AnyLift: Scaling Motion Reconstruction from Internet Videos via 2D Diffusion

Anonymous CVPR submission

Paper ID 5

Abstract

Reconstructing 3D human motion and human-object interactions (HOI) from Internet videos is a fundamental step toward building large-scale datasets of human behavior. Existing methods struggle to recover globally consistent 3D motion under dynamic cameras, especially for motion types underrepresented in current motion-capture datasets, and face additional difficulty recovering coherent human-object interactions in 3D. We introduce a two-stage framework leveraging 2D diffusion that reconstructs 3D human motion and HOI from Internet videos. In the first stage, we synthesize multi-view 2D motion data for each domain, leveraging 2D keypoints extracted from Internet videos to incorporate human motions that rarely appear in existing MoCap datasets. In the second stage, a camera-conditioned multi-view 2D motion diffusion model is trained on the domain-specific synthetic data to recover 3D human motion and 3D HOI in the world space. We demonstrate the effectiveness of our method on Internet videos featuring challenging motions such as gymnastics, as well as in-the-wild HOI videos, and show that it outperforms prior work in producing realistic human motion and human-object interaction.

1. Introduction

Large-scale 3D human motion and human-object interaction (HOI) data are essential for a wide range of applications in computer vision, computer graphics, and robotics. While high-quality motion capture (MoCap) datasets have been widely used for these purposes, they remain limited in scale and diversity. Estimating 3D human motion directly from videos offers a scalable alternative. However, existing methods struggle to reconstruct global motion in the world coordinate frame under dynamic cameras, as well as human-object interactions involving dynamic object movement.

In this work, we adopt the formulation of learning 2D motion priors for 3D reconstruction [6] and present a unified two-stage framework that reconstructs both 3D human motion and 3D human-object interactions (HOI) from dynamic-camera videos. In Stage 1, we synthesize multi-view 2D motions that serve as training data for the subsequent stage. In Stage 2, we train a camera-conditioned multi-view 2D diffusion model on the synthesized data to reconstruct 3D motion directly from single-view 2D keypoints. For human motions that are underrepresented in existing MoCap datasets, we leverage Internet videos to generate synthetic multi-view data by learning single-view 2D motion priors and introducing a hybrid data-source training strategy that mitigates limited viewpoint coverage. For HOI, we focus on the reconstruction of everyday interactions and synthesize category-specific multi-view 2D trajectories by reprojecting existing HOI MoCap sequences under diverse camera trajectories. Given the resulting synthetic data, our Stage-2 diffusion model predicts consistent multi-view 2D motions from a single-view input.

To summarize, our work makes the following contributions. First, we propose a camera-trajectory-conditioned 2D motion diffusion model that enables the use of dynamic-camera videos for training and motion reconstruction from unconstrained monocular inputs. Second, we introduce a hybrid data-source training strategy with a decomposed motion representation to address the challenge of limited camera-view coverage during training. Third, we present a unified framework for reconstructing world-grounded human motion and human-object interactions from in-the-wild videos.

2. AnyLift

Formulation. Our goal is to estimate world-coordinated 3D human and human-object interaction (HOI) motion sequences $\tau = (\mathcal{H}, \mathcal{O})$ from single-view 2D keypoint sequences $\mathbf{X} \in \mathbb{R}^{T\times K\times 2}$ captured under dynamic cameras. Here, $\mathcal{H}$ denotes the human motion, and $\mathcal{O}$ denotes the object motion, which is included only for HOI reconstruction. T and K represent the number of frames and keypoints, respectively. We adopt the SMPL model [9] to parameterize the human pose, where each frame t is represented by $\mathcal{H}_t = (\mathbf{r}_t, \phi_t, \mathbf{\Theta}_t)$, consisting of the root translation $\mathbf{r}_t$, global orientation ϕ_t, and body pose parameters $\mathbf{\Theta}_t$.

