

BYND-3D: Metric 3D Hand Tracking from a Single Head-Mounted Camera

Technical Report · September 2026

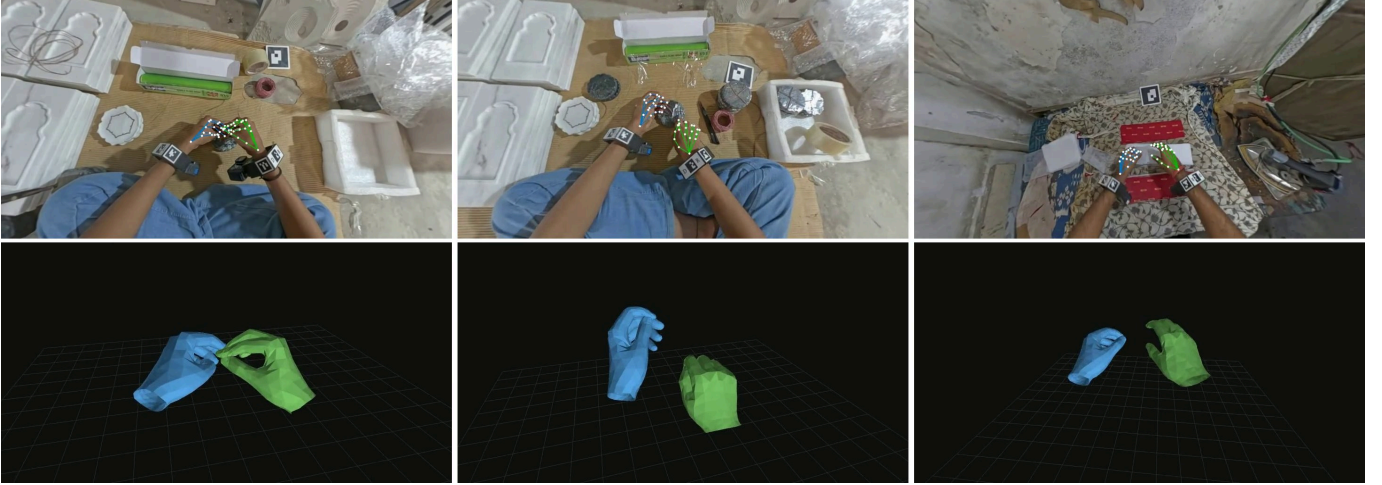


Figure 1. Output of our production setup on three workplace recordings (packing, two sessions; ironing, one session). Top: the camera view with the estimated hand joints. Bottom: the recovered 3D hands on a 50 mm grid. The production setup differs from the configurations measured in Section 4; these recordings have no ground truth and are not part of the measurements.

Abstract

Recording human hands in 3D is a bottleneck for learning manipulation from human demonstrations. We present BYND-3D, a system that recovers both hands in metric 3D, 21 joints per hand (also exported in Meta’s 25-joint hand layout), from video captured by a single head-mounted camera. We compare it with three recent methods, ACE-Ego-Hand, HandFlow and EgoForce, which we ran from their released code and weights and scored with the same code, on two test sets built from the public HOT3D dataset: EgoForce’s released validation protocol (Project Aria glasses), and a Quest 3 protocol we defined on four participants held out from all of our development and scored once. On the validation protocol BYND-3D’s mean 3D joint error is 25.6 mm, against 31.3, 35.0 and 47.2 mm. On the Quest 3 protocol it is 44.0 mm, against 117.2, 70.2 and 58.5 mm. The two results use two BYND-3D configurations, each fixed before its test was scored. Under the comparison rules fixed before scoring, all six of these differences are statistically clear over recording sequences; the smallest, 5.7 mm against ACE-Ego-Hand on the validation protocol, is not clear after correcting for our three scorings of that data (Section 5). The Quest 3 configuration is also clearly lower than all three methods on the validation protocol (26.8 mm), although its 4.5 mm margin over ACE-Ego-Hand is likewise not clear after that correction; the validation configuration scores 53.9 mm on Quest 3, lower than EgoForce but not clearly. A multi-camera version of BYND-3D is in development.

