# Bimanual 3D Hand Motion and Articulation Forecasting in Everyday Images

Aditya Prakash[1] Richard Li[2] David Forsyth[1] Saurabh Gupta[1]

[1] University of Illinois Urbana-Champaign
[2] Massachusetts Institute of Technology
https://bit.ly/ForeHand4D

**Abstract.** We tackle the problem of forecasting bimanual 3D hand motion and articulation from a single image in everyday settings. To address the lack of 3D hand annotations in diverse settings, we design an annotation pipeline consisting of a diffusion model to lift 2D hand keypoint sequences to 4D hand motion. For the forecasting model, we adopt a diffusion loss to account for the multimodality in hand motion distribution. Extensive experiments on 6 datasets show the benefits of training with our imputed labels (14% improvement) and the effectiveness of our lifting (45% better) & forecasting (16.4% gain) models, over the best baselines, especially in zero-shot generalization to everyday images.

**Keywords:** 3D Pose Forecasting · Egocentric Hand-object Interaction · Diffusion models with weak supervision

## 1 Introduction

This paper develops ForeHand4D, a system for *forecasting bimanual 3D hand motion* from a single everyday RGB image as input. ForeHand4D can operate on *diverse everyday images* to generate the *full articulation* of the hand in *3D* for *both* hands over *long* time horizons while only requiring *a single RGB image*. This expands capability along several axes: generalization, prediction horizon & completeness of output; thereby improving the utility of such models for downstream human-robot interaction, AR/VR coaching or animation applications. *E.g.*, accurate forecasts for what a human's hand might do next can prevent a co-operating robot from colliding with it & better yet offer more meaningful assistance. Or an AR coaching app can show how a human should modify their grasp of a tennis racket. Or generated motions can offer suggestions for an animator to pick from. Fig. 1 shows sample outputs from ForeHand4D, including on images from EgoExo4D not used for training in any way. Generations from ForeHand4D are more coherent than the baseline ($1^{\text{st}}$ row) and different samples capture different possible interactions from the same starting point ($2^{\text{nd}}$ row).

Generating hand motion is difficult because of the complex ways in which hands interact with one another and the surrounding environment. Because hands can do many different things in the future (*i.e.* output is multi-modal), training a regressor is not suitable. Therefore, we adopt a diffusion loss for training our

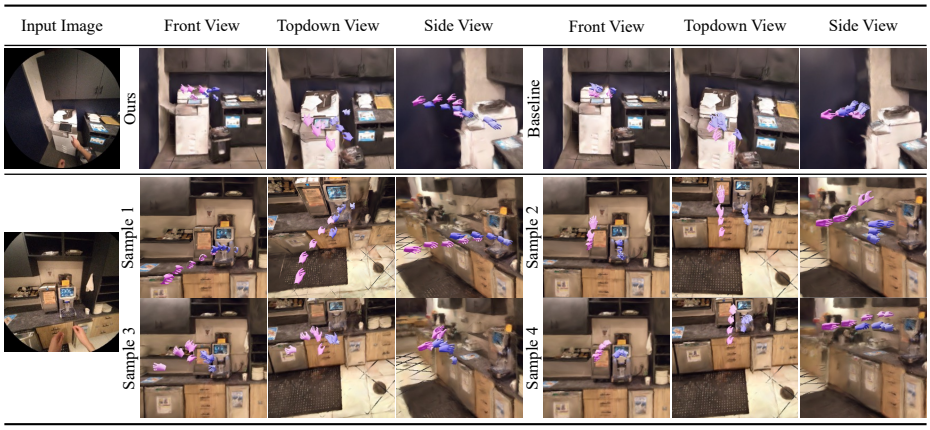


**Fig. 1:** ForeHand4D forecasts bimanual 3D hand motion from a single RGB image input. (top) The predicted motions (shown from 3 views) highlight object interactions, *e.g.* reaching for the paper on the table beside the printer, compared to the baseline where the hand is in air. (bottom) ForeHand4D can also generate multiple motions from the same input, *e.g.* reaching for the coffee machine/cabinet, leftward grabbing action or pointing to the screen. Left hand shown in pink, right hand in blue. Color saturation decreases as time proceeds, *i.e.*, further out timesteps are denoted by lighter shades. The 3D scene is only for visualization & these images are not seen during training.

models. We find that this successfully mitigates problems arising due to multi modality and leads to a large improvement in forecasting performance on a suite of lab datasets as shown in Tab. 5, representing the first experimental results on this challenging problem.

However, there is a mismatch between the data we can train such a diffusion models on (lab datasets where *complete* 3D ground truth, *i.e.* MANO [45] parameters, are available) *vs*. the data we would like this diffusion model to work on (everyday images outside of lab settings, that may have some 2D annotations but no 3D labels). Because a diffusion model needs complete ground truth for training (the forward diffusion process adds noise to the ground truth before denoising), prior techniques [38, 54, 60] that leverage weak supervision via reprojection losses are not applicable since MANO parameters are not available. Generating 3D pseudo-labels from available 2D annotations is the obvious solution but existing methods, *e.g.* EasyMocap [50], that directly optimize the MANO parameters using 2D reprojection loss, are not effective since they are highly sensitive to initialization, optimization objective & hyper-parameter tuning. Our innovation is to develop a *learned lifting model* that lifts available 2D annotations into complete 3D annotations. This increases the diversity of data for training the forecasting model, and thereby its performance on held-out datasets (Tab. 5).

Our overall pipeline is shown in Fig. 2. We first use 3D annotated lab datasets to develop the lifting diffusion model, $L$. This model takes as input 2D hand keypoints & camera parameters (intrinsics & extrinsics) across all timesteps in a sequence to output the corresponding 3D hand articulation & placement. We use $L$ to lift 2D annotations on diverse datasets into 3D ground truth. We

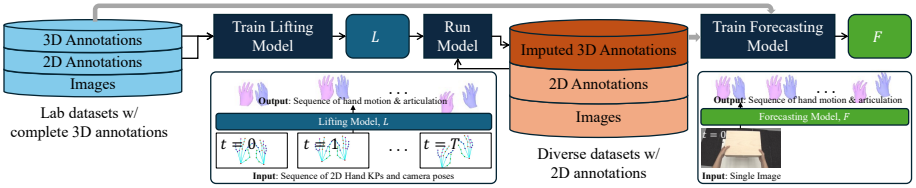


**Fig. 2: Overall Training Pipeline.** We first use the 2D & 3D annotations in lab datasets to train a lifting diffusion model, $L$ that maps 2D keypoints sequences to 3D MANO hands. We then run $L$ on diverse datasets with 2D annotations to generate 3D annotations. Finally, the forecasting model $F$ is trained on lab & diverse datasets with complete 3D supervision.

then use these imputed 3D labels, alongside true 3D ground truth labels on lab datasets to train our forecasting diffusion model, $F$. Because 2D annotations are more readily available on diverse datasets, this increases the data diversity for training the forecasting model. Experiments reveal that training on diverse data, enabled by our method, substantially improves the predictions of learned models on both everyday images from EgoExo4D (zero-shot generalization) & lab datasets (Tab. 5), and outperforms LatentAct [42] (adapted & retrained to work in our setting) by 16.4%. It also improves over alternative ways of injecting weak supervision via auxiliary 2D forecasting heads. Also, our lifting model generates more accurate 3D labels, 65.3% better than the recent HaWoR method [72] (Tab. 2). Code & models will be released upon publication.

