EGOFOUND3R: END-TO-END EGOCENTRIC HAND RECONSTRUCTION IN WORLD SPACE WITH POINT-WISE INTERACTION ATTRIBUTES

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric video has become a primary source of supervision for embodied models, and its value rests on recovering hand motion in world coordinates, which camera motion and hand occlusion make difficult. Existing reconstruction pipelines typically separate hand and scene estimation, leave interaction attributes to separate task-specific models, and invoke several models per video, so no prior reconstruction model estimates these attributes and throughput becomes a practical constraint on large-scale annotation. We therefore introduce EgoFound3R, a unified end-to-end model that estimates world-space hand geometry in a metric scale shared with the scene, and predicts point-wise interaction attributes, including visibility, contact, and distance. The model integrates three designs: (i) structured hand prompts that transfer pretrained geometric priors to world-space hand reconstruction; (ii) an explicit hand representation that decodes hand geometry and interaction attributes; and (iii) a shared-parameter multi-rate design that lowers inference cost. Together, these designs predict hand geometry and point-wise attributes in one pass. On OakInk-v2, TACO, and HOI4D, EgoFound3R reduces the mean per-joint position error (MPJPE) by 43.2%, 22.4%, and 11.6% over previous methods and predicts point-wise contact and distance alongside the geometry in the same pass, while attaining approximately $6\times$ higher throughput.

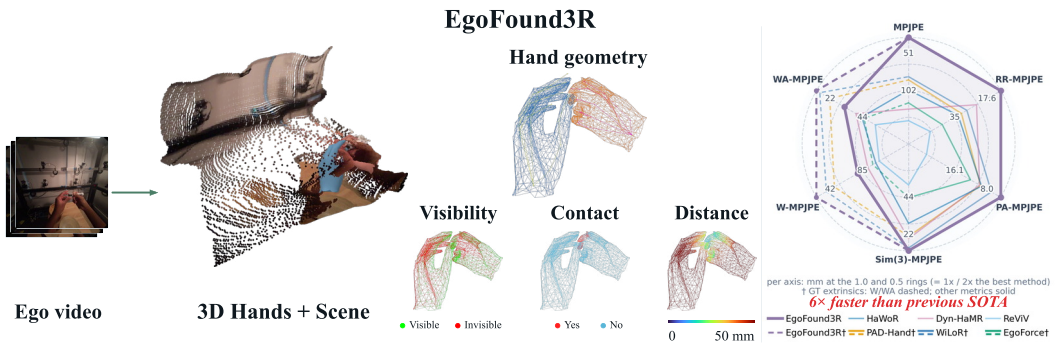


Figure 1: **EgoFound3R at a glance.** End-to-end world-space hand reconstruction with point-wise interaction attributes; the radar panel summarizes Tab. 1.

1 INTRODUCTION

Egocentric video captures manipulation as it is performed, and it is the primary channel through which embodied models acquire human manipulation skill: dexterous manipulation, world modeling, and vision-language-action policies increasingly learn from first-person recordings (Grauman et al., 2022; Paliwal et al., 2026; Li et al., 2026b), which now span 10^5 to 10^6 hours (Zheng et al., 2026; Li et al., 2026a; Wu et al., 2026; Ye et al., 2026a). In particular, they suit these models because the hand, the objects it acts on, and the wearer's motion appear in one view, and because such data can be collected at scale without instrumenting the environment. World-space hand geometry is a primary target of that supervision: hand placement and its attributes are defined relative to the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

scene rather than to the camera, so they stay fixed while the wearer moves. However, recovering this geometry is difficult: camera egomotion and hand–object occlusion corrupt the monocular cues, and the observations themselves provide no absolute scale. In addition, the scale of current collections makes throughput a first-order constraint: a method that processes a fraction of a recording per day determines how much supervision the data can provide. Errors in hand placement or interaction state then propagate into the trained model, where they are difficult to detect.

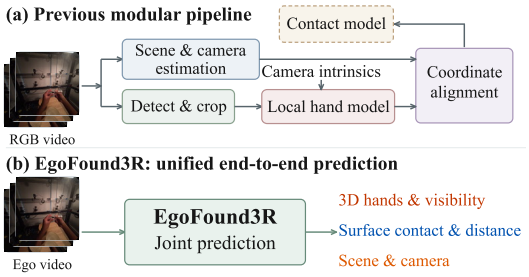


Figure 2: Pipeline vs. end-to-end.

Recent estimators have improved hand pose accuracy in both camera and world space, but existing pipelines chain them in separate stages, as illustrated in Fig. 2(a), leaving both constraints unresolved. Specifically, camera-space estimators recover local hand geometry, world-space methods such as HaWoR (Zhang et al., 2025) and Dyn-HaMR (Yu et al., 2025) add camera tracking or motion completion, and contact is delegated to additional predictors or post-hoc optimization. The stages keep separate representations of the same evidence, so geometric priors are not shared, and errors accumulate across stages and over long sequences. Repeated visual encoding also lowers throughput, and since annotation is the bottleneck, this cost directly limits how much of a recording becomes supervision. Interaction attributes such as visibility, contact, and distance are likewise not estimated by any reconstruction model; they are left to separate predictors that take the hand as given, so errors in the hand or its placement propagate into the attributes rather than being corrected. Together, these limitations motivate a single model that places hands in world coordinates, predicts hand geometry and interaction attributes in the same pass, and does so at the throughput that annotation at scale demands.

Such a model can be built on geometry backbones pretrained on large-scale 3D data, which already encode the scene context that disambiguates an occluded hand (Chen et al., 2026b; Lyu et al., 2026). We therefore introduce **EgoFound3R**, an end-to-end model that estimates world-space hand geometry and predicts point-wise visibility, contact, and distance in one pass. The model integrates three designs: **(i) Structured hand prompts.** An explicit hand representation is conditioned on a frozen VGGT-Ω (Wang et al., 2026b) backbone through structured prompts. Only the prompts, adapters, and prediction heads are trained, so inference needs only monocular video, with no depth, calibration, object template, or detector-and-crop stage. **(ii) Explicit hand representation.** Geometry and attributes are decoded from the same point features: the model regresses root-relative joints and vertices together with point-wise visibility, contact, and distance, and an observation-constrained root estimate places the hand in the scene. **(iii) Shared-parameter multi-rate design.** Hand observations need dense coverage while global scene aggregation does not, and global aggregation is the dominant cost in prior pipelines; we therefore run it at a configurable lower rate, reducing redundant computation without duplicating parameters.

We evaluate EgoFound3R on OakInk-v2, TACO, and HOI4D, where it reduces the mean per-joint position error (MPJPE) by 43.2%, 22.4%, and 11.6% relative to previous methods and predicts point-wise contact and distance alongside the geometry in the same pass, while attaining approximately $6\times$ higher throughput than WiLoR (Potamias et al., 2025), the previous state-of-the-art hand-reconstruction method.

Our contributions are threefold:

- We present EgoFound3R, an end-to-end model that reconstructs world-space hands and predicts point-wise visibility, contact, and distance in one pass, replacing the chained pipeline of hand estimation, camera tracking, and contact prediction with a single model.
- We show that interaction attributes can be predicted inside the reconstruction model: visibility, contact, and distance are decoded point-wise from the same features as the geometry, without separate task-specific models.
- At annotation scale, EgoFound3R runs $6\times$ faster than WiLoR while improving world-space hand accuracy on all three datasets, enabled by its shared-parameter multi-rate design.

2 Related Work

Geometric priors and unified reconstruction. DUSt3R (Wang et al., 2024), VGGT (Wang et al., 2025a), π^3 (Wang et al., 2026c), and Depth Anything 3 (Lin et al., 2026) establish feed-forward geometric reconstruction. VGGT-Ω (Wang et al., 2026b), CUT3R (Wang et al., 2025b), and LingBot-Map (Chen et al., 2026a) extend this line toward scalable or persistent scene representations. These priors have also been used for articulated reconstruction. Human3R (Chen et al., 2026b) introduces human prompts and ReViV (Lyu et al., 2026) jointly predicts hand joints, cameras, and scene depth. EgoFound3R adds interaction attributes to this joint reconstruction, decoding hand geometry and attributes from shared full-image features without a separate detection or crop stage.

Egocentric data for embodied models. Dexterous manipulation, world-modeling, vision-language-action, and world-action systems such as EgoScale (Zheng et al., 2026), LingBot-VA (Li et al., 2026a), LingBotVLA2 (Wu et al., 2026), and DreamZero (Ye et al., 2026a) are increasingly trained on egocentric data, where annotation throughput becomes a practical constraint. EgoFound3R addresses this constraint by producing world-space hand geometry, scene context, and point-wise visibility, contact, and distance in one model.

Hand geometry and motion. HaMeR (Pavlakos et al., 2024) and WiLoR (Potamias et al., 2025) estimate camera-space hand geometry; EgoForce (Millerdurai et al., 2026) incorporates forearm crops and calibrated geometry for metric placement. World-space methods combine hand and camera estimates, including HaWoR (Zhang et al., 2025) and Dyn-HaMR (Yu et al., 2025). Generative refinement further constrains motion: PAD-Hand (Ismayilzada et al., 2026) applies physics-aware diffusion to WiLoR estimates, and StableHand (Zeng et al., 2026) uses quality-aware flow matching with hand and camera conditioning. Explicit representations such as CMR (Chen et al., 2021) and MobRecon (Chen et al., 2022) regress hand-mesh coordinates directly. EgoFound3R extends this direction by attaching point-wise interaction attributes to the same explicit representation.

Contact and interaction data. Contact estimation relies on geometric or visual evidence: ContactOpt (Grady et al., 2021) and S^2Contact (Tse et al., 2022) classify contact from hand-object geometry, InteractVLM (Dwivedi et al., 2025) transfers image-based interaction cues, and EgoPHI (Ilic et al., 2026) predicts dense contact and force from additional geometry. Beyond classification, WHOLE (Ye et al., 2026b) and EgoGrasp (Fu et al., 2026) estimate world-space hand–object interaction. For large-scale annotation, Do as I Do (Paliwal et al., 2026) builds on HaWoR, EgoInfinity (Wang et al., 2026a) refines WiLoR in a static-camera setting, and Open-AoE (Li et al., 2026b) extracts supervision at scale. Because these systems chain several learned models per video, cost accumulates with each added stage. EgoFound3R replaces the chain with one end-to-end model that predicts hand geometry and interaction attributes and requires no object meshes at inference.

3 Method

3.1 Problem Formulation

Given a monocular egocentric RGB video, we aim to recover hand geometry in world space together with point-wise interaction attributes, and to place the hands in the surrounding scene. The setting is difficult in three respects: the camera moves with the wearer, so camera motion and hand motion are entangled; the hands are small and frequently occluded by the objects they hold; and a single monocular view carries no absolute scale. The two sides, however, do not behave alike in time: hands change quickly and demand dense observation, whereas the scene evolves slowly while still providing the scale and camera motion that place them. As illustrated in Fig. 3(1, 2), EgoFound3R therefore keeps both in a single feed-forward model and assigns them different temporal rates, in place of the detector, hand model, camera tracker, and contact predictor that prior pipelines run in sequence. We write H for the hand-observation times, $G \subseteq H$ for the sparser anchors at which the scene is updated, and R for the requested output times. Hand root translation, joints, and surface points with their attributes are predicted at every $h \in H$, while the anchors $g \in G$ carry depth and camera motion that place them, and hand and camera outputs are returned at every time in R.

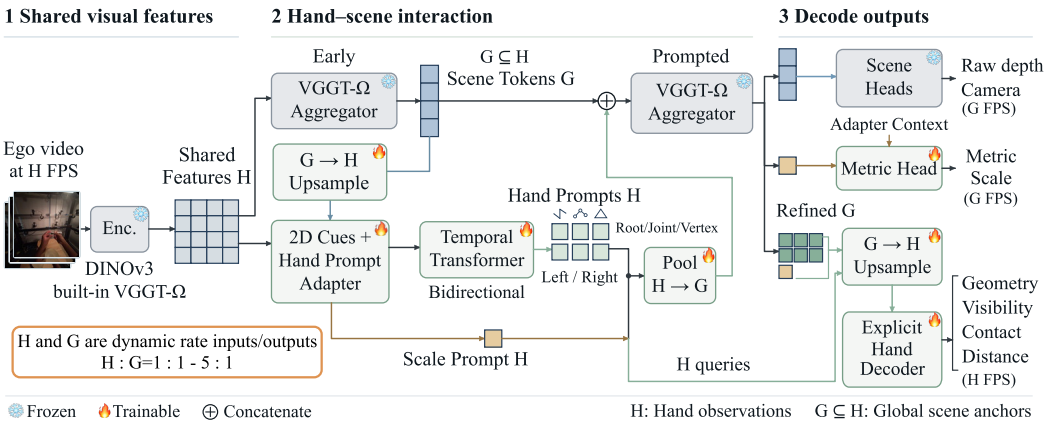


Figure 3: **Overview of EgoFound3R.** Egocentric video input is encoded at the hand rate H, and the built-in DINOv3 features of the frozen VGGT-Ω backbone provide the shared visual features, from which an adapter builds structured root, joint, and surface prompts. A bidirectional temporal Transformer links the two hands across H, and pooling neighboring prompts yields global scene anchors $G \subseteq H$ at an adjustable H:G ratio. The prompted VGGT-Ω aggregator drives the explicit hand head that decodes hand geometry, visibility, contact, and distance, and the scene and metric heads that place the hands in world coordinates.

3.2 Hand Prompts that Reuse Pretrained Geometry

The first design question is what a hand model should look at. A hand crop contains little beyond the hand, which is exactly the evidence that occlusion and camera motion corrupt, whereas a geometry backbone pretrained on large-scale 3D data already encodes the surrounding scene that disambiguates the hand. We therefore keep such a backbone frozen—VGGT-Ω (Wang et al., 2026b), whose built-in DINOv3 features (Siméoni et al., 2026) provide the shared visual representation—and let the hands read from it through learned queries, as shown in Fig. 3(1), instead of regressing them from a crop in isolation. The queries are organized by the question they ask: root queries locate the hand, joint queries describe its articulation, and surface queries describe what it touches, with a bidirectional temporal Transformer allowing the two hands to exchange information over the clip. Detection is one more question of the same form, so the model estimates where and when each hand is visible from 2D heatmaps and in-view scores computed on the same features, and no external detector is required. The queries are the only hand-specific inputs to the frozen backbone, whose geometric priors are reused rather than overwritten, and only the prompts, adapters, and prediction heads are trained; Appendix A details the token layout.

3.3 Cross-Rate Interaction between Hands and Scene

Grounding the hands in the scene raises a second design question: how often must that scene context be refreshed? Re-aggregating the scene at the hand rate is a major source of cost in prior pipelines, while the hands still require dense observation, so we decouple the two rates and expose their ratio as a single user-facing knob, as shown in Fig. 4(a, b). A stride $s \in \{1, \ldots, 5\}$ subsamples the hand times H into the anchors G, and the hand prompts observed since the previous update are pooled into each anchor, so that the scene receives a compact summary of the hand motion it has missed:

$$p_g = p_{a_g} + \gamma f\big(\mathrm{LN}(\bar{p}_g - p_{a_g}), \mathrm{LN}(p_{a_g})\big)\,. \tag{1}$$

Here $\bar{p}_g$ is the pooled summary of the neighboring hand prompts, p_{a_g} the anchor prompt before pooling, p_g the updated anchor prompt, f a residual projection, γ a scalar gate, and LN layer normalization. The update is residual and zero-initialized, so training starts from the dense behavior and larger strides only thin the aggregation. The anchors then re-enter the frozen aggregator, and a second bridge returns the aggregated context to the hand rate, so evidence flows in both directions. Because every stride reuses the same adapters and phase embedding, the ratio acts as a computation knob rather than defining a new model.