Overview. We propose AnyLift, a unified framework that reconstructs both 3D human motion and human-object interactions (HOI) from monocular videos captured by dynamic cameras. An overview of AnyLift is shown in Fig. 1. AnyLift follows a two-stage pipeline: (1) *multi-view 2D synthetic data generation*, where we prepare training data with diverse

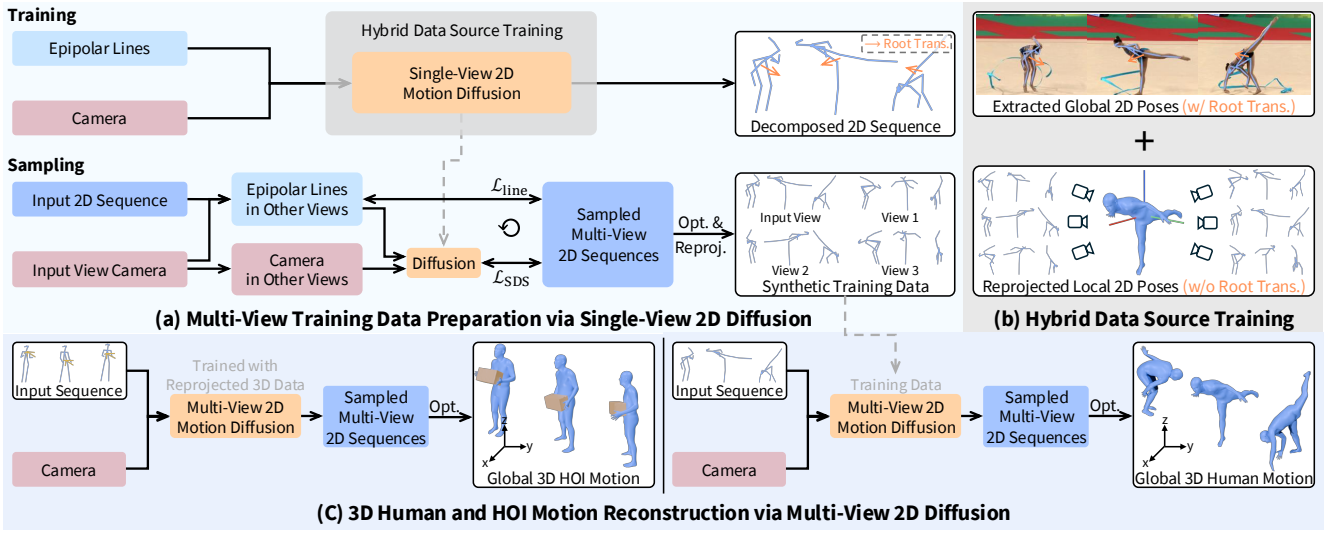

Figure 1. **Overview of AnyLift.** (a) We first train a single-view 2D motion diffusion model conditioned on camera trajectories and epipolar lines to synthesize multi-view 2D training data. (b) During training, we employ a hybrid data source strategy that enhances viewpoint coverage by combining global 2D pose sequences from videos with locally reprojected poses. (c) Finally, we train a multi-view 2D motion diffusion model to reconstruct consistent world-coordinated 3D human and HOI motions from real-world videos.

camera trajectories (Sec. 2.1); and (2) *multi-view 2D motion diffusion*, where we learn to generate consistent multi-view 2D motion from single-view 2D inputs using the synthesized data, and subsequently reconstruct world-coordinated 3D motion (Sec. 2.2). For **human motion** such as gymnastics and martial arts that are rarely represented in existing motion-capture datasets, we leverage Internet videos to extract 2D keypoints and camera trajectories, and train a conditional single-view 2D diffusion model to synthesize multi-view training data. For **human-object interactions (HOI)**, we follow the same two-stage pipeline but generate synthetic multi-view data by reprojecting existing 3D HOI motion-capture sequences [1, 4, 5, 19, 20].

