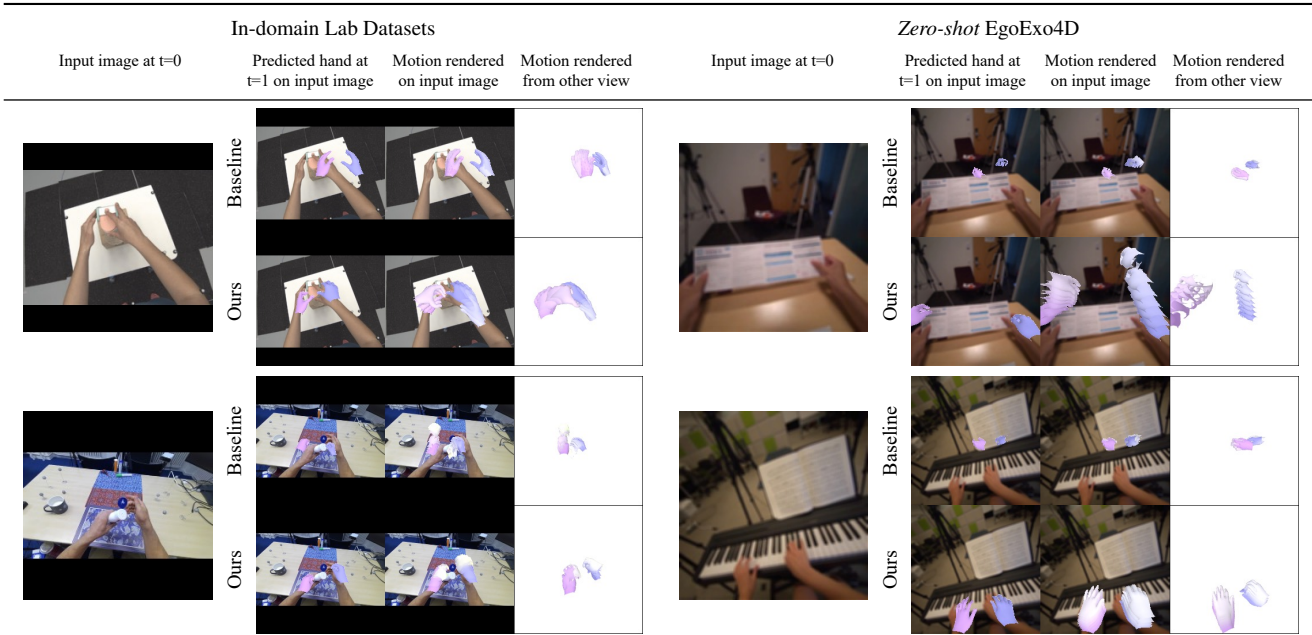# Bimanual 3D Hand Motion and Articulation Forecasting in Everyday Images

Aditya Prakash   David Forsyth   Saurabh Gupta

University of Illinois, Urbana-Champaign

https://bit.ly/ForeHand4D

Figure 1. ForeHand4D forecasts bimanual 3D hand motion from single RGB image input: (left) on 2 lab datasets (ARCTIC, H2O), (right) *zero-shot* forecasts on challenging EgoExo4D. 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. We render the predicted motion on the input image & from another view. Our predictions span longer trajectories, are smoother & better placed in the scene compared to the baseline, especially on everyday images from EgoExo4D.

Abstract

We tackle the problem of forecasting bimanual 3D hand motion & 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 across 6 datasets show the benefits of training on diverse data with imputed labels (14% improvement) and effectiveness of our lifting (42% better) & forecasting (16.4% gain) models, over the best baselines, especially in zero-shot generalization to everyday images.

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 output 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, and completeness of output; thereby improving the utility of such models for downstream human robot interaction & AR/VR applications. Fig. 1 shows sample outputs from ForeHand4D, including on images from EgoExo4D not used for training in any way.