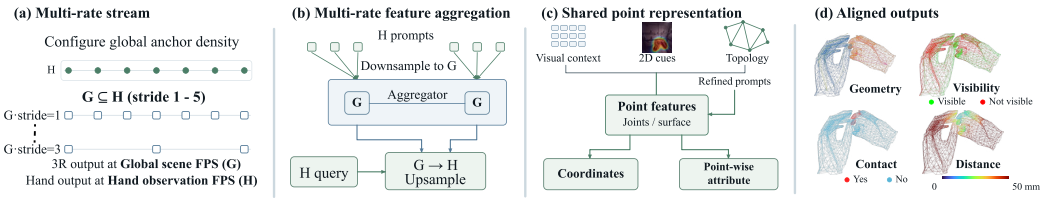


Figure 4: **Cross-rate prompting and point decoding.** (a, b) Hand prompts run at rate H while global anchors $G \subseteq H$ aggregate them at a configurable lower rate. (c, d) A shared point representation decodes geometry and point-wise attributes.

3.4 Explicit Hand Geometry with Interaction Attributes

The hand features now carry scene context and 2D hand evidence, but they still have to be turned into geometry and interaction attributes. The attributes we target are defined on individual hand points—whether a point is visible, whether it touches, and how far it lies from the contact surface—so we decode point coordinates directly instead of describing the hand through pose and shape coefficients, as shown in Fig. 4(c, d). Root, joint, and surface queries read the same point features, which keeps placement, articulation, and contact aligned; the surface branch additionally inherits features from associated joints and mixes over incident faces, so the decoded points respect the connectivity of the hand. Each point then receives three attributes from a small prediction head: visibility, contact, and distance to the contact surface. Contact merges object contact and opposite-hand contact into one per-point state that never names the partner, so no object template or object pose is required. The hand stays supervised through MANO topology (Romero et al., 2017), yet its geometry is decoded as point coordinates rather than pose or shape parameters.

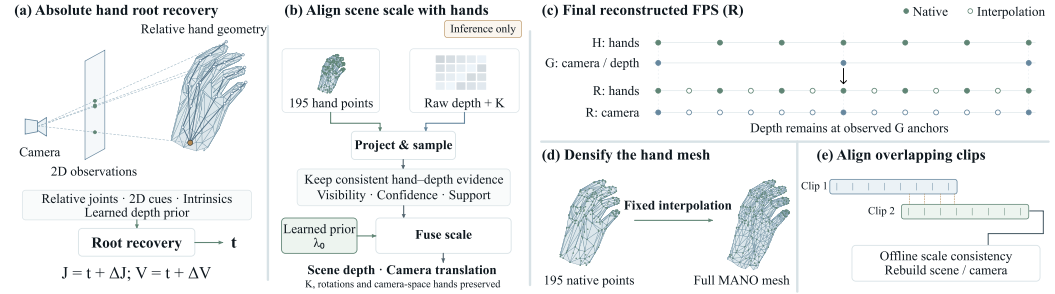


Figure 5: **Root recovery, scale alignment, and outputs.** (a) Root recovery from 2D rays, relative joints, and a learned depth prior. (b) Inference-time scale alignment, with camera-space hands and rotations fixed. (c) Outputs at requested times R; (d) 195→MANO densification; (e) scale-consistent clip stitching.

3.5 Metric Hand Placement

Decoding hand geometry does not yet place the hand in the camera, and the depth of a monocular scene is defined only up to an unknown factor. Both are questions of scale, and the hands answer both. As shown in Fig. 5(a), the hand root is recovered from the observed 2D rays and the predicted relative joints under the wrist constraint, with a learned depth prior stabilizing the solve:

$$J_j = t + \Delta J_j, \qquad V_v = t + \Delta V_v, \tag{2}$$

where t is the root translation, ΔJ_j and ΔV_v the predicted offsets of joint j and vertex v, and J_j and V_v the resulting point coordinates, which the predicted cameras finally express in world space.

The same evidence sets the metric scale of the scene, as shown in Fig. 5(b): a metric head predicts it from scene summaries together with hand features, a learned prior keeps it well defined when hand evidence is weak, and at inference the agreement between the projected hand vertices and the raw depth refines it further. Calibration rescales scene depth and camera translation only, so rotations,

intrinsics, and the camera-space hands stay unchanged and the hand branch is never recomputed; Appendix J reports the scene and camera outputs.

3.6 Training and Outputs at Arbitrary Times

The designs above form a single model that is trained end-to-end on a mixture of egocentric datasets, with the backbone frozen so that only the prompts, adapters, and prediction heads are optimized; Appendix B states the resulting objectives. The network runs at the hand rate, but applications need outputs at their own times and at the mesh resolution their downstream use requires. We therefore treat the requested times R as an interface and recover them from the network outputs by deterministic interpolation, as shown in Fig. 5(c), and the same stage densifies the native 195-vertex geometry to the full 778-vertex MANO mesh, as shown in Fig. 5(d). Dense outputs thus cost no additional network computation, and the predicted points are preserved by construction. Long videos are processed as overlapping clips that must share one scale before stitching, which the same stage restores by rescaling their reconstructions to agree on the shared frames, as shown in Fig. 5(e). Because interpolation uses future frames and the alignment compares whole clips, EgoFound3R is an offline method, and both operations remain outside the feed-forward network.

4 Experiments

4.1 Evaluation Setting

We train on seven egocentric datasets: H2O (Kwon et al., 2021), HOT3D Aria (Banerjee et al., 2025), OakInk-v2 (Zhan et al., 2024), TACO (Liu et al., 2024), HOI4D (Liu et al., 2022), ARCTIC (Fan et al., 2023), and STERA-10M (Palanisamy et al., 2026). Evaluation uses 2,378 windows sampled at a fixed ratio per dataset, each a 2 s clip and held out from training; the main tables report OakInk-v2, TACO, and HOI4D, where the full metric suite is available, and the remaining datasets appear in the qualitative comparisons. Our model uses the default stride $s = 5$; we defer the temporal and scene studies to Appendices C and J. To bound evaluation cost, Dyn-HaMR and InteractVLM are each evaluated on a 100-window subset paired with an EgoFound3R control on the same windows. Geometry is reported in millimeters as mean per joint position error (MPJPE) and mean per vertex position error (MPVPE) under six alignments: unaligned (Raw), root-relative (RR), per-frame aligned (PA), similarity-aligned ($\mathrm{Sim}(3)$), world (W), and world-aligned (WA). Contact and visibility use precision (P), recall (R), and F1, and distance uses the mean absolute error (MAE, mm) of the point-to-surface distance, reported over all valid points (MAE(All)), predicted contacts (MAE(Pred)), and ground-truth contacts (MAE(GT)). See Appendix D for details.

4.2 Hand Geometry and Interaction

Hand geometry. Tab. 1 reports world-space hand reconstruction, comparing EgoFound3R with camera-space estimators (WiLoR (Potamias et al., 2025) and its refinements PAD-Hand (Ismayilzada et al., 2026) and EgoForce (Millerdurai et al., 2026)), world-space pipelines (HaWoR (Zhang et al., 2025) and Dyn-HaMR (Yu et al., 2025), the latter on matched 100-window subsets), and the joint reconstruction system ReViV (Lyu et al., 2026). EgoFound3R attains the lowest raw and root-relative joint error on every subset, reducing MPJPE by 43.2%, 22.4%, and 11.6% relative to the strongest baseline on OakInk-v2, TACO, and HOI4D; the ordering survives the removal of the root translation, and on the common windows the model also attains the lowest world and world-aligned error among the methods that predict their own camera poses. The three benchmarks differ in capture rigs and object vocabularies, so the consistent ordering across them reflects the design rather than a single benchmark, and the baselines fail as their designs predict: Camera-space estimators recover shape once the prediction is aligned per frame, but their raw and world errors stay high because they never estimate where the camera sits in the scene. World-space pipelines place the hands but drift as sequences grow: on the matched subsets Dyn-HaMR's world error rises several-fold from OakInk-v2 to HOI4D, HaWoR loses articulation on OakInk-v2, where its raw error is more than twice its value on the other two datasets, and ReViV is the weakest overall, with per-frame aligned errors several times ours on every subset. Our single-pass design targets this failure mode: hands and scene are decoded by one model, so the hand trajectory inherits one scale across the clip instead of the drift of chained stages; Appendices E and F report the vertex-level and temporal results.

Table 1: **Joint reconstruction** (MPJPE, mm); each cell lists OakInk-v2, TACO, and HOI4D left to right, with † marking methods placed with ground-truth extrinsics and ‡ marking EgoFound3R with ground-truth intrinsics and predicted extrinsics. For EgoFound3R† the camera-space columns are identical to the default row and are left blank. ^{100w} subsets are unranked, W/WA are ranked within each extrinsics group.

	OakInk-v2 · TACO · HOI4D																	
Method	Raw↓			RR↓			PA↓			Sim(3)↓			W↓			WA↓		
WiLoR†	91.24	70.40	81.51	27.66	28.33	27.75	**7.78**	9.49	9.62	**16.89**	25.95	24.14	42.14	51.14	**49.41**	**17.04**	26.14	24.57
PAD-Hand†	107.16	72.97	75.35	32.23	30.30	29.77	9.63	11.19	10.12	21.37	29.21	24.79	49.01	57.06	51.92	21.62	29.60	25.04
EgoForce†	123.57	126.63	147.77	47.75	44.53	52.78	11.09	11.93	12.92	32.62	47.84	49.22	80.59	105.52	142.58	33.10	46.65	55.76
EgoFound3R†	–			–			–			–			**27.49**	**36.68**	63.01	17.89	**24.06**	**23.30**
Dyn-HaMR^{100w}	96.39	127.12	245.06	17.91	27.47	26.49	7.84	11.62	11.89	13.76	29.27	30.47	36.16	77.88	173.62	17.19	31.46	65.71
HaWoR	174.73	74.28	53.46	45.85	24.19	28.47	9.49	10.55	10.49	32.19	27.71	27.85	100.46	86.38	171.09	37.47	35.34	55.96
ReViV	253.68	91.19	357.46	80.60	62.48	82.12	44.14	40.14	39.66	58.05	54.08	58.04	123.25	93.25	182.33	57.75	54.26	68.74
EgoFound3R	51.81	54.64	47.28	**17.77**	18.54	**16.63**	8.02	**8.13**	**7.99**	17.87	**24.15**	**23.29**	**30.88**	**40.03**	159.70	**18.77**	**25.16**	49.84
EgoFound3R‡	**39.45**	**50.44**	**43.43**	17.84	**18.52**	16.67	8.04	**8.13**	8.00	18.01	24.63	23.39	31.56	41.78	**159.18**	18.94	25.75	**49.77**

Table 2: **Joint-level contact prediction**. Superscripts mark 100-window subsets. InteractVLM is shown for completeness and does not enter the ranking.

	OakInk-v2			TACO			HOI4D		
Method	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
InteractVLM^{100w}	0.180	0.403	0.202	0.388	0.705	0.472	0.313	0.537	0.281
S^{2}Contact	0.327	0.081	0.079	0.519	0.104	0.146	0.444	0.265	0.284
ContactOpt	0.334	0.196	0.154	0.557	0.197	0.259	0.518	0.477	0.435
EgoFound3R	**0.510**	**0.665**	**0.493**	**0.654**	**0.777**	**0.697**	**0.612**	**0.835**	**0.675**

Table 3: **Visibility prediction** at 21 joints on OakInk-v2, TACO, and HOI4D. HVD predicts joint visibility only; vertex-level results are reported in Appendix G.

	OakInk-v2			TACO			HOI4D		
Method	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
HVD	0.823	**0.808**	0.792	0.878	0.834	0.853	**0.917**	0.769	0.831
EgoFound3R	**0.919**	0.747	**0.802**	**0.917**	**0.850**	**0.880**	0.892	**0.884**	**0.885**

Contact prediction. Tab. 2 reports joint-level contact prediction, comparing EgoFound3R with InteractVLM (Dwivedi et al., 2025), S^{2}Contact (Tse et al., 2022), and ContactOpt (Grady et al., 2021). EgoFound3R predicts contact together with the hand geometry, on the same points and from the same scene-aware features rather than by a separate classifier, and it is the only compared method that also predicts a per-point distance to the contact surface. EgoFound3R attains the highest precision, recall, and F1 on all three subsets. The baselines fail in opposite directions: the geometric classifiers keep precision moderate but recover few contacting joints, whereas InteractVLM trades precision for recall. Decoding contact on the points that carry the geometry, from features that already encode the surrounding scene, improves both sides at once. The compared systems use different contact conventions and the two geometric baselines additionally receive ground-truth geometry; Appendix G states the protocol and the sensitivity analysis.

Visibility. The same head predicts point visibility, which the compared pipelines do not report: as shown in Tab. 3, EgoFound3R surpasses the joint-only predictor HVD (Hara et al., 2026), in joint-visibility F1 on all three datasets, with higher precision on two subsets and higher recall on two others, and it also predicts vertex-level visibility; Appendix G reports the vertex-level results.

4.3 World-Space Qualitative Comparison

Fig. 6 shows two clips rendered in the reconstructed world frame from two fixed viewpoints, so the comparison includes where the cameras place the hands rather than a single projected view. Camera-space estimators recover the hand shape but offset it relative to the object surface, and both world-space pipelines drift away from it instead of recovering the hand motion; ReViV predicts joints only and provides no hand surface. EgoFound3R reproduces the ground-truth layout with its own predicted cameras; Appendix L shows additional results.

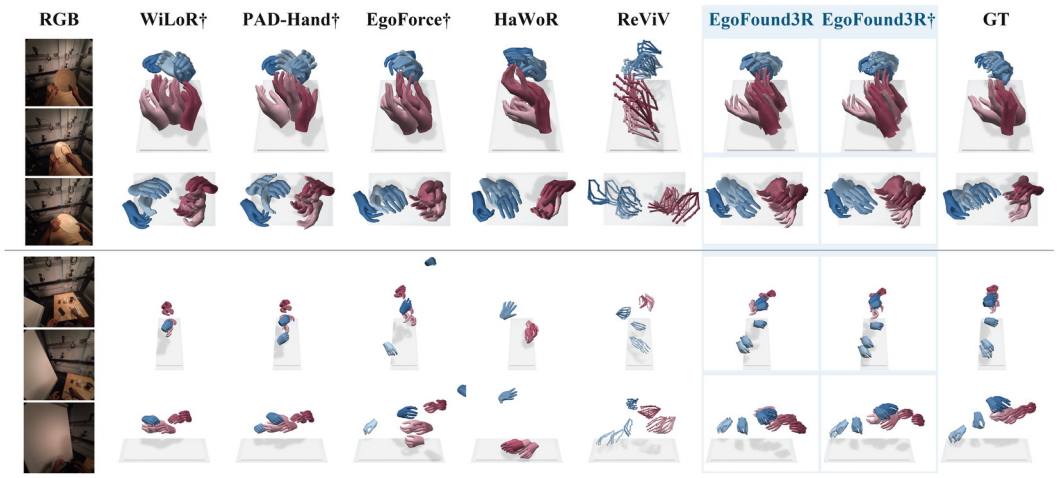


Figure 6: **World-space qualitative comparison.** Rows: two viewpoints of each clip; columns: the compared methods, rendered in the reconstructed world frame. Hands are pink (left) and blue (right), darker over time; † marks methods placed with ground-truth extrinsics.