1. Introduction

Robot learning from human video needs the 3D position of each hand, in metric units, at every frame. A single head-mounted camera is the cheapest and least intrusive way to record such video, but it leaves depth ambiguous, sees hands at the distorted edge of wide-angle lenses, and loses hands when they leave the view or are hidden by objects.

BYND-3D is built for this setting. This report measures it against three recent research methods on two test sets built from the public HOT3D dataset [1], under rules cho-

sen to limit selective reporting: every scored hand counts, a missed hand is charged as an error, all methods are scored by the same code, and the comparison rule and the wording of each possible outcome were fixed before the final scoring of each test.

2. BYND-3D

BYND-3D takes video from one head-mounted camera, including wide-angle lenses, and outputs, for each frame, the 3D positions of 21 joints for each hand it finds, in millimetres in the camera’s coordinate frame, together with a hand mesh. It also exports each hand in Meta’s 25-joint

layout: the same 21 joints plus four metacarpal joints, which are derived from the other 21. All measurements in this report use the 21 standard joints. It is offline: its estimate for a frame also uses later frames of the same video. This report describes results, not the method.

3. Evaluation protocol

Metric. Mean per-joint position error over 21 joints, in millimetres, in the camera frame, with no root or Procrustes alignment, averaged over all scored hands. A hand that a method does not report is charged as a prediction at the camera centre, an error of roughly the hand's distance from the camera, so leaving out hard hands is heavily penalised.

EgoForce HOT3D validation. The validation split and hand-visibility rule released with EgoForce [2]: 31 sequences from 4 participants recorded with Project Aria glasses, 190,983 scored hands. The scorer is EgoForce's, changed so that a missing prediction is charged instead of being replaced by the ground truth; its temporal filter and failure exclusion are off. With those settings on and errors averaged per batch, as in EgoForce's paper, our run of its released weights reproduces the published 43.9 mm.

Our Quest 3 protocol. All annotated HOT3D-Clips [1] Quest 3 clips of the four participants who were never recorded with Aria: 450 clips of 150 frames in 57 sequences, 125,912 scored hands, from one headset camera that is the same for every clip. Each clip is processed on its own. We defined this protocol and no other group has published results on it. The four participants were held out from all of our development, and the test was scored once.

Methods. ACE-Ego-Hand [3], HandFlow [4] and EgoForce [2], run by us from their authors' released code and weights. EgoForce receives hand boxes and hand sides derived from the ground truth, as in its own evaluation: the boxes released with it on the validation protocol, and boxes we built from the annotations to match them on Quest 3. The other methods, including BYND-3D, find the hands themselves. On Quest 3, EgoForce runs without its forearm input, the setting that scored better for it on our development participants; with its validation setting it scores 96.9 mm there.

Two BYND-3D configurations. BYND-3D was run in two configurations, A and B, both fixed before scoring. A was the pre-registered headline on the validation protocol, with B as its secondary. B was chosen as the Quest 3 head-

line after results on development participants, with A as its secondary.

Statistics. Differences are paired over recording sequences, with 95% intervals from 20,000 bootstrap resamples. A difference counts as clear only if the whole interval lies on one side of zero. The rows, the rule and the wording of each possible outcome were written down and fingerprinted before the final scoring of each test.

4. Results

Table 1. Mean 3D joint error (mm, lower is better), all scored hands, misses charged. Headline results in bold; the other BYND-3D cell of each column is that test's pre-registered secondary.