2.1. Multi-View 2D Synthetic Data Generation

Conditional Single-View 2D Motion Diffusion. We begin by training a conditional single-view 2D motion diffusion model for each motion category. The conditioning terms include camera trajectories and epipolar lines. Camera trajectories provide awareness of global viewpoint motion over time, allowing the model to learn the 2D root translation under dynamic cameras. We represent the camera trajectories as a sequence of extrinsic parameters $\mathbf{C} = \{\mathbf{C}_t\}_{t=1}^{T}$, where each $\mathbf{C}_t \in \mathbb{R}^{4\times3}$ is normalized by removing the camera transformation of the initial frame. Epipolar lines encode pairwise geometric constraints between views, encouraging the model to learn cross-view consistency. Each epipolar line $\boldsymbol{l} = (a, b, c)^{\mathrm{T}}$ is defined by the 2D line equation $ax + by + c = 0$. For every frame, we assign an epipolar line to each keypoint, passing through the keypoint and its corresponding epipole, resulting in a condition matrix $\mathbf{L}_t \in \mathbb{R}^{K\times3}$. During training, we simulate several fixed epipoles based on sampled camera extrinsics, while at inference time the epipole is determined by the relative camera transformation between paired views.

Following the DDPM framework [3], we adopt a forward diffusion process that progressively adds noise to clean 2D motions over N steps:

$$q(\mathbf{X}_n|\mathbf{X}_{n-1}) = \mathcal{N}(\mathbf{X}_n; \sqrt{1-\beta_n}\mathbf{X}_{n-1}, \beta_n\mathbf{I}), \quad (1)$$

where $n \leq N$ denotes the diffusion step, and β_n is the variance schedule controlling noise magnitude. The reverse denoising process is learned by a network $\mathbf{X}_\theta$, which learns to iteratively denoise samples across N steps starting from $\mathbf{X}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ conditioned on the camera trajectories $\mathbf{C}$ and epipolar lines $\mathbf{L}$. We reparameterize the prediction objective to directly estimate the clean sample $\mathbf{X}_0$. The network is optimized with an L_1 reconstruction loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{X}_0,n}\left\|\mathbf{X}_0 - \mathbf{X}_\theta(\mathbf{X}_n, n, \mathbf{C}, \mathbf{L})\right\|_1. \quad (2)$$

In line with prior works [15], we adopt a Transformer-based backbone [16] for the denoising network $\mathbf{X}_\theta$. The conditioning inputs $\mathbf{C}$ and $\mathbf{L}$ are concatenated with the noisy keypoint sequence $\mathbf{X}_n$ along the feature dimension and embedded through an MLP encoder before feeding into the backbone.

Following Li et al. [6], we add line matching loss to encourage the 2D keypoints to align with their corresponding epipolar lines by minimizing the 2D point-line distance:

$$\mathcal{L}_{\text{line}} = \sum_{t=1}^{T} \langle \mathbf{L}_t, (\hat{\mathbf{X}}_t, \mathbf{1}) \rangle, \quad (3)$$

where $\hat{\mathbf{X}}_t$ denotes the 2D keypoints at frame t after the denoising process.

Hybrid Data Source Training. Unlike standard datasets such as AIST++ [7], which provide multi-view videos with uniformly distributed cameras, Internet videos of specific motion categories are usually captured from a few forward-facing angles, resulting in limited viewpoint coverage. To mitigate this limitation, we introduce a hybrid training strategy that combines two complementary sources of 2D motion data: (1) global 2D keypoints extracted from Internet videos, and (2) local 2D projections $\mathbf{X}^{\text{proj}}$ obtained by reprojecting reconstructed 3D motions from off-the-shelf estimators [13].

However, including $\mathbf{X}^{\text{proj}}$ in training biases the model toward learning motion patterns with limited global translation. To address this, we decompose each 2D motion $\mathbf{X}$ into root translation $\mathbf{X}^{\text{r}} \in \mathbb{R}^{T\times 2\times 2}$—represented by the two hip joints—and local pose $\mathbf{X}^{\text{l}} \in \mathbb{R}^{T\times(K-2)\times 2}$. The global 2D motion $\mathbf{X}^{\text{g}}$ is then recovered by adding the average root translation (computed across the two hip joints) back to the local pose. With this representation, the diffusion loss is computed as in Eq. (2), while the line-matching loss is applied to the global 2D motion $\mathbf{X}^{\text{g}}$ defined in Eq. (3). During training, $\mathbf{X}^{\text{proj}}$ is generated by projecting reconstructed 3D motions through randomly sampled camera viewpoints from the training set, augmented with a small set of predefined camera trajectories to increase viewpoint diversity. We compute the diffusion loss only for the local pose:

$$\mathcal{L}^{\text{proj}} = \mathbb{E}_{\mathbf{X}_0,n}\left\|\mathbf{M}\odot\mathbf{X}_0 - \mathbf{M}\odot\mathbf{X}_\theta(\mathbf{X}_n^{\text{proj}}, n, \mathbf{C}, \mathbf{L})\right\|_1, \quad (4)$$

where $\mathbf{M}$ is mask that excludes the two hip joints from loss computation. The line matching loss is not applied to $\mathbf{X}^{\text{proj}}$.

Multi-View 2D Motion Data Synthesis. Leveraging the learned 2D motion prior, we employ score distillation sampling [12] with multi-view consistency loss to prepare multi-view training data. Given a single-view sequence, we optimize $V-1$ additional 2D keypoint sequences from viewpoints evenly distributed along a circular ring around the input camera, resulting in a set of sequences $\{\mathbf{X}_v\}_{v=1}^{V}$. The gradient of the SDS loss, which encourages each $\mathbf{X}_v$ to conform to the learned diffusion prior, is computed as:

$$\nabla_{\mathbf{X}_v}\mathcal{L}_{\text{SDS}} = \mathbb{E}_{n,\epsilon}\Big[w(n)\big(\epsilon_\theta(\mathbf{X}_{v,n}, n, \mathbf{C}, \mathbf{L}) - \epsilon\big)\Big], \quad (5)$$

where the weight $w(n)$ is determined by the noise level n.

For two different views u and v, we compute the epipolar lines $\mathbf{L}^{u\to v}$ in view v using the 2D keypoint sequence $\mathbf{X}_u$ and the relative camera transformation between the two views. We then apply line matching loss to enforce the 2D keypoints $\mathbf{X}_v$ satisfy the corresponding constraints:

$$\mathcal{L}_{\text{line}}^{u\to v} = \sum_{t=1}^{T}\big\langle \mathbf{L}_t^{u\to v}, (\mathbf{X}_{v,t}^{\text{g}}, \mathbf{1})\big\rangle, \quad (6)$$

where $\mathbf{X}_{v,t}^{\text{g}}$ represents the recovered global motion from the decomposed representation.

After obtaining roughly consistent multi-view 2D pose sequences via SDS optimization, we recover 3D joint positions by minimizing multi-view reprojection errors. The recovered 3D joints are then used to fit SMPL parameters [9] using VPoser [11], producing full-body 3D motion sequences. Finally, we reproject the fitted 3D motions into four evenly distributed cameras to generate geometrically consistent multi-view 2D training data.

2.2. Multi-View 2D Motion Diffusion

We train a multi-view 2D motion diffusion model using the data synthesized in Sec. 2.1 to generate multi-view 2D motion sequences from a single-view input.

Data and Condition Representation. We train the multi-view diffusion model on global 2D sequences. Camera trajectories are represented in the same way as in Sec. 2.1.

Model Architecture. We extend the single-view 2D motion diffusion model introduced in Sec. 2.1. The camera condition is embedded in the same way. The transformer backbone is further augmented with cross-view attention layers to enhance multi-view awareness following MVLift [6].

HOI Motion Reconstruction For human-object interactions, we train the multi-view diffusion model on specific object categories (e.g., boxes and tables). The objects are represented by a set of manually designed 2D keypoints $\mathbf{O} \in \mathbb{R}^{T\times M\times 2}$, where M denotes the number of keypoints. The corresponding 3D positions of these keypoints on the canonical object mesh are denoted as $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{M}$. The object 2D keypoints $\mathbf{O}$ are concatenated with the human keypoints $\mathbf{X}$ to form a unified representation. During training, we randomly mask out a subset of $\mathbf{O}$ to handle partial occlusions and potential tracking failures.