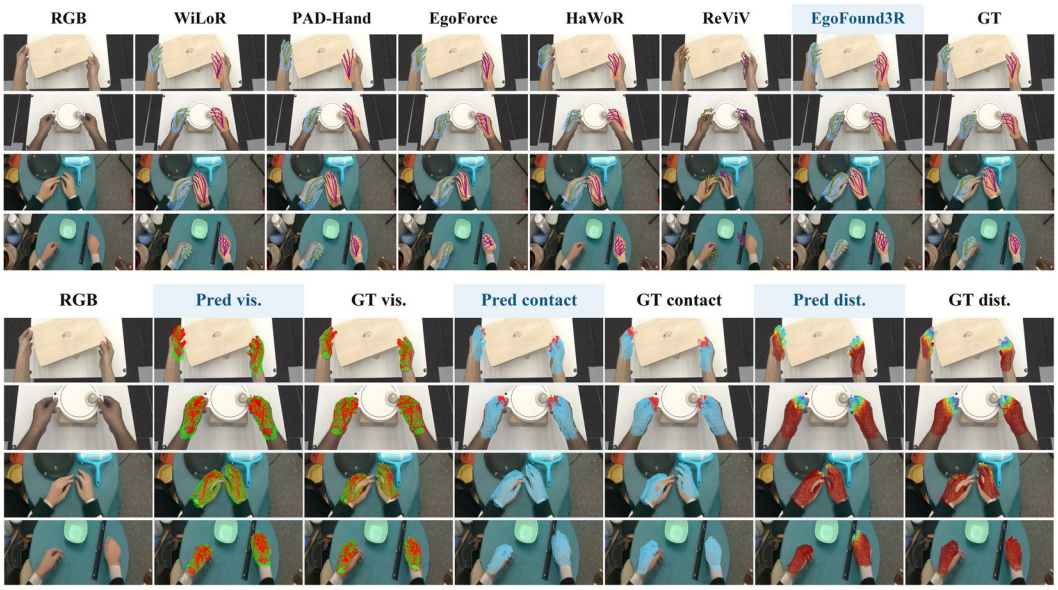


Figure 7: **Camera-space qualitative comparison.** Top: hand geometry (left/right hands in pink/blue). Bottom: visibility, contact, and distance for EgoFound3R (*Pred*) and the ground truth (*GT*); green/red = visible/occluded, red/light blue = contact/no contact, and distance is a blue-to-red 0–50 mm map.

4.4 Camera-Space Qualitative Comparison

Fig. 7 shows four held-out windows, two from ARCTIC and two from TACO; both halves are 2D overlays on the observed pixels, with hand geometry on top and visibility, contact, and distance below. On all four windows EgoFound3R keeps the hand aligned with the object surface and the attribute maps follow the ground truth, consistent with Tab. 1 and Tab. 2. In these windows every compared method is displaced from the annotated hand and misestimates its pose and size, whereas our overlay tracks the annotation. The attribute heads read the same points that carry the geometry; Appendix K shows additional results.

4.5 Core Design Ablation

We ablate the design choices on TACO; Appendix I defines each variant and repeats the study on the three evaluation datasets. (i) **Model capacity.** Reducing the hand module to 0.05B and 0.1B parameters raises raw error monotonically, and the reduced variants are also the weakest in the contact and visibility columns and among the weakest in distance, so the representation benefits from capacity rather than saturating early. (ii) **Pretrained backbone.** Removing the frozen VGGT-Ω backbone (w/o Ω) raises raw error by about half over the default and costs the distance columns, which supports reusing pretrained geometric priors rather than relearning them from the supervision alone. (iii) **Hand–scene exchange.** Restricting the exchange to one direction (1-way) degrades geometry and attributes alike, and it costs more contact F1 than removing the backbone does, indicating that hand evidence must also shape the scene representation rather than only read from it. (iv) **MANO parameters.** Replacing the explicit point representation with MANO parameter regression degrades every column relative to the default, and it is the weakest of the three structural variants in geometry, visibility, and distance, which is consistent with decoding attributes on the same points that carry the geometry. (v) **Multi-rate default.** S5 is the default configuration of EgoFound3R, and the stride ablation for the other rates is reported in Appendix I.

Table 4: **Core design ablation** on TACO; the F1 columns are measured at 21 joints, and Sk is EgoFound3R at stride k.

Variant	Raw↓	RR↓	W↓	WA↓	Contact F1↑	Vis. F1↑	MAE(All)↓	MAE(Pred)↓	MAE(GT)↓
0.05B	109.71	36.74	81.37	42.66	0.542	0.800	41.47	45.30	49.54
0.1B	96.90	30.55	61.66	35.81	0.579	0.801	49.27	52.15	58.61
MANO	86.66	24.29	57.45	35.15	0.643	0.854	45.46	48.21	51.97
w/o Ω	82.59	23.49	53.38	32.33	0.661	0.856	36.19	40.42	42.52
1-way	78.82	23.70	54.16	32.14	0.628	0.861	32.79	35.52	37.29
S5	**54.64**	**18.54**	**40.03**	**25.16**	**0.697**	**0.880**	**26.26**	**28.09**	**30.47**

4.6 Inference Efficiency

At annotation scale, the cost of a single pass matters as much as its accuracy.

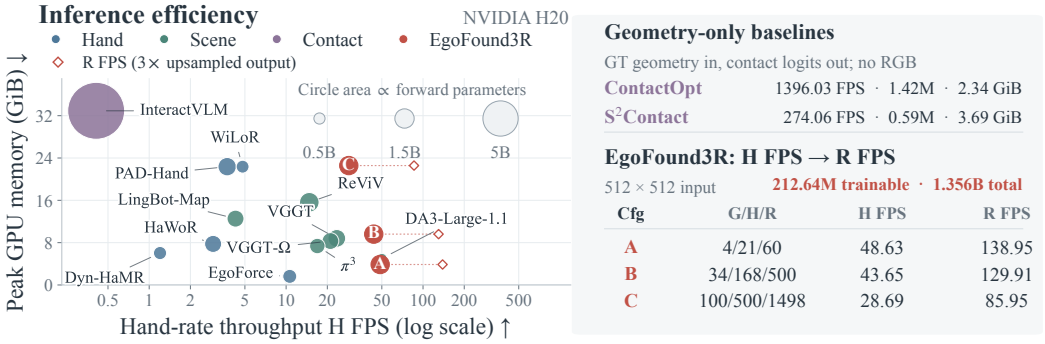


Figure 8: **Inference efficiency.**

We report the hand-rate throughput (H FPS), the hand-observation frames processed per second and therefore the rate single-rate baselines report, together with the dense output rate (R FPS) that 3× temporal upsampling yields. As shown in Fig. 8 and Tab. 6, EgoFound3R processes 28.69 H FPS at 512 × 512, 5.97× the 4.81 FPS of WiLoR and faster than every compared hand pipeline. One forward pass determines throughput, not the sum of chained stages. See Appendix C for details.

5 Conclusion

We presented EgoFound3R, a unified end-to-end model that recovers hands in world space and predicts point-wise visibility, contact, and distance in a single pass. The model combines structured prompts that carry pretrained geometric context into hand decoding, an explicit hand representation that decodes geometry and attributes, and a shared-parameter multi-rate design that lowers inference cost. On OakInk-v2, TACO, and HOI4D, it reduces MPJPE by 43.2%, 22.4%, and 11.6% over the strongest baseline and runs 6× faster than the previous state of the art.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure, namely for discussing and refining the research methodology and experimental design, for developing the conceptual framing of the paper and the hypotheses it tests, for implementing and debugging parts of the method and of the training and evaluation code, for assisting with the translation and rephrasing of our own drafts into English, and for interpreting experimental results. We have not used generative AI tools to generate synthetic datasets or to formulate or prove mathematical claims, and the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure, including drafting and editing the manuscript, brainstorming, proposing its structure and title, retrieving and summarizing related literature, formatting references, and creating or modifying figures. We have reviewed all AI-assisted work: generated code was tested against our evaluation pipeline, and the reported quantities were checked against our evaluation logs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The training configuration (optimizer, schedule, batch size, input resolution, and compute) is listed in Tab. 5, the dataset splits with clip counts in Tab. 7, and the loss terms together with the notation used for them in Appendix B. The evaluation protocol—window sampling, the six alignments, the aggregation of per-point contact and visibility scores, and the ranking rules—is specified in Appendix D, the matched-subset controls in Appendix H, and the scene and camera protocol in Appendix J. The efficiency measurements and the configurations behind Fig. 8 are listed in Tab. 6 and Appendix C, and the architecture and multi-rate ablations, including the definitions of the compared variants, in Appendix I. Video demonstrations of the qualitative results are provided in the supplementary material.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a.

Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2026b.

Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J Black, and Dimitrios Tzionas. Interactvlm: 3d interaction reasoning from 2d foundational models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Rolandos Alexandros Potamias, Taku Komura, Shuo Yang, Zheng Liu, and Bo Zhao. Egograsp: World-space hand-object interaction estimation from egocentric videos. *arXiv preprint arXiv:2601.01050*, 2026.

Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ryosei Hara, Masashi Hatano, Rintaro Yanagi, Atsushi Hashimoto, Takuma Yagi, and Mariko Isogawa. Hand visibility detector: Per-keypoint visibility estimation for hands. *arXiv preprint arXiv:2608.11574*, 2026.

Andela Ilic, Rachel Schuchert, Yijing Jiang, and Christian Holz. Egophi: Estimating contact and force from egocentric vision. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2026.

Elkhan Ismayilzada, Yufei Zhang, and Zijun Cui. Pad-hand: Physics-aware diffusion for hand motion recovery. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a.

Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, et al. Open-aoe: An open egocentric manipulation dataset and toolchain for embodied learning. *arXiv preprint arXiv:2607.14183*, 2026b.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2026.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Xiaozhong Lyu, Gen Li, Zhiyin Qian, Xucong Zhang, Marc Pollefeys, and Siyu Tang. ReViV: Reconstructing the viewer and the view in 4D from monocular egocentric video. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2026.

Christen Millerdurai, Shaoxiang Wang, Yaxu Xie, Vladislav Golyanik, Didier Stricker, and Alain Pagani. Egoforce: Forearm-guided camera-space 3d hand pose from a monocular egocentric camera. In *Proc. of the SIGGRAPH Conference Papers*, 2026.

Senthil Palanisamy, Abhishek Anand, Satpal Singh Rathore, Pratyush Patnaik, and Shubhanshu Khatana. Mobileego anywhere: Open infrastructure for long horizon egocentric data on commodity hardware. *arXiv preprint arXiv:2605.05945*, 2026.

Bhawna Paliwal, Haritheja Etukuru, William Liang, Pieter Abbeel, Nur Muhammad Mahi Shafiullah, and Jitendra Malik. Do as i do: Dexterous manipulation data from everyday human videos. *arXiv preprint arXiv:2606.19333*, 2026.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)*, 36(6), 2017.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S^2Contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022.

Gaotian Wang, Kejia Ren, Andrew Morgan, Yiting Chen, Howard H Qian, Podshara Chanrungmaneekul, and Kaiyu Hang. Egoinfinity: A web-scale 4d hand-object interaction data engine for any-view robot retargeting and video-to-action robot learning. *arXiv preprint arXiv:2606.17385*, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2026c.

Wei Wu, Fangjing Wang, Fan Lu, He Sun, Shi Liu, Yunnan Wang, Yibin Yan, Yong Wang, Shuailei Ma, Xinyang Wang, Yibin Liu, Shuai Yang, Tianxiang Zhou, Kejia Zhang, Lei Zhou, Cheng Su, Nan Xue, Bin Tan, Han Zhang, Youchao Zhang, Fei Liao, Xing Zhu, Yujun Shen, and Kecheng Zheng. From foundation to application: Improving vla models in practice. *arXiv preprint arXiv:2607.06403*, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026a.

Yufei Ye, Jiaman Li, Ryan Rong, and C. Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026b.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Huajian Zeng, Chaohua Yao, Yuantai Zhang, Jiaqi Yang, Rolandos Alexandros Potamias, and Xingxing Zuo. Stablehand: Quality-aware flow matching for world-space dual-hand motion estimation from egocentric video. *arXiv preprint arXiv:2605.18553*, 2026.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

APPENDIX CONTENTS

A Detailed Architecture and Information Flow 14
B Training Objectives and Supervision 15
C Inference Efficiency and Temporal Rates 17
D Evaluation Metrics and Ranking Rules 18
E Additional Hand Geometry Evaluation 19
F Temporal Evaluation 19
G Contact and Visibility Evaluation 20
H Matched-Subset Controls for 100-Window Baselines 23
I Architecture and Multi-Rate Ablations 25
J Scene and Camera Evaluation 28
K Camera-Space Qualitative Results 29
L World-Space Qualitative Results 33

A DETAILED ARCHITECTURE AND INFORMATION FLOW

Fig. 9 expands the design-level description of Sec. 3 with the token layout, the localization and decoding branches, and the auxiliary heads. The paragraphs below follow the order of the main text and record the details that it leaves out: what the prompts carry, how the two rates exchange information, and what the decoding heads read. They are the choices that turn the design-level description of the main text into a concrete model. We keep the notation of Sec. 3: H denotes the hand-observation times, $G \subseteq H$ the anchors, s the stride, and R the requested output times.

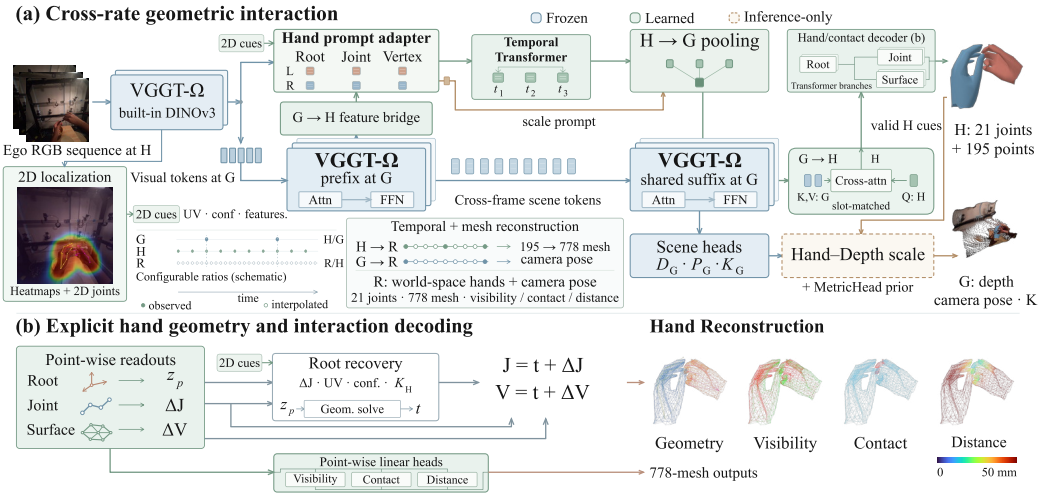


Figure 9: **Detailed technical framework.** Full token layout, geometric interaction, explicit decoding, and auxiliary branches.