Method	EgoForce HOT3D val	Our Quest 3 protocol
BYND-3D (A)	25.6	53.9
BYND-3D (B)	26.8	44.0
ACE-Ego-Hand [3]	31.3	117.2
HandFlow [4]	35.0	70.2
EgoForce [2]	47.2	58.5

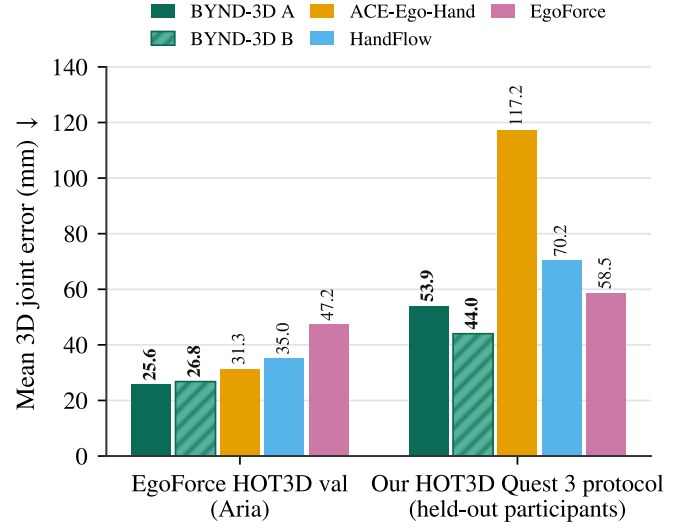


Figure 2. Mean 3D joint error on the two test sets. Of the four methods we ran, BYND-3D has the lowest error on both. The Quest 3 protocol is our own, and EgoForce receives hand boxes derived from the ground truth.

On the validation protocol configuration A's error is 25.6 mm: 5.7 mm lower than ACE-Ego-Hand, 9.4 mm lower than HandFlow and 21.6 mm lower than EgoForce (Table 2, Figure 3). It is lower on 23, 23 and 30 of the 31 sequences, and each difference stays below zero when any one of the four participants is left out. Configuration B scores 26.8 mm there and is also clearly lower than all three methods.

On our Quest 3 protocol every method has a higher error than on validation; the camera, lens, participants and scored hands all differ, so the two columns are not directly comparable. Configuration B's error is 44.0 mm: 14.5 mm lower than EgoForce, 26.2 mm lower than HandFlow and 73.2 mm lower than ACE-Ego-Hand, lower on 47, 52 and 56 of the 57 sequences, and each difference holds with

any one participant left out. Much of ACE-Ego-Hand's gap is depth scale: it places Quest 3 hands too far from the camera (by about 30% on our development participants and, in a check made after scoring, about 21% on the test). Configuration A scores 53.9 mm on Quest 3; its difference to EgoForce, -4.6 mm $[-11.8, +2.8]$, is not statistically clear.

Table 2. BYND-3D's error minus each method's (mm), with 95% paired bootstrap intervals over sequences and the number of sequences where BYND-3D is lower. Negative favours BYND-3D. Every interval shown lies below zero. Configuration A on Quest 3, that test's pre-registered secondary, is not shown: against EgoForce its difference is -4.59 $[-11.83, +2.84]$ · 36/57, not clear; its differences to ACE-Ego-Hand and HandFlow were not pre-registered.

vs	A on EgoForce HOT3D val (headline)	B on EgoForce HOT3D val (secondary)	B on our Quest 3 protocol (headline)
ACE-Ego-Hand	-5.73 $[-10.98, -0.94]$ · 23/31	-4.53 $[-8.96, -0.15]$ · 21/31	-73.16 $[-82.73, -64.00]$ · 56/57
HandFlow	-9.37 $[-12.99, -5.81]$ · 23/31	-8.18 $[-12.23, -4.15]$ · 20/31	-26.22 $[-34.20, -19.22]$ · 52/57
EgoForce	-21.61 $[-27.08, -16.33]$ · 30/31	-20.41 $[-25.47, -15.41]$ · 28/31	-14.54 $[-20.70, -8.23]$ · 47/57

