This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MoCapAnything: Unified 3D Motion Capture for Arbitrary Skeletons from Monocular Videos

Kehong Gong*,1, Zhengyu Wen*,2, Weixia He2, Mingxi Xu2, Qi Wang2, Ning Zhang2,
Zhengyu Li2, Dongze Lian2, Wei Zhao2, Xiaoyu He2, Mingyuan Zhang†,2
[1]Huawei International Pte. Ltd., [2]Huawei Central Media Technology Institute
* Equal Contributions, † Corresponding Author

