motion D3D-HOI clips.

TABLE S4
COMPARISON OF REVOLUTE PIVOT LINE DISTANCE ERROR IN CM AND ACCURACY WITHIN 10 CM.

Method	Syn-10	Syn-38	RBO
Direct regression	6.49 / 0.64	5.70 / 0.68	9.34 / 0.53
FAMOS	4.35 / 0.69	4.96 / 0.67	11.14 / 0.46
PhyVGGT (Ours), Eq. (9)	**3.16 / 0.83**	**3.36 / 0.82**	**8.76 / 0.55**

Table S4 evaluates revolute hinge pivot estimation across the 290 joints of Syn-10, 510 joints of Syn-38 and 357 clips of RBO. Solving the positive-definite linear system in Eq. (9) around the selected Manhattan axis reduces median line-to-line distance error to 3.16 cm, 3.36 cm and 8.76 cm, with 83%, 82% and 55% within 10 cm, against 6.49 cm, 5.70 cm and 9.34 cm for Direct regression and 4.35 cm, 4.96 cm and 11.14 cm for FAMOS, since eliminating axis angular tilt prevents lever-arm errors from shifting the recovered hinge line. End-to-end single-pass inference takes 1.16 s per 12-frame clip and 6.79 s per 48-frame scene on a dedicated NVIDIA A100 GPU.

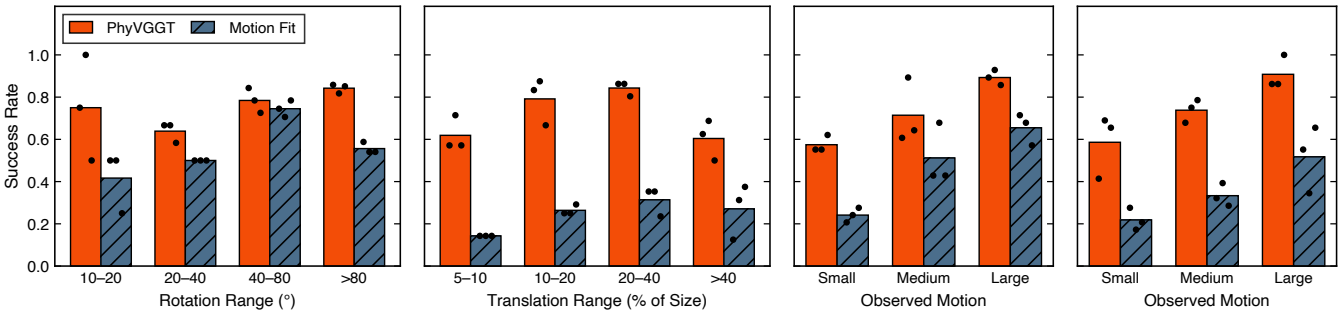
S4. ADDITIONAL QUALITATIVE RESULTS

Figures S3, S4 and S5 present additional single-pass feed-forward qualitative results of PhyVGGT across PartNet-Mobility test clips, egocentric Arti4D test interactions and third-person D3D-HOI test clips, showing segmented 3D parts across frames 1, 6 and 12 alongside predicted (orange) and ground-truth (green dashed) joint axes and pivots.

TABLE S5

EMPIRICAL MEASUREMENT OF VIEWING-RAY DISPLACEMENT UNDERESTIMATION ACROSS FOUR FROZEN 3D FOUNDATION MODELS. VALUES REPORT THE RATIO OF POINT-MAP PART DISPLACEMENT TO GROUND-TRUTH DISPLACEMENT ON 158 SYNTHETIC VALIDATION CLIPS WITH 95% BOOTSTRAP CONFIDENCE INTERVALS.

