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

Hongjie Li*,† Heng Yu* Jiaman Li Hong-Xing Yu
Ehsan Adeli C. Karen Liu Jiajun Wu

Stanford University

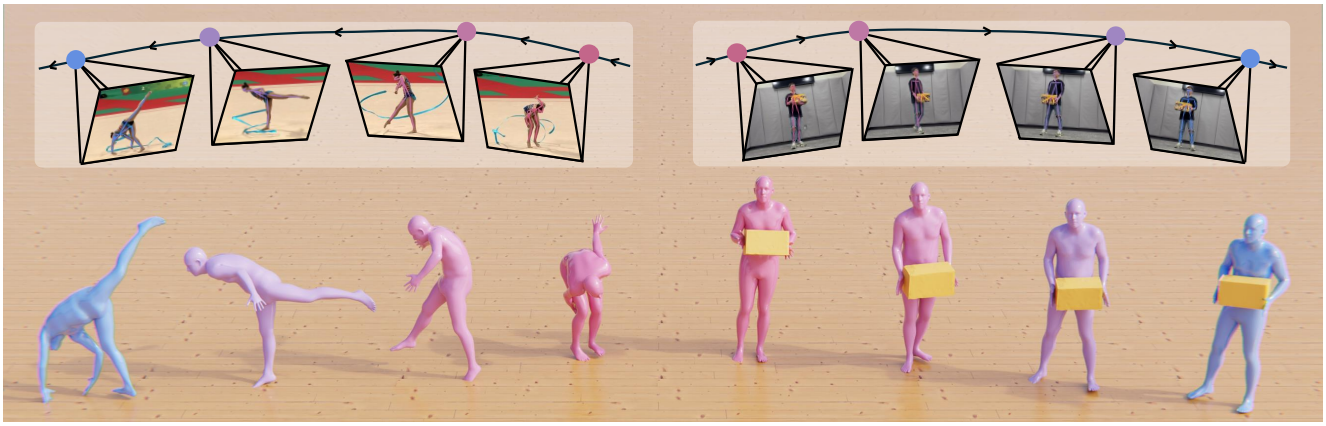

Figure 1. **Human and human-object interaction (HOI) motions lifted by our approach.** Trained on 2D keypoints and corresponding camera trajectories, our framework AnyLift reconstructs world-coordinated 3D human motion and HOI from monocular videos captured by dynamic cameras. We demonstrate its effectiveness on human motion reconstruction from Internet gymnastics videos (left) and on HOI reconstruction from captured real-world videos (right). Please refer to our *project page* for video results.

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. Such data enable realistic character animation, simulation of human behavior in virtual environments, and motion imitation or policy learning for humanoid robots. While high-quality motion capture (MoCap) datasets have been widely used for these purposes, they remain limited in scale and diversity. Acquiring MoCap data requires controlled environments, specialized hardware, and professional actors, making it infeasible to capture the vast diversity of motions observed in everyday life. Estimating 3D human motion directly from videos offers a scalable alternative. However, while recent progress has been made in lifting 2D pose sequences to 3D motion, existing methods still struggle to reconstruct global motion in the world coordinate frame under dynamic camera settings, especially for motion categories that are rarely represented in existing MoCap datasets, as well as human-object interactions involving dynamic object movement.

Existing approaches to human motion estimation can be broadly grouped into two categories. The first relies on large-scale MoCap datasets [26] to train networks with 3D supervision [34, 36]. Although these models achieve high

* Equal contribution. † Work was done while H. Li was a visiting student at Stanford University. H. Li is now with Peking University.

accuracy for in-distribution motion, they generalize poorly to actions that are underrepresented in MoCap data, such as gymnastics, martial arts, or other dynamic, real-world movements, since collecting such 3D data is prohibitively difficult. The second direction seeks to reduce dependence on 3D data by leveraging 2D keypoints extracted from domain-specific videos [9, 18]. For example, MVLift [18] reconstructs 3D motion using only 2D inputs within a multi-stage generative framework. While effective under static cameras, this approach assumes fixed viewpoints during both training and inference, whereas real-world videos commonly feature moving cameras and limited view coverage. These constraints make it difficult to scale to diverse in-the-wild videos where camera motion varies significantly across sequences. Furthermore, extending these frameworks to reconstruct world-grounded human-object interaction from in-the-wild videos remains an open problem.

In this work, we adopt the formulation of learning 2D motion priors for 3D reconstruction [18], as it scales beyond the limited motion diversity of MoCap datasets and naturally incorporates HOI within the same reconstruction framework as human motion. Specifically, we 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. Although human motion and HOI reconstruction share the same two-stage structure, they differ in how the synthetic multi-view data are generated. 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 reconstruction of everyday interactions. Although existing 3D HOI MoCap datasets contain only a limited set of objects, HOI motions often exhibit consistent interaction patterns across objects within the same category. Therefore, we 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, enabling faithful reconstruction of global 3D human motion and HOI in world coordinates.

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. Related Work

Human Motion Reconstruction from Video. Estimating 3D human motion from monocular videos has been widely studied using the SMPL [25] body model, with advances in both image-based [8, 12, 14, 16, 49] and video-based methods [11, 33, 36, 41]. Recent works [13, 36, 37, 47, 48] focus on recovering global trajectories to obtain human motion in the world coordinate system. However, these methods are typically trained with 3D motion capture datasets such as AMASS [26], which limits their generalization to motion types underrepresented in MoCap. In contrast, our approach learns 2D motion priors directly from Internet videos, allowing reconstruction of diverse motions such as gymnastics and martial arts that are rarely captured in MoCap collections.

