From Lifting to Forecasting for Bimanual 3D Hand Motion

Aditya Prakash[1] Richard Li[2] David Forsyth[1] Saurabh Gupta[1]
[1]University of Illinois Urbana-Champaign [2]Massachusetts Institute of Technology
https://ap229997.github.io/projects/forehand4d/

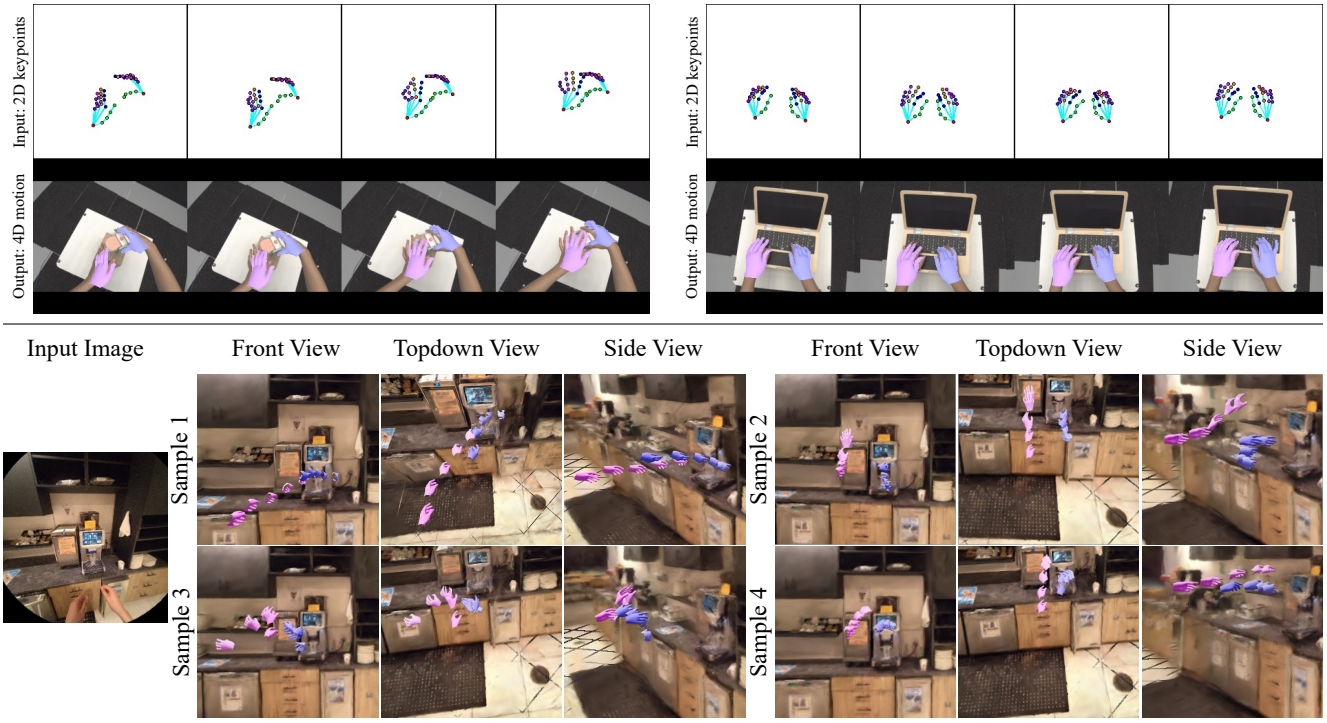

Figure 1. (*top*) LiftHand4D outputs 4D MANO parameters from 2D keypoints & camera parameter inputs. We show 4 frames with the MANO mesh rendered onto the image for visualization (images are not used as input). (*bottom*) ForeHand4D forecasts bimanual 4D hand motion from a single RGB image. It can also 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.

Abstract

3D hands move in diverse ways in everyday tasks. To enable learning hand motions from diverse settings, we propose a pipeline to extract 3D hand motion & articulation from everyday videos. We then use these extracted labels to train a model that can forecast 3D hand motion & articulation from a single image. 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.