Inference on Real-World Videos. During inference, we extract 2D human keypoint sequences using ViTPose [18] and estimate camera motion with MegaSaM [8]. The object keypoints are tracked using DELTA [10]. The SMPL parameters $\mathcal{H} = (\mathbf{r}, \boldsymbol{\phi}, \boldsymbol{\Theta})$ are obtained through the final optimization process described in Sec. 2.1. For objects, we also start with obtaining the 3D object keypoints $\mathbf{Q} \in \mathbb{R}^{T\times M\times 3}$ by minimizing the multi-view reprojection error. Using the reconstructed $\mathbf{Q}$ and the predefined canonical keypoints $\mathbf{P}$ on the object mesh, we then estimate the object pose $\mathcal{O}_t = \{\mathbf{r}_t, \mathbf{t}_t, s\}$, where $\mathbf{r}_t \in \mathbb{R}^6$ is the 6D rotation [21], $\mathbf{t}_t \in \mathbb{R}^3$ is the translation, and $s \in \mathbb{R}^+$ is the scale factor.

3. Experiments

3.1. Human Motion Reconstruction

We compare our method with two categories of baselines: methods that do not rely on 3D motion data during training (SMPLify [2] and MVLift [6]), and methods trained with 3D motion ground truth (WHAM [14] and GVHMR [13]).

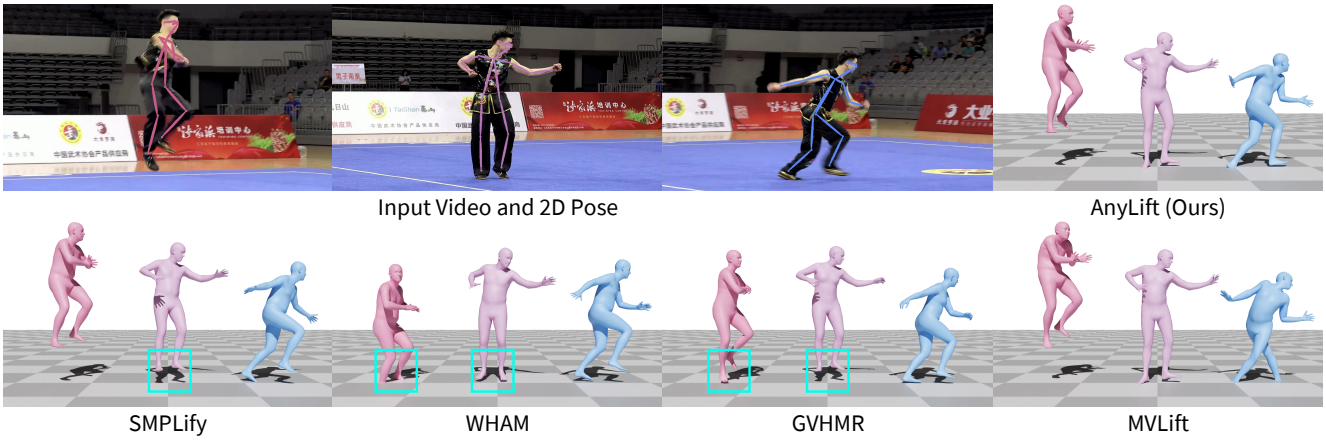

Figure 2. **Qualitative comparison of human motion reconstruction on our collected Internet videos.** AnyLift produces more plausible motions, mitigating the root trajectory errors, inaccurate local body pose, and self-penetration artifacts observed in baselines.