Human-Object Interaction Reconstruction from Video. Reconstructing human-object interactions from videos has been less explored compared to human motion alone. Recent work has jointly modeled human motion, object motion, and contact [44, 45]. CHORE [44] estimates human and object poses from a single image. VisTracker [45] reconstructs human, object, and contact trajectories from a single RGB camera by conditioning neural field representations on SMPL fits and performing visibility-aware temporal aggregation. While robust to occlusion, it assumes a static camera and focuses on relative human-object motion rather than world-coordinate trajectories under moving cameras. Our work instead reconstructs both human and object motion in the world coordinate system, enabling interaction reconstruction under dynamic camera settings.

Weakly-Supervised Motion Estimation. To mitigate dependence on 3D motion capture data, several works train only on 2D poses or weak supervision. Traditional 2D-to-3D lifting methods [3, 20, 27, 30, 43, 50] regress 3D joint positions from 2D inputs using MLPs, temporal convolutions, or transformers. However, they remain constrained by paired 2D-3D training data and struggle to generalize beyond their capture domains. Recent weakly supervised methods such as ElePose [42] and MAS [9] train solely on in-domain 2D sequences, relying on reprojection or diffusion-based priors to recover plausible 3D poses. Despite removing the need for 3D supervision, these methods are restricted to local pose estimation and cannot infer global trajectories in the world frame. MVLift [18] extends this weakly supervised setting by learning a generative lifting model from 2D inputs to recover global 3D motion, yet it remains limited to static-camera scenarios and struggles when viewpoint coverage is limited during training. Our framework advances this line

of research by learning camera-conditioned 2D motion priors and introducing a hybrid training strategy that together enable reconstruction of world-coordinated human motion from unconstrained Internet videos.

Multi-View 2D Content Generation. Large-scale 3D collections such as Objaverse [4] have spurred multi-view generation for objects, where diffusion models synthesize view-consistent images for 3D reconstruction [22–24, 35]. These ideas have recently extended to multi-view video generation that jointly models spatial and temporal consistency for 4D content [15, 39]. Inspired by prior works on multi-view generation, MVLift [18] reformulates 3D motion reconstruction as multi-view 2D motion generation from single-view inputs, learning entirely from 2D keypoints without relying on 3D annotations. We build on this formulation and extend it to handle dynamic cameras, employ hybrid training to address limited viewpoint coverage, and reconstruct human-object interactions from in-the-wild videos.

3. 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 [25] to parameterize the human pose, where each frame t is represented by $\mathcal{H}_t = (\mathbf{r}_t, \boldsymbol{\phi}_t, \boldsymbol{\Theta}_t)$, consisting of the root translation $\mathbf{r}_t$, global orientation $\boldsymbol{\phi}_t$, and body pose parameters $\boldsymbol{\Theta}_t$.

Preliminary. MVLift [18] reconstructs world-coordinated 3D motion sequences from single-view 2D pose inputs. The framework enforces multi-view consistency through four stages. Stages 1-3 form a multi-view training data synthesis pipeline, where 2D motion diffusion and multi-view optimization are used to generate consistent multi-view 2D pose sequences. The final stage trains a multi-view 2D motion diffusion model on these synthetic data to directly produce consistent multi-view 2D motions from a single 2D sequence input. Together, these stages enable 3D motion reconstruction without requiring any 3D supervision. While MVLift effectively reconstructs human motion from static, single-view videos, it cannot leverage Internet videos with dynamic cameras and limited viewpoint coverage during training. For human-object interaction (HOI) reconstruction, MVLift assumes access to precomputed 2D human and object keypoints reprojected from motion-capture data, without addressing reconstruction from real-world videos. These limitations motivate our new framework.

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. 2. AnyLift follows a two-stage pipeline: (1) *multi-view 2D synthetic data generation*, where we prepare training data with diverse camera trajectories (Sec. 3.1); 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. 3.2, Sec. 3.3). 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, 7, 17, 51, 52].

Built upon MVLift, our method introduces several key innovations. First, to accommodate dynamic cameras, we design a single-view 2D diffusion model conditioned on both camera trajectories and epipolar lines. Second, Internet videos of specific motion categories are typically recorded from limited forward-facing views, resulting in insufficient viewpoint coverage. We propose a hybrid training strategy that combines 2D motions extracted from videos and local 2D poses reprojected from reconstructed 3D motions obtained using off-the-shelf estimators. Third, we develop a multi-view 2D motion diffusion model conditioned on camera trajectories for human motion and extend it to HOI reconstruction from single-view 2D inputs. Finally, we further extend our HOI reconstruction approach to handle category-specific object keypoints tracked from real-world videos, enhancing its generalization at inference to diverse real-world videos beyond models trained with object-specific reprojected MoCap data in MVLift.

3.1. 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 $\mathbf{C}_t \in \mathbb{R}^{4\times 3}$ 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