Motion component	π^3	VGGT	StreamVGGT	STream3R
Prismatic joint, along viewing ray (γ)	0.61 [0.52, 0.78]	0.69 [0.60, 0.82]	0.66 [0.52, 0.88]	0.60 [0.45, 0.81]
Prismatic joint, across viewing ray	0.95	0.99	1.05	0.95
Revolute joint, along viewing ray	0.87 [0.77, 0.98]	0.92 [0.81, 1.02]	0.89 [0.81, 0.97]	0.84 [0.74, 0.93]
Revolute joint, across viewing ray	0.98	0.97	1.05	0.96
Prismatic joint, along ray, single-frame input	0.62 [0.47, 0.82]	0.69 [0.63, 0.80]	0.70 [0.64, 0.81]	0.70 [0.63, 0.77]
Context-frame depth bias coefficient	0.18 to 0.22	0.17 to 0.20	-0.06 to -0.05	-0.05 to -0.04



(a) Synthetic Revolute (b) Synthetic Prismatic (c) Arti4D (d) D3D-HOI

Fig. S2. **Joint accuracy at 10° stratified by observed motion amplitude across four development benchmarks.** Bars show means and dots show individual training runs, comparing PhyVGGT against continuous Motion estimate (*Motion Fit* in legend) under small rotations and translations.