Forecasting hand motion is difficult because of the complex ways in which hands interact with one another and
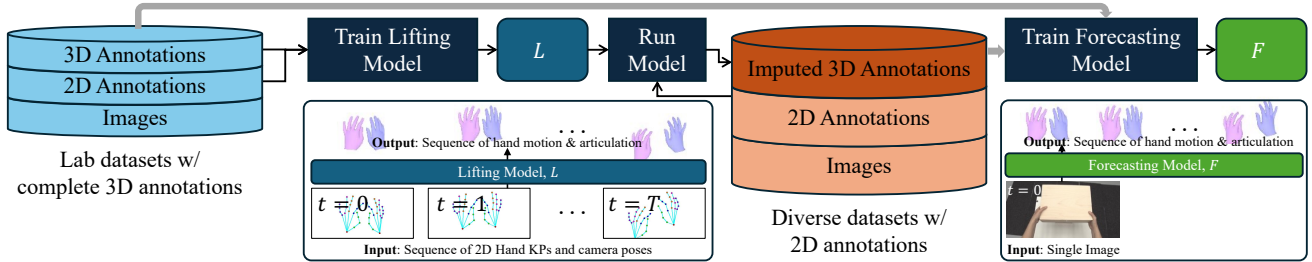
Figure 2. Overall Training Pipeline. We first use the 2D & 3D annotations in lab datasets to train a lifting diffusion model, L that maps 2D keypoints sequences to 3D MANO hands. We then run L on diverse datasets with 2D annotations to generate 3D annotations. Finally, the forecasting model F is trained on lab & diverse datasets with complete 3D supervision.

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 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. 3, 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 [43] 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 [35, 51, 56] 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 [47] ,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. 3).

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 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. 3), and outperforms LatentAct [40] (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 [66] (Tab. 1). Code & models will be released upon publication.

2. Related Work

3D Hand Prediction from Images and Videos. Given image or video input, many recent works make *predictions* (and not forecasts) for hands presents in them from egocentric [13, 14, 24, 25] and exocentric observations [18, 19, 28, 31, 32, 37, 44, 60, 63]. HaMeR [35] is a high-performing recent work that makes 3D hand predictions from single images, while Dyn-HaMR [61] makes temporally consistent 3D hand predictions from video input. Predictions on videos are made via feed forward model [59], test-time optimization [66] or hybrid approaches [21].

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, 40] produce hand motion, articulation & contact maps, Bao *et al.* [4] forecast the 3D wrist location, while Liu *et al.* [29] forecast only the 2D wrist location. On the input side, [4, 6, 15, 29, 57] only use RGB images as input, while [7, 9, 40, 65, 68] 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, 57], single hand-object interaction [4, 15, 40] or bimanual interactions [7, 9, 29, 65, 68]. 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.

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 [41], probabilistic models [48], directly optimizing 3D poses [1],
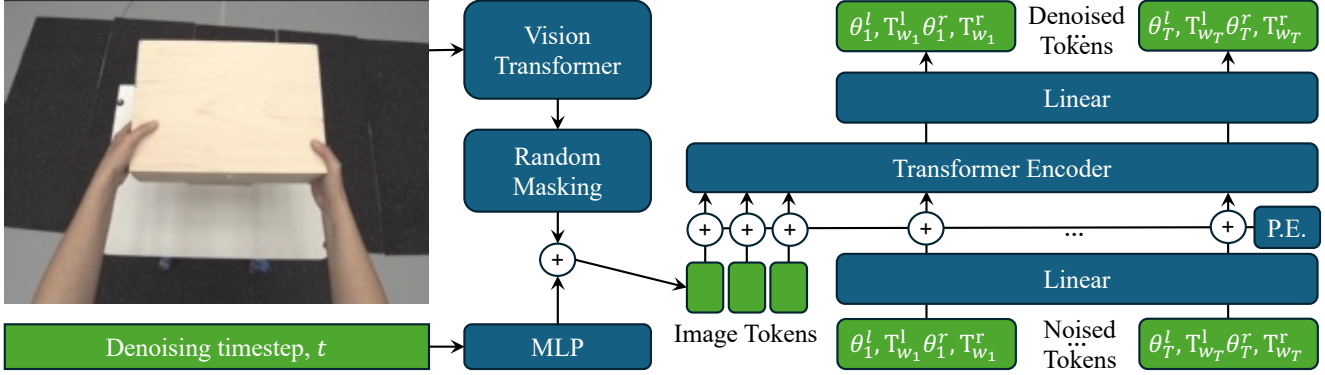
Figure 3. Architecture for Forecasting Model. We modify MDM [49] to condition on images features extracted from a ViT backbone. Each input & output token is 198-dimensional: 2 hands × (16 (joints) × (6 (6D rotation for each joint) + 3 (wrist translation))).