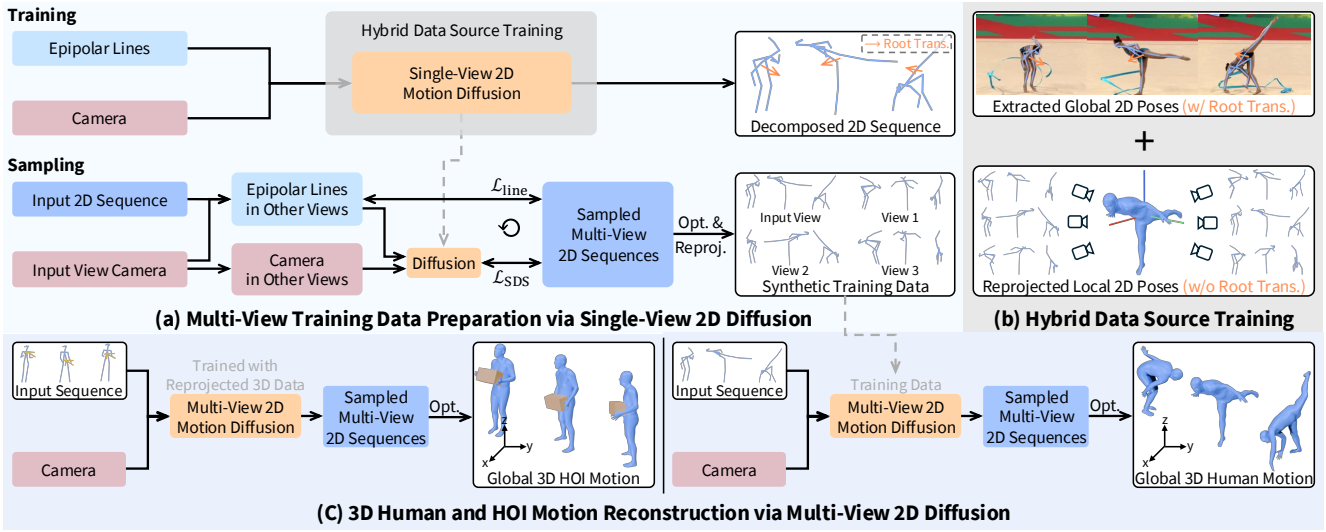

Figure 2. **Overview of AnyLift.** (a) We first train a single-view 2D motion diffusion model conditioned on camera trajectories and epipolar lines to synthesize multi-view 2D training data. (b) During training, we employ a hybrid data source strategy that enhances viewpoint coverage by combining global 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.

$\mathbf{L}_t \in \mathbb{R}^{K\times 3}$. 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 [6], 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 [38], we adopt a Transformer-based backbone [40] 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. [18], 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++ [19], which provide multi-view videos with uniformly distributed camera viewpoints, Internet videos of specific motion categories like gymnastics 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) 2D keypoints extracted from real Internet videos, and (2) 2D projections $\mathbf{X}^{\text{proj}}$ obtained by reprojecting reconstructed 3D motions from off-the-shelf estimators [34]. Since these estimators are generally reliable only for local pose estimation, we use only their local 2D projections with the root joint aligned to the image center, while discarding global translation.

However, including $\mathbf{X}^{\text{proj}}$ in training biases the model toward learning motion patterns with limited global translation, since these sequences lack root movement. 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 a binary mask that excludes the two hip joints from the 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 [31] with multi-view consistency loss to prepare multi-view training data. Specifically, 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 that the 2D keypoints $\mathbf{X}_v$ satisfy the corresponding geometric 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. Unlike MVLift [18], we apply the line-matching loss only between adjacent views and between each view and the input view during score distillation sampling for computational efficiency.

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 [25] using VPoser [29], 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.

3.2. Multi-View 2D Motion Diffusion

We train a multi-view 2D motion diffusion model using the data synthesized in Sec. 3.1 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 keypoint sequences. Camera trajectories are represented in the same way as described in Sec. 3.1. During inference, we extract 2D human keypoint sequences using ViTPose [46] and estimate camera motion with MegaSaM [21].

Model Architecture. We extend the single-view 2D motion diffusion model introduced in Sec. 3.1. 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 [18].

3.3. 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. The human keypoints and camera motions are extracted in the same way as for human motion reconstruction, while the object keypoints are tracked using DELTA [28]. The 3D mesh of each object is captured using a free 3D scanner, providing high-fidelity geometry for subsequent reconstruction and alignment.

The SMPL parameters $\mathcal{H} = (\mathbf{r}, \boldsymbol{\phi}, \boldsymbol{\Theta})$ are obtained through the final optimization process described in Sec. 3.1. 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 [53], $\mathbf{t}_t \in \mathbb{R}^3$ is the translation, and $s \in \mathbb{R}^+$ is a global scale factor, as detailed in Sec. B.2.

4. Experiments

We evaluate AnyLift on human motion reconstruction and human-object interaction reconstruction. Training and evaluation are conducted using both MoCap datasets with 3D ground truth and our collected video-only datasets.

4.1. Datasets and Evaluation Metrics

Human Motion Reconstruction. For human motion lifting, we evaluate our method on the **AIST++** [19] dataset and two newly collected video-only datasets of **gymnastics** and **martial arts**. We train and test AnyLift on these datasets separately. AIST++ provides multi-view RGB videos with corresponding 3D dance motions. We conduct evaluations under two settings: (1) following MVLift [18], we use its processed version of the AIST++ dataset for training and testing, where only one camera view is available for each

sequence; and (2) since AIST++ does not include dynamic camera videos, we synthesize additional training and evaluation data by projecting the 3D ground truths with simulated moving cameras to assess AnyLift's capability under dynamic cameras.

The collected gymnastics and martial arts videos feature highly dynamic motions involving complex body movements, typically captured with moving cameras and limited view coverage. Such sequences are rarely present in existing MoCap datasets. Evaluating these challenging cases allows us to assess AnyLift 's ability to handle scenarios that are difficult for methods relying on MoCap data. We process the collected videos using ViTPose [46] to obtain 2D pose sequences and estimate camera poses with MegaSaM [21].

We follow the evaluation protocol of MVLift [18]. We project the reconstructed 3D motion to 2D and calculate the 2D joint position error ($\mathbf{J}_{2D}$) and the 2D joint position error with the 2D root joint translated to the image center ($\mathbf{J}^{C}_{2D}$). To assess motion realism, we train a 2D motion feature extractor for each dataset and compute the Fréchet Inception Distance (**FID**) on 2D motion features. For datasets with 3D ground truth, we further evaluate performance using the root translation error ($\mathbf{T}_{root}$), the mean per-joint position error (**MPJPE**), and the Procrustes-aligned mean per-joint position error (**PA-MPJPE**). In addition, we assess the foot sliding score (**FS**) following He et al. [5].

HOI Motion Reconstruction. We evaluate AnyLift 's capability for HOI reconstruction on the **BEHAVE** dataset [1]. We train separate models for three object categories: *box*, *chair*, and *table*. For training data, we use interaction sequences involving these objects from the BEHAVE training set, together with additional data from InterCap [7], HODome [51], IMHD [52], and OMOMO [17]. Following the evaluation protocol used for AIST++, we perform experiments under two settings: (1) static-camera setting using the original camera poses from the BEHAVE test set, and (2) synthetic dynamic-camera setting obtained via simulated camera motion. Object keypoint sequences are obtained by reprojecting 3D HOIs from the BEHAVE dataset. We also provide an alternative optimization-based method for extracting object 2D keypoints (Sec. C.2), with qualitative results available on our project page.