1. Introduction

Hands interact in complex ways with one another and the surrounding environment. Predictive models for hand motion

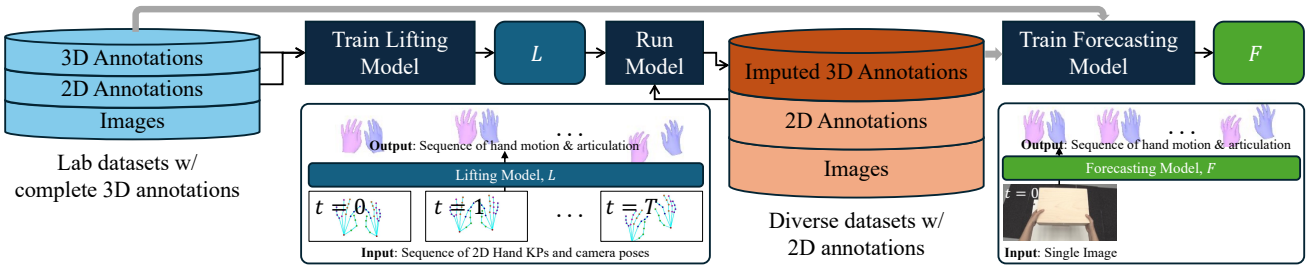

Figure 2. Overall Training Pipeline. We first use the 2D & 3D annotations in lab datasets to train a lifting diffusion model, L that maps 2D keypoints sequences to 3D MANO hands. We then run L on diverse datasets with 2D annotations to generate 3D annotations. Finally, the forecasting model F is trained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

can enable assistive tools, *e.g*. an AR tennis coaching app can visually show adjustments to the racket grip for a better serve. Given their success in modeling human motion [19, 51], diffusion models are a natural fit for diverse, multi-modal hand motions as well. However, diffusion models need *complete* ground truth (*i.e*. 3D hand motion & articulation over time) for training (to add noise to in the forward diffusion process), which is not available in diverse settings. Focusing on this gap, we first design LiftHand4D, a pipeline to extract full 3D hand motion & articulation from diverse videos (Fig. 1 top). These labels are used to train ForeHand4D, a single-image 3D hand motion & articulation forecaster (Fig. 1 bottom).

Prior works [38, 53, 59] side step the need for 3D labels by leveraging weak supervision via reprojection losses on more readily available 2D labels, but such a strategy is not applicable to diffusion models, what do you add noise to?). Generating 3D pseudo-labels from available 2D annotations is the obvious solution but existing methods, *e.g*. EasyMocap [49], that directly optimize the MANO parameters using 2D reprojection loss are not effective (highly sensitive to initialization, optimization objective & hyper-parameter tuning). Our innovation is to develop a *learned lifting model*, LiftHand4D, that lifts available 2D annotations over time into complete 3D annotations (Fig. 1 bottom). The lifting model is made possible by a conditioning module that properly encodes the 2D annotations under a moving camera.

Using the imputed labels from LiftHand4D, we develop ForeHand4D, a system for *forecasting bimanual 3D hand motion* from a single everyday RGB image as input. Because 2D annotations are more readily available on diverse datasets, this increases the data diversity for training the forecasting model. 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 in diverse settings. Fig. 1 (bottom) shows sample outputs from ForeHand4D, including on images from the held-out EgoExo4D dataset [17] not used for training in any way. Generations from ForeHand4D are coherent and different samples capture different possible interactions from the same starting point. Our overall pipeline is shown in Fig. 2.

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 [43] (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. At the same time, the lifting model generates more accurate 3D labels, 65.3% better than the recent HaWoR method [71] (Tab. 2); and the way 2D keypoints and camera parameters are encoded matters (Tab. 4). Code and models will be released upon publication.

2. Related Work

3D Hand Prediction. Given image or video, recent works predict 3D hands in egocentric [14, 15, 27, 28, 57, 62] & exocentric settings [20, 21, 31, 34, 35, 37, 40, 46, 64, 68]. HaMeR [38] is a high-performing recent work that predicts 3D hands from single images, while Dyn-HaMR [66] makes temporally consistent 3D hand predictions from videos. Predictions on videos are made via feed-forward model [63], test-time optimization [71] or hybrid methods [24].