MLP [30, 33], convolutional [33, 36, 50], graph-based [5, 10, 53], transformer [27, 45, 64, 69], diffusion [23, 26] & normalizing flows [52]. These works operate on different types of inputs, *e.g.* static 2D pose [1, 30, 52], 2D pose estimated from image [33, 48], seqence of 2D keypoints [5, 23, 45, 53, 64, 69] or multi-view 2D poses [26]. These ideas have also been extended to estimated 3D hand poses using 2D keypoints [71] or 2D/2.5D heatmaps [19, 22]. 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 [35, 51, 56]: the predicted 3D is differentiable 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 $\nabla_{\mathbf{x}} p(\mathbf{x})$ 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 $\nabla_{\mathbf{x}} p(\mathbf{x})$ but only a projection of it), we don't know what point $\mathbf{x}$ in space is this partial supervision for. Recent works have explored training diffusion models on corrupted or partial data [2, 11] using EM [20] or aggressive masking [55]. However, our setting is different because we don't quite have partial ground truth, but rather a projection of the 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 [42], consisting of the shape β, articulation θ & 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 $\Phi_t = \{(\theta_t^l, {}_cT_{w_t}^l, \theta_t^r, {}_cT_{w_t}^r)\}$ for all timesteps in the prediction horizon, where l & r superscripts denote left & right hand. We do not predict β (we use the mean β shape

from the MANO model when β is not available).

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 [13], H2O [24], H2O-3D [19], 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 [12] backbone to encode the image I. We modify the diffusion model from MDM [49] for our setting. Specifically, we change the conditioning to provide image features as input. Following [49], we use a transformer encoder for the denoising (Fig. 3) and 6D representation [70] 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 & articulation 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
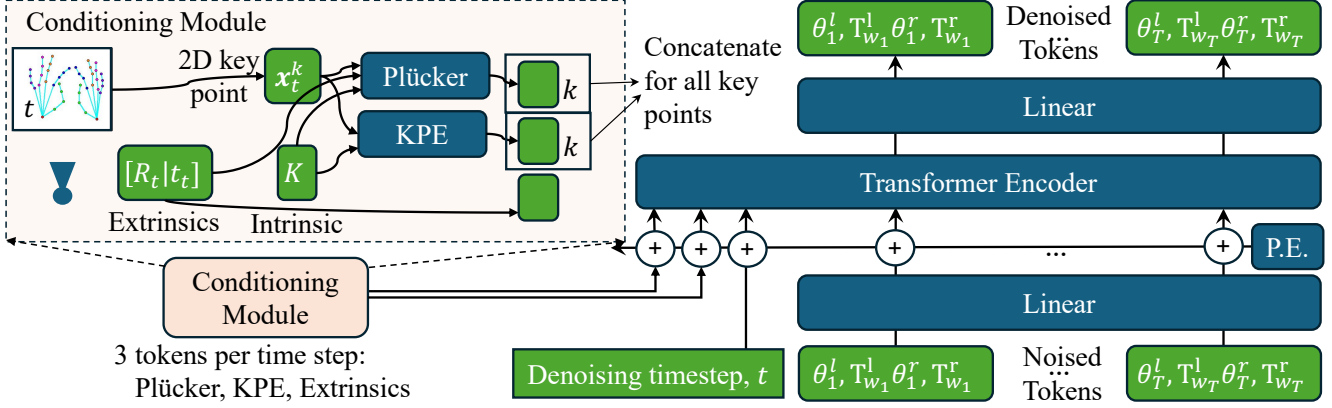
Figure 4. Architecture for Lifting Model. We modify MDM [49] to condition on a sequence of 2D hand keypoints & camera parameters. The conditioning module combines different input representations: 3D pose (rotation, translation) of camera, Plücker rays [67] & KPE [38].

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 kth 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 [67]. **(3) KPE encoding [38]**: 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 for Training the Forecasting Model, F

We impute MANO labels on diverse datasets by running the lifting model L on 2D annotations in diverse datasets (AssemblyHands [34] & HoloAssist [54]). 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 [35]) 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 [35]. Following [49], 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.

4. Experiments

Our experiments test (a) the difficulty of the forecasting task, (b) how well does our model forecast hand motion & articulation from a single image w.r.t. related past methods, (c) the effectiveness of a diffusion head over a regression head, d) does incorporating diverse 2D labeled data help performance, (e) how does our lifting approach compare to other ways of injecting 2D supervision, (f) the quality of our imputed labels, and (g) what design decisions matter?