Hand prompts and localization. The frozen VGGT-Ω backbone (Wang et al., 2026b) is built on a DINOv3 encoder (Siméoni et al., 2026), which embeds every hand-rate frame, and its aggregator converts the anchor frames into cross-frame scene tokens. An attention bridge copies that context into the hand features, which are organized as root, joint, and surface queries. Two small heads localize the hands inside the model: one predicts joint heatmaps together with in-frame offsets, and one predicts whether each hand is in view. Their outputs are detached from the gradient and reused

as queries. A bidirectional temporal Transformer links the two hands across the clip, and a learnable scale prompt carries the metric evidence that the decoding heads consume.

Cross-rate interaction. A stride $s \in \{1, \ldots, 5\}$ selects the anchor times $G \subseteq H$. Each anchor gathers the valid hand prompts of its stride block and pools them with learned attention weights that also observe the position of a prompt inside the block, and the pooled summary is added to the anchor through the gated residual update of Eq. (1). Because the residual is zero-initialized and single-element blocks return the anchor unchanged, the dense configuration is recovered exactly. The anchors re-enter the frozen aggregator, where only the adapters are trained, and a slot-matched cross-attention step with the hand prompts as queries and the aggregated global tokens as keys and values returns the context to the hand rate, as shown in Fig. 9(a); the same layout carries information back from the hands to the anchors, so H→G and G→H exchanges share one interface. A single phase embedding is shared across strides and read at continuous, cyclically wrapped positions, which is why the ratio changes aggregation density rather than parameter count.

Explicit decoding and attribute heads. As shown in Fig. 9(b), root, joint, and surface queries read the same point features. Joint and surface queries regress offsets relative to the root, ΔJ_j and ΔV_v in Eq. (2), the surface branch inherits features from the joints it is attached to and mixes over the incident faces of the MANO mesh (Romero et al., 2017), and the root query additionally predicts a depth prior z_p. The root translation is then recovered by a geometry solve rather than regressed directly: a wrist-anchored, confidence-weighted robust fit reconciles the observed 2D cues, the predicted relative joints, the depth prior, and the hand intrinsics K_H. Three point-wise linear heads decode visibility, contact, and distance from the same features, with a sigmoid for the two classification targets and a non-negative decode for distance, and contact merges object contact and opposite-hand contact without identifying the partner.

Scene heads and the metric head. The scene heads predict anchor depth D_G, camera poses P_G, and intrinsics K_G. The metric head is a clip-level decoder: it reads scene summaries, detached hand features, in-view probabilities, and the scene scale prompt, and predicts the scene scale together with a raw-to-meter prior $\lambda_0 = \hat{\rho}/\rho_{\text{raw}}$, where ρ_{raw} is computed from the valid raw scene points and $\hat{\rho}$ denotes the predicted scale. Because the hand and scene heads read different grids, K_H is estimated as a clip-level consensus over per-frame estimates with interpolated local residuals and is detached when the root is solved, whereas scene back-projection keeps the native anchor intrinsics K_G. At inference, visible hand vertices are projected into the raw depth maps to check the hand–depth agreement, and the resulting factor rescales scene depth and camera translation only, leaving rotations, intrinsics, and the camera-space hands untouched. Because world-space hand positions are composed with the camera trajectory, they follow the rescaled scene rather than being predicted again.

Requested output times. The network runs at the hand rate; a parameter-free stage produces the requested times R by interpolating cameras and hands separately and by densifying the native 195 vertices to the 778-vertex MANO mesh with the fixed level-1 upsampling of Mesh Graphormer (Lin et al., 2021), which reproduces the measured points exactly. Overlapping clips are stitched by rescaling them to a common scale from the raw depths at shared frames. Both operations read future frames and therefore stay outside the feed-forward network, so the deployment remains offline.

B Training Objectives and Supervision

We summarize the representative loss terms used to train the unified model. The training pipeline draws from the seven egocentric datasets listed in the main text and combines their available masked supervision over 2D observations, hand and scene geometry, cameras, scale, point attributes, and temporal motion. Contact and distance targets follow from point-to-triangle distances between the annotated hand points and the annotated object or opposite-hand surfaces. For each term, only finite targets covered by its supervision mask contribute to the loss; valid task losses are aggregated into the multi-task objective. We write p for a hand query point (a joint or a hand-surface point) and $\mathcal{Q}$ for the valid points of a term, i.e. the points whose target is finite and whose supervision mask is valid. A hat denotes a predicted quantity, the confidence weight α is used by the geometry and contact terms, and each term below carries its own balancing weight.

Training recipe. During training, we sample the supported temporal strides in fixed proportions.

Table 5: **Training configuration.** Only the hand prompts, adapters, and prediction heads are optimized; the VGGT-Ω backbone stays frozen.

Setting	Value
Optimizer	AdamW ($\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$), weight decay 0.01
Learning rate	2.5×10^{-4}
Schedule	linear warmup over 100 steps, then cosine decay to 10^{-5}
Batch size	8 per device on 8× NVIDIA H20 (effective 64)
Resolution	256×256
Training steps	3,200
Gradient clipping	1.0
Precision	bfloat16 with DeepSpeed ZeRO-2

Explicit hand geometry. For a query point $p \in \mathcal{Q}$, the model predicts a root-relative offset $\hat{\delta}_p$ and a log-confidence s_p, and δ_p is the annotated offset of the same point. Offsets are normalized by the average valid hand-point scale a of the instance, with $\hat{a}$ its predicted counterpart, and α weights the confidence term, giving the confidence-weighted robust regression loss

$$\mathcal{L}_{\text{geom}} = \frac{1}{|\mathcal{Q}|} \sum_{p \in \mathcal{Q}} \left[\exp(s_p) \left\| \frac{\hat{\delta}_p}{\hat{a}} - \frac{\delta_p}{a} \right\|_1 - \alpha s_p \right]. \tag{3}$$

The same form is applied to joints and the native surface points, while the root is recovered from the metric scene and image observations.

Visibility and contact. Visibility is supervised with a masked binary cross-entropy, where $\hat{v}_p$ is the predicted visibility logit and $v_p \in \{0, 1\}$ the annotated visibility of point p:

$$\mathcal{L}_{\text{vis}} = \frac{1}{|\mathcal{Q}|} \sum_{p \in \mathcal{Q}} \text{BCEWithLogits}(\hat{v}_p, v_p). \tag{4}$$

For contact, the classification error is coupled with a learned point confidence:

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{Q}|} \sum_{p \in \mathcal{Q}} \left[\exp(c_p)\ \text{BCEWithLogits}(\hat{z}_p, y_p) - \alpha c_p \right], \tag{5}$$

where $\hat{z}_p$ is the predicted contact logit, $y_p \in \{0, 1\}$ the annotated contact of the same point, c_p the predicted log-confidence, and α the confidence weight introduced above. These losses are evaluated at both joints and surface points when the corresponding labels are available.

Distance. The distance head predicts a log-distance $\hat{\ell}$. To preserve resolution near contact while limiting the influence of distant points, we regress

$$\mathcal{L}_{\text{dist}} = \frac{1}{|\mathcal{Q}|} \sum_{p \in \mathcal{Q}} \text{SmoothL1}\Big(\hat{\ell}_p, \log(1 + 1000 d_p) \Big), \tag{6}$$

where d_p is the annotated point-to-surface distance of point p in meters and the factor 1000 expresses the target in millimeters.

2D cues and scene supervision. The full-image localization branch uses a weighted heatmap BCE $\mathcal{L}_{\text{heatmap}}$ and a Smooth-L_1 offset loss $\mathcal{L}_{\text{offset}}$; the predicted 3D points $\hat{X}$ are additionally constrained by the observed 2D coordinates u through the projection Π, and the three terms are weighted by λ_h, λ_o, and λ_r:

$$\mathcal{L}_{\text{2D}} = \lambda_h \mathcal{L}_{\text{heatmap}} + \lambda_o \mathcal{L}_{\text{offset}} + \lambda_r \left\| \Pi(\hat{X}) - u \right\|_1. \tag{7}$$

The scene and camera branches provide masked depth, camera-intrinsics, and camera-motion supervision: the depth term uses a confidence-weighted robust error, camera translation is compared after scene-scale normalization, and camera rotation uses a geodesic rotation loss. We write $\mathcal{L}_{\text{3R}}$ for this scene-and-camera objective; the geometry term keeps unit weight, so the full training loss is

$$\mathcal{L} = \mathcal{L}_{\text{geom}} + \lambda_{\text{vis}} \mathcal{L}_{\text{vis}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}} + \lambda_{\text{2D}} \mathcal{L}_{\text{2D}} + \lambda_{\text{3R}} \mathcal{L}_{\text{3R}}. \tag{8}$$

C Inference Efficiency and Temporal Rates

Table 6: **Inference efficiency on one NVIDIA H20.** The $G/H/R$ column lists the anchor, hand, and output rates; H FPS and R FPS are in frames per second, VRAM is in GiB, and parameter counts use B/M; timings follow the released configuration of each system.

Method	$G/H/R$	H FPS↑	R FPS↑	VRAM (GiB)↓	Total params. (B/M)
EgoFound3R configurations					
EgoFound3R-A (512^2)	4/21/60	48.63	138.95	3.87	1.356B
EgoFound3R-B (512^2)	34/168/500	43.65	129.91	9.61	1.356B
EgoFound3R-C (512^2)	100/500/1498	28.69	85.95	22.57	1.356B
Contact methods: GT hand + object mesh + object 6DoF					
ContactOpt	–	1396.03	–	2.34	1.42M
S^2Contact	–	274.06	–	3.69	0.59M
Multimodal contact method					
InteractVLM	–	0.41	–	32.92	13.05B
Camera-space hand methods					
PAD-Hand	–	3.71	–	22.36	1.36B
WiLoR	–	4.81	–	22.36	0.67B
EgoForce	–	10.63	–	1.63	0.79B
World-space hand methods					
Dyn-HaMR	–	1.20	–	6.03	0.70B
HaWoR	–	2.93	–	7.76	1.17B
ReViV	–	14.77	–	15.63	1.54B
Scene methods					
DA3-Large-1.1	–	50.15	–	5.02	0.41B
VGGT	–	23.46	–	8.81	1.26B
VGGT-Ω	–	21.05	–	8.30	1.14B
π^3	–	16.90	–	7.36	0.96B
LingBot-Map	–	4.27	–	12.54	1.16B

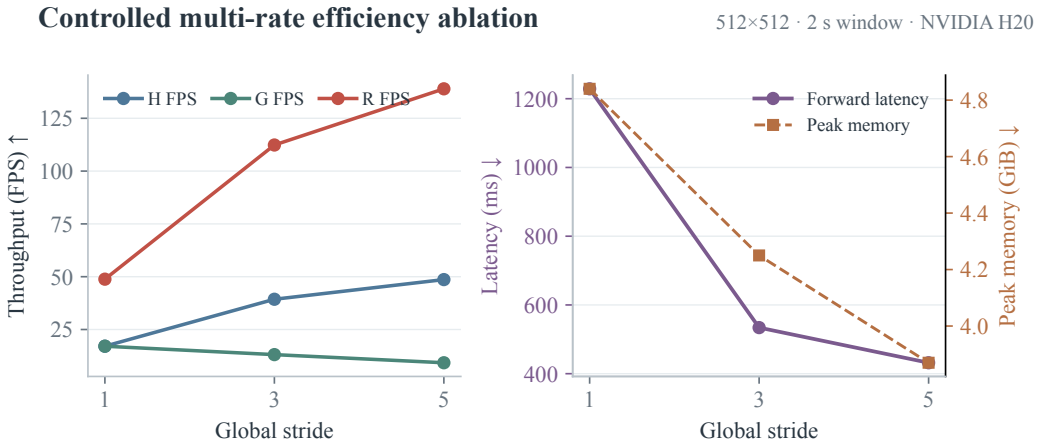


Figure 10: **Hand-rate efficiency across global strides.** Only the global stride $s \in \{1, 3, 5\}$ changes, with parameters, 512×512 input, and 2 s windows held fixed, so the panels isolate its effect on hand-rate throughput (H FPS), the anchor and output rates, forward latency, and peak memory.

Tab. 6 lists the configurations underlying the main-text efficiency plot; its $G/H/R$ column reports the anchor, hand, and output rates of each configuration. Fig. 10 ablates the global stride at fixed parameters, resolution, and window length, reporting hand-rate throughput together with latency and peak memory; reconstruction accuracy across strides is reported in Appendix I. Timings use one

NVIDIA H20 with 500-frame measurements and follow the original measurement reports; PAD-Hand includes its WiLoR prerequisite. ContactOpt and S^2Contact consume ground-truth hand geometry, object meshes, and object 6DoF poses, so they are reported separately from the throughput comparison.

D Evaluation Metrics and Ranking Rules

This section expands the metric summary given in the experimental setup: we state how each geometry variant aligns prediction and ground truth before measuring the error, and how contact and visibility scores are computed. Let $\hat{X}$ and X denote predicted and ground-truth points of a window, respectively, and let N be the number of valid points. Geometry metrics are reported in millimeters for joints (mean per joint position error (MPJPE)) and vertices (mean per vertex position error (MPVPE)), and each variant fixes the alignment applied before the error is measured.

Evaluation protocol. The common evaluation uses 2,378 randomly drawn windows, each a 2 s (60-frame) clip whose frames are kept at the native frame rate of its dataset. Windows are sampled at a fixed ratio per dataset from six datasets (H2O, HOT3D Aria, ARCTIC, OakInk-v2, TACO, and HOI4D) and are kept disjoint from the training split; clips are non-overlapping within a sequence, and the sampling ratio bounds evaluation cost. To bound evaluation cost, Dyn-HaMR and InteractVLM are each evaluated on a separate, randomly sampled subset of 100 windows and paired with an EgoFound3R control computed on exactly the same windows; these pairs are ranked only within the pair, as reported in Appendix H.

Table 7: **Dataset splits.** Train counts source sequences in the training partition; eval counts the 2 s windows sampled from the test partition under the common protocol. STERA-10M contributes training data only.

Dataset	Train sequences	Eval clips
OakInk-v2	565	400
TACO	839	400
HOI4D	1,431	461
H2O	138	283
HOT3D Aria	136	400
ARCTIC	226	434
STERA-10M	all	–

Raw and root-relative error. Raw is the unaligned error, computed after expressing both prediction and ground truth in the coordinate system of the window without any similarity alignment. RR (root-relative) subtracts the root translation, here the wrist, from both prediction and ground truth before evaluation, so that only the relative articulation of joints or vertices is measured.