Lifting 3D pose from 2D keypoints. Several works in human pose literature estimate 3D pose from 2D keypoints in different ways, *e.g*. linear models [44], probabilistic models [50], directly optimizing 3D poses [1], MLP [33, 36], convolutional [36, 39, 52], graph-based [5, 11, 55], transformer [30, 47, 69, 74], diffusion [26, 29] & normalizing flows [54]. These works operate on different types of inputs, *e.g*. static 2D pose [1, 33, 54], 2D pose estimated from image [36, 50], sequence of 2D keypoints [5, 26, 47, 55, 69, 74] or multi-view 2D poses [29]. These ideas have also been extended to estimated 3D hand poses using 2D keypoints [58, 77], 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.

4D Hand Forecasting. Prior works have looked at forecasting specific aspects of hand motion in different settings.

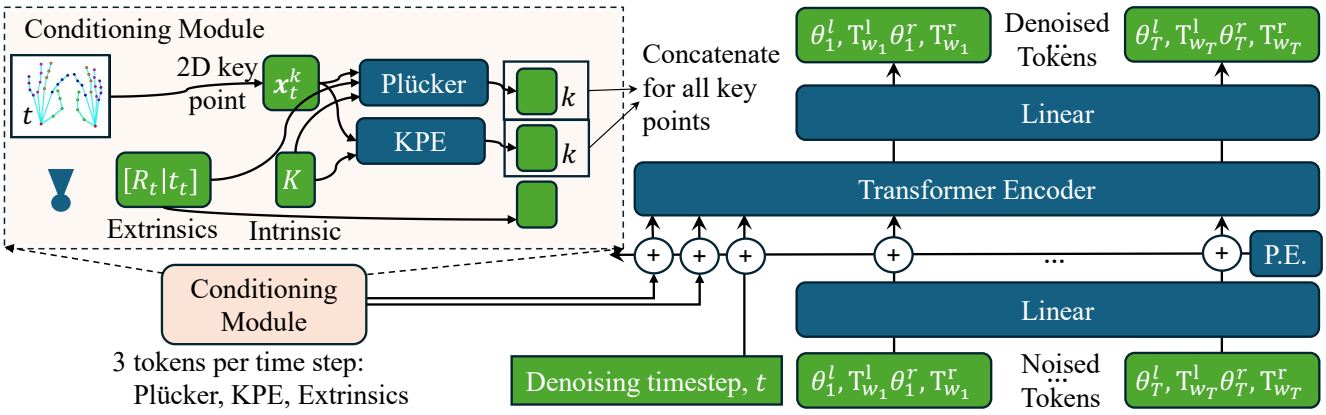

Figure 3. Architecture for Lifting Model. We modify MDM [51] to condition on a sequence of 2D hand keypoints & camera parameters. The conditioning module combines different input representations: 3D pose (rotation, translation) of camera, Plücker rays [72] & KPE [41].

Works differ in what they output and from what input. On the output side: [7, 43] 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, 60] only use RGB images as input, while [7, 10, 43, 70, 73] 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, 60], single hand-object interaction [4, 16, 43, 75] or bimanual interactions [7, 10, 32, 70, 73]. 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.

Learning from Incomplete 3D Ground Truth. Prior works often inject weak supervision into 3D regression models via a 2D reprojection loss [38, 53, 59]: 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. Method

We use MANO hand representation [45], 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 $\Phi_t = \{(\theta^l_t, {}_cT^l_{w_t}, \theta^r_t, {}_cT^r_{w_t})\}$ for all timesteps in the prediction horizon, where l & r superscripts denote left & right hand. We do not predict β (we use the mean β shape from the MANO model when β is not available).

Diffusion models require *complete* 3D annotations for training. These 3D labels are only available in carefully constructed lab datasets (*e.g*. ARCTIC [14], H2O [27], H2O-3D [21], HOT3D [3], & DexYCB [8]). Thus, we develop a *lifting model, L* to lift 2D key point annotations to complete 3D annotations (Sec. 3.1). Our forecasting model F is realized using a transformer & trained using a diffusion loss (Sec. 3.3). It 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. 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. 3).