Following the human motion lifting evaluation, we assess human motion accuracy using $\mathbf{T}_{root}$, **MPJPE**, and **PA-MPJPE**. For object motions, we evaluate the object translation error ($\mathbf{T}^{O}_{root}$) and the object mean per-joint position error (**O-MPJPE**), which respectively quantify the translation and keypoint accuracy of reconstructed object movements.

4.2. Human Motion Reconstruction

Baselines. We compare our method with two categories of baselines: methods that do not rely on 3D motion data during training, and methods trained with 3D motion ground truth. SMPLify [2] estimates SMPL [25] parameters by optimizing 2D reprojection objectives without requiring any training. In contrast, WHAM [36] and GVHMR [34] leverage AMASS [26] together with several other datasets for training. We further adapt MVLift [18] to the dynamic camera setting by first running its original pipeline and then applying the estimated camera motion to its outputs.

Method	$\mathbf{J}_{2D}$	$\mathbf{J}^{C}_{2D}$	FID	$\mathbf{T}_{root}$	MPJPE	PA-MPJPE	FS
SMPLify [2]	24.3	14.7	2.4	90.3	174.7	132.8	1.548
WHAM [36]	75.5	22.1	3.1	164.3	104.8	75.1	0.579
GVHMR [34]	106.4	20.3	2.9	143.0	**97.6**	**64.4**	0.547
MVLift [18]	17.5	14.3	2.2	67.6	110.7	79.2	**0.471**
AnyLift (ours)	**16.6**	**13.3**	**2.1**	**64.9**	108.0	82.3	0.475
SMPLify [2]	24.7	14.9	2.7	90.9	175.2	134.6	1.530
MVLift [18]	18.0	14.9	2.1	64.9	122.1	94.3	0.487
AnyLift (ours)	**16.7**	**13.7**	**2.0**	**64.2**	**109.3**	**83.0**	**0.446**

Table 1. **Quantitative evaluation on the AIST++ dataset [19]** under (1) static-camera setup (upper) and (2) dynamic-camera setup (lower). AnyLift achieves competitive 3D joint accuracy and improved root translation estimation while maintaining robustness under dynamic camera.

Method	Gymnastics				Martial Arts			
	$\mathbf{J}_{2D}$	$\mathbf{J}^{C}_{2D}$	FID	FS	$\mathbf{J}_{2D}$	$\mathbf{J}^{C}_{2D}$	FID	FS
SMPLify [2]	39.2	16.0	11.2	0.463	48.0	13.2	5.8	0.397
WHAM [36]	88.6	21.7	16.4	0.245	87.4	22.3	10.7	0.212
GVHMR [34]	71.5	18.8	13.0	**0.061**	66.3	15.9	6.0	**0.056**
MVLift [18]	33.1	17.0	11.2	0.188	24.6	12.0	4.6	0.145
AnyLift (ours)	**21.6**	**11.4**	**10.9**	0.152	**15.1**	**9.8**	**3.6**	0.136

Table 2. **Quantitative evaluation on our collected Internet videos.** AnyLift outperforms all baselines across most metrics, demonstrating the plausibility of our method on Internet videos.

Method	Ground Contact	Motion Quality
vs. SMPLify [2]	84.2%	85.0%
vs. WHAM [36]	75.6%	74.2%
vs. GVHMR [34]	75.5%	61.7%
vs. MVLift [18]	61.3%	65.4%
vs. AnyLift w/o Hybrid	65.6%	66.3%

Table 3. **Human study on reconstructed human motions from our collected Internet videos.** Participants prefer our reconstruction results for their better ground contact and motion quality.

Results. We report quantitative results on the AIST++ dataset [19] in Tab. 1. Under the original MVLift evaluation setting with static cameras (upper), AnyLift outperforms all baselines across most metrics, including the original MVLift implementation. In terms of 3D joint position errors, our method achieves comparable accuracy to WHAM and GVHMR, which require training on AMASS, while showing a substantial improvement in root translation accuracy. Under the synthetic dynamic-camera setting (lower), AnyLift significantly outperforms both SMPLify and MVLift across all metrics. Notably, AnyLift maintains similar accuracy on 3D joint position errors between the two settings, high-