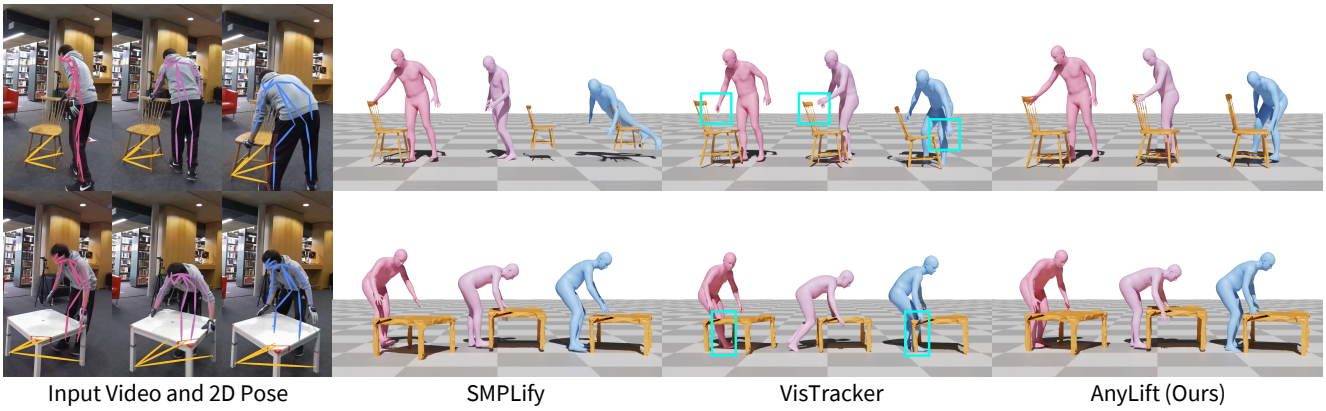

Figure 3. **Qualitative comparison of HOI reconstruction on the BEHAVE [1] dataset.** We show results on two object categories, *chair* and *table*. AnyLift produces coherent and physically plausible human-object interactions with accurate contact and minimal penetration.

We show qualitative comparisons in Fig. 2. SMPLify produces 3D poses with abrupt and unrealistic changes due to depth ambiguity arising from optimization based solely on 2D joint positions. Although WHAM and GVHMR predict plausible local poses, they often yield implausible root trajectories, leading to noticeable penetration with the ground plane, as shown in the two examples. MVLift yields severely distorted leg poses in the example. In contrast, AnyLift reconstructs stable and plausible human motions with accurate global trajectories and consistent body poses across frames.

3.2. HOI Motion Reconstruction

We compare against two representative baselines. SMPLify [2] reconstructs 3D human and object motions by optimizing SMPL parameters and object poses to match 2D keypoints without any training. VisTracker [17] jointly tracks 3D humans, objects, and contacts from monocular videos using a visibility-aware object pose network.

Qualitative comparisons on the BEHAVE dataset are presented in Fig. 3. We show two object categories: *chair* and *table*. SMPLify fails to produce reasonable human-object interaction motions due to depth ambiguity from relying solely on 2D joints and incorrectly models the relative transformation between the human and the object. VisTracker also struggles to generate plausible interactions; in the chair example, even with minimal or no occlusion, it fails to capture correct hand-chair contact, while in the table example, it exhibits severe object penetration. In contrast, our approach produces plausible human-object interactions with accurate contact and minimal penetration under both categories.

4. Conclusion

We presented AnyLift, a unified framework for reconstructing world-grounded human motion and human-object interactions (HOI) from Internet and in-the-wild videos with dynamic cameras. We addressed the problem using a two-stage framework that first synthesizes multi-view 2D motion data and then trains a camera-conditioned multi-view diffusion model on the generated data to reconstruct globally consistent 3D motion and interactions in the world coordinate frame. We demonstrated the effectiveness of AnyLift on collected Internet videos and captured HOI videos.

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 3, 4

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[4] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: joint markerless 3d tracking of humans and objects in interaction from multi-view rgb-d images. *International Journal of Computer Vision (IJCV)*, 132(7):2551–2566, 2024. 2

[5] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6), 2023. 2

[6] Jiaman Li, C Karen Liu, and Jiajun Wu. Lifting motion to the 3d world via 2d diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3

[7] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[8] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 1, 3

[10] Tuan Duc Ngo, Peiye Zhuang, Chuang Gan, Evangelos Kalogerakis, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. Delta: Dense efficient long-range 3d tracking for any video. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[11] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[12] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[13] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *ACM SIGGRAPH Asia Conference Proceedings*, 2024. 3

[14] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[15] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[17] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[19] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[20] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

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