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 $\boldsymbol{x}^k_t = (x^k_t, y^k_t, 1)$ denote the 2D location of the k^{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 $\boldsymbol{x}^k_t$ is given by $\lambda \boldsymbol{R}_t^{-1}\boldsymbol{K}^{-1}\boldsymbol{x}^k_t - \boldsymbol{K}^{-1}\boldsymbol{t}_t$. We represent this ray using the 6D Plücker representation [72]. **(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^k_t, y^k_t), we estimate the angles $\phi_x = \tan^{-1}((x^k_t - p_x)/f_x)$ and $\phi_y = \tan^{-1}((y^k_t - p_y)/f_y)$ and compute sinusoidal encodings.

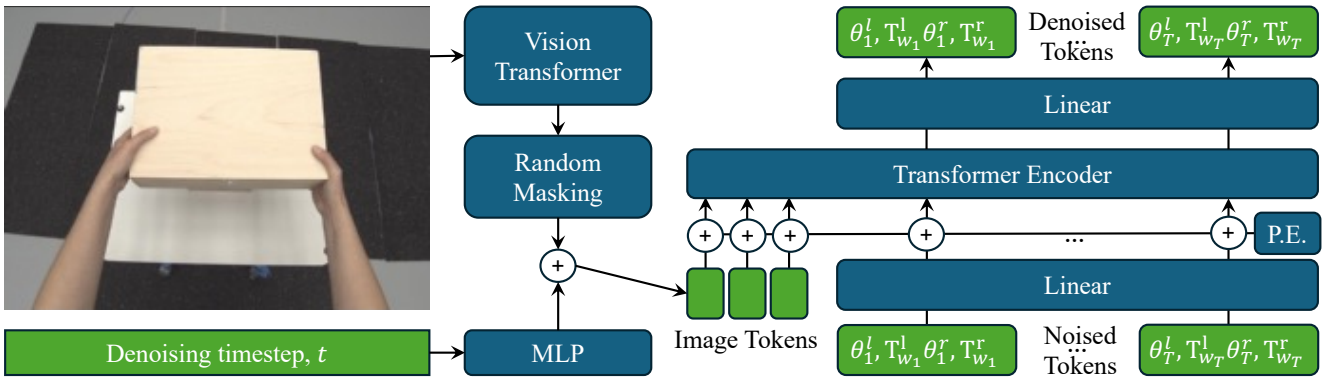

Figure 4. Architecture for Forecasting Model. We modify MDM [51] to condition on images features extracted from a ViT backbone. Each input & output token is 198-dimensional: 2 hands $\times$ (16 (joints) $\times$ (6 (6D rotation for each joint) + 3 (wrist translation))).

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.

3.2. 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 [56]). 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.

3.3. 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 [51] for our setting. Specifically, we change the conditioning to provide image features as input. Following [51], we use a transformer encoder for the denoising (Fig. 4) and 6D representation [76] 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 $t = 0$.

3.4. 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 [51], 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

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, 48, 51, 61, 67] 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.

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 & time horizon (in seconds) and we train a single model over all these variable factors. Note that EgoExo4D is only used for testing the *zero-shot generalization* performance of different models (not for training), and the *lifting model is not trained on AssemblyHands*.

Dataset Name	View	Lab/Wild	Labels	#Motions	#Objs	Horizon(s)	Role (L)	Role (F)
ARCTIC [14]	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 [56]	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