Metrics. We adopt metrics from human motion literature [16, 17, 46, 49, 58, 62] 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
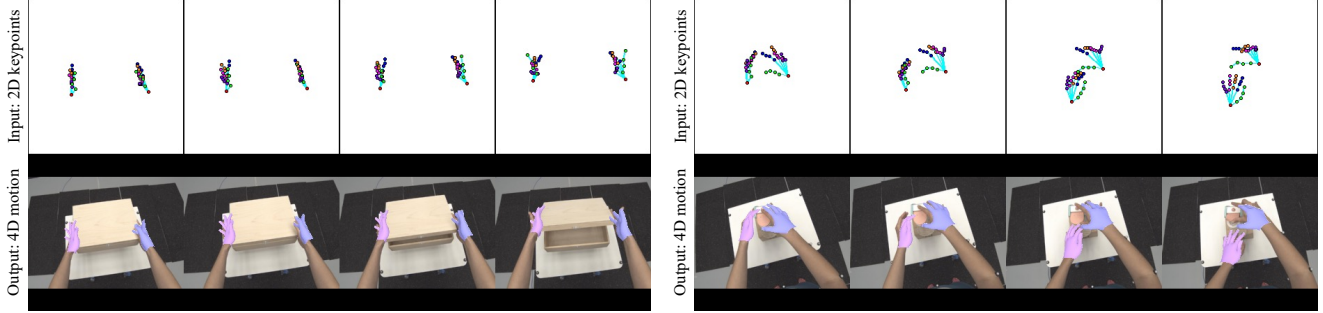
Figure 5. 4D hand predictions from the lifting model, that predicts 3D MANO parameters from 2D keypoints & camera parameter inputs. We show 4 frames with the MANO mesh rendered onto the image for visualization (images are not used as input).

Method	Assembly Hands			
	M	**M-G**	**M-F**	**MR**
HaMeR [35]	29.7	6.8	16.1	36.9
EasyMocap [47]	36.3	7.7	16.8	15.4
HaWoR [66]	61.6	6.1	46.3	131.1
(Ours) Lifting	**9.6**	**5.1**	**10.0**	**11.7**
Ours + 2D refinement	3.2	2.7	2.5	3.7

Table 1. Our Lifting model is better than other methods for estimating 3D MANO labels for video trajectories on Assembly.

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 2. Analysis of input representation for Lifting model (Sec. 3.2). Different ways of encoding camera parameters help with Plücker rays being the most effective.

4.1. Lifting Results

We start by evaluating the lifting model & quality of imputed labels (Tab. 1, Tab. 2). We evaluate using 3D keypoint labels on AssemblyHands (not used for training the lifting model).

Comparisons to existing pseudo-labeling approaches. We consider 3 alternatives: (a) predictions from HaMeR [35], a high-performing 3D hand pose estimator, (b) EasyMocap [47]: optimizes 3D MANO to conform to given 2D hand keypoints with temporal smoothing & pose regularization, and (c) HaWoR [66], a recent method that uses a data-driven motion priors to reconstruct the 3D hand motions in the world-frame. Tab. 1 shows that our lifting model produces the most accurate pseudo MANO labels across all metrics. These can be refined further using a 2D reprojection loss with the input 2D keypoints. For fair comparisons, we modified EasyMocap to use ground truth camera poses in global motion initialization, and provide the same keypoints (as tracks) to HaWoR that are used by our lifting model.

Ablations for design of the lifting model. In Tab. 2 we see that conditioning only on 2D keypoint trajectories performs poorly. Injecting camera extrinsics (rotation, translation) & intrinsics via angular encoding of 2D keypoints (KPE), in Row 2, helps quite a bit. Using the Plücker rays in Row 3 also provides benefits. Our final model that uses all the different encodings together, performs the best.

Qualitative Visualizations. Fig. 5 shows qualitative examples of the 2D to 3D lifting achieved by our model. The lifting model accurately places and articulates the hands.

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 compute the mean pairwise L2 distance (in the MANO space) among the motions. As in [16], 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).

Datasets. We use 5 lab datasets: H2O [24], H2O-3D [19], ARCTIC [13] Ego, HOT3D [3], & DexYCB [8] with complete 3D annotations (*i.e.* MANO labels) but limited data diversity. For diverse images, we include HoloAssist [54] & AssemblyHands [34] (*i.e.* incomplete annotations). The 2D keypoints on HoloAssist are estimated using HaMeR [35]. 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: a held-out test split that is not used for imputing labels, (c) EgoExo4D: zero-shot generalization (not used for training in any way).