Aligned error. PA (Procrustes-aligned) computes one similarity transform per frame by Procrustes analysis and reports the residual after alignment, which removes global rotation, translation, and scale from each frame independently. $\mathrm{Sim}(3)$ instead estimates a single similarity transform per window and applies it to all frames of that window, which preserves within-window motion while removing window-level scale and rotation.

World-coordinate error. W (world-coordinate) evaluates hand points in the world frame, after the predictions are mapped there with the camera poses: a method that predicts its own trajectory uses its predicted cameras, whereas a camera-space method without one falls back to the ground-truth extrinsics and is marked † in the tables. Both metrics then fit one Sim(3) from the predicted to the ground-truth hand points: W on the first two valid frames, applying that transform to the whole window, and WA on all valid hand points of the window, which removes global drift of the reconstructed hand trajectory while retaining local motion. For W and WA, methods marked with † in the tables use ground-truth extrinsics and are ranked separately from methods that use predicted extrinsics; ‡ marks our model when it receives ground-truth camera intrinsics while still predicting extrinsics; superscripts identify the 100-window subsets, and local metrics are ranked jointly.

Contact and visibility. Contact is evaluated per hand point, that is, per joint or per vertex, before any aggregation over points. Precision is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, recall is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, and F1 is their harmonic mean $2PR/(P+R)$; all three are unitless scores in $[0,1]$ and are computed with the same labels for joints, native vertices, and interpolated vertices. Scores are computed per window and then averaged over windows, so the reported F1 is the mean of the per-window F1 values rather than the harmonic mean of the reported precision and recall. The threshold sweep of Tab. 13 is the one exception: its AP and best-F1 pool the points of a dataset, while its F1@0.5 column repeats the window-averaged value. Visibility uses the same three definitions on the binary per-point visibility label. Distance is the point-to-surface distance from a hand point to the annotated object or opposite-hand surface, and the tables report its mean absolute error (MAE, mm) over all valid query points (MAE(All)), over points with predicted contact (MAE(Pred)), and over points with ground-truth contact (MAE(GT)). The best and second-best values at the displayed precision are marked in bold and with an underline, shaded light blue and amber, and ties share the same marking. A dash denotes an undefined metric. Rows evaluated on the 100-window subsets are ranked only within their matched groups, as described in Appendix H. The tables use these symbols as column headers: Raw, RR, PA, $\mathrm{Sim}(3)$, W, and WA for joints and vertices, and P, R, and F1 for contact and visibility.

E Additional Hand Geometry Evaluation

Tab. 8 reports hand geometry at 21 joints (MPJPE) together with native 195-vertex and interpolated 778-vertex surfaces (MPVPE) on OakInk-v2, TACO, and HOI4D, complementing the main-text joint comparison. Across resolutions the ranking follows the joint comparison: EgoFound3R leads the unaligned, root-relative, and world metrics. The 195 points are the fixed level-1 MANO sampling released with Mesh Graphormer (Lin et al., 2021), a subset of the 778 MANO vertices selected by a one-hot matrix; our 778-vertex results are recovered from them with the matching upsampling matrix from the same release.

Table 8: **Hand geometry** on OakInk-v2, TACO, and HOI4D in mm at 21 joints and 195/778 vertices; marks follow Table 1.

	21 joints: MPJPE						195 vertices: MPVPE						778 vertices: MPVPE					
Method	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓
OakInk-v2																		
WiLoR†	91.24	27.66	**7.78**	**16.89**	42.14	**17.04**	91.16	26.00	**7.55**	**16.90**	41.84	**17.06**	91.46	26.78	**7.44**	**16.82**	42.02	**16.97**
PAD-Hand†	107.16	32.23	9.63	21.37	49.01	21.62	106.47	30.03	9.32	21.15	47.23	21.40	106.98	30.79	9.12	21.02	47.41	21.26
EgoForce†	123.57	47.75	11.09	32.62	80.59	33.10	123.63	45.07	10.95	33.19	81.00	33.65	123.95	45.82	10.64	32.93	81.19	33.40
EgoFound3R†	–	–	–	–	**27.49**	17.89	–	–	–	–	**27.79**	18.07	–	–	–	–	**28.53**	18.76
Dyn-HaMR^{100w}	96.39	17.91	7.84	13.76	36.16	17.19	95.78	16.87	7.67	13.73	35.84	17.20	95.94	17.22	7.56	13.62	35.90	17.06
HaWoR	174.73	45.85	9.49	32.19	100.46	37.47	172.35	42.56	9.54	32.32	100.62	37.78	173.68	44.08	9.24	32.19	100.83	37.59
EgoFound3R	51.81	**17.77**	8.02	17.87	**30.88**	**18.77**	51.40	**17.11**	7.96	18.05	**31.32**	**18.98**	51.64	**18.37**	8.98	18.75	**31.99**	**19.63**
EgoFound3R‡	**39.45**	17.84	8.04	18.01	31.56	18.94	**39.28**	17.18	7.98	18.21	32.03	19.17	**39.53**	18.44	9.01	18.88	32.58	19.80
TACO																		
WiLoR†	70.40	28.33	9.49	25.95	51.14	26.14	69.91	26.94	9.87	26.11	50.71	26.29	70.13	27.30	**9.72**	26.02	50.52	26.20
PAD-Hand†	72.97	30.30	11.19	29.21	57.06	29.60	72.45	29.03	10.84	28.74	54.03	29.15	72.73	29.32	10.64	28.64	54.50	29.06
EgoForce†	126.63	44.53	11.93	47.84	105.52	46.65	126.58	41.99	12.18	48.41	106.20	47.28	126.83	42.56	11.89	48.06	106.11	46.92
EgoFound3R†	–	–	–	–	**36.68**	**24.06**	–	–	–	–	**37.47**	**24.45**	–	–	–	–	**38.18**	**25.37**
Dyn-HaMR^{100w}	127.12	27.47	11.62	29.27	77.88	31.46	127.79	26.04	11.55	29.49	78.23	31.82	127.94	26.33	11.46	29.45	78.09	31.73
HaWoR	74.28	24.19	10.55	27.71	86.38	35.34	73.75	23.02	10.99	27.91	86.78	35.81	74.05	23.26	10.71	27.74	86.63	35.53
EgoFound3R	54.64	18.54	**8.13**	**24.15**	**40.03**	**25.16**	53.76	**18.36**	**8.67**	**24.49**	**40.69**	**25.55**	54.03	**19.91**	10.19	**25.39**	**41.36**	**26.42**
EgoFound3R‡	**50.44**	**18.52**	**8.13**	24.63	41.78	25.75	**50.61**	**18.36**	**8.67**	25.03	42.68	26.20	**51.54**	19.99	10.31	25.98	43.22	27.11
HOI4D																		
WiLoR†	81.51	27.75	9.62	24.14	**49.41**	24.57	81.34	26.40	9.61	24.25	**48.78**	24.64	82.08	26.69	**9.51**	**24.14**	**48.80**	**24.56**
PAD-Hand†	75.35	29.77	10.12	24.79	51.92	25.04	74.91	28.68	9.83	24.65	50.73	24.93	75.80	28.92	9.71	24.54	50.68	24.83
EgoForce†	147.77	52.78	12.92	49.22	142.58	55.76	147.13	50.29	12.93	50.26	143.30	56.49	147.64	50.75	12.56	49.81	142.96	56.10
EgoFound3R†	–	–	–	–	63.01	**23.30**	–	–	–	–	63.02	**23.56**	–	–	–	–	63.62	24.72
Dyn-HaMR^{100w}	245.06	26.49	11.89	30.47	173.62	65.71	245.09	25.53	12.28	30.99	173.69	66.73	245.58	25.77	12.01	30.74	173.51	66.03
HaWoR	53.46	28.47	10.49	27.85	171.09	55.96	53.11	26.53	10.38	27.73	170.80	56.69	53.30	27.33	10.19	27.67	170.58	56.31
EgoFound3R	47.28	**16.63**	**7.99**	**23.29**	159.70	49.84	48.46	**16.20**	**8.29**	**23.60**	160.03	50.68	49.72	**18.06**	10.16	24.75	160.98	50.85
EgoFound3R‡	**43.43**	16.67	8.00	23.39	**159.18**	**49.77**	**44.20**	16.24	8.31	23.78	**159.74**	**50.65**	**45.50**	18.09	10.17	24.91	**160.62**	**50.80**

F Temporal Evaluation

Velocity (MPJVE/MPMVE/MPVVE, in mm/s) and acceleration (MPJAE/MPMAE/MPVAE, in m/s^2) errors summarize how faithfully the recovered trajectories follow the ground-truth motion

at 21 joints, 195 vertices, and the full 778-vertex MANO mesh. Tab. 9 compares EgoFound3R with the hand baselines on the three evaluated subsets, and Tab. 10 repeats the comparison on the matched 100-window subsets shared with Dyn-HaMR. ReViV predicts joints only, so its 195- and 778-vertex entries are unavailable, and our temporal values come from the default-inference rerun of the stride-5 configuration.

Table 9: **Temporal consistency** on OakInk-v2, TACO, and HOI4D. Velocity (MPJVE/MPMVE/MPVVE) and acceleration (MPJAE/MPMAE/MPVAE) errors at 21 joints, 195 vertices, and 778 vertices are in mm/s and m/s^2; marks follow Table 1.

	21 joints		195 vertices		778 vertices	
Method	MPJVE↓	MPJAE↓	MPMVE↓	MPMAE↓	MPVVE↓	MPVAE↓
OakInk-v2						
WiLoR†	116.81	4.66	110.07	4.38	112.47	4.49
PAD-Hand†	124.78	4.29	118.80	4.08	120.99	4.16
EgoForce†	258.44	12.11	242.73	11.35	247.57	11.58
HaWoR	59.59	2.07	56.20	1.96	57.23	1.99
ReViV	165.72	6.43	–	–	–	–
EgoFound3R	**56.55**	**1.17**	**54.09**	**1.12**	**54.29**	**1.12**
TACO						
WiLoR†	138.28	5.44	130.00	5.11	132.79	5.23
PAD-Hand†	136.35	4.56	129.16	4.33	131.18	4.41
EgoForce†	453.68	21.99	424.57	20.56	433.80	21.02
HaWoR	93.13	2.67	87.76	2.52	89.10	2.56
ReViV	125.87	3.02	–	–	–	–
EgoFound3R	**92.00**	**1.87**	**87.10**	**1.76**	**87.53**	**1.77**
HOI4D						
WiLoR†	158.49	5.99	148.44	5.61	151.50	5.73
PAD-Hand†	145.21	4.97	137.02	4.70	139.47	4.79
EgoForce†	544.59	26.48	510.65	24.80	520.77	25.31
HaWoR	**140.11**	4.73	**131.19**	4.42	**133.78**	4.52
ReViV	448.65	18.25	–	–	–	–
EgoFound3R	152.69	**4.34**	144.58	**4.08**	147.84	**4.20**

Table 10: **Matched 100-window temporal consistency.** Dyn-HaMR and EgoFound3R are evaluated on the same windows; errors are in mm/s and m/s^2 and marks follow Table 1.

	21 joints		195 vertices		778 vertices	
Method	MPJVE↓	MPJAE↓	MPMVE↓	MPMAE↓	MPVVE↓	MPVAE↓
OakInk-v2						
Dyn-HaMR^{100w}	132.68	6.04	130.90	5.95	131.84	6.00
EgoFound3R^{100w,Dyn}	**63.85**	**1.32**	**61.40**	**1.27**	**61.69**	**1.27**
TACO						
Dyn-HaMR^{100w}	339.05	13.55	334.51	13.36	336.74	13.47
EgoFound3R^{100w,Dyn}	**99.81**	**2.00**	**93.68**	**1.88**	**94.24**	**1.89**
HOI4D						
Dyn-HaMR^{100w}	439.65	18.90	436.66	18.79	438.56	18.87
EgoFound3R^{100w,Dyn}	**204.38**	**6.78**	**191.63**	**6.29**	**196.89**	**6.48**

G Contact and Visibility Evaluation

Contact at three point sets. Tab. 11 reports contact classification at 21-joint, 195-vertex, and 778-vertex resolutions. InteractVLM uses its original 100-window subsets; the corresponding matched EgoFound3R controls are reported separately in Appendix H. Our ground truth marks a point as in contact when its distance to the object or the opposite-hand surface is below 18 mm at joints and 14 mm at the 195 vertices of the hand surface, the fingertips included. The three compared systems define contact differently, so none of them shares this criterion: ContactOpt and S^{2}Contact treat a point as covered at a 2 mm surface tolerance, which is a narrower band than ours, while

InteractVLM inherits annotated vertex labels. The two geometric baselines are additionally given ground-truth hand and object geometry.

Table 11: **Contact prediction** on OakInk-v2, TACO, and HOI4D at 21 joints, 195 vertices, and 778 vertices. Superscripts mark 100-window subsets, and marks follow Table 1.

	21 joints			195 vertices			778 vertices		
Method	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
OakInk-v2									
InteractVLM^{100w}	0.180	0.403	0.202	0.173	0.408	0.205	0.168	0.412	0.201
S^{2}Contact	0.327	0.081	0.079	0.401	0.180	0.158	0.397	0.198	0.173
ContactOpt	0.334	0.196	0.154	0.334	0.234	0.178	0.338	0.248	0.185
EgoFound3R	**0.510**	**0.665**	**0.493**	**0.525**	**0.611**	**0.479**	**0.516**	**0.329**	**0.320**
TACO									
InteractVLM^{100w}	0.388	0.705	0.472	0.348	0.743	0.451	0.356	0.750	0.458
S^{2}Contact	0.519	0.104	0.146	0.590	0.210	0.285	**0.592**	0.234	0.308
ContactOpt	0.557	0.197	0.259	0.571	0.249	0.311	0.573	0.241	0.303
EgoFound3R	**0.654**	**0.777**	**0.697**	**0.636**	**0.719**	**0.661**	0.516	**0.265**	**0.320**
HOI4D									
InteractVLM^{100w}	0.313	0.537	0.281	0.256	0.556	0.261	0.249	0.553	0.261
S^{2}Contact	0.444	0.265	0.284	0.437	0.340	0.333	0.426	0.344	0.330
ContactOpt	0.518	0.477	0.435	0.506	0.548	0.463	0.496	**0.538**	**0.449**
EgoFound3R	**0.612**	**0.835**	**0.675**	**0.608**	**0.807**	**0.656**	**0.531**	0.466	0.445

Vertex-level visibility. Tab. 12 reports visibility at the native 195 vertices and the interpolated 778-vertex mesh, which the compared pipelines do not provide; HVD predicts joint visibility only, so no ranking is shown.

Table 12: **Vertex-level visibility** at the native 195 vertices and the interpolated 778-vertex mesh.

Dataset	Method	195 vertices			778 vertices		
		P↑	R↑	F1↑	P↑	R↑	F1↑
OakInk-v2	EgoFound3R	0.556	0.705	0.595	0.555	0.690	0.588
TACO	EgoFound3R	0.437	0.818	0.567	0.439	0.805	0.565
HOI4D	EgoFound3R	0.572	0.789	0.660	0.581	0.781	0.663

Threshold and definition sensitivity. Tab. 11 reports one operating point under one ground-truth definition. Because the compared systems were trained under different contact conventions, we isolate the two effects that can be separated without retraining: the decision threshold applied to each method's scores, and the distance tolerance that defines the ground truth. No model or prediction is modified, and every number below is post-processing of the same cached per-point scores and distances.