## 2 Related Work

**3D Hand Prediction**. Given image or video input, many recent works make *predictions* (and not forecasts) for hands present in them from egocentric [14, 15, 27, 29, 58, 63] and exocentric observations [20, 21, 31, 34, 35, 37, 40, 47, 66, 69]. HaMeR [38] is a high-performing recent work that predicts 3D hands from single images, while Dyn-HaMR [67] makes temporally consistent 3D hand predictions from videos. Predictions on videos are made via feed-forward model [64], test-time optimization [72] or hybrid methods [24].

**4D Hand Forecasting**. Prior works have looked at forecasting specific aspects of hand motion in different settings. Works differ in what they output and from what input. On the output side: [7, 42] produce hand motion, articulation & contact maps, [4, 23] forecast the 3D wrist location, while Liu *et al*. [32] forecast only the 2D wrist location. On the input side, [4, 6, 16, 32, 61] only use RGB images as input, while [7, 10, 42, 71, 74] are conditioned on privileged information in the form of articulating 3D objects or 3D contact points on objects. Papers also tackle different settings: full body forecasting [6, 61], single hand-object interaction [4, 16, 42, 76] or bimanual interactions [7, 10, 32, 71, 74]. Thus, past work addresses individual aspects of the problem, but none as comprehensively as ours: they either produce rich 3D output from stronger input (3D object models) or use RGB images but predict only coarse 2D/3D results.

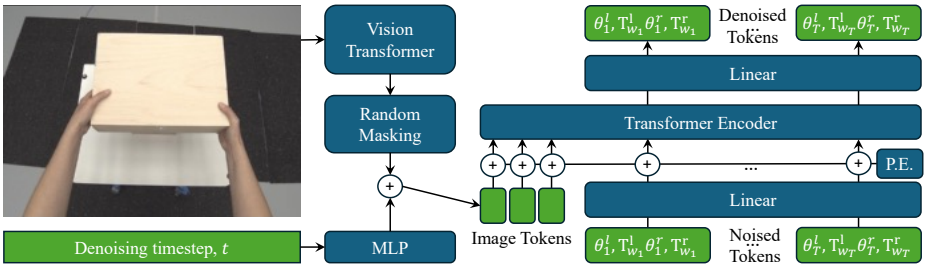


**Fig. 3: Architecture for Forecasting Model**. We modify MDM [52] to condition on images features extracted from a ViT backbone. Each input & output token is 198-dimensional: 2 hands $\times$ (16 (joints) $\times$ (6 (6D rotation for each joint) + 3 (wrist translation))).

**3D pose from 2D keypoints**. Several works in the human pose literature have explored estimating 3D pose from 2D keypoints using different approaches, *e.g.* linear models [44], probabilistic models [51], directly optimizing 3D poses [1], MLP [33, 36], convolutional [36, 39, 53], graph-based [5, 11, 56], transformer [30, 48, 73, 75], diffusion [26, 28] & normalizing flows [55]. These works operate on different types of inputs, *e.g.* static 2D pose [1, 33, 55], 2D pose estimated from image [36, 51], sequence of 2D keypoints [5, 26, 48, 56, 73, 75] or multi-view 2D poses [28]. These ideas have also been extended to estimated 3D hand poses using 2D keypoints [59, 78], 2D/2.5D heatmaps [21, 25] or depth [9]. Building on top of these works, we design a diffusion-based lifting model to estimate 3D hand poses from 2D keypoints to scale up 3D hand annotations for diverse settings.

**Learning from Incomplete 3D Ground Truth**. Prior works often inject weak supervision into 3D regression models via a 2D reprojection loss [38, 54, 60]: the predicted 3D is differentiably rendered or projected into 2D and encouraged to match the 2D annotations. However, this doesn't naturally extend to diffusion models that need to know the score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ at different locations $\mathbf{x}$. While we can render a denoised 3D shape and compute partial supervision on it using the reprojection loss (*i.e.* we don't know $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ but only a projection of it), we don't know what point $\mathbf{x}$ in space is this the partial supervision for. Recent works explore training diffusion models on corrupted or partial data [2, 12] using EM [2, 22] or aggressive masking [12]. Our setting is different because we don't quite have partial ground truth, but rather a projection of a 3D shape into 2D.

## 3 4D Hand Forecasting

Given a single RGB image $I$ showing a hand object scenario, the task is to forecast the 3D hand motion for both hands. We use MANO hand representation [46], consisting of the shape $\beta$, articulation $\theta$ & global wrist pose ${}_cT_w$, where $c$ is the world frame located at the camera center. The goal is to learn a function $F(I)$ that takes the image $I$ as input and predicts $\Phi_t = \{(\theta^l_t, {}_cT^l_{w_t}, \theta^r_t, {}_cT^r_{w_t})\}$ for all timesteps in the prediction horizon, where $l$ & $r$ superscripts denote left & right

hand. We do not predict $\beta$ (we use the mean $\beta$ shape from the MANO model when $\beta$ is not available).

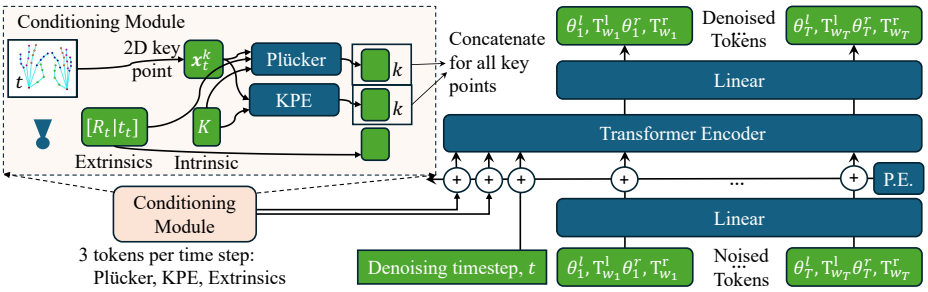


**Fig. 4: Architecture for Lifting Model**. We modify MDM [52] to condition on a sequence of 2D hand keypoints & camera parameters. The conditioning module combines different input representations: 3D pose (rotation, translation) of camera, Plücker rays [70] & KPE [41].