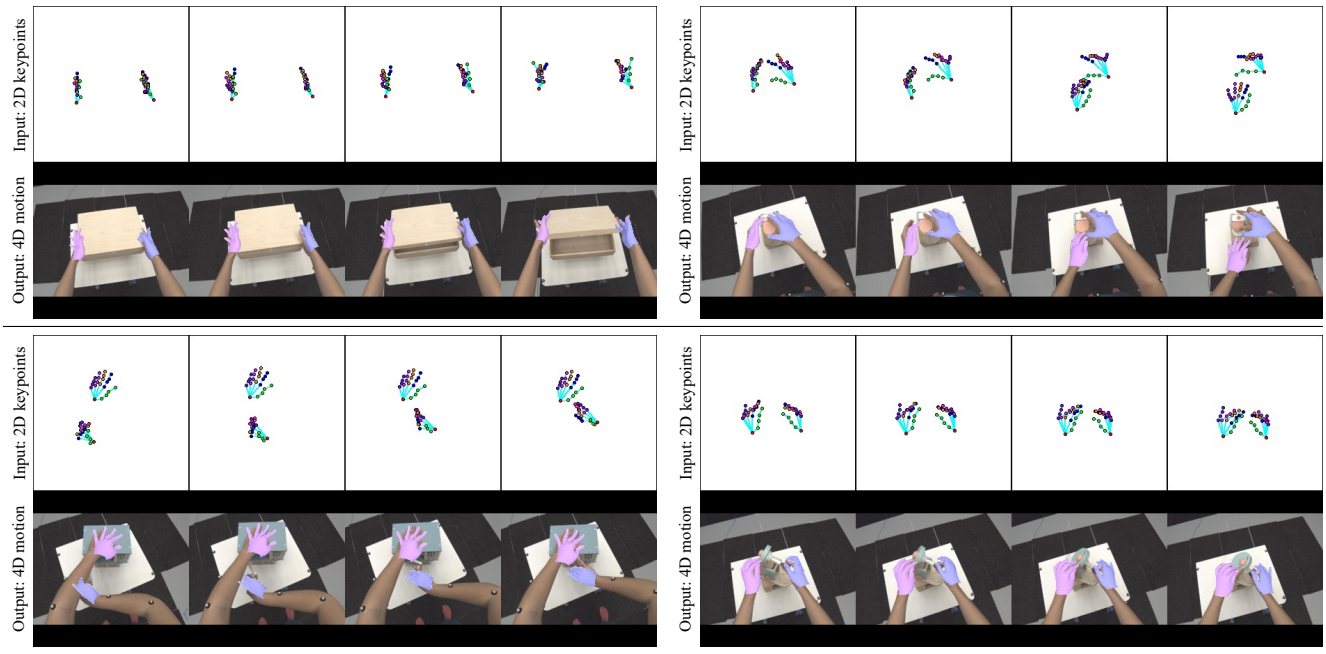

Figure 5. 4D hand predictions from the lifting model, that outputs 3D MANO parameters from 2D keypoints & camera parameter inputs. We show 4 frames with the MANO mesh rendered onto the image for visualization (images are not used as input).

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 [18], 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 [27], H2O-3D [21], ARCTIC [14] Ego, HOT3D [3], & DexYCB [8] with complete 3D annotations (*i.e.* MANO labels) but limited data diversity. For diverse images, we include HoloAssist [56] & 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 (not used for training F), (c) EgoExo4D [17]: zero-shot generalization (whole dataset not used for training).

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 [71], a recent method that uses a data-driven motion priors to reconstruct the 3D hand motions in the world-frame

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 [71], 3) an optimization-based method EasyMocap [49]. 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 [71]	61.6	6.1	46.3	131.1
EasyMocap [49]	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 [49]	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 [49]	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**

using videos, (c) EasyMocap [49] 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.

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).

Table 4. Camera inputs for Lifting model (Sec. 3.1). Different ways of encoding camera parameters help, with Plücker rays being the most effective (this ablation uses ground truth 2D keypoints).

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**

Qualitative Visualizations. In Fig. 5 (2D to 3D lifting), our lifting model accurately places and articulates the hands.

4.2. Forecasting Results

Due to lack of prior work on this problem, we adapt recent LatentAct [43] for our setting & 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, 42]). 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 [43]** 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).

Static pose results. The large values of M & MR for static pose methods indicate that there is indeed a significant hand

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 [43], 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 [43] (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**

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**

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 [43] (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.

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.

Diversity & multi-modality. We compute these standard motion quality metrics [18, 51] 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.

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.