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 [35])	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 [40] (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 sup.)	15.3	6.7	14.7	14.8	30.0	8.6	21.6	25.5	29.2	**12.9**	21.7	20.9
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 (3D sup.)	**14.8**	**5.9**	**13.3**	**12.4**	27.9	8.4	18.1	18.1	24.0	13.0	20.9	19.6
(Ours) ForeHand4D (3D + 2D sup.)	17.1	6.5	15.3	13.5	**20.3**	**8.4**	15.4	**16.5**	**18.8**	13.2	18.9	**13.5**

Table 3. Baseline comparisons. Our ForeHand4D model improves M & MR by 36.02% compared to static pose methods, indicating significant hand movement in our setting. Compared to the transformer regressor baseline (Row 4 vs Row 6), we see improvements in 11/12 metrics. Adding weak supervision from 2D labels leads to further gains in M & MR, especially in the zero-shot generalization to EgoExo4D.

Method	Diversity	Multimodality
Reference (ground truth distribution)	*39.16*	*N/A*
Transformer Regressor	187.82	N/A
LatentAct [40] (adapted for our task)	302.45	13.04
(Ours) ForeHand4D	**41.14**	**19.64**

Table 4. ForeHand4D produces multimodal output that better matches the diversity of the ground truth trajectories on ARCTIC.

4.2. Forecasting Results

Since there is no prior work that tackles this problem, we adapt recent work LatentAct [40] 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 [35] (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 2 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 [35, 39]). 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.

• **LatentAct [40]** 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 Fore-Hand4D 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).

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.

Diffusion outperforms transformer-based regressor. Comparing Row 6 vs Row 4 in Tab. 3 we see improvements in 11/12 metrics. Gains are particularly large in M & MR, especially in zero-shot setting. Moreover, we can sample different plausible forecasts from our ForeHand4D (Fig. 6).

Injecting 2D supervision improves performance on novel datasets. Comparing Rows 6 & 7 in Tab. 3, we see that injection of 2D supervision leads to large improvement in metrics on Assembly & EgoExo4D. Notably, M, M-F, & MR improve by 10 – 30%. The Transformer Regressor baseline benefits less from these additional 2D labels (Row 4 vs Row 5) and overall ForeHand4D outperforms it on 7/8 metrics (Row 4, 5 vs Row 7). This suggests a) the utility of incorporating weak supervision in ForeHand4D, and b) the effectiveness of our proposed scheme in doing so.

Also, we find that injecting 2D supervision mildly hurts performance on in-domain datasets (Row 6 vs Row 7, Row 4 vs Row 5) 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: 80 objects *vs*. 300 objects and from many different viewpoints & cameras. ForeHand4D may also suffer because the imputed labels are not perfect (Tab. 1). Nevertheless, injecting 2D labels helps by a lot on datasets without complete 3D annotations.