Our forecasting model $F$ is realized using a transformer & trained using a diffusion loss (Sec. 3.1). A diffusion loss means we need *complete* 3D annotations for training. Thus, we can only use carefully constructed lab datasets with complete 3D annotations (*e.g.* ARCTIC [15], H2O [27], H2O-3D [21], HOT3D [3], & DexYCB [8]) for training $F$. This severely limits the diversity of data that $F$ is exposed to, and thereby its generalization capabilities. To mitigate this limitation, we develop a *lifting model,* $L$ to lift 2D key point annotations to complete 3D annotations (Sec. 3.2). Our final forecasting model is trained on the union of 3D lab datasets, and 2D in-the-wild datasets lifted to 3D, as shown in Fig. 2.

### 3.1 Forecasting Model, $F$

Since temporal forecasts are multimodal, we adopt a conditional diffusion model to represent $F$, *i.e.* given the inputs & noisy versions of the desired outputs, $F$ predicts the noise that was added to the outputs. $F$ uses a ViT [13] backbone to encode the image $I$. We modify the diffusion model from MDM [52] for our setting. Specifically, we change the conditioning to provide image features as input. Following [52], we use a transformer encoder for the denoising (Fig. 3) and 6D representation [77] for rotation. All the 3D poses are represented in the camera coordinate frame at $t = 0$ and the predictions are also done in the camera frame at $t = 0$.

### 3.2 Lifting Model, $L$

The lifting model takes as input 2D hand keypoints & camera parameters over a sequence to output the 3D hand placement (*i.e.* wrist translation and rotation) & articulation (*i.e.* joint angles) in MANO representation in the camera frame

from the first frame. It is realized using a conditional diffusion model with a transformer backbone (Fig. 4).

**Conditioning Module.** Because 2D hand keypoints and the camera parameters are intertwined, we concatenate different representations to use as conditioning to the diffusion model: **(1) Extrinsics**: 6D rotation representation and 3D translation. **(2) Plücker rays**: These encode the camera rays joining the camera center with the 2D keypoints in the image. Specifically, let $\boldsymbol{x}_t^k = (x_t^k, y_t^k, 1)$ denote the 2D location of the $k^{\text{th}}$ hand keypoint at time step $t$ in homogeneous coordinates, $\boldsymbol{K}$ denote the camera intrinsic parameters, and $\boldsymbol{R}_t, \boldsymbol{t}_t$ denote the camera rotation and translation, such that $\boldsymbol{K}[\boldsymbol{R}_t|\boldsymbol{t}_t]\boldsymbol{X}$ maps a world point $\boldsymbol{X}$ into the camera frame. The ray joining the camera center to $\boldsymbol{x}_t^k$ is given by $\lambda \boldsymbol{R}_t^{-1}\boldsymbol{K}^{-1}\boldsymbol{x}_t^k - \boldsymbol{K}^{-1}\boldsymbol{t}_t$. We represent this ray using the 6D Plücker representation [70]. **(3) KPE encoding [41]**: This captures the location of each 2D keypoint in the field of view of the camera (with principal point $(p_x, p_y)$ and focal length $(f_x, f_y)$). For each $(x_t^k, y_t^k)$, we estimate the angles $\phi_x = \tan^{-1}((x_t^k - p_x)/f_x)$ and $\phi_y = \tan^{-1}((y_t^k - p_y)/f_y)$ and compute sinusoidal encodings.

**Training Data.** For training the lifting model, we render out 3D hand and camera trajectories in the lab datasets into 2D hand keypoints trajectories. We also introduce augmentations in the camera trajectories to increase diversity in data for training. Because of these augmentations and not using any visual information, the lifting model generalizes very well to datasets not seen during training.

### 3.3 Using Lifting Model, $L$, to Impute 3D Labels

We impute MANO labels by running the lifting model $L$ on 2D annotations in diverse datasets (AssemblyHands [37] & HoloAssist [57]). Rather than directly using the 3D output from the lifting model, we adjust the 3D output to get it to better conform to the 2D annotations. Concretely, we pass the complete 3D predictions from the lifting model to the differentiable MANO model, $\mathcal{M}$, to get the 3D hand joints, which are then projected into the image with known intrinsics to get 2D keypoints. We optimize the reprojection loss on 2D keypoints labels (either available in datasets like AssemblyHands or estimated from off-the-shelf model [38]) using gradient descent for 1000 iterations with a learning rate of 0.01 & gradient norm clipped to 1 for regularization. For datasets with 3D keypoint labels (but no MANO labels), we also add a L2 loss on 3D keypoints.

### 3.4 Implementation Details

The denoiser in both $F$ & $L$ is implemented as a transformer encoder with 16 layers, 4 heads, latent dimension of 1024 & dropout of 0.1. The ViT backbone for computing image features is initialized from [38]. Following [52], we use 1000 steps for denoising with cosine noise schedule. We also mask out the conditioning tokens (image features for $F$ and 2D keypoints + camera parameters for $L$) with probability 0.1 to simulate noise in diverse settings. For augmentation, we

add pixel-level noise & image scaling for forecasting and we jitter & scale 2D keypoints for lifting. Both $F$ & $L$ predict normalized translation value (using mean & standard deviation across all the wrist translations in the training dataset). For $L$, we find the combination of extrinsics, plucker rays & KPE to work the best. The predictions span 256 timesteps. Both $F$ & $L$ are trained in a mixed-dataset setting across 4 NVIDIA L40S or 4 A40 GPUs. Since different datasets have varying length sequences, we mask out the extra timesteps.

**Table 1: Datasets used in this work.** We train jointly on all the datasets using the available MANO labels & incomplete supervision from HoloAssist (2D keypoint labels are from off-the-shelf HaMeR [38]) & AssemblyHands. Different datasets have varying number of objects, sequences and time horizon (in seconds) and we train a single model over all these variable factors. Note that EgoExo4D is not used for training in any way and is only used for testing the *zero-shot generalization* performance of different models.

| Dataset Name | View | Lab/Wild | Labels | #Motions | #Objs | Horizon(s) | Role ($L$) | Role ($F$) |
|---|---|---|---|---|---|---|---|---|
| ARCTIC [15] | Ego | Lab | MANO | 4499 | 11 | 0.5-5 | train | train (MANO), test |
| H2O [27] | Ego | Lab | MANO | 534 | 8 | 1-7.27 | train | train (MANO), test |
| H2O3D [21] | Exo | Lab | MANO | 57 | 10 | 0.8-4.87 | train | train (MANO) |
| HOT3D [3] | Ego | Lab | MANO | 6000 | 33 | 5 | train | train (MANO), test |
| DexYCB [8] | Exo | Lab | MANO | 5743 | 20 | 1.77-1.97 | train | train (MANO), test |
| HoloAssist [57] | Ego | Wild | 2D Kps | 7461 | 120 | 0.53-8.5 | – | train ($L$(2D Kps) |
| AssemblyHands [37] | Ego | Lab | 3D + 2D Kps | 2134 | 101 | 0.53-8.5 | test | train ($L$(2D + 3D Kps)), test |
| EgoExo4D [17] | Ego | Wild | 3D Kps | 53 | – | 0.53-6.57 | – | zero-shot testing |

## 4 Experiments

We focus on (a) how does our lifting approach compare to other ways of injecting 2D supervision, (b) the quality of our imputed labels, (c) how well does our model forecast hand motion from a single image w.r.t. related past methods, (d) the effectiveness of a diffusion head over a regression head, (e) does incorporating diverse 2D labeled data help?