Sweep and reading. Tab. 13 sweeps each method's scores from 0.01 to 0.99. AP summarizes ranking quality independently of the operating point, and best-F1 reports the best attainable point together with the threshold that attains it. Under this sweep the proposed model attains the highest AP and best-F1 on the three evaluation datasets, and its best-F1 thresholds (0.39–0.51) lie close to the conventional 0.5 decision point, whereas the two geometric baselines peak at thresholds as low as 0.09.

Agreement on the predicted positives. Tab. 14 narrows the ground truth to a 2 mm surface band, which matches the coverage criterion of the two geometric baselines. That band is computed on the 195 vertices. The ordering is preserved on the three evaluation datasets. As shown in Fig. 11, the positives predicted by the two geometric baselines are not confined to the narrow band that defines their coverage: their distances spread over tens of millimeters, with a median of 10–21 mm across the three evaluation datasets and a quarter to nearly half of the points beyond 25 mm. The proposed model keeps a tighter distribution, with a median of 10–12 mm and 57–65% of its positives inside the 14 mm contact band, which is consistent with the distance regression used to supervise its attribute head.

Table 13: **Decision-threshold sweep for contact prediction** on the 195 vertices. AP and best-F1 pool the points of a dataset over 99 thresholds, with the best-F1 threshold in parentheses; marks follow Table 1.

Dataset	Method	AP↑	best-F1↑	F1@0.5↑
OakInk-v2	ContactOpt	0.308	0.411 (0.21)	0.178
	S^2Contact	0.395	0.449 (0.09)	0.158
	EgoFound3R	**0.603**	**0.617** (0.39)	**0.479**
TACO	ContactOpt	0.475	0.542 (0.15)	0.311
	S^2Contact	0.512	0.558 (0.09)	0.285
	EgoFound3R	**0.714**	**0.691** (0.43)	**0.661**
HOI4D	ContactOpt	0.482	0.589 (0.35)	0.463
	S^2Contact	0.399	0.515 (0.26)	0.333
	EgoFound3R	**0.735**	**0.714** (0.51)	**0.656**
100-window paired subsets				
OakInk-v2	InteractVLM^{100w}	0.184	0.275 (0.71)	0.205
	EgoFound3R^{100w}	**0.559**	**0.594** (0.42)	**0.442**
TACO	InteractVLM^{100w}	0.397	0.494 (0.90)	0.451
	EgoFound3R^{100w}	**0.725**	**0.712** (0.42)	**0.691**
HOI4D	InteractVLM^{100w}	0.249	0.371 (0.82)	0.261
	EgoFound3R^{100w}	**0.700**	**0.693** (0.48)	**0.670**

Table 14: **Narrow-band contact comparison** on the 195 vertices at the 0.5 operating point. The two column groups differ only in the labeling tolerance; the 2 mm band is defined on the 195 vertices only.

		our GT			2 mm band		
Dataset	Method	P↑	R↑	F1↑	P↑	R↑	F1↑
OakInk-v2	ContactOpt	0.334	0.234	0.178	0.080	0.335	0.096
	S^2Contact	0.401	0.180	0.158	**0.117**	0.331	0.115
	EgoFound3R	**0.525**	**0.611**	**0.479**	0.103	**0.741**	**0.164**
TACO	ContactOpt	0.571	0.249	0.311	0.101	0.358	0.140
	S^2Contact	0.590	0.210	0.285	**0.127**	0.376	0.171
	EgoFound3R	**0.636**	**0.719**	**0.661**	0.108	**0.836**	**0.188**
HOI4D	ContactOpt	0.506	0.548	0.463	0.100	0.657	0.157
	S^2Contact	0.437	0.340	0.333	0.086	0.390	0.125
	EgoFound3R	**0.608**	**0.807**	**0.656**	**0.101**	**0.865**	**0.174**
100-window paired subsets							
OakInk-v2	InteractVLM^{100w}	0.173	0.408	0.205	0.031	0.426	0.055
	EgoFound3R^{100w}	**0.441**	**0.638**	**0.442**	**0.090**	**0.774**	**0.151**
TACO	InteractVLM^{100w}	0.348	0.743	0.451	0.052	0.758	0.094
	EgoFound3R^{100w}	**0.654**	**0.755**	**0.691**	**0.111**	**0.877**	**0.196**
HOI4D	InteractVLM^{100w}	0.256	0.556	0.261	0.035	0.589	0.055
	EgoFound3R^{100w}	**0.631**	**0.794**	**0.670**	**0.095**	**0.874**	**0.168**

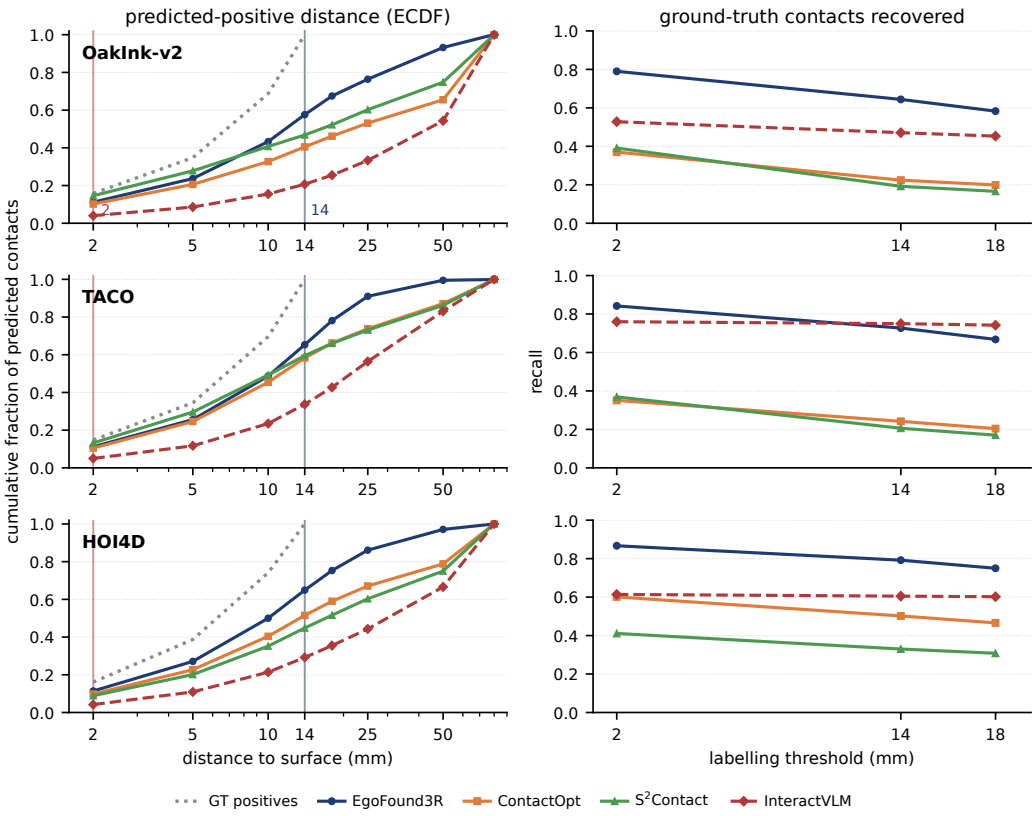


Figure 11: **Where each method's predicted contacts lie.** Predicted-positive distances on the 195 vertices, one row per dataset. Left: cumulative fraction of a method's positives within a given distance to the surface, ground truth in gray, with vertical lines marking the 2 mm coverage band of the two geometric baselines and our 14 mm contact band. Right: recall of ground-truth contacts as the labelling threshold varies, with reference lines at 2, 14, and 18 mm.

H Matched-Subset Controls for 100-Window Baselines

Tab. 15 and Tab. 16 use the original Dyn-HaMR hand subsets and InteractVLM contact subsets, respectively. Each baseline and its EgoFound3R controls are evaluated on the same windows. Results are ranked within each matched group, separately from the common-window evaluation.

Table 15: **Matched 100-window hand geometry.** Dyn-HaMR and EgoFound3R are evaluated on the same windows; errors are in mm.

	21 joints: MPJPE						195 vertices: MPVPE						778 vertices: MPVPE					
Method	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓
OakInk-v2																		
Dyn-HaMR^{100w}	96.39	**17.91**	**7.84**	**13.76**	36.16	**17.19**	95.78	**16.87**	**7.67**	**13.73**	35.84	**17.20**	95.94	**17.22**	**7.56**	**13.62**	35.90	**17.06**
EgoFound3R^{100w,Dyn}	50.88	21.17	8.50	20.30	**30.20**	20.88	49.07	20.97	8.54	20.49	**30.21**	21.02	49.17	22.25	9.53	21.11	**30.99**	21.63
EgoFound3R^{‡,100w,Dyn}	**43.29**	21.18	8.51	21.15	32.68	21.88	**42.73**	20.97	8.54	21.33	32.71	22.06	**42.94**	22.28	9.55	21.93	33.40	22.63
TACO																		
Dyn-HaMR^{100w}	127.12	27.47	11.62	29.27	77.88	31.46	127.79	26.04	11.55	29.49	78.23	31.82	127.94	26.33	11.46	29.45	78.09	31.73
EgoFound3R^{100w,Dyn}	**56.15**	**20.80**	**8.74**	**24.53**	**40.70**	**25.49**	**55.18**	**20.37**	9.02	**24.70**	**41.43**	**25.82**	**55.59**	**21.96**	**10.47**	**25.58**	**42.09**	**26.62**
EgoFound3R^{‡,100w,Dyn}	59.47	**20.80**	**8.74**	24.90	42.10	25.93	59.76	**20.37**	**9.01**	25.13	43.20	26.34	60.91	22.04	10.58	26.06	43.64	27.19
HOI4D																		
Dyn-HaMR^{100w}	245.06	26.49	11.89	30.47	**173.62**	65.71	245.09	25.53	12.28	30.99	**173.69**	66.73	245.58	25.77	12.01	30.74	**173.51**	**66.03**
EgoFound3R^{100w,Dyn}	50.92	**17.73**	**8.71**	26.25	205.48	65.61	52.06	16.98	**8.65**	26.27	208.31	66.75	53.15	18.81	10.49	27.38	209.55	66.45
EgoFound3R^{‡,100w,Dyn}	**40.88**	17.74	**8.71**	**26.16**	206.30	**65.53**	**41.73**	**16.97**	**8.65**	**26.23**	209.41	**66.65**	**42.88**	**18.79**	**10.48**	**27.37**	210.35	66.37

Table 16: **Matched 100-window contact comparison.** InteractVLM and EgoFound3R are evaluated on exactly the same windows, and the best value in each column is marked in bold.

Method	21 joints			195 vertices			778 vertices		
	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
OakInk-v2									
InteractVLM^{100w}	0.180	0.403	0.202	0.173	0.408	0.205	0.168	**0.412**	0.201
EgoFound3R^{100w,InteractVLM}	**0.381**	**0.762**	**0.431**	**0.441**	**0.638**	**0.442**	**0.457**	0.295	**0.298**
TACO									
InteractVLM^{100w}	0.388	0.705	0.472	0.348	0.743	0.451	0.356	**0.750**	**0.458**
EgoFound3R^{100w,InteractVLM}	**0.670**	**0.821**	**0.730**	**0.654**	**0.755**	**0.691**	**0.582**	0.320	0.388
HOI4D									
InteractVLM^{100w}	0.313	0.537	0.281	0.256	0.556	0.261	0.249	**0.553**	0.261
EgoFound3R^{100w,InteractVLM}	**0.653**	**0.796**	**0.691**	**0.631**	**0.794**	**0.670**	**0.458**	0.410	**0.418**

I Architecture and Multi-Rate Ablations

We report local and world-space hand geometry, contact, visibility, and distance on OakInk-v2, TACO, and HOI4D: Tab. 17 for hand geometry, Tab. 18 for contact and visibility, and Tab. 19 for distance.

0.05B and 0.1B (reduced capacity). The hand module is reduced to about 0.05B and 0.1B parameters while the rest of the design is left unchanged, so the two variants show how much of the accuracy depends on the capacity of the hand decoder.

MANO (parameter regression). Instead of predicting point coordinates, the decoder regresses MANO rotation parameters and obtains the mesh through the MANO model, which separates the value of the explicit point representation from that of the underlying mesh topology.

w/o Ω (no pretrained backbone). The frozen VGGT-Ω backbone is removed together with the features it provides, so the hand and scene heads are trained from the supervision alone; the variant measures what the pretrained geometric priors contribute.

1-way (one-way exchange). The cross-rate interface is restricted to one direction: backbone features still reach the hands, but the hand prompts no longer update the anchors, which tests whether hand evidence must also shape the scene representation.

Sk (temporal stride). The full model runs at temporal stride k, that is, with one global anchor per k hand frames; S1 is the densest and S5 the default, shaded in the tables, so the variants vary aggregation density at a fixed parameter count.

Table 17: **Hand-geometry ablation** on OakInk-v2, TACO, and HOI4D. Errors are in mm; Sk is EgoFound3R at stride k.