Limitations. 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 forecasted

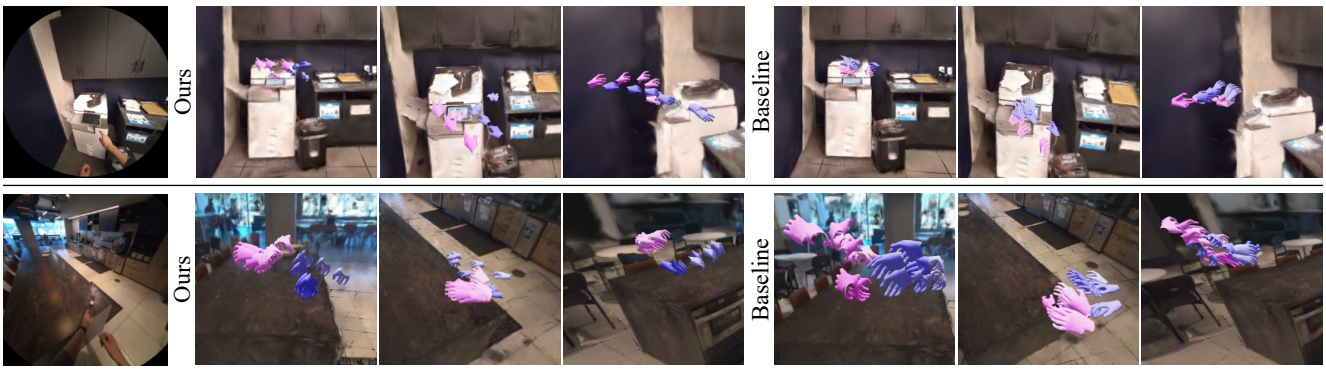

Figure 6. Qualitative comparison: ForeHand4D vs. Transformer Regressor (3D + 2D sup.) *Zero-shot* forecasts on unseen Aria images from [65]. Our predictions span longer trajectories, are smoother, better placed in the scene & more plausible compared to the best baseline. The predicted motions highlight 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).