**Metrics**. We adopt metrics from human motion literature [18, 19, 49, 52, 62, 68] that measure the accuracy and quality of predicted motions. For accuracy, we use: (a) Mean Per Joint Position Error (M) in 3D (in cm), averaged over time, keypoints & 2 hands, (b) Mean Relative-Root Position Error (MR) in 3D (in cm) that measures the translation between the root joint of left & right hand. We also include two variants of M: (c) M-G (predictions are *globally aligned* to the ground truth before computing M) & (d) M-F (predictions are aligned to the ground truth at the *first timestep*). By doing some form of alignment, M-G & M-F focus on the accuracy of the predicted articulation. Lower is better.

For quality (forecasting task only): (a) Diversity: measures the variance over a set of motions across the full dataset of predicted or ground truth motions. Specifically, we represent motions by stacking MANO parameters over time and compute the mean pairwise L2 distance among the motions. As in [19], the predicted diversity should be comparable to that of the ground truth motions. (b) Multimodality: measures the variation within forecasted motions from the

same input, computed as the mean pairwise L2 distance between 5 samples per input (higher is better).

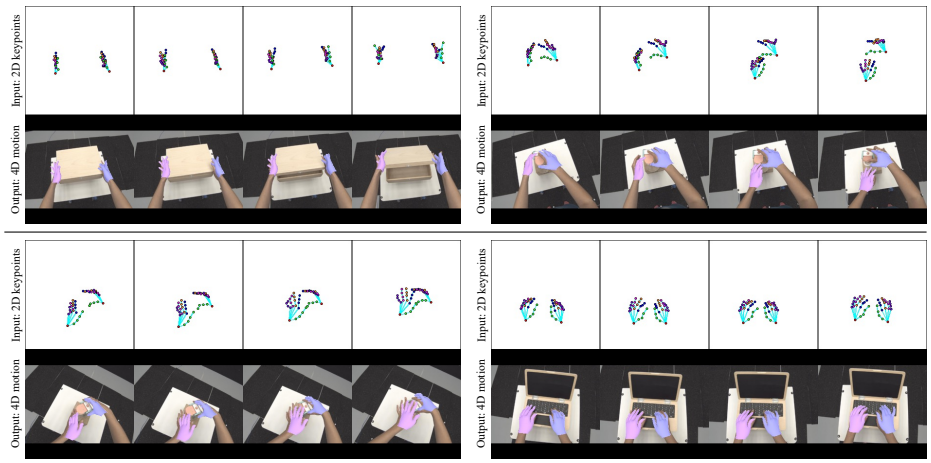


**Fig. 5: 4D hand predictions from the lifting model,** that outputs 3D MANO parameters from 2D keypoints & camera parameter inputs. We show 4 frames with the MANO mesh rendered onto the image for visualization (images are not used as input).

**Datasets**. We use 5 lab datasets: H2O [27], H2O-3D [21], ARCTIC [15] Ego, HOT3D [3], & DexYCB [8] with complete 3D annotations (*i.e.* MANO labels) but limited data diversity. For diverse images, we include HoloAssist [57] & AssemblyHands [37] (*i.e.* incomplete annotations). The 2D keypoints on HoloAssist are estimated using HaMeR [38]. AssemblyHands contain 3D & 2D keypoints but no MANO labels. We evaluate in 3 settings: (a) In-domain datasets: held-out test splits from training datasets (generalization to novel instances), (b) AssemblyHands [37]: a held-out test split that is not used for imputing labels, (c) EgoExo4D [17]: zero-shot generalization (not used for training in any way). More details about these datasets are provided in Tab. 1.

### 4.1 Lifting Results

We start by evaluating the lifting model & quality of imputed labels (Tab. 2, Tab. 4). We evaluate using 3D keypoint labels on AssemblyHands (not used for training the lifting model).

**Comparisons to existing pseudo-labeling approaches**. We consider: (a) predictions from HaMeR [38], a high-performing single-image 3D hand pose estimator, (b) HaWoR [72], a recent method that uses a data-driven motion priors to reconstruct the 3D hand motions in the world-frame using videos, (c) EasyMocap [50] that optimizes 3D MANO to conform to given 2D keypoints with temporal smoothing & pose regularization. Since EasyMocap & our method require 2D keypoints as input, we use 2D keypoint predictions from HaMeR (*vs*.

ground truth 2D keypoints) for both. We also modify EasyMocap to use ground truth camera poses in global motion initialization.

Tab. 2 shows that our lifting model produces the most accurate 3D labels across all metrics. These can be refined further via our reprojection loss to better conform to the input 2D keypoints, which improves performance further (last row in Tab. 2). Tab. 3 shows that our method outperforms EasyMocap even when EasyMocap has access to ground truth 2D keypoints. This suggests that together with a strong 2D keypoint predictor like HaMeR, our lifting model can serve as an alternative to existing annotation pipelines that use human labeling together with EasyMocap.

**Table 2:** Our Lifting model is better than other methods for estimating 3D MANO labels from videos on Assembly. We compare to: 1) a single-frame feed-forward method HaMeR [38], 2) a video-based feed-forward method HaWoR [72], 3) an optimization-based method EasyMocap [50]. The top-part of the table shows that our proposed lifting model outperforms all other methods across all metrics. Our output can be further improved via 2D refinement (last row), which leads to further gains. Here, both EasyMocap & our method make use of 2D hand keypoints predicted by HaMeR.

| **Method** | **Assembly Hands** | | | |
|---|---|---|---|---|
| | **M** | **M-G** | **M-F** | **MR** |
| HaMeR [38] | 29.7 | 6.8 | 16.1 | 36.9 |
| HaWoR [72] | 61.6 | 6.1 | 46.3 | 131.1 |
| EasyMocap [50] | 44.5 | 8.0 | 31.9 | 24.1 |
| (Ours) Lifting | **13.1** | **5.2** | **11.1** | **17.8** |
| *(Ours) Lifting + 2D Refinement* | *4.2* | *3.3* | *3.6* | *5.4* |

**Table 3:** Our lifting model alone is better than EasyMocap, even when EasyMocap has access to ground truth keypoints (row 1 *vs*. row 5). Our full method (lifting + 2D refinement) only suffers a minor degradation in performance when switching from ground truth to predicted 2D keypoints (row 6 *vs*. row 3).

| **Method** | **Assembly Hands** | | | |
|---|---|---|---|---|
| | **M** | **M-G** | **M-F** | **MR** |
| **Using ground truth 2D keypoints** | | | | |
| EasyMocap [50] | 36.3 | 7.7 | 16.8 | 15.4 |
| (Ours) Lifting | 9.6 | 5.1 | 10.0 | 11.7 |
| (Ours) Lifting + 2D refinement | **3.2** | **2.7** | **2.5** | **3.7** |
| **Using 2D keypoints predicted from HaMeR [38]** | | | | |
| EasyMocap [50] | 44.5 | 8.0 | 31.9 | 24.1 |
| (Ours) Lifting | 13.1 | 5.2 | 11.1 | 17.8 |
| (Ours) Lifting + 2D refinement | **4.2** | **3.3** | **3.6** | **5.4** |