	21 joints: MPJPE						195 vertices: MPVPE						778 vertices: MPVPE					
Variant	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓	Raw↓	RR↓	PA↓	Sim(3)↓	W↓	WA↓
OakInk-v2																		
0.05B	112.47	35.90	16.76	33.35	65.09	35.51	110.83	35.50	16.15	33.45	65.83	35.66	111.13	36.33	16.20	33.46	66.14	35.63
0.1B	98.34	31.74	15.24	30.74	61.31	32.39	98.33	30.92	15.27	30.67	59.29	32.32	98.69	31.64	15.36	30.73	59.56	32.32
MANO	73.39	21.02	8.74	24.13	45.73	25.71	73.27	19.87	8.55	24.30	45.46	25.92	73.12	21.01	9.36	24.70	46.05	26.27
w/o Ω	86.13	22.06	9.92	23.63	44.03	24.88	85.25	20.59	9.54	23.69	43.87	24.95	86.07	21.67	10.32	24.17	44.39	25.40
1-way	79.62	22.01	9.96	22.94	41.28	24.30	80.46	21.44	9.98	23.00	40.84	24.43	80.88	22.47	10.65	23.41	41.40	24.81
S1	53.15	**17.68**	**7.97**	**17.50**	**28.13**	**17.91**	52.36	17.11	7.97	**17.71**	**28.45**	**18.13**	52.46	**18.33**	**8.96**	**18.39**	**29.17**	**18.79**
S2	53.14	17.82	8.02	18.01	30.37	18.81	52.72	**17.08**	**7.96**	18.18	30.75	18.99	52.96	**18.33**	8.98	18.88	31.41	19.65
S3	51.97	17.78	8.02	17.99	30.70	18.85	51.64	17.10	**7.96**	18.18	31.09	19.05	51.90	18.35	8.98	18.88	31.77	19.71
S4	**51.72**	17.78	8.02	17.91	30.68	18.80	**51.38**	17.10	**7.96**	18.10	31.09	19.00	**51.63**	18.35	8.98	18.80	31.76	19.65
S5	51.81	17.77	8.02	17.87	30.88	18.77	51.40	17.11	**7.96**	18.05	31.32	18.98	51.64	18.37	8.98	18.75	31.99	19.63
TACO																		
0.05B	109.71	36.74	15.15	42.08	81.37	42.66	110.05	33.75	14.38	42.10	83.13	42.84	110.63	35.05	15.30	42.42	82.54	43.09
0.1B	96.90	30.55	13.14	35.07	61.66	35.81	93.16	29.92	13.53	34.95	60.85	35.78	93.04	30.89	14.17	35.34	61.53	36.14
MANO	86.66	24.29	10.68	34.41	57.45	35.15	86.25	23.33	10.80	34.49	57.54	35.29	86.18	24.47	11.96	35.06	57.95	35.78
w/o Ω	82.59	23.49	10.11	31.38	53.38	32.33	81.91	22.06	10.50	31.76	54.42	32.76	82.17	23.33	11.71	32.38	54.53	33.32
1-way	78.82	23.70	10.07	31.21	54.16	32.14	78.33	22.25	10.31	31.20	54.16	32.19	78.69	23.61	11.48	31.86	54.46	32.77
S1	67.42	22.55	8.51	27.90	47.10	28.54	68.98	19.25	8.75	27.94	46.22	28.59	69.70	21.01	10.50	28.79	46.54	29.44
S2	53.45	**18.53**	**8.12**	24.25	40.23	25.18	52.76	**18.35**	**8.66**	24.59	40.88	25.59	53.17	19.92	10.22	25.51	41.56	26.48
S3	**53.34**	**18.53**	**8.12**	24.23	40.39	25.18	**52.61**	18.36	**8.66**	24.56	40.98	25.58	**53.01**	19.93	10.21	25.49	41.66	26.47
S4	53.86	18.54	**8.12**	24.20	40.28	25.18	53.08	18.36	8.67	24.54	40.92	25.58	53.42	19.93	10.21	25.45	41.60	26.46
S5	54.64	18.54	8.13	**24.15**	**40.03**	**25.16**	53.76	18.36	8.67	**24.49**	**40.69**	**25.55**	54.03	**19.91**	**10.19**	**25.39**	**41.36**	**26.42**
HOI4D																		
0.05B	157.49	32.62	14.54	41.65	155.50	58.89	152.87	31.79	15.20	42.94	159.34	59.95	152.03	33.17	15.88	43.22	158.96	59.79
0.1B	140.38	31.89	13.76	40.07	177.85	55.65	136.61	30.75	14.88	40.69	176.65	56.59	135.81	32.01	15.47	40.99	176.79	56.39
MANO	97.37	25.36	10.71	40.45	169.67	57.57	97.08	24.25	10.74	40.99	170.05	58.42	96.93	25.72	11.97	41.49	171.27	58.39
w/o Ω	92.76	26.09	10.37	37.26	153.96	56.10	92.65	23.57	10.75	38.08	158.38	56.92	93.81	25.36	12.29	38.75	159.89	57.02
1-way	99.21	23.90	10.52	39.40	158.57	56.71	98.46	23.07	10.99	40.03	158.08	57.51	98.47	24.80	12.39	40.60	159.69	57.60
S1	52.05	17.04	8.03	23.49	**151.04**	**49.58**	53.91	16.43	8.29	23.72	**151.48**	**50.33**	55.12	18.33	10.19	24.89	**152.16**	**50.55**
S2	48.90	16.63	**7.99**	23.49	158.35	50.17	50.14	**16.19**	**8.28**	23.81	158.46	50.99	51.44	**18.06**	10.17	24.96	159.47	51.18
S3	49.60	16.64	**7.99**	23.47	158.96	49.95	50.89	16.20	**8.28**	23.79	159.48	50.78	52.23	18.08	10.18	24.95	160.39	50.97
S4	48.40	**16.61**	**7.99**	23.44	158.51	50.16	49.63	**16.19**	8.29	23.75	158.72	50.99	50.92	**18.06**	10.17	24.91	159.71	51.17
S5	**47.28**	16.63	**7.99**	**23.29**	159.70	49.84	**48.46**	16.20	8.29	**23.60**	160.03	50.68	**49.72**	**18.06**	**10.16**	**24.75**	160.98	50.85

Table 18: **Contact and visibility ablation** on OakInk-v2, TACO, and HOI4D. P, R, and F1 denote contact precision, recall, and F1, and VP, VR, and VF1 the corresponding visibility scores.

	21 joints						195 vertices						778 vertices					
Variant	P↑	R↑	F1↑	VP↑	VR↑	VF1↑	P↑	R↑	F1↑	VP↑	VR↑	VF1↑	P↑	R↑	F1↑	VP↑	VR↑	VF1↑
OakInk-v2																		
0.05B	0.541	0.283	0.298	0.820	0.736	0.760	0.600	0.251	0.277	0.678	0.170	0.244	0.373	0.165	0.175	0.695	0.161	0.235
0.1B	0.549	0.358	0.345	0.843	0.708	0.752	**0.620**	0.170	0.205	**0.720**	0.192	0.272	0.325	0.194	0.185	**0.733**	0.186	0.267
MANO	0.512	0.595	0.474	**0.927**	**0.807**	**0.856**	0.537	0.582	**0.485**	0.566	**0.775**	**0.642**	0.530	0.175	0.206	0.567	**0.760**	**0.636**
w/o Ω	0.535	0.606	0.480	0.925	0.796	0.844	0.552	0.557	0.463	0.554	0.770	0.630	0.276	0.250	0.203	0.553	0.757	0.625
1-way	**0.595**	0.489	0.451	0.923	0.802	0.848	0.592	0.492	0.449	0.561	0.772	0.637	0.311	0.254	0.212	0.561	0.757	0.632
S1	0.511	**0.673**	**0.498**	0.921	0.744	0.800	0.524	0.602	0.473	0.552	0.706	0.592	**0.552**	**0.331**	**0.329**	0.550	0.692	0.585
S2	0.509	0.670	0.494	0.920	0.747	0.802	0.518	**0.613**	0.475	0.557	0.703	0.594	0.528	0.319	0.313	0.555	0.689	0.587
S3	0.513	0.667	0.497	0.919	0.747	0.802	0.522	0.612	0.478	0.556	0.704	0.595	0.513	0.326	0.318	0.555	0.690	0.588
S4	0.511	0.665	0.495	0.919	0.747	0.802	0.521	0.611	0.477	0.557	0.704	0.595	0.521	0.329	0.320	0.555	0.690	0.588
S5	0.510	0.665	0.493	0.919	0.747	0.802	0.525	0.611	0.479	0.556	0.705	0.595	0.516	0.329	0.320	0.555	0.690	0.588
TACO																		
0.05B	0.610	0.523	0.542	0.852	0.761	0.800	0.609	0.447	0.493	**0.602**	0.260	0.350	0.333	0.162	0.188	**0.612**	0.249	0.341
0.1B	0.590	0.611	0.579	0.871	0.750	0.801	0.600	0.476	0.506	0.591	0.298	0.381	0.357	0.147	0.177	0.597	0.288	0.373
MANO	0.668	0.665	0.643	**0.921**	0.801	0.854	0.642	0.586	0.592	0.458	0.741	0.562	0.462	0.163	0.218	0.463	0.719	0.559
w/o Ω	0.665	0.695	0.661	0.913	0.813	0.856	0.640	0.608	0.599	0.447	0.758	0.557	0.347	0.163	0.196	0.451	0.741	0.554
1-way	**0.705**	0.600	0.628	0.907	0.825	0.861	**0.645**	0.613	0.609	0.459	0.749	0.564	0.370	0.202	0.236	0.462	0.732	0.562
S1	0.639	**0.800**	**0.698**	0.912	**0.857**	**0.881**	0.598	**0.782**	**0.665**	0.432	**0.821**	0.563	0.364	0.258	0.274	0.434	**0.808**	0.562
S2	0.655	0.776	0.696	0.917	0.850	0.879	0.636	0.719	0.661	0.436	0.818	**0.567**	**0.516**	**0.277**	**0.330**	0.439	0.805	**0.565**
S3	0.655	0.776	0.697	0.917	0.850	0.879	0.635	0.719	0.661	0.436	0.819	**0.567**	0.514	**0.277**	**0.330**	0.439	0.806	**0.565**
S4	0.655	0.777	0.697	0.917	0.850	0.880	0.636	0.719	0.661	0.436	0.819	**0.567**	0.513	0.272	0.326	0.439	0.805	**0.565**
S5	0.654	0.777	0.697	0.917	0.850	0.880	0.636	0.719	0.661	0.437	0.818	**0.567**	**0.516**	0.265	0.320	0.439	0.805	**0.565**
HOI4D																		
0.05B	0.612	0.443	0.448	0.825	0.762	0.783	0.612	0.503	0.475	0.606	0.332	0.411	**0.561**	0.034	0.051	0.644	0.328	0.417
0.1B	0.546	0.636	0.526	0.833	0.744	0.777	**0.639**	0.426	0.422	**0.626**	0.348	0.428	0.478	0.063	0.081	**0.666**	0.345	0.438
MANO	0.600	0.676	0.593	0.892	0.812	0.844	0.638	0.550	0.528	0.613	0.685	0.641	0.507	0.163	0.195	0.624	0.672	0.640
w/o Ω	0.606	0.699	0.597	0.902	0.781	0.833	0.624	0.678	0.594	0.589	0.697	0.631	0.350	0.263	0.244	0.601	0.683	0.631
1-way	**0.630**	0.645	0.582	**0.914**	0.787	0.841	0.605	0.696	0.592	0.605	0.693	0.641	0.443	0.148	0.169	0.620	0.680	0.643
S1	0.606	**0.845**	0.675	0.892	**0.885**	**0.886**	0.597	**0.820**	0.654	0.570	**0.793**	0.660	0.493	0.442	0.404	0.578	**0.784**	0.662
S2	0.612	0.834	0.675	0.891	0.884	0.885	0.606	0.809	**0.656**	0.573	0.789	**0.661**	0.515	0.458	0.433	0.582	0.781	**0.663**
S3	0.613	0.835	**0.676**	0.891	0.884	0.885	0.606	0.809	**0.656**	0.573	0.789	**0.661**	0.511	**0.468**	0.439	0.581	0.781	**0.663**
S4	0.613	0.835	**0.676**	0.892	0.884	0.885	0.607	0.807	**0.656**	0.573	0.789	**0.661**	0.521	0.467	0.441	0.581	0.781	**0.663**
S5	0.612	0.835	0.675	0.892	0.884	0.885	0.608	0.807	**0.656**	0.572	0.789	0.660	0.531	0.466	**0.445**	0.581	0.781	**0.663**

Table 19: **Distance MAE ablation** on OakInk-v2, TACO, and HOI4D at 21 joints, 195 vertices, and 778 vertices. Errors are in mm; MAE(All/Pred/GT) use all valid, predicted-contact, and ground-truth-contact points.

	21 joints			195 vertices			778 vertices		
Variant	MAE(All)↓	MAE(Pred)↓	MAE(GT)↓	MAE(All)↓	MAE(Pred)↓	MAE(GT)↓	MAE(All)↓	MAE(Pred)↓	MAE(GT)↓
OakInk-v2									
0.05B	45.01	47.86	55.45	44.29	46.07	55.38	43.88	25.49	54.74
0.1B	39.91	41.60	45.88	39.61	41.08	47.73	39.35	25.20	46.76
MANO	34.73	36.28	39.00	35.32	37.76	41.37	34.66	17.35	41.13
w/o Ω	32.79	30.73	31.78	32.24	31.80	33.21	32.28	28.37	32.47
1-way	31.17	28.30	29.59	31.10	28.64	31.03	30.86	23.54	30.00
S1	21.96	22.96	24.10	21.65	22.73	24.15	21.27	**12.00**	23.91
S2	21.83	22.66	24.01	21.57	22.14	24.10	21.29	12.86	23.93
S3	21.31	21.90	23.27	21.10	21.68	23.34	20.85	13.06	23.20
S4	**21.20**	**21.82**	**23.01**	**20.99**	**21.65**	**23.10**	**20.74**	12.95	**22.97**
S5	21.29	22.04	23.21	21.06	21.78	23.32	20.79	12.94	23.18
TACO									
0.05B	41.47	45.30	49.54	40.81	44.41	49.73	40.87	26.78	49.45
0.1B	49.27	52.15	58.61	46.88	50.39	57.48	46.00	17.83	56.79
MANO	45.46	48.21	51.97	46.16	51.21	54.07	45.31	12.78	53.51
w/o Ω	36.19	40.42	42.52	36.31	41.14	43.77	35.96	19.43	43.33
1-way	32.79	35.52	37.29	32.80	36.89	38.57	32.65	18.41	38.26
S1	27.34	28.13	30.06	27.78	29.21	31.54	27.65	18.57	30.96
S2	25.13	27.16	29.37	25.01	27.60	30.19	24.63	10.99	29.81
S3	**25.09**	**27.07**	**29.29**	**24.96**	**27.53**	**30.12**	**24.56**	11.01	**29.72**
S4	25.52	27.48	29.73	25.40	27.94	30.60	24.94	10.94	30.17
S5	26.26	28.09	30.47	26.12	28.51	31.38	25.57	**10.78**	30.91
HOI4D									
0.05B	120.58	119.53	129.15	117.58	119.51	128.37	115.66	**12.06**	127.52
0.1B	104.80	111.40	120.17	103.23	115.80	119.63	101.29	19.07	118.74
MANO	52.05	52.19	57.18	52.96	52.95	57.71	52.09	16.03	57.45
w/o Ω	38.01	38.09	41.15	38.30	38.24	41.19	38.24	25.03	41.14
1-way	56.20	58.68	64.31	55.95	57.95	64.38	55.01	20.74	63.90
S1	20.84	18.50	18.71	22.08	19.22	19.43	22.22	15.93	19.39
S2	18.59	16.82	16.37	19.60	17.15	16.88	19.76	14.62	16.91
S3	18.53	16.69	16.14	19.54	17.02	16.58	19.71	14.87	16.61
S4	18.20	16.34	**15.84**	19.16	16.53	**16.23**	19.33	14.33	16.26
S5	**18.07**	**16.27**	15.85	**19.04**	**16.43**	16.24	**19.17**	13.99	**16.25**

J SCENE AND CAMERA EVALUATION

Camera metrics. ATE (absolute trajectory error, m) is measured after a window-level Sim(3) alignment, and rotation error (degrees) is measured relative to the first valid frame.

Depth metrics and scale protocol. Depth uses AbsRel (absolute relative error), RMSE (root mean square error, m), and δ_1, the fraction of valid pixels satisfying $\max(\hat{d}/d, d/\hat{d}) < 1.25$. A model that returns depth up to scale receives one per-window ratio $\mathrm{median}(\mathrm{GT}/\mathrm{pred})$ before the metrics are computed, whereas a metric-scale model keeps scale one; the last column of Tab. 20 reports the metric gauge s^d, that ratio aggregated in log space, so that 1 means metric scale.

Comparison. Tab. 20 is restricted to egocentric methods that reconstruct hands together with the surrounding scene, namely HaWoR, Dyn-HaMR, ReViV, and EgoFound3R. Dedicated geometry models predict cameras and dense shape without hands or interaction attributes, so they are not part of this comparison. EgoFound3R attains the lowest camera error of the four methods and, among the two that return depth, also the lowest AbsRel and RMSE and the highest δ_1; its depth is evaluated at its own scale, whereas ReViV requires a per-window scale fit. HaWoR and Dyn-HaMR return no depth, so their depth and gauge entries are not applicable.