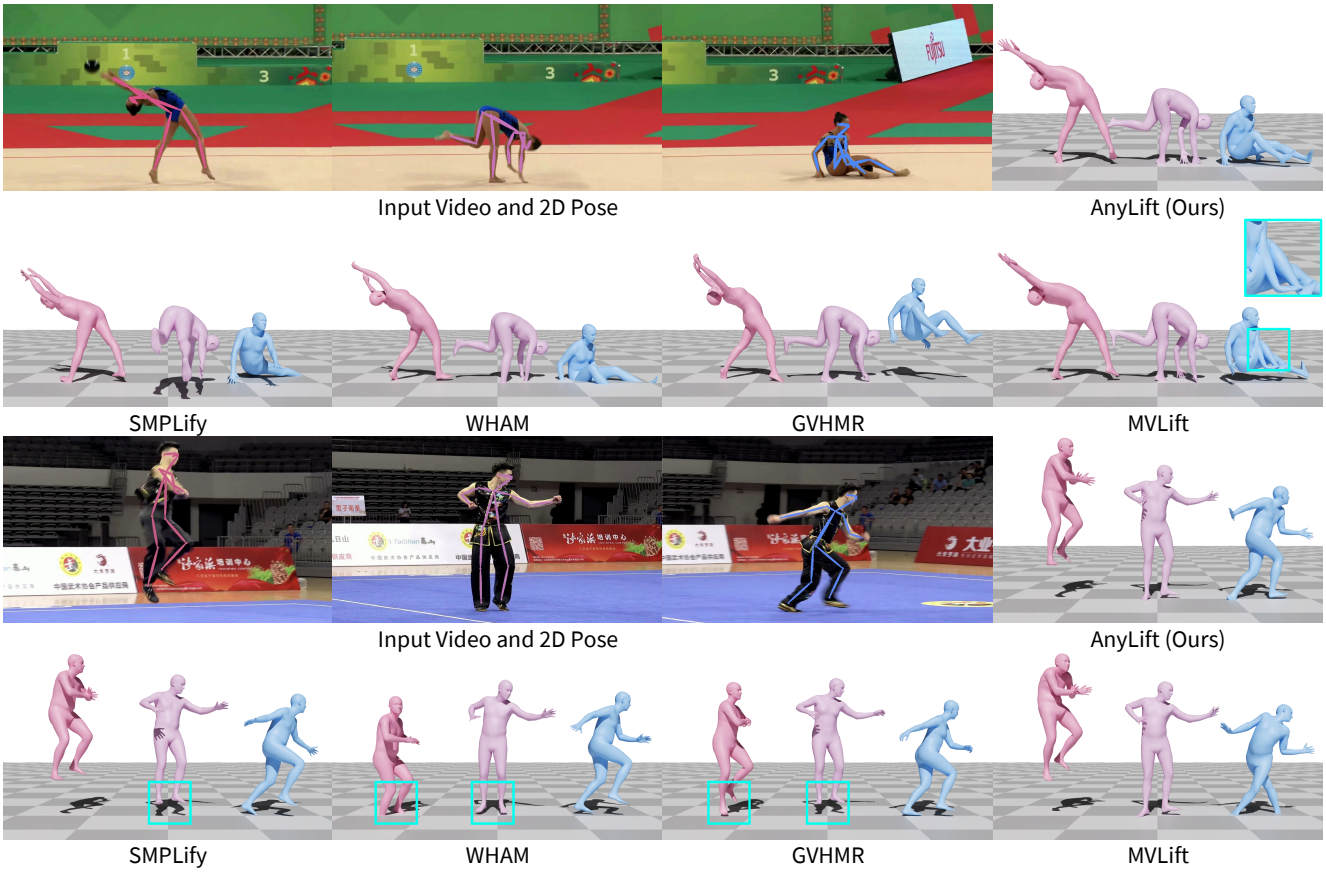

Figure 3. **Qualitative comparison of human motion reconstruction on our collected Internet videos.** AnyLift produces more plausible motions, mitigating the root trajectory errors, inaccurate local body pose, and self-penetration artifacts observed in baselines.

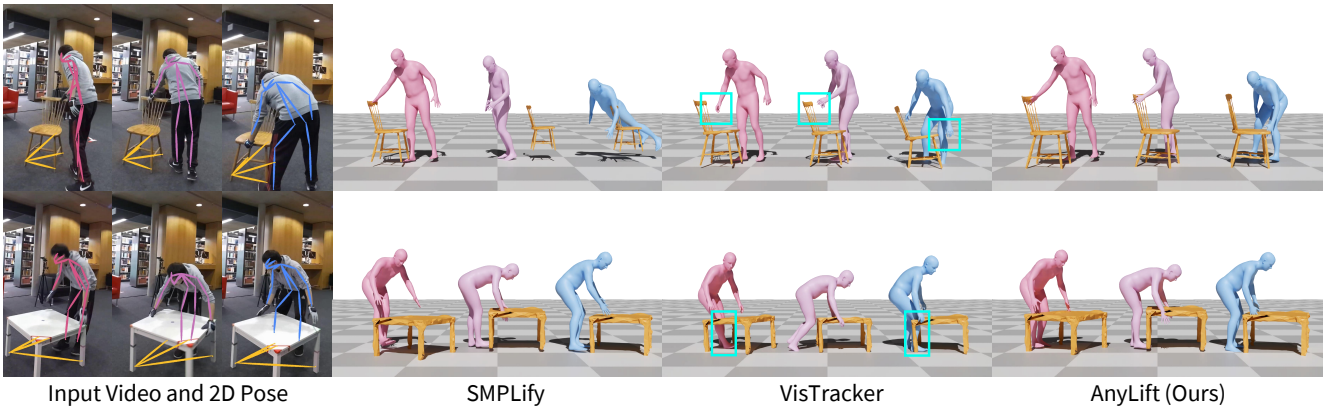

Figure 4. **Qualitative comparison of HOI reconstruction on the BEHAVE [1] dataset.** We show results on two object categories, *chair* and *table*. AnyLift produces coherent and physically plausible human-object interactions with accurate contact and minimal penetration.

lighting its robustness to real-world camera motion. The quantitative results on our collected Internet videos are presented in Tab. 2. AnyLift achieves superior performance over all baseline methods across most metrics.

We show qualitative comparisons in Fig. 3. 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 suffers from self-penetration (first example) and yields severely distorted leg poses in the second example. In contrast, AnyLift reconstructs stable and plausible human motions with accurate

Method	Box					Chair					Table				
	T_{root}	MPJPE	PA-MPJPE	T^{O}_{root}	O-MPJPE	T_{root}	MPJPE	PA-MPJPE	T^{O}_{root}	O-MPJPE	T_{root}	MPJPE	PA-MPJPE	T^{O}_{root}	O-MPJPE
SMPLify [2]	78.12	114.56	86.64	795.06	207.77	64.86	102.76	80.75	438.98	145.35	70.06	109.31	86.10	385.60	124.86
VisTracker [45]	51.72	54.40	50.40	143.59	359.50	52.37	72.72	67.84	90.18	134.95	65.18	85.51	82.73	177.17	540.96
AnyLift (ours)	**24.61**	**42.68**	**35.62**	**95.38**	**32.98**	**22.48**	**53.58**	**43.85**	**75.93**	**34.54**	**26.05**	**48.34**	**39.06**	**77.53**	**51.28**
SMPLify [2]	82.21	126.12	89.85	669.64	185.48	63.37	107.22	81.73	555.70	165.90	77.19	119.42	78.03	514.62	149.91
AnyLift (ours)	**29.99**	**43.60**	**39.02**	**96.76**	**33.96**	**23.87**	**57.82**	**45.25**	**79.16**	**44.60**	**28.09**	**54.60**	**42.91**	**88.93**	**56.97**

Table 4. **Quantitative evaluation on the BEHAVE dataset [1]** under (1) static-camera setup (upper) and (2) dynamic-camera setup (lower). AnyLift outperforms all baselines across object categories and achieves robust performance under dynamic-camera conditions.

global trajectories and consistent body poses across frames.