**Table 4: Analysis of camera inputs for Lifting model** (Sec. 3.2). Different ways of encoding camera parameters help with Plücker rays being the most effective. Note that we consider ground truth 2D keypoints in this ablation.

| Method | Assembly Hands | | | |
|---|---|---|---|---|
| | M | M-G | M-F | MR |
| No camera poses | 26.3 | 8.2 | 16.1 | 15.5 |
| Extrinsics + KPE | 17.9 | 7.6 | 16.4 | 15.8 |
| Plücker rays | 14.0 | 5.5 | 12.5 | 12.6 |
| Extrinsics + Plücker + KPE | **9.6** | **5.1** | **10.0** | **11.7** |

**Table 5: Comparisons to Methods based on the State-of-the-Art**. Our ForeHand4D model improves M & MR by 36.02% compared to static pose methods, indicating significant hand movement in our testbed. ForeHand4D also outperforms LatentAct [42], a recent VAE-based approach, adapted to our setting, across all metrics. Finally, ForeHand4D also outperforms a Transformer-based Regressor trained jointly on 3D and 2D labels, where the 2D labels are injected using a re-projection loss following the current SOTA practice.

| Method | In-domain datasets | | | | AssemblyHands | | | | EgoExo4D (Zero-shot) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | M-G | M-F | MR | M | M-G | M-F | MR | M | M-G | M-F | MR |
| Static Pose (trained in our setting) | 23.3 | 8.4 | 15.4 | 16.8 | 29.8 | 8.9 | 16.1 | 26.2 | 28.8 | 13.5 | 19.2 | 18.9 |
| Static Pose (from HaMeR [38]) | 26.8 | 8.5 | 15.5 | 18.5 | 31.9 | 9.0 | **14.2** | 38.8 | 32.7 | 13.6 | **18.3** | 29.8 |
| LatentAct [42] (adapted for our task) | 17.7 | 7.4 | 16.2 | 17.6 | 21.3 | 9.1 | 17.2 | 22.8 | 26.4 | 13.5 | 19.5 | 49.9 |
| Transformer Regressor (3D + 2D sup.) | **15.1** | 6.9 | 14.6 | 14.9 | 27.5 | 8.5 | 18.4 | 19.3 | 28.9 | **13.0** | 21.3 | 19.0 |
| (Ours) ForeHand4D | 17.1 | **6.5** | **15.3** | **13.5** | **20.3** | **8.4** | 15.4 | **16.5** | **18.8** | 13.2 | 18.9 | **13.5** |

**Ablations for lifting model.** In Tab. 4 we see that conditioning only on 2D keypoints performs poorly. Injecting camera extrinsics (rotation, translation) & intrinsics via angular encoding of 2D keypoints (KPE), in Row 2, helps quite a bit. Using Plücker rays in Row 3 also provides benefits. Our final model that uses all the different encodings together, performs the best. To maximally isolate the impact of these different strategies, these ablations use ground truth 2D keypoints (instead of predicted 2D keypoints as in Tab. 2 & Tab. 3).
**Qualitative Visualizations.** Fig. 5 shows examples of 2D to 3D lifting achieved by our model. The lifting model accurately places and articulates the hands.

### 4.2 Forecasting Results

Since there is no prior work that tackles this problem, we adapt recent work LatentAct [42] to work in our setting and construct several baselines:
• **Static Pose Baseline** assumes a stationary hand & uses 3D hand pose estimates on the input image as the forecast. We consider 2 variants: training a pose predictor in our setting and using outputs from off-the-shelf HaMeR [38] (a high performing model trained on 10 datasets).
• **Transformer Regressor** uses the same architecture as our model but directly regresses the future hand motion & articulation. We consider 3 variants. **3D sup.** is trained with the same 3D ground truth as ForeHand4D. **3D sup.** +

**2D sup.** also uses 2D supervision via a reprojection loss on the predicted 3D (following [38, 43]). This is only possible because the model directly regresses the 3D output. We jointly train on 5 datasets with 3D labels & 2 datasets with 2D labels. Finally, **3D sup. + 2D sup. + Imputed 3D.** uses additional 3D imputed labels on top of the previous *3D sup. + 2D sup.*

• **LatentAct [42]** takes an image, text, contact point & an interaction codebook (represented as the latent space of a VQVAE) as input to predict future 3D hand & contact trajectory for a single hand. We adapt LatentAct to take only a single image as input and retrain it in our setting.

• **(Ours) ForeHand4D** We consider 2 variants of our ForeHand4D model, based on the supervision used. **3D sup.** is only trained on 5 datasets with 3D labels. **3D sup. + 2D sup.** is our final model that is trained jointly on 5 datasets with 3D labels & the imputed 3D labels from our lifting model.

• We compare to other ways of using 2D labels with a diffusion model. This amounts to attaching another head to make 2D predictions so that the image backbone also gets gradients from 2D labels. We consider: **2D Regression Head** & **2D Diffusion Head** (denoising is done for 2D keypoints).

**Table 6: How to best incorporate 2D labels when training a diffusion model?** Labels from lifting model (Row 4) generalize better than alternatives based on attaching auxiliary 2D forecasting heads either diffusion (Row 3) or regression (Row 2).

| Method | In-domain datasets | | | | AssemblyHands | | | | EgoExo4D (Zero-shot) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | M-G | M-F | MR | M | M-G | M-F | MR | M | M-G | M-F | MR |
| No Additional 2D Supervision | **14.8** | **5.9** | **13.3** | **12.4** | 27.9 | **8.4** | 18.1 | 18.1 | 24.0 | **13.0** | 20.9 | 19.6 |
| Inject 2D Sup. via a 2D Regression Head | 15.5 | 6.4 | 14.0 | 13.8 | 26.8 | 8.5 | 16.6 | 16.9 | 25.9 | **13.0** | 21.2 | 16.6 |
| Inject 2D Sup. via a 2D Diffusion Head | 16.2 | 6.1 | 13.8 | 15.0 | 31.9 | **8.4** | 18.2 | 19.8 | 24.8 | 13.2 | **18.3** | 16.6 |
| (Ours) Inject 2D Sup. via Imputed Labels | 17.1 | 6.5 | 15.3 | 13.5 | **20.3** | **8.4** | **15.4** | **16.5** | **18.8** | 13.2 | 18.9 | **13.5** |

**Static pose results**. The large values of M & MR for static pose methods indicate that there is indeed a significant hand movement across timesteps since M & MR are translation-focused metrics. Our ForeHand4D model leads to gains of 36.02% on M & MR across all settings. HaMeR scores the highest on M-F in EgoExo4D / Assembly, likely because it is trained across diverse images from 10 datasets.