Comparison with LatentAct. We see benefits of 16.4% using our ForeHand4D model with MR gaining the most. This is likely due to LatentAct requiring additional inputs, in the form of contact points & text to better place the predicted
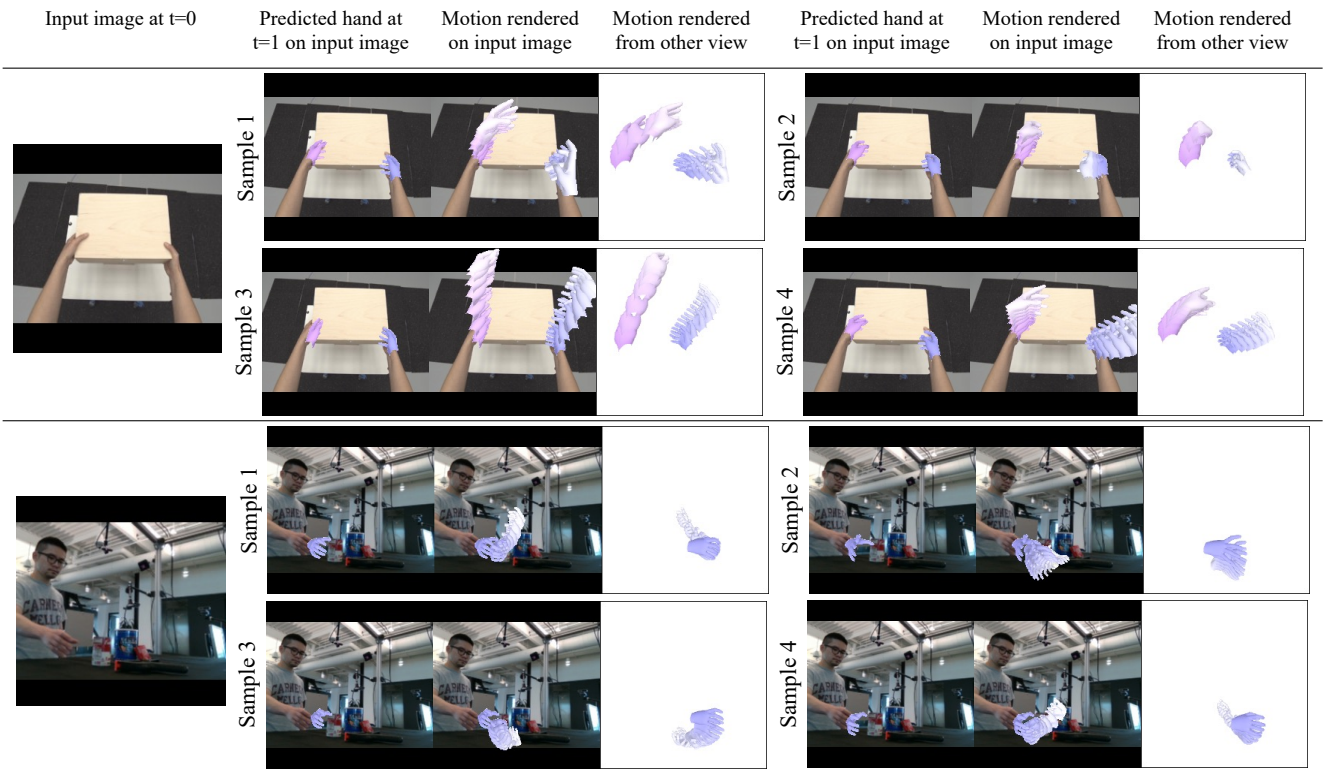
Figure 6. Different forecasts from the same input image. We show 4 samples for 2 input images from ForeHand4D, indicating different modes of interaction with the object: (top) the box lifted in different directions, (bottom) the right hand moves towards different objects.

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**

Table 5. Labels from lifting model (Row 4) better incorporate 2D supervision than alternatives based on attaching auxiliary 2D forecasting heads based on diffusion (Row 3) or regression (Row 2).

motion in 3D space, which are not available in our setting.