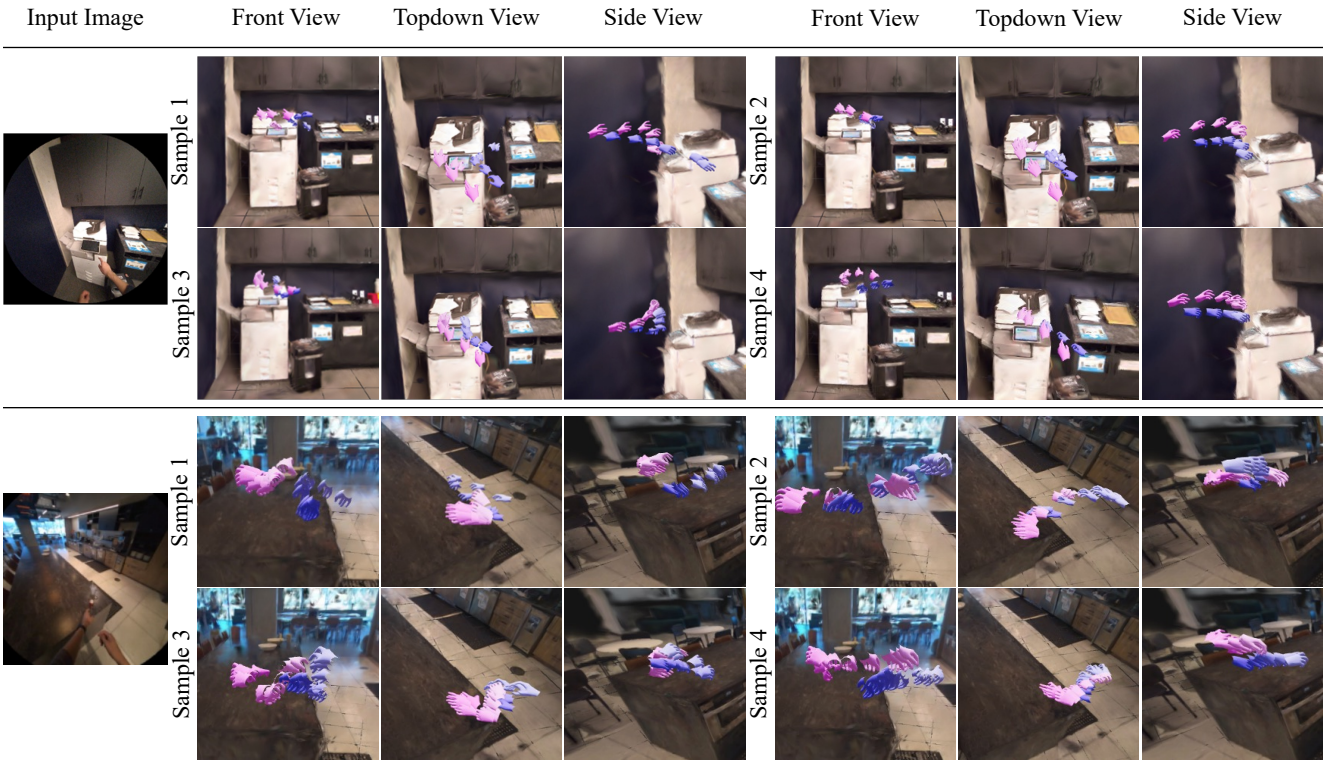

Figure 7. Different forecasts from the same input. We show 4 samples for 2 images from ForeHand4D, indicating different modes of interaction like reaching & grabbing towards different objects in the scene, *e.g.* coffee machine, paper on the table, printer. We show the predicted motion in the 3D scene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

hand being incorrectly placed. Moreover, errors in hand forecasts compound over time (details in supplementary).

5. Conclusion

We present a pipeline for lifting 2D hand keypoints to 3D MANO annotations and using these imputed labels for forecasting bimanual 3D hand motion from a single image in diverse settings. Our predictions span longer horizon, are smoother, better placed and capture multiple interaction modes, especially in zero-shot generalization settings. For future work, incorporating context , *e.g.* past frames or intent and predicting object motion are important directions.

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

[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. 3

[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. 2

[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. 3

[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. 3

[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] Wencan Cheng, Hao Tang, Luc Van Gool, and Jong Hwan Ko. 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. 2

[10] 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. 3

[11] 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. 2

[12] 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

[13] 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. 4

[14] 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

[15] 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

[16] 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. 3

[17] Kristen Grauman 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. 2, 5

[18] 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

[19] 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. 2, 4

[20] 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

[21] 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

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

[23] Masashi Hatano, Zhifan Zhu, Hideo Saito, and Dima Damen. The invisible egohand: 3d hand forecasting through egobody pose estimation. *arXiv preprint arXiv:2504.08654*, 2025. 3

[24] 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

[25] 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. 2

[26] 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. 2

[27] 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

[28] 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

[29] 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. 2

[30] 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. 2

[31] 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

[32] 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. 3

[33] 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. 2

[34] 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

[35] 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

[36] 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. 2

[37] 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. 2, 4, 5

[38] 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, 7

[39] 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. 2

[40] 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

[41] 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. 3

[42] 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

[43] 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, 3, 6, 7

[44] 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

[45] 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

[46] 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

[47] 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. 2

[48] 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

[49] Qing Shuai, Qi Fang, Junting Dong, Sida Peng, Di Huang, Hujun Bao, and Xiaowei Zhou. Easymocap - make human motion capture easier. Github, 2021. 2, 6

[50] 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

[51] 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. 2, 3, 4, 7

[52] 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. 2

[53] 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

[54] 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. 2

[55] 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. 2

[56] 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

[57] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. 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. 2

[58] Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. Handbooster+: Boosting 3d hand-mesh reconstruction from data synthesis to progressive multi-hypothesis aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025. 2

[59] 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

[60] 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. 3

[61] 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

[62] Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J. Black. Predicting 4d hand trajectory from monocular videos. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2025. 2

[63] 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

[64] 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

[65] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 7, 8

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

[67] 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

[68] 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

[69] 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. 2

[70] 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. 3

[71] 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, 6

[72] 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. 3

[73] 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. 3

[74] 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. 2

[75] Bohan Zhou, Yi Zhan, Zhongbin Zhang, and Zongqing Lu. Megohand: Multimodal egocentric hand-object interaction motion generation. *arXiv preprint arXiv:2505.16602*, 2025. 3

[76] 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. 4

[77] 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. 2