Human Perceptual Study. We conduct a human perceptual study using the two-alternative forced choice (2AFC) method. A total of 300 participants are recruited via the Prolific platform. In each trial, participants are presented with reconstruction results from different methods and asked to choose the one that exhibits better ground contact and motion quality. As shown in Tab. 3, participants consistently prefer our results over the baselines, aligning well with the quantitative metrics and qualitative comparisons.

4.3. HOI Motion Reconstruction

Baselines. 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 [45] jointly tracks 3D humans, objects, and contacts from monocular videos using a visibility-aware object pose network.

Results. Quantitative results on the BEHAVE dataset are shown in Tab. 4. Across all object categories (*box*, *chair*, *table*) and both static and dynamic camera settings, AnyLift consistently outperforms the baselines in all metrics. SMPLify, which relies purely on optimization based on 2D reprojection objectives, struggles to recover accurate 3D motion and translation, often converging to suboptimal local minima when 2D evidence is ambiguous. VisTracker suffers from temporal jitter and ambiguity when handling symmetric objects, as its visibility-aware module may yield inconsistent pose predictions across frames. These issues lead to large errors in object-related metrics (T^{O}_{root} and O-MPJPE). In contrast, AnyLift jointly predicts human and object dynamics within a unified framework, achieving stable and accurate reconstruction even under challenging dynamic-camera conditions. The small performance gap between static and dynamic settings further demonstrates its strong generalization to realistic camera motion.

Qualitative comparisons on the BEHAVE dataset are presented in Fig. 4. We show two object categories: *chair* and *table*. Similar to human motion reconstruction, 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 coherent and plausible human-object interactions with accurate contact and minimal penetration under both categories. We further present qualitative results on real-world videos captured in natural environments in Fig. 1.

Method	Gymnastics			Martial Arts		
	J_{2D}	J^{C}_{2D}	FID	J_{2D}	J^{C}_{2D}	FID
AnyLift w/o Hybrid	36.1	18.7	11.2	25.2	12.7	4.3
AnyLift (ours)	**21.6**	**11.4**	**10.9**	**15.1**	**9.8**	**3.6**

Table 5. **Ablation study on our collected Internet videos.** Performance drops across all metrics without incorporating local 2D poses from diverse viewpoints.

4.4. Ablation Study

We conduct an ablation study on our collected Internet video datasets, which are captured with limited camera views. We remove the hybrid training strategy that incorporates local poses estimated by GVHMR and train AnyLift solely on 2D motion sequences extracted from the videos. The quantitative results are presented in Tab. 5. Our full model achieves superior performance across all metrics compared to the ablated variants. Together with the comparison to GVHMR [34] in Tab. 2, these results demonstrate that combining estimated local 2D poses from diverse viewpoints with global 2D motions directly extracted from videos is a more effective design choice.

5. 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 our approach on newly collected Internet videos featuring complex human motions, as well as on our captured in-the-wild HOI videos.

Acknowledgement. We thank Chen Geng, Yanzhe Lyu, Yichao Zhou, and Yongqian Peng for fruitful discussions. We also extend our sincere thanks to Yazhou Zhang, Cyrus Zhou, and Di Fan for their support in data collection. This work is in part supported by the Stanford Institute for Human-Centered AI (HAI) and ONR MURI N00014-22-1-2740.

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. 3, 6, 7, 8, S1, S2

[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. 6, 8

[3] 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 *International Conference on Computer Vision (ICCV)*, 2019. 2

[4] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[5] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[7] 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. 3, 6

[8] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[9] Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H Bermano. Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. S1, S3

[11] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[12] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[13] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and motion estimation from in-the-wild videos. In *International Conference on 3D Vision (3DV)*, 2024. 2

[14] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[15] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[16] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[17] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6), 2023. 3, 6

[18] 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. 2, 3, 4, 5, 6

[19] 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. 4, 5, 6, S1

[20] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[21] 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. 5, 6

[22] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[23] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations (ICLR)*, 2024.

[24] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[25] 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. 2, 3, 5, 6

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 6

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[28] 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. 5, S2, S3

[29] 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. 5

[30] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[31] 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. 5

[32] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. S2

[33] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[34] 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. 1, 4, 6, 8

[35] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[36] 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. 1, 2, 6

[37] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[38] 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. 4

[39] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[40] 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. 4

[41] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[42] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[43] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[44] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[45] 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. 2, 8

[46] 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. 5, 6

[47] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[48] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[49] 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. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10):12287–12303, 2023. 2

[50] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[51] 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. 3, 6

[52] 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. 3, 6

[53] 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. 5, S1, S2

A. Overview

In this supplementary material, we provide additional details on implementation (Sec. B) and video data processing (Sec. C). We highly recommend viewing our *project page* for compelling demonstrations.

B. Additional Implementation Details

B.1. Details on Model Training

We train the single-view 2D motion diffusion model for 300,000 steps and the multi-view diffusion model for 120,000 steps. The Adam optimizer[10] is used with a learning rate of 1×10^{-4}, a batch size of 64, and no weight decay. All training is performed on a single NVIDIA L40S GPU. Model checkpoints are saved every 20,000 steps, and the final model is selected based on validation performance. During the training process of multi-view diffusion, we randomly mask out a subset of input keypoints to handle partial occlusions and potential tracking failures.