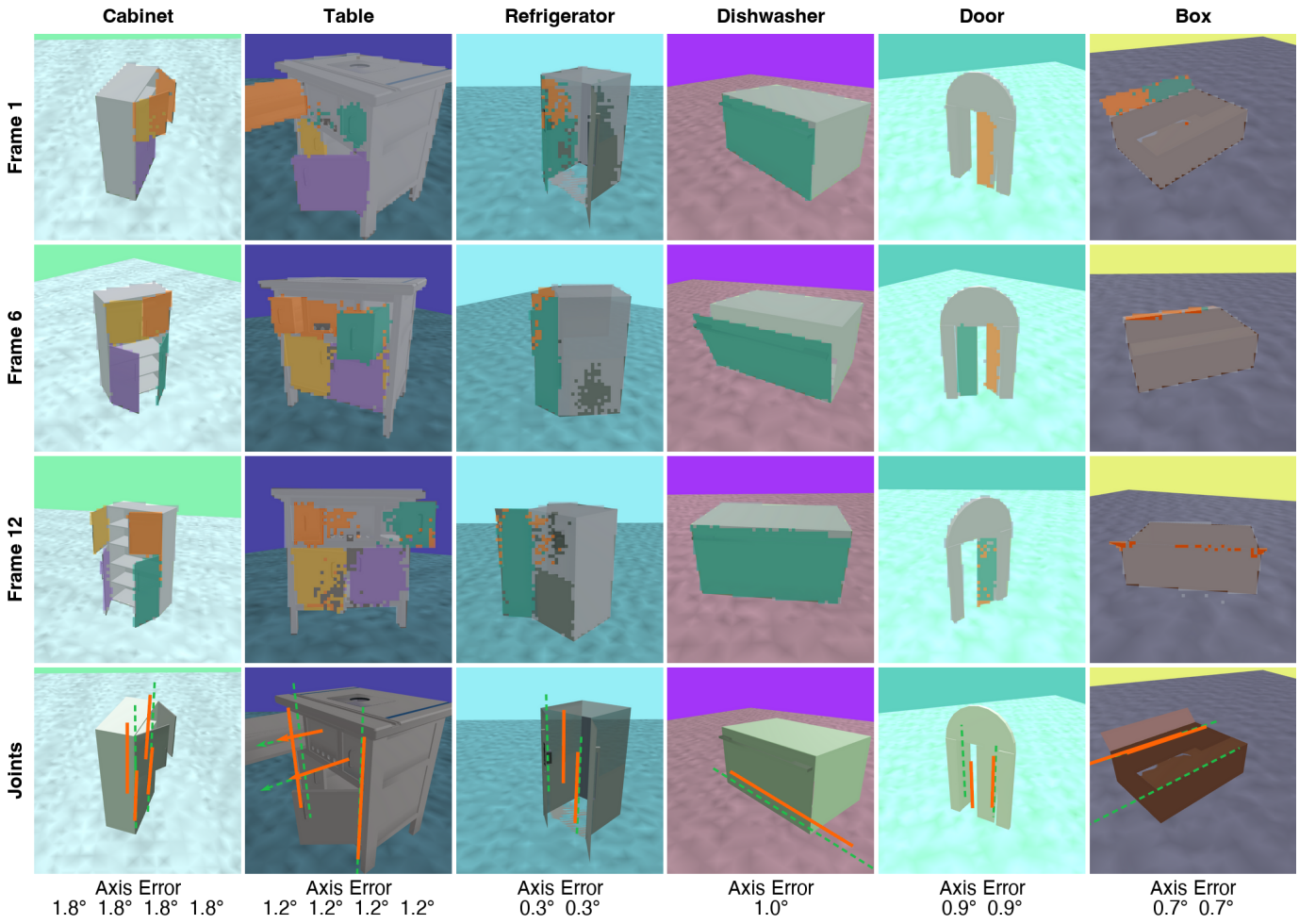


Fig. S3. **Qualitative results on PartNet-Mobility synthetic test videos.** Colours denote predicted movable parts and static base across frames, solid orange lines and arrows denote predicted revolute and prismatic joints, green dashed lines denote ground truth, and numbers indicate axis angular errors.

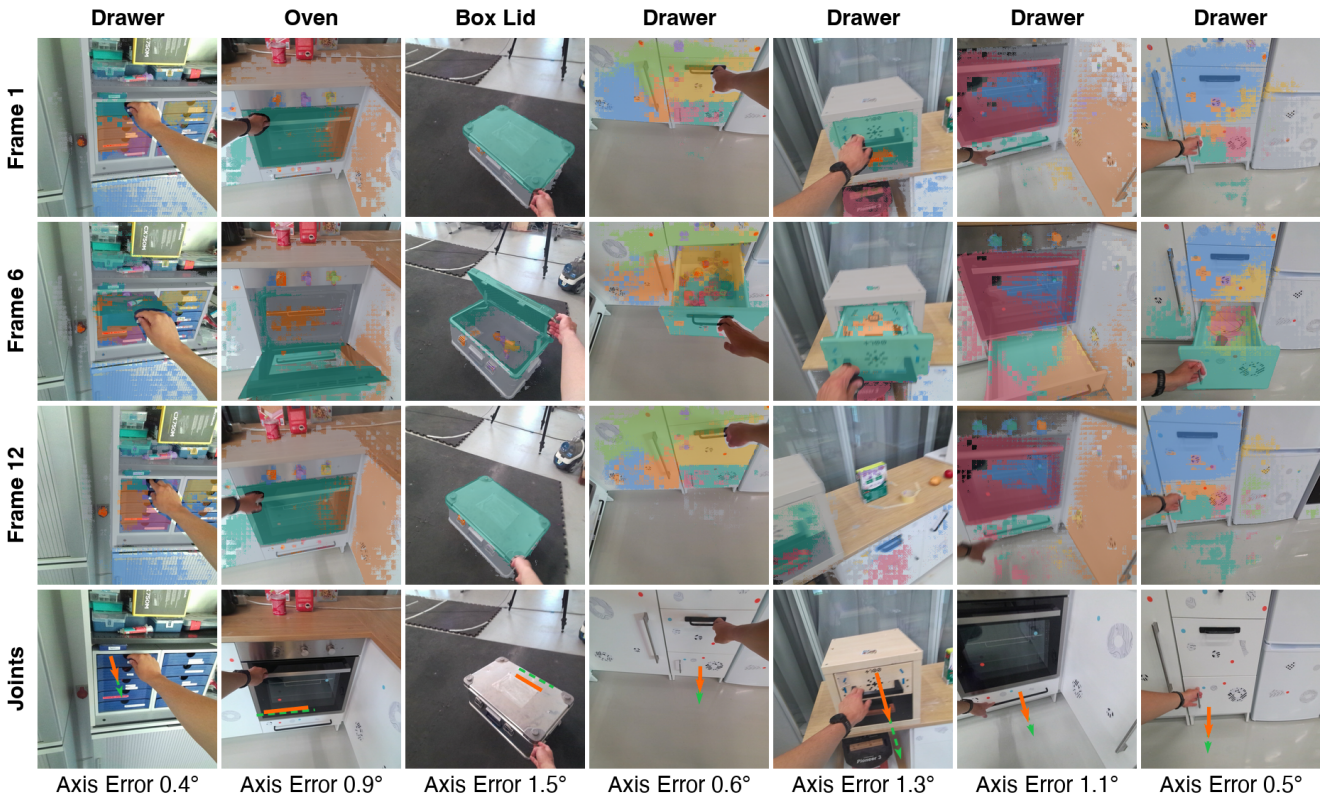


Fig. S4. **Zero-shot qualitative results on real egocentric Arti4D test videos.** All predictions are produced zero-shot in a single 1.2 s forward pass from uncalibrated RGB frames.