**Comparison with LatentAct**. ForeHand4D outperforms the recent VQVAE-based LatentAct [42] (adapted to our task) across all metrics. We see benefits of 16.4% using our ForeHand4D model with MR gaining the most. This is likely due to LatentAct requiring additional inputs, *i.e.*, contact points & text, to better place the predicted motion in 3D space, which are not available in our setting.

**Peformance *vs*. Transformer Regressor.** ForeHand4D also outperforms on 11/12 metrics, a Transformer-based regression approach that is trained using current SOTA practice of jointly training on 3D and 2D labels, where the 2D labels are incorporated via a re-projection loss. Gains are particularly large in M & MR in the zero-shot setting.

**Injecting 2D supervision improves performance on novel datasets for ForeHand4D.** Comparing Rows 1 & 4 in Tab. 6, we see that 2D supervision leads to large improvement in metrics on Assembly & EgoExo4D. Notably, M,

M-F, & MR improve by 10 – 30%. We also find that injecting 2D supervision mildly hurts performance on in-domain datasets (Row 1 vs Row 4) for both models. We believe this is because the same model now has to learn a much broader distribution than what is tested in the in-domain datasets: ForeHand4D may also suffer because the imputed labels are not perfect (Tab. 2). Nevertheless, injecting 2D labels helps by a lot on datasets without complete 3D annotations (*e.g.* Assembly, EgoExo4D). We did not see a clear benefit by injecting imputed 3D labels on top of the 2D reprojection loss in Transformer regressor.

**Table 7: Breakdown of translation and articulation metrics in forecasting**. **(Row 1):** Static GT Articulation at $t = 0$ + GT wrist pose at $t = t$ **(Row 2):** GT Articulation at $t = t$ + Static GT wrist pose at $t = 0$ **(Row 3):** Full static GT pose, *i.e.* Static GT Articulation at $t = 0$ + Static GT wrist pose at $t = 0$ **(Row 4):** Full static predicted pose (model trained on our datasets), *i.e.* Static Predicted Articulation at $t = 0$ + Static Predicted wrist pose at $t = 0$. **(Row 5):** Same as Row 4, but predictions come from off-the-shelf HaMeR [38]. These highlight that translation constitutes a significant part of the metrics & EgoExo4D involves much more dexterous actions.

| Method | Articulation | Wrist Pose | In-domain datasets | | | | AssemblyHands | | | | EgoExo4D (Zero-shot) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | M | M-G | M-F | MR | M | M-G | M-F | MR | M | M-G | M-F | MR |
| | GT at $t = 0$ | GT at $t = t$ | 4.5 | 3.0 | 4.6 | 2.3 | 5.0 | 3.6 | 5.9 | 3.4 | 15.1 | 10.5 | 25.1 | 2.6 |
| | GT at $t = t$ | GT at $t = 0$ | 13.7 | 6.7 | 13.6 | 12.0 | 13.4 | 7.4 | 13.3 | 15.5 | 23.0 | 13.3 | 18.5 | 19.1 |
| | GT at $t = 0$ | GT at $t = 0$ | 15.4 | 8.2 | 15.2 | 12.0 | 14.1 | 8.6 | 13.9 | 15.5 | 22.7 | 13.5 | 18.2 | 19.1 |
| Predictor (trained on our datasets) | Pred at $t = 0$ | Pred at $t = 0$ | 23.3 | 8.4 | 15.4 | 16.8 | 29.8 | 8.9 | 16.1 | 26.2 | 28.8 | 13.5 | 19.2 | 18.9 |
| Predictions from HaMeR [38] | Pred at $t = 0$ | Pred at $t = 0$ | 26.8 | 8.5 | 15.5 | 18.5 | 31.9 | 9.0 | 14.2 | 38.8 | 32.7 | 13.6 | 18.3 | 29.8 |

**Proposed lifting scheme *vs*. alternatives.** In Tab. 6, for the diffusion model, the gains are marginal when injecting supervision via a 2D regression head or 2D diffusion head, obvious ways that are typically used with non-diffusion based models. However, imputing labels via our lifting model is quite effective & greatly improves M & MR.

**Analysis of forecasting metrics**. In Tab. 7, we analyze the impact of translation & articulation on the metrics by using ground truth (GT) articulation & wrist poses in different ways: **Row 1**: This measures how much the hand articulation changes over the motion. **Row 2**: This measures how much the wrist translates with respect to the first time step. **Row 3**: This considers changes in both articulation and translation as the motion progresses. **Row 4**: Pose predictions for the given frame, copied over as the forecast for all future frames. Here predictions are coming from a model hand pose predictor trained on our datasets. **Row 5**: Same as Row 4, but predictions come from off-the-shelf HaMeR [38].

This analysis highlights that translation constitutes a significant part of the metrics and EgoExo4D involves much more dexterous actions compared to lab datasets. For evaluations involving GT poses, the M and M-F values should ideally be the same (as is the case with in-domain lab datasets) since the pose at first timestep is the same. However, that is not the case with AssemblyHands and EgoExo4D since they often contain invalid or missing labels for several joints due to which SVD does not converge during the procrustus alignment.

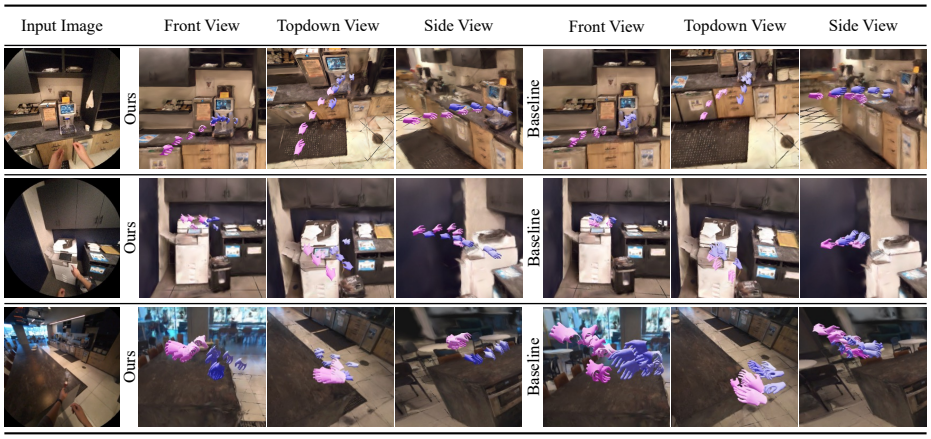


**Fig. 6: Qualitative comparison: ForeHand4D vs. Transformer Regressor (3D + 2D sup.)** *Zero-shot* forecasts on unseen Aria images from [65]. Our predictions span longer trajectories, are smoother, better placed in the scene & more plausible compared to the best baseline. The predicted motions highlight object interactions, *e.g.* reaching for coffeemachine or the table. Left hand is shown in pink, right hand in blue. Color saturation decreases as time proceeds, *i.e.* further away timesteps in future are in lighter shades. We show the predicted motion in the 3D scene from 3 different views (3D scene is only for visualization & these images are not seen during training).