Table 20: **Scene estimation on TACO** for hand-centric world-space methods. Depth follows the scale-aligned protocol defined above; ^{100w} rows are unranked.

Method	ATE (m)↓	Rot (°)↓	AbsRel↓	RMSE (m)↓	δ_1 ↑	s^d
HaWoR	0.0064	1.22	–	–	–	–
Dyn-HaMR^{100w}	0.0064	1.26	–	–	–	–
ReViV	0.0068	1.65	0.2649	0.3577	0.5829	0.6401
EgoFound3R	**0.0055**	**0.99**	**0.1318**	**0.2015**	**0.8469**	**1.1364**

Ablation. Tab. 21 reports the TACO scene ablation with the variants defined in Appendix I. The w/o Ω and 1-way rows report the same uncoupled scene export, so their scene and camera columns are identical and are not independent scene predictions; their hand, contact, and distance metrics are evaluated separately. Dense aggregation attains the lowest camera error and the default configuration the second lowest, so the multi-rate design costs little on the scene side.

Table 21: **Scene ablation on TACO.**

Variant	ATE (m)↓	Rot (°)↓	AbsRel↓	RMSE (m)↓	δ_1 ↑
0.05B	0.0086	1.86	**0.117**	0.232	**0.857**
0.1B	0.0074	1.83	0.122	0.256	0.825
MANO	0.0084	1.60	0.132	0.256	0.806
w/o Ω	0.0069	1.15	0.178	0.242	0.728
1-way	0.0069	1.15	0.178	0.242	0.728
S1	**0.0043**	**0.82**	0.136	0.284	0.805
S2	0.0057	1.08	0.168	0.291	0.763
S3	0.0057	1.07	0.169	0.292	0.759
S4	0.0056	1.00	0.166	0.290	0.768
S5	0.0055	0.99	0.132	**0.202**	0.847

K CAMERA-SPACE QUALITATIVE RESULTS

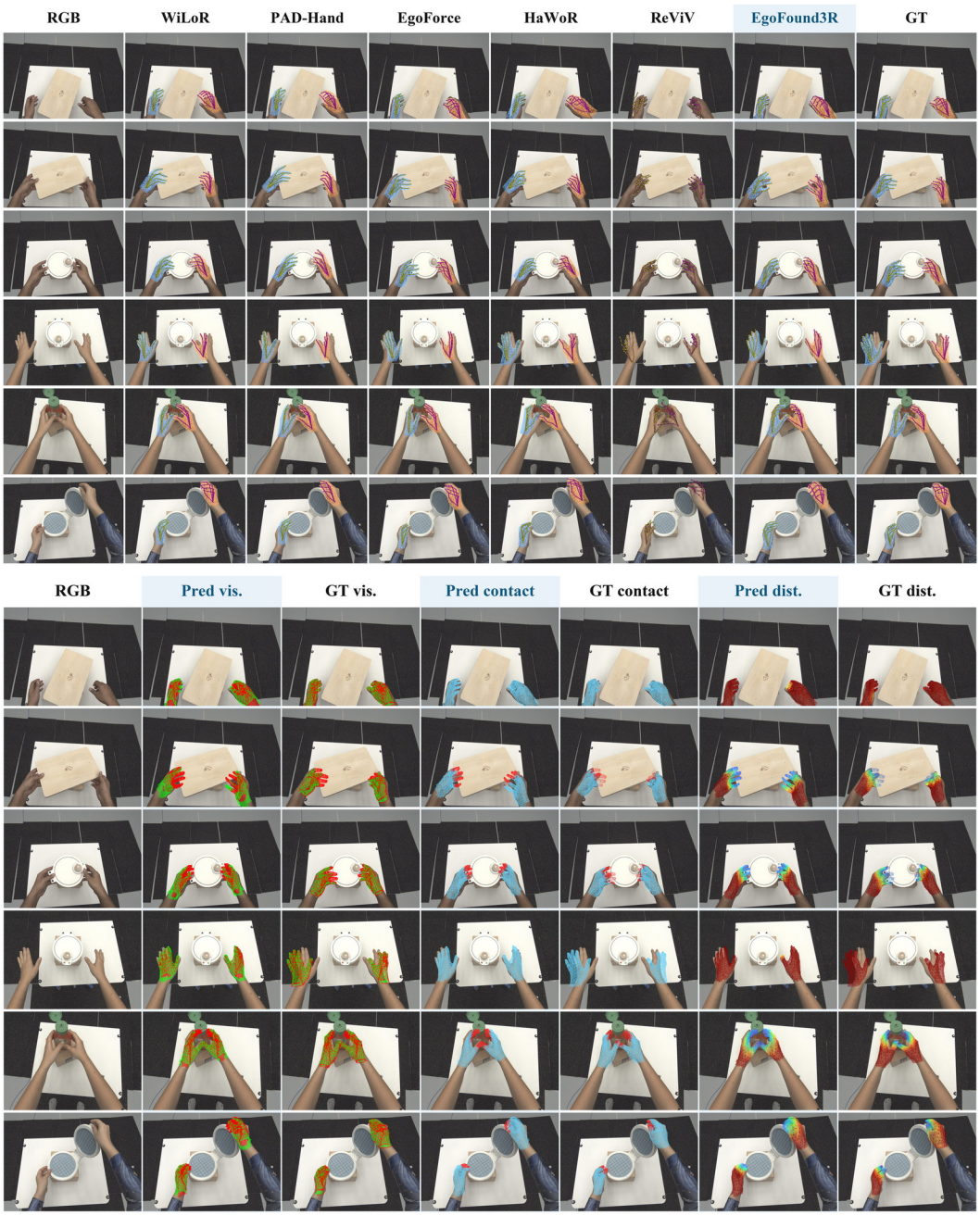


Figure 12: **Qualitative comparison on ARCTIC.** Top: hand geometry. Bottom: per-point visibility, contact, and distance for EgoFound3R (*Pred*) and the ground truth (*GT*).

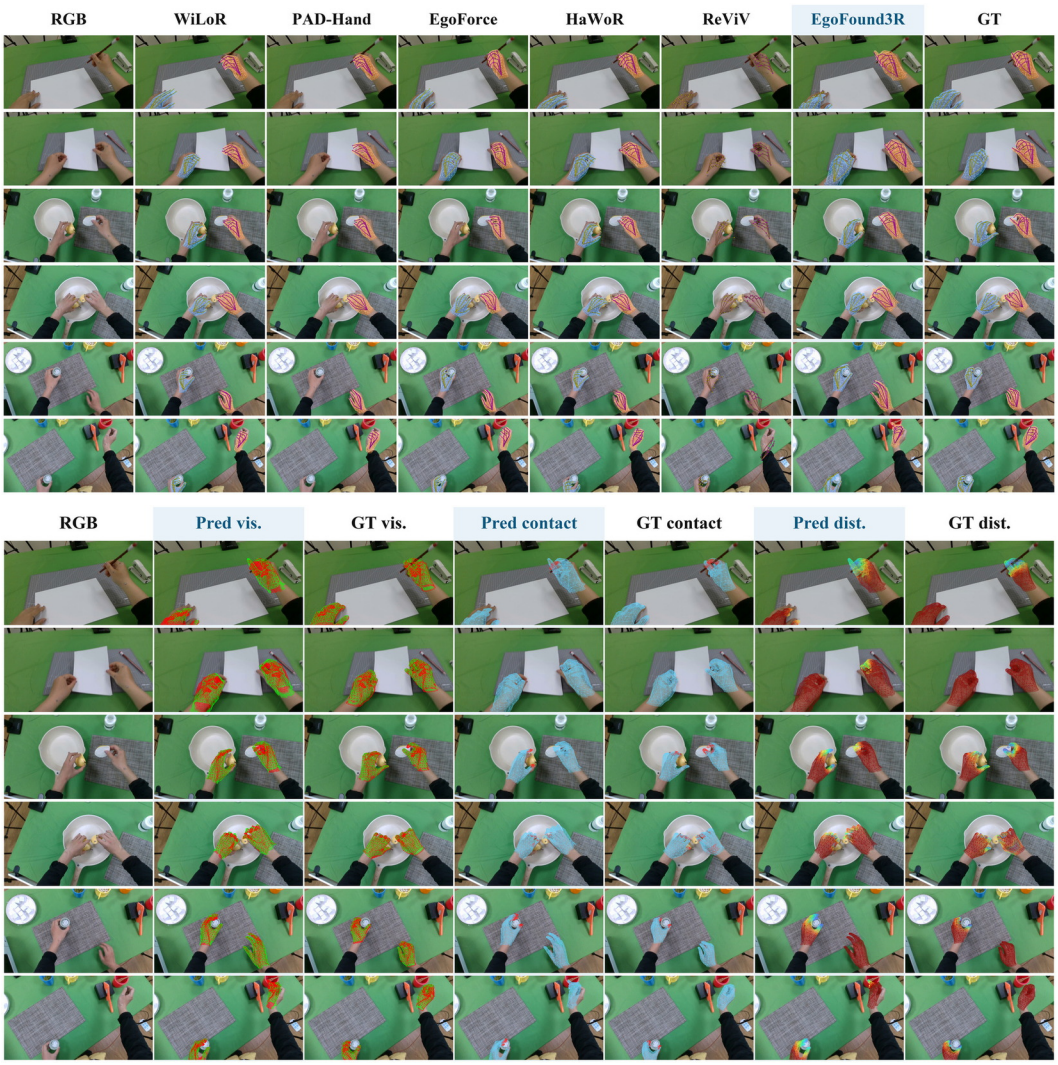


Figure 13: **Qualitative comparison on OakInk-v2.** Top: hand geometry. Bottom: per-point visibility, contact, and distance for EgoFound3R (*Pred*) and the ground truth (*GT*).

Layout. Figs. 12 and 13 give one dataset per figure, with the hand geometry on top and the attributes of the same windows, in the same row order, below; the two panels of the 19-window TACO block are kept apart in Figs. 14 and 15 because the pair does not fit on one page. Columns follow the tables: the first column is the unmodified RGB input, the next three are the camera-space estimators WiLoR, PAD-Hand, and EgoForce, then the world-space pipelines HaWoR and ReViV, then EgoFound3R, and the last column is the ground-truth annotation.

Reading the panels. Each panel overlays the prediction on the pixels of the input frame without additional 2D refinement or fitting, so the comparison does not depend on camera extrinsics. The attribute panels show visibility, contact, and point-to-surface distance for EgoFound3R (*Pred*) and for the ground truth (*GT*) on the same points as the quantitative tables.

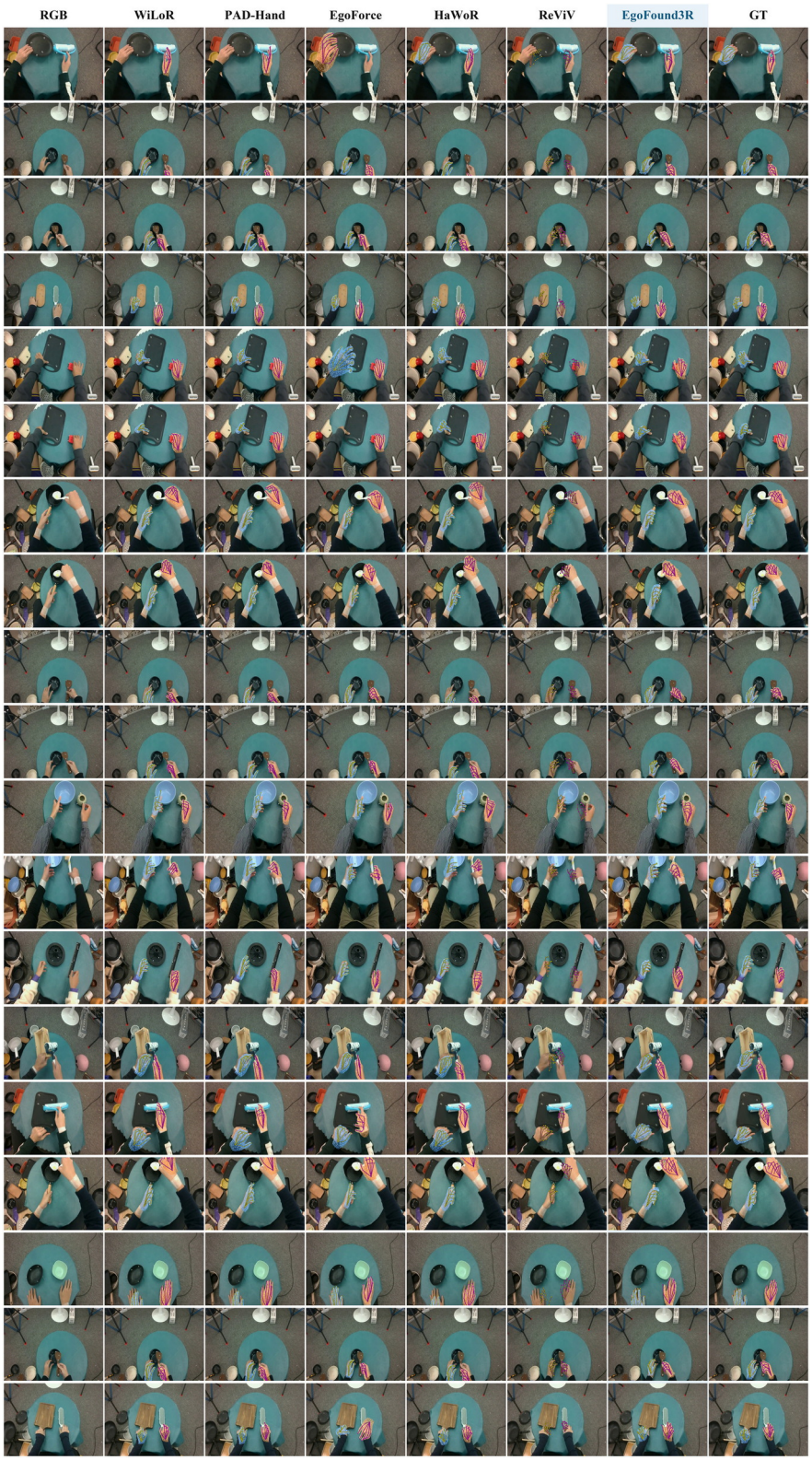


Figure 14: **Qualitative hand geometry on TACO.** Columns as in Fig. 12; the matching attribute panel of the same windows is shown separately below.



Figure 15: **Per-point attributes on TACO.** Rows follow the same windows and order as Fig. 14; columns show the RGB input and the per-point visibility, contact, and distance for EgoFound3R (*Pred*) and the ground truth (*GT*).

L WORLD-SPACE QUALITATIVE RESULTS

Figs. 16 and 17 show additional results for the world-space comparison in Fig. 6. Both figures keep the layout, columns, hand colors, time shading, and extrinsics convention of Fig. 6: two fixed viewpoints per clip, one column per method, and the sampled RGB frames on the left.



Figure 16: **World-space qualitative comparison, part 1.**



Figure 17: **World-space qualitative comparison, part 2.**