Motion diversity and multi-modality. We compute these standard motion quality metrics [16, 49] on ARCTIC, which contains ground truth MANO labels. As reported in Tab. 4, the diversity score for the ground truth distribution is 39.16. The predicted motion distribution of ForeHand4D (diversity = 41.14) is signficantly closer to the ground truth distribution compared to LatentAct (diversity = 302.45) & tranformer regressor (diversity = 187.82). We also observe better muti-modality score for our model (19.64 *vs*. 13.04 for LatentAct). Fig. 6 shows examples of multiple forecasts from the same input: (top) box lifted in different directions, (bottom) right hand moves towards different objects.
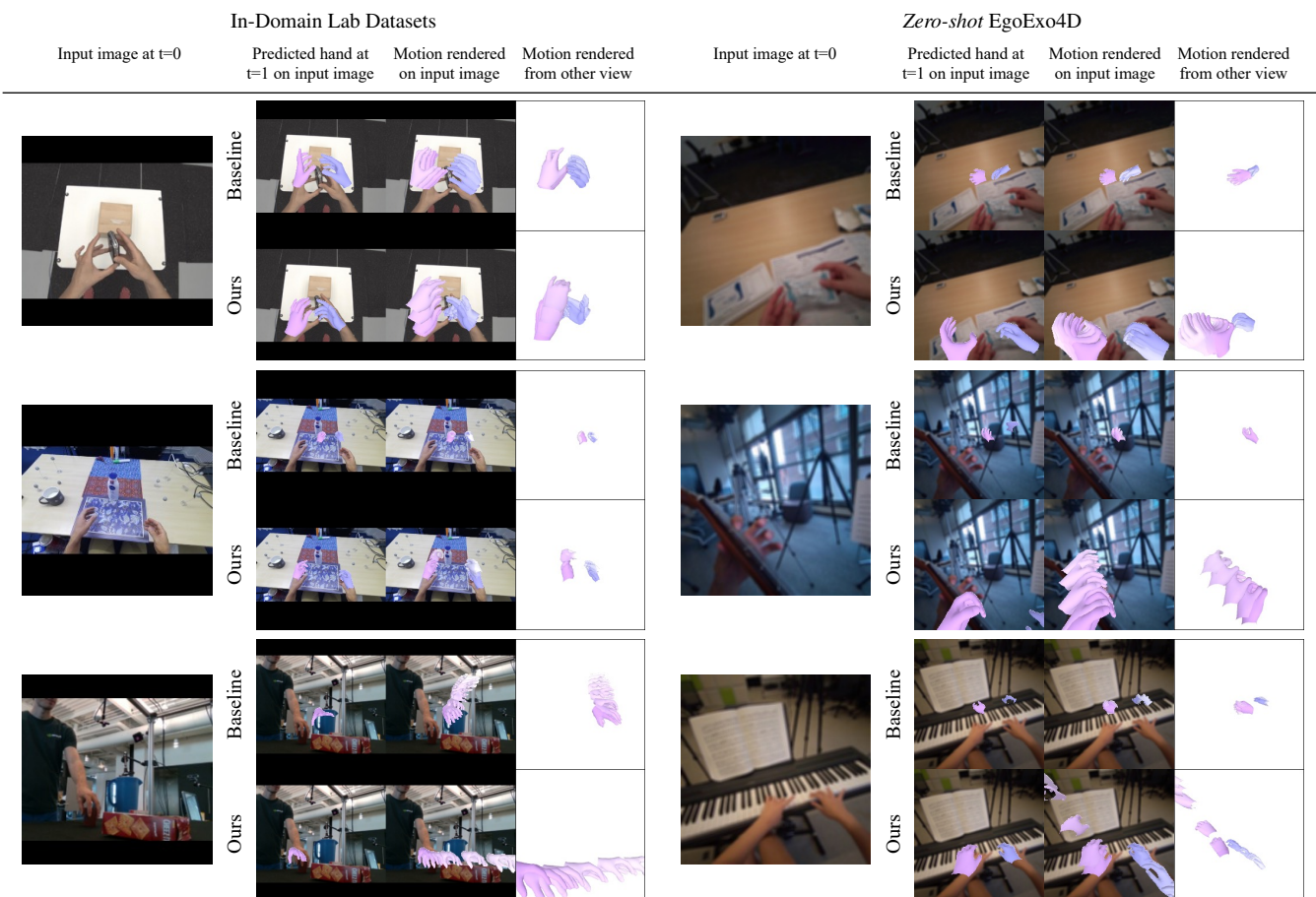
Comparison of proposed lifting scheme against alternatives. In Tab. 5, performance does not improve by much upon injecting supervision via a 2D regression head or 2D diffusion head. However, imputing labels via our lifting model is quite effective & improves M & MR by large amounts.

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 and the Transformer Regressor (3D + 2D sup.) baseline in Fig. 7. Our motion predictions span longer trajectories, are smoother & better placed in the scene compared to the baseline, with the predictions being signifi-

In-Domain Lab Datasets				Zero-shot EgoExo4D			
Input image at t=0	Predicted hand at t=1 on input image	Motion rendered on input image	Motion rendered from other view	Input image at t=0	Predicted hand at t=1 on input image	Motion rendered on input image	Motion rendered from other view

Figure 7. Qualitative comparison: ForeHand4D vs. Transformer Regressor (3D + 2D sup.) (Left) forecasts on 3 lab datasets (ARCTIC, H2O, DexYCB), (Right) *zero-shot* forecasts on challenging EgoExo4D. Left hand in pink, right hand in blue. Color saturation decreases as time proceeds, *i.e.* further away timesteps in future are in lighter shades. We render motions in camera frame & another view. Our predictions span longer trajectories, are smoother, better placed in the scene & significantly more plausible on zero-shot generalization to EgoExo4D.
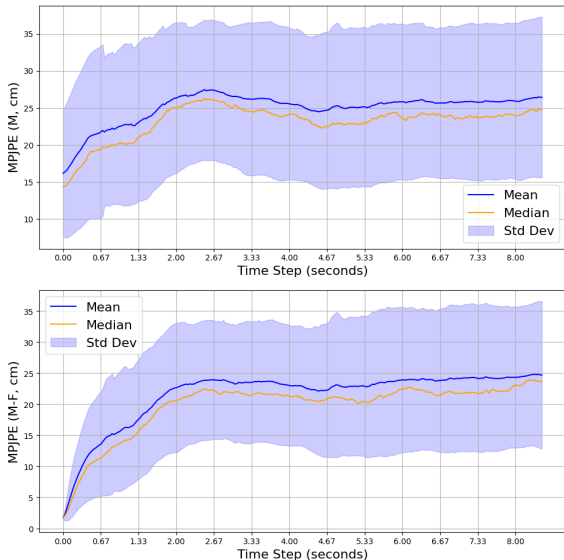
Figure 8. Performance trends over time. Forecasting gets harder for longer prediction horizons for both (top) M & (bottom) M-F.

cantly more plausible on the novel EgoExo4D dataset. More visualizations & analysis are provided in the supplementary.