For the hybrid data-source training, we use a data ratio of 2:1 between 2D keypoints extracted from real Internet videos and local 2D projections reprojected from reconstructed 3D motions.

B.2. Details on Object Motion Optimization

We provide details on how we recover the 3D object motion from our sampled multi-view 2D results. We 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 $\mathbf{P} \in \mathbb{R}^{M\times 3}$ 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 [53], $\mathbf{t}_t \in \mathbb{R}^3$ is the translation, and $s \in \mathbb{R}^+$ is a global scale factor. We denote the mapping from $\mathbf{r}_t$ to the rotation matrix as $\mathbf{R}_t = \mathrm{Rot}(\mathbf{r}_t) \in \mathrm{SO}(3)$.

We first initialize each frame independently by aligning the canonical keypoints $\mathbf{P}$ with the observed 3D keypoints $\mathbf{Q}_t$ through a rigid transformation and scale estimation as:

$$\hat{\mathbf{r}}_t, \hat{\mathbf{t}}_t = \arg\min_{\mathbf{r}_t,\mathbf{t}_t} \sum_{i=1}^{M} \left\| \hat{s}\,\mathrm{Rot}(\mathbf{r}_t)\mathbf{p}_i + \mathbf{t}_t - \mathbf{q}_{i,t} \right\|_2, \quad \text{(S1)}$$

where the global scale is computed from the distance ratio of a reference keypoint pair in the observed and canonical spaces:

$$\hat{s} = \frac{\| \mathbf{q}_{1,t} - \mathbf{q}_{2,t} \|_2}{\| \mathbf{p}_1 - \mathbf{p}_2 \|_2}. \quad \text{(S2)}$$

This initialization yields per-frame pose estimates without temporal coupling. After initialization, we further optimize the object pose sequence with the object fitting loss:

$$\mathcal{L}_{\text{fit}}^{\text{obj}} = \frac{1}{TM} \sum_{t=1}^{T} \sum_{i=1}^{M} m_i \left\| s\,\mathrm{Rot}(\mathbf{r}_t)\mathbf{p}_i + \mathbf{t}_t - \mathbf{q}_{i,t} \right\|_2, \quad \text{(S3)}$$

where $m_i \in \{0, 1\}$ is a point-wise visibility mask shared across all frames.

Additionally, we apply a regularization loss on consecutive rotations to ensure temporal smoothness:

$$\mathcal{L}_{\text{smooth}}^{\text{obj}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| \mathbf{r}_t - \mathbf{r}_{t+1} \right\|_2. \quad \text{(S4)}$$

The optimization loss for object motion is defined as:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{fit}}^{\text{obj}} + \lambda_{\text{smooth}}^{\text{obj}} \mathcal{L}_{\text{smooth}}^{\text{obj}}. \quad \text{(S5)}$$

The Adam optimizer [10] is employed for all optimization stages. To reconstruct the 3D keypoint sequence from the sampled multi-view 2D motions, we run 500 optimization iterations with a learning rate of 0.01. During the object pose fitting stage, we perform 2,000 iterations with a learning rate of 0.05.

B.3. Details on Predefined Camera Trajectory

To enrich camera trajectory diversity, we predefine six camera movement modes: zoom in, zoom out, move left, move right, rotate clockwise, and rotate counterclockwise. Each mode generates a sequence of camera transformations simulating the corresponding motion pattern. For each sequence, we randomly determine whether the camera keeps moving in one direction throughout the sequence or moves back to its original position after reaching the maximum displacement. In addition, with a certain probability, the camera tracks the pelvic joint to keep the human roughly centered in the view.

During training, the camera configuration is dynamically determined at each step. For each 3D motion sequence, we randomly choose to use either a camera trajectory extracted from real Internet videos (70%) or one of the predefined camera trajectories (30%) to reproject it into 2D for training.

B.4. Details on Experimental Setup

We provide details on how we synthesize additional training and evaluation data by projecting the 3D ground-truth motions using simulated moving cameras on the AIST++ [19] and BEHAVE [1] datasets. Following our hybrid data-source training, we reproject the 3D data using camera viewpoints sampled from a large camera motion base estimated from our gymnastic and martial arts videos, and we also incorporate a small set of predefined camera trajectories (Sec. B.3) with a ratio of 7:3 between the two sources.

We ensure that camera motions used for training do not overlap with those in the testing set. The estimated camera motion base is inherently divided into separate training and testing subsets to enforce this separation. For the predefined cameras, the range of motion, the decision of whether to return to the origin, and whether to track the human pelvic joint are all randomized, ensuring that no camera trajectory in the testing phase exactly matches any trajectory seen during training.