**Diversity & multi-modality**. We compute these standard motion quality metrics [19, 52] on ARCTIC, which contains ground truth MANO labels. The diversity score for the ground truth distribution is 39.16. The predicted motion distribution of ForeHand4D (diversity = 41.14) is a lot closer to the ground truth distribution compared to LatentAct (diversity = 302.45) & transformer regressor (diversity = 187.82). We also observe better mutimodality score for our model (19.64 *vs*. 13.04 for LatentAct). Fig. 7 shows examples of multiple forecasts from the same input, *e.g.* reaching & grabbing motions towards different objects.

**Performance trends over time**. In Fig. 8, we see M (MPJPE) does not start from 0. This is because the model finds it hard to precisely predict the hand translation in the given frame (likely due to scale ambiguity in predicting metric 3D from a single image). M-F, where we factor out this imperfection by aligning to the ground truth hand in the first frame, shows a clear increasing trend.

**Qualitative comparisons.** We visualize the predicted motions for both our model & Transformer Regressor (3D + 2D sup.) in Fig. 6 on Aria images from [65]. Our motion predictions span longer trajectories, are smoother & better placed in the scene compared to the baseline. The predicted motions also highlight object interactions, *e.g.* reaching for coffeemachine or the table. See supplementary for more visualizations and analysis.

**Failure modes.** In egocentric videos, hands often go out of view as camera moves. This makes lifting hard since there is no input 2D keypoints for the model to use. For forecasting, heavy occlusions in the input image often leads to

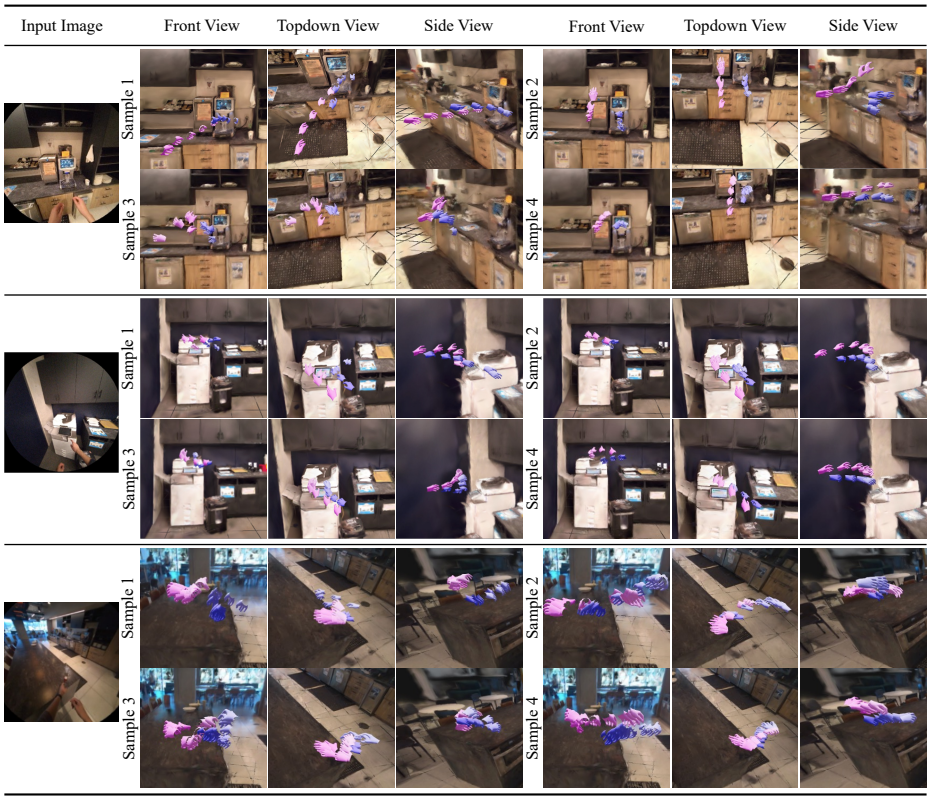


**Fig. 7: Different forecasts from the same input**. We show 4 samples for 2 images from ForeHand4D, indicating different modes of interaction like reaching & grabbing towards different objects in the scene, *e.g.* coffee machine, paper on the table, printer. We show the predicted motion in the 3D scene from 3 different views (3D scene is only for visualization & these images are not seen during training).

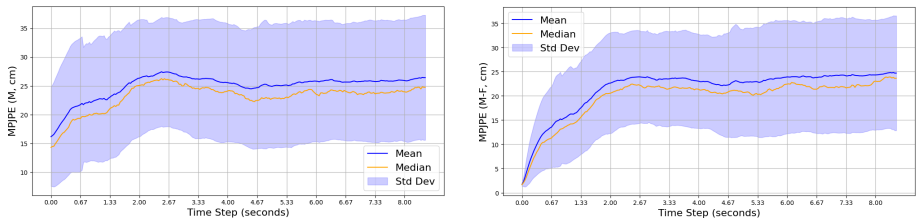


**Fig. 8: Performance trends over time.** Forecasting gets harder for longer prediction horizons for both (top) M & (bottom) M-F (Standard deviation is computed per time-step after aggregating errors from 5 generated motions for each sample).

forecasted hand being incorrectly placed. As we are forecasting future motion, errors in the near timesteps compound over time (as seen in Fig. 8).

## 5 Conclusion

We present a system for forecasting bimanual 3D hand motion from a single image in everyday settings. Our forecasting model consists of a conditional diffusion model trained with different types of supervision: 3D labels in lab datasets & imputed 3D labels from diverse datasets using our lifting model. Our predictions span longer horizon, are smoother, better placed & capture multiple interaction modes, especially in zero-shot generalization settings. While we consider single image inputs for generality, incorporating context, *e.g.* past frames or intent, as additional inputs to the forecasting model could be useful. Lastly, object motion is also an important aspect of interaction & is relevant for future work.

**Acknowledgements:** We thank Arjun Gupta, Shaowei Liu, Anand Bhattad & Kashyap Chitta for feedback on the draft, and David Forsyth for useful discussion. This material is based upon work supported by NSF (IIS2007035), NASA (80NSSC21K1030), DARPA (Machine Common Sense program), Amazon Research Award, NVIDIA Academic Hardware Grant, and the NCSA Delta System (supported by NSF OCI 2005572 and the State of Illinois). Richard Li was supported by the NSF Institute for Artificial Intelligence and Fundamental Interactions (Grant No. PHY-2019786) and the Felicis Scholars program.