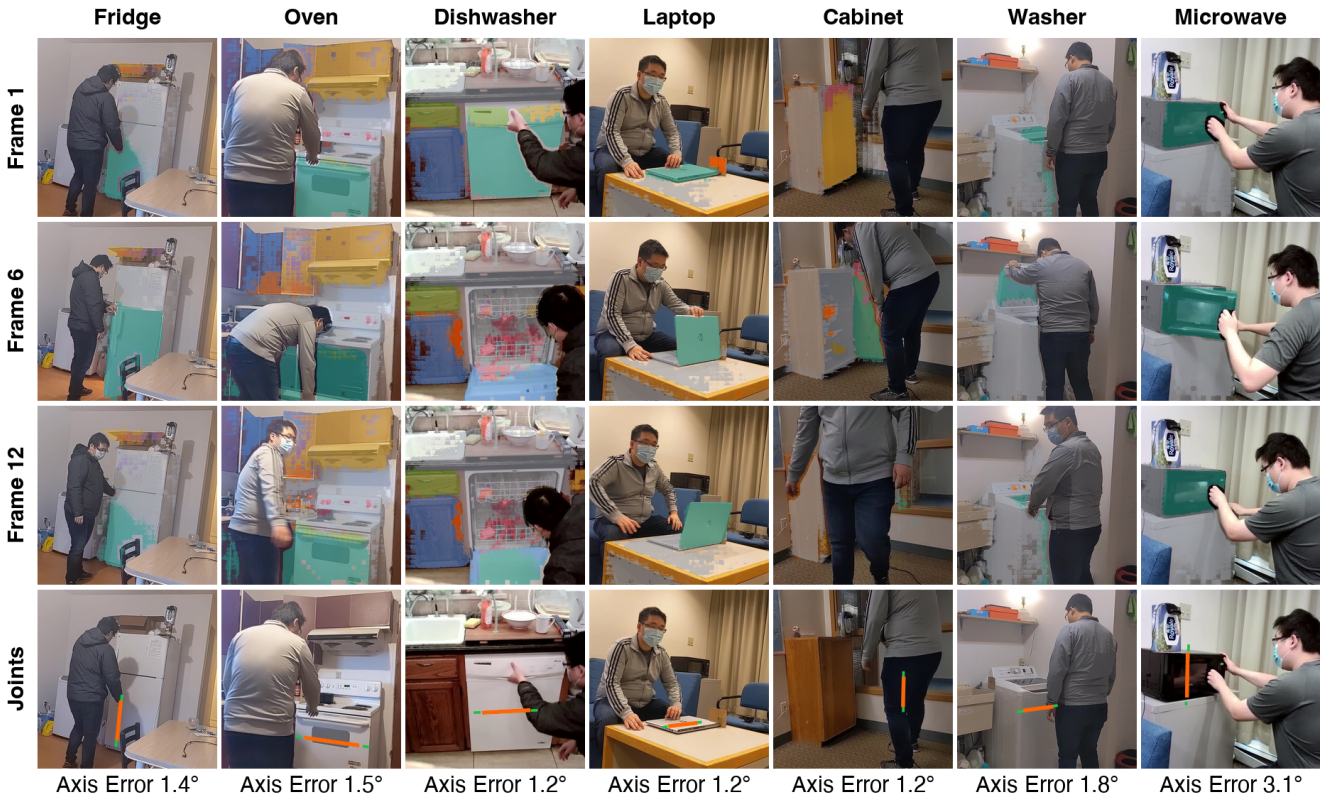


Fig. S5. **Zero-shot qualitative results on real third-person D3D-HOI test videos.** Since D3D-HOI annotates 3D axis direction without depth, both predicted and ground-truth axes are rendered through the part centroid.