5. Conclusion, Limitations, and Future Work

We present a system for forecasting bimanual 3D hand motion & articulation 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.

Zero-shot predictions on novel datasets are challenging for all models. In Fig. 7, we see some cases on EgoExo4D where the hands are not well placed. 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.

References

[1] Ijaz Akhter and Michael J. Black. 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, 3

[2] Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv: 2406.09598*, 2024. 3, 5

[4] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. 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. 2

[5] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat-Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[6] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[7] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5

[9] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[10] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[11] Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam R. Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5

[14] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, Fei Li, Zheng Liu, Feng Lu, Karim Abou Zeid, Bastian Leibe, Jeongwan On, Seungryul Baek, Aditya Prakash, Saurabh Gupta, Kun He, Yoichi Sato, Otmar Hilliges, Hyung Jin Chang, and Angela Yao. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[15] Alexey Gavryushin, Florian Redhardt, Gaia Di Lorenzo, Luc Van Gool, Marc Pollefeys, Kaichun Mo, and Xi Wang. SIGHT: single-image conditioned generation of hand trajectories for hand-object interaction. *arXiv*, 2503.22869, 2025. 2

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2021–2029, 2020. 4, 5, 7

[17] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 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. 2

[19] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. 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. 2, 3, 5

[20] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009. 3

[21] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J. Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrusaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics*, 43, 2024. 2

[22] Umar Iqbal, Pavlo Molchanov, Thomas M. Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[23] Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H. Bermano. 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. 3

[24] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 5

[25] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[26] Jiaman Li, C. Karen Liu, and Jiajun Wu. Lifting motion to the 3d world via 2d diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[27] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[29] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[30] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[31] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 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. 2

[32] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. 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. 2

[33] Francesc Moreno-Noguer. 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. 3

[34] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 4, 5

[35] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 5, 6

[36] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 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. 3

[37] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024. 2

[38] Aditya Prakash, Arjun Gupta, and Saurabh Gupta. Mitigating perspective distortion-induced shape ambiguity in image crops. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 4

[39] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 6

[40] Aditya Prakash, Benjamin E Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet S. Sawhney. 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. 2, 6

[41] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 2

[42] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. 3

[43] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 2017. 2

[44] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 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. 2

[45] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. 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. 3

[46] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[47] Qing Shuai, Qi Fang, Junting Dong, Sida Peng, Di Huang, Hujun Bao, and Xiaowei Zhou. Easymocap - make human motion capture easier. Github, 2021. 2, 5

[48] Edgar Simo-Serra, Ariadna Quattoni, Carme Torras, and Francesc Moreno-Noguer. 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. 2, 3

[49] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3, 4, 7

[50] Denis Tomè, Chris Russell, and Lourdes Agapito. 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. 3

[51] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2626–2634, 2017. 2, 3

[52] Bastian Wandt, James J. Little, and Helge Rhodin. 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. 3

[53] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[54] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 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. 4, 5

[55] Sirui Xie, Zhisheng Xiao, Diederik P. Kingma, Tingbo Hou, Ying Nian Wu, Kevin P. Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. EM distillation for one-step diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[56] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *Advances in neural information processing systems*, 29, 2016. 2, 3

[57] Ji Yang, Youdong Ma, Xinxin Zuo, Sen Wang, Minglun Gong, and Li Cheng. 3d pose estimation and future motion prediction from 2d images. *Pattern Recognition*, 124:108439, 2022. 2

[58] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[59] Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J Black. Predicting 4d hand trajectory from monocular videos. *arXiv preprint arXiv:2501.08329*, 2025. 2

[60] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 2

[61] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dynhamr: Recovering 4d interacting hand motion from a dynamic camera. *arXiv preprint arXiv:2412.12861*, 2024. 2

[62] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[63] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 2023. 2

[64] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. 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. 3

[65] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *arXiv preprint arXiv:2409.09300*, 2024. 2

[66] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 5

[67] Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 4

[68] Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. *arXiv preprint arXiv:2412.05066*, 2024. 2

[69] Ce Zheng, Sijie Zhu, Matías Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[70] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[71] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single RGB images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3