## References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
2. Bai, W., Wang, Y., Chen, W., Sun, H.: An expectation-maximization algorithm for training clean diffusion models from corrupted observations. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
3. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., Engel, J.J., Hodan, T.: Introducing hot3d: An egocentric dataset for 3d hand and object tracking. arXiv: 2406.09598 (2024)
4. Bao, W., Chen, L., Zeng, L., Li, Z., Xu, Y., Yuan, J., Kong, Y.: Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
5. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T., Yuan, J., Magnenat-Thalmann, N.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
6. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)

7. Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2hoi: Text-guided 3d motion generation for hand-object interaction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
8. Chao, Y., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Wyk, K.V., Iqbal, U., Birchfield, S., Kautz, J., Fox, D.: Dexycb: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
9. Cheng, W., Tang, H., Gool, L.V., Ko, J.H.: Handdiff: 3d hand pose estimation with diffusion on image-point cloud. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
10. Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., Tekin, B.: Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
11. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
12. Daras, G., Shah, K., Dagan, Y., Gollakota, A., Dimakis, A., Klivans, A.R.: Ambient diffusion: Learning clean distributions from corrupted data. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (NeurIPS) (2023)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
14. Fan, Z., Ohkawa, T., Yang, L., Lin, N., Zhou, Z., Zhou, S., Liang, J., Gao, Z., Zhang, X., Zhang, X., Li, F., Liu, Z., Lu, F., Zeid, K.A., Leibe, B., On, J., Baek, S., Prakash, A., Gupta, S., He, K., Sato, Y., Hilliges, O., Chang, H.J., Yao, A.: Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
15. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
16. Gavryushin, A., Redhardt, F., Lorenzo, G.D., Gool, L.V., Pollefeys, M., Mo, K., Wang, X.: SIGHT: single-image conditioned generation of hand trajectories for hand-object interaction. arXiv **2503.22869** (2025)
17. Grauman, K., et al.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
18. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
19. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2021–2029 (2020)
20. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

21. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
22. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer (2009)
23. Hatano, M., Zhu, Z., Saito, H., Damen, D.: The invisible egohand: 3d hand forecasting through egobody pose estimation. arXiv preprint arXiv:2504.08654 (2025)
24. Hewitt, C., Saleh, F., Aliakbarian, S., Petikam, L., Rezaeifar, S., Florentin, L., Hosenie, Z., Cashman, T.J., Valentin, J., Cosker, D., Baltrusaitis, T.: Look ma, no markers: holistic performance capture without the hassle. ACM Transactions on Graphics **43** (2024)
25. Iqbal, U., Molchanov, P., Breuel, T.M., Gall, J., Kautz, J.: Hand pose estimation via latent 2.5d heatmap regression. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
26. Kapon, R., Tevet, G., Cohen-Or, D., Bermano, A.H.: MAS: multi-view ancestral sampling for 3d motion generation using 2d diffus,ion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
27. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
28. Li, J., Liu, C.K., Wu, J.: Lifting motion to the 3d world via 2d diffusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
29. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
30. Li, W., Liu, H., Tang, H., Wang, P., Gool, L.V.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
31. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
32. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
33. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
34. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops) (2022)
35. Moon, G., Yu, S., Wen, H., Shiratori, T., Lee, K.M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single RGB image. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
36. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
37. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In: Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12999–13008 (2023)
38. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3D with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
39. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
40. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. arXiv preprint arXiv:2409.12259 (2024)
41. Prakash, A., Gupta, A., Gupta, S.: Mitigating perspective distortion-induced shape ambiguity in image crops. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
42. Prakash, A., Lundell, B.E., Andreychuk, D., Forsyth, D., Gupta, S., Sawhney, H.S.: How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
43. Prakash, A., Tu, R., Chang, M., Gupta, S.: 3d hand pose estimation in everyday egocentric images. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
44. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3d human pose from 2d image landmarks. In: Proceedings of the European Conference on Computer Vision (ECCV) (2012)
45. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG) (2017)
46. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics **36**(6), 245:1–245:17 (2017)
47. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3D hand and body motion capture by regression and integration. Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (2021)
48. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-STMO: pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
49. Shin, S., Kim, J., Halilaj, E., Black, M.J.: WHAM: reconstructing world-grounded humans with accurate 3d motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
50. Shuai, Q., Fang, Q., Dong, J., Peng, S., Huang, D., Bao, H., Zhou, X.: Easymocap - make human motion capture easier. Github (2021), `https://github.com/zju3dv/EasyMocap`
51. Simo-Serra, E., Quattoni, A., Torras, C., Moreno-Noguer, F.: A joint model for 2d and 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
52. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
53. Tomè, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

54. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2626–2634 (2017)
55. Wandt, B., Little, J.J., Rhodin, H.: Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
56. Wang, J., Yan, S., Xiong, Y., Lin, D.: Motion guided 3d pose estimation from videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
57. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., Joshi, N., Pollefeys, M.: Holoassist: an egocentric human interaction dataset for interactive AI assistants in the real world. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023)
58. Wen, Y., Pan, H., Yang, L., Pan, J., Komura, T., Wang, W.: Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
59. Xu, H., Li, H., Wang, Y., Liu, S., Fu, C.: Handbooster+: Boosting 3d hand-mesh reconstruction from data synthesis to progressive multi-hypothesis aggregation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2025)
60. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. Advances in neural information processing systems **29** (2016)
61. Yang, J., Ma, Y., Zuo, X., Wang, S., Gong, M., Cheng, L.: 3d pose estimation and future motion prediction from 2d images. Pattern Recognition **124**, 108439 (2022)
62. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
63. Ye, Y., Feng, Y., Taheri, O., Feng, H., Tulsiani, S., Black, M.J.: Predicting 4d hand trajectory from monocular videos. In: Proceedings of the International Conference on 3D Vision (3DV) (2025)
64. Ye, Y., Feng, Y., Taheri, O., Feng, H., Tulsiani, S., Black, M.J.: Predicting 4d hand trajectory from monocular videos. arXiv preprint arXiv:2501.08329 (2025)
65. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
66. Yi, B., Ye, V., Zheng, M., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. arXiv preprint arXiv:2410.03665 (2024)
67. Yu, Z., Zafeiriou, S., Birdal, T.: Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. arXiv preprint arXiv:2412.12861 (2024)
68. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: GLAMR: global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
69. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence **45** (2023)

70. Zhang, J.Y., Lin, A., Kumar, M., Yang, T., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
71. Zhang, J., Zhang, Y., An, L., Li, M., Zhang, H., Hu, Z., Liu, Y.: Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. arXiv preprint arXiv:2409.09300 (2024)
72. Zhang, J., Deng, J., Ma, C., Potamias, R.A.: HaWoR: World-space hand motion reconstruction from egocentric videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
73. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
74. Zhang, W., Dabral, R., Golyanik, V., Choutas, V., Alvarado, E., Beeler, T., Habermann, M., Theobalt, C.: Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. arXiv preprint arXiv:2412.05066 (2024)
75. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
76. Zhou, B., Zhan, Y., Zhang, Z., Lu, Z.: Megohand: Multimodal egocentric hand-object interaction motion generation. arXiv preprint arXiv:2505.16602 (2025)
77. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
78. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single RGB images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)