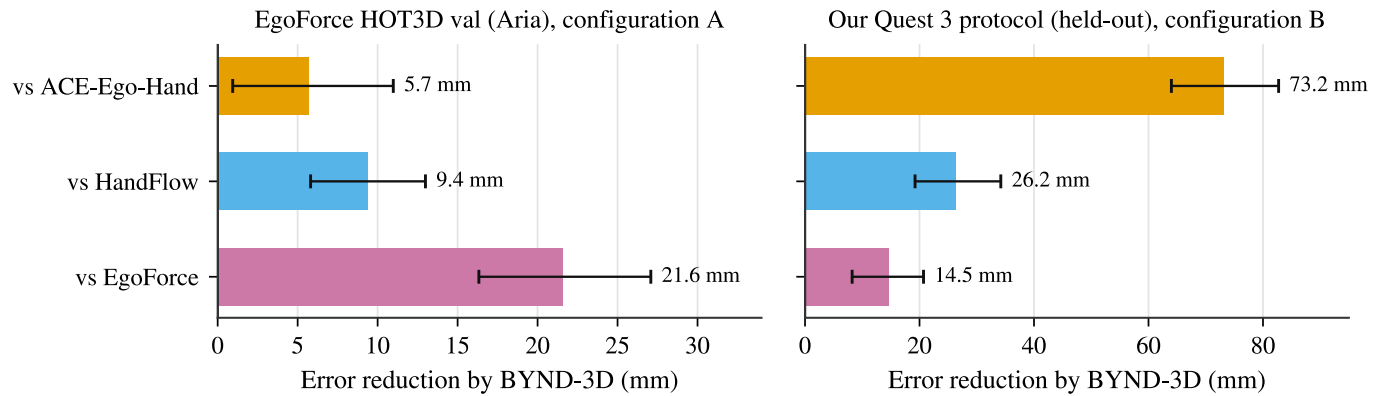


Figure 3. How much lower BYND-3D's error is than each method's, for the headline configuration of each test, with 95% paired bootstrap intervals over recording sequences. Every interval lies above zero, the pre-registered criterion for a clear difference. Against ACE-Ego-Hand on the validation protocol the interval reaches zero after correcting for our three scorings of that data ($[-0.03, 12.3]$ mm) and when participants are resampled (Section 5). The two panels use different scales.