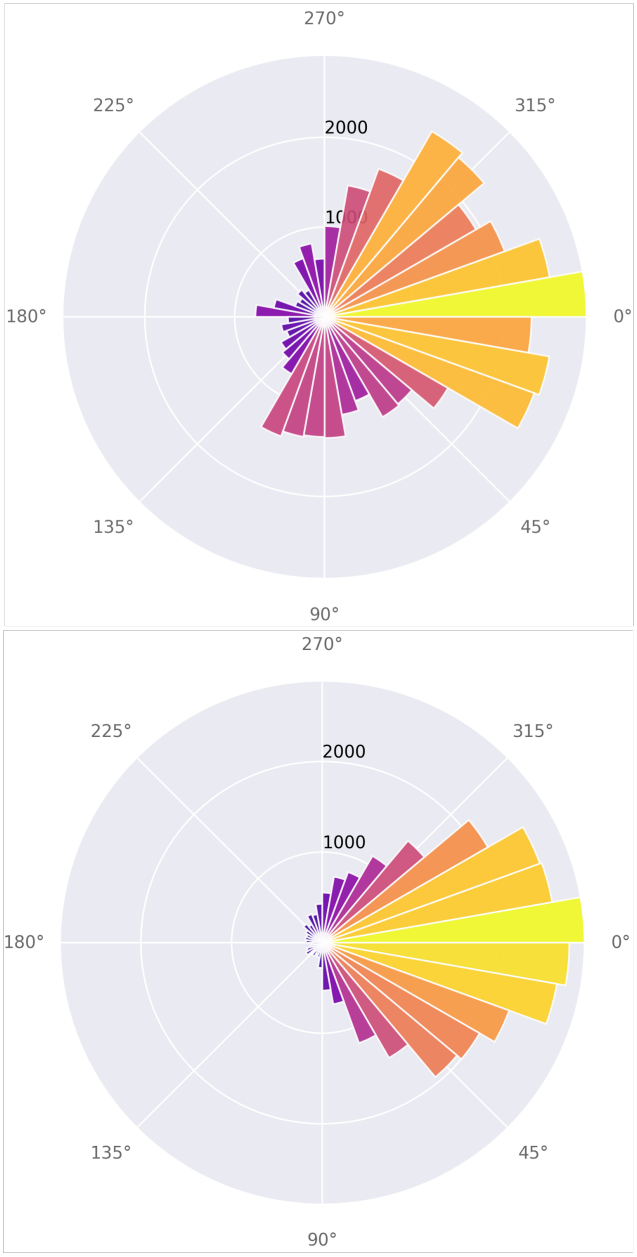

Figure S1. **Facing direction distributions of estimated humans** in the gymnastics (upper) and martial arts (lower) videos under the camera coordinate system. The angular axis indicates the facing direction and the radial axis represents number of frames.

C. Additional Details on Data Processing

C.1. Details on Internet Video Processing

We provide details on how we process the manually collected Internet videos. We first filter out low-quality videos, such as those containing advertisements or long idle segments. The remaining videos are then divided into 10-second clips, resulting in roughly 4,000 gymnastics clips and 5,000 martial arts clips. In subsequent processing, we further filter out these clips to handle failures in human tracking and 2D pose estimation. Segments with low pose confidence score or severe 2D jittering are discarded. After the filtering process, we obtain around 1,600 gymnastics clips and 3,000 martial arts clips that are used for training and evaluation. The processed data are then split into training and testing sets with a 9:1 ratio.

As shown in Fig. S1, we plot the facing directions of the estimated humans in the gymnastics and martial arts videos in the camera coordinate system using polar plots, where the angular axis represents the facing direction and the radial axis corresponds to the number of frames. The results show that the facing directions in both datasets are mostly concentrated within the same semicircle, which supports our observation that the viewpoint coverage of online videos is severely limited.

C.2. Details on HOI Video Processing

Processing Videos from the BEHAVE Dataset. The BEHAVE dataset [1] involves complex object motions, particularly frequent rotations, which makes off-the-shelf tracking methods [28] unreliable. To obtain more reliable 2D keypoints, we avoid tracking in the image plane and instead recover the per-frame object pose by aligning the object mesh to the observed object mask using a point sampling, projection-based method.

Concretely, for each object, we start from its template mesh and pre-select a small set of mesh vertices as semantic keypoints. In each frame, we first obtain an object segmentation mask using Grounded-SAM [32], and randomly sample 5,000 2D points $\mathcal{Q} = \{\mathbf{q}_j\}$ from the foreground pixels. We also randomly sample 5,000 3D surface points $\mathcal{X} = \{\mathbf{x}_i\}$ from the object mesh. For projection, we fix a pinhole camera model with focal length $f_x = f_y = 1000$ and principal point $(c_x, c_y) = (\frac{W}{2}, \frac{H}{2})$ determined by the image width W and height H. With these intrinsics and an unknown $\mathrm{SE}(3)$ pose $(\mathbf{R}, \mathbf{t})$ of the object in the camera coordinate system, each 3D point is projected to the image as

$$\mathbf{p}_i = \Pi(\mathbf{R}\mathbf{x}_i + \mathbf{t}), \tag{S6}$$

where $\Pi(\cdot)$ denotes the pinhole projection. Let $\mathcal{P} = \{\mathbf{p}_i\}$ be the set of projected points. We estimate $(\mathbf{R}, \mathbf{t})$ by minimizing a symmetric 2D Chamfer distance between $\mathcal{P}$ and $\mathcal{Q}$:

$$\mathcal{L}_{\text{chamfer}} = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p}_i \in \mathcal{P}} \min_{\mathbf{q}_j \in \mathcal{Q}} \|\mathbf{p}_i - \mathbf{q}_j\|_2^2 + \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{q}_j \in \mathcal{Q}} \min_{\mathbf{p}_i \in \mathcal{P}} \|\mathbf{q}_j - \mathbf{p}_i\|_2^2. \tag{S7}$$

In implementation, we parameterize $\mathbf{R}$ using a continuous 6D rotation representation [53] and initialize $\mathbf{t}$ heuristically

from the 2D mask bounding box and the object's 3D extent, which stabilizes optimization under large rotations. For the first frame of each sequence, we perform 200 random restarts of $\mathbf{R}$ and retain the solution with the lowest Chamfer loss. For each subsequent frame, we use the optimized pose from the previous frame as initialization, allowing the optimizer to refine the pose smoothly over time. After convergence, we apply the estimated $(\mathbf{R}, \mathbf{t})$ to the predefined semantic keypoints and project them into the image, producing temporally consistent 2D object keypoints on BEHAVE RGB videos. The Adam optimizer [10] is adopted in this optimization process.

Processing Captured Real-World Videos. For the videos we captured, we manually select a small set of visible object keypoints in the first RGB frame and track them across the sequence using DELTA [28]. The reconstructed human-object interactions and tracked object keypoints are visualized on our project webpage.