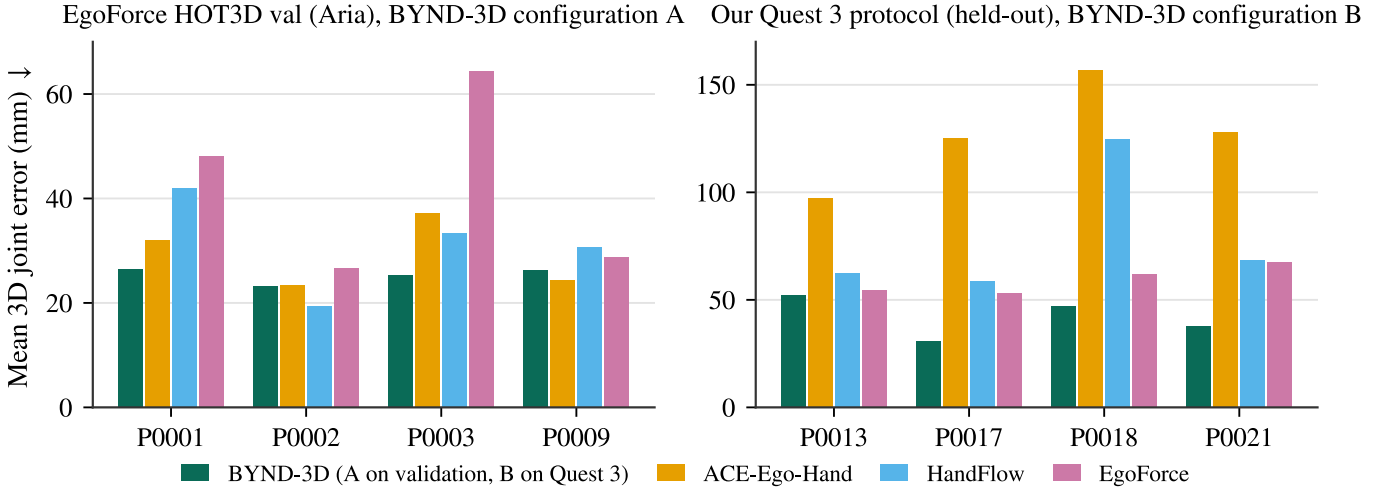


Figure 4. Mean 3D joint error per participant, for the headline configuration of each test. BYND-3D is not the lowest for every person: on validation HandFlow is lower for P0002 and ACE-Ego-Hand for P0009, and BYND-3D and ACE-Ego-Hand are level for P0002 (23.2 and 23.4 mm); on Quest 3, BYND-3D and EgoForce are close for P0013 (52.3 and 54.5 mm). P0001 contributes 16 of the 31 validation sequences.

5. Limitations

- The validation set is public and this is the third time we scored it. The two earlier scorings, and early tuning on 11 of its 31 sequences, informed the design, including the choice of ACE-Ego-Hand as the main comparison; the intervals are not adjusted for this. A correction for three looks, applied after scoring, gives $[-12.3, +0.03]$ for configuration A and $[-10.0, +0.9]$ for configuration B against ACE-Ego-Hand; both include zero. The Quest 3 test was scored once.
- In checks made after scoring, about 70% of the validation gap to ACE-Ego-Hand comes from three sequences where that method fails; on the other 28 sequences the difference is -1.9 mm $[-4.7, +1.3]$. Resampling participants instead of sequences widens the validation intervals against ACE-Ego-Hand and HandFlow past zero; on Quest 3 all three stay below zero. Each test has only four participants, and on Quest 3 the margin over EgoForce is not clear for two of the four participants taken alone.
- BYND-3D's lead is in absolute 3D position. With the wrist position removed, ACE-Ego-Hand and HandFlow have lower error on both test sets (validation 11.6 and 15.7 against 16.1 mm; Quest 3 20.7 and 25.7 against 32.6 mm), and on validation also after aligning the whole hand (6.3 and 5.8 against 6.5 mm). On validation hands that both systems locate on their own, BYND-3D and HandFlow are level (-0.6 mm); most of that gap is hands HandFlow's detector misses. BYND-3D's validation lead over ACE-Ego-Hand depends on its offline processing. Some of these comparisons were made after scoring.
- Inputs differ by method: ACE-Ego-Hand and HandFlow were run on an undistorted view of each image, as their released code expects (on Quest 3, stored as a high-quality compressed video); BYND-3D and EgoForce read the original camera images. BYND-3D is offline: its estimate for a frame also uses later frames of the same sequence on validation and of the same 150-

frame clip on Quest 3. Neither protocol penalises reporting a hand that is not there.

- The Quest 3 protocol is our own. Configuration B was chosen after seeing results on development participants. All Quest 3 test recordings are full resolution; on lower-resolution recordings of development participants BYND-3D was clearly less accurate than EgoForce, so these results do not cover such recordings.
- BYND-3D's training data includes HOT3D annotations of other participants, one of whom appears in 2 of the 31 validation sequences, and we cannot rule out that data used to train BYND-3D includes frames from both test sets. The four Quest 3 test participants were never used in our development. HandFlow trained on 29 of the 31 validation sequences; ACE-Ego-Hand also trains on HOT3D, with unknown overlap.

6. Next steps

A single view can only estimate depth. We are extending BYND-3D to head-mounted rigs with six or more cameras and will evaluate that version on new multi-camera data.

References

1. P. Banerjee et al. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. CVPR, 2025.
2. C. Millerdurai et al. EgoForce: Forearm-Guided Camera-Space 3D Hand Pose from a Monocular Egocentric Camera. arXiv:2605.12498, 2026.
3. Y. Liu et al. ACE-Ego-Hand: Repurposing Video Diffusion Models for Occlusion-Robust Egocentric 3D Hand Motion Recovery. arXiv:2608.20308, 2026.
4. M. Xu et al. HandFlow: Fully Generative 4D Hand Recovery with Flow Matching. arXiv:2607.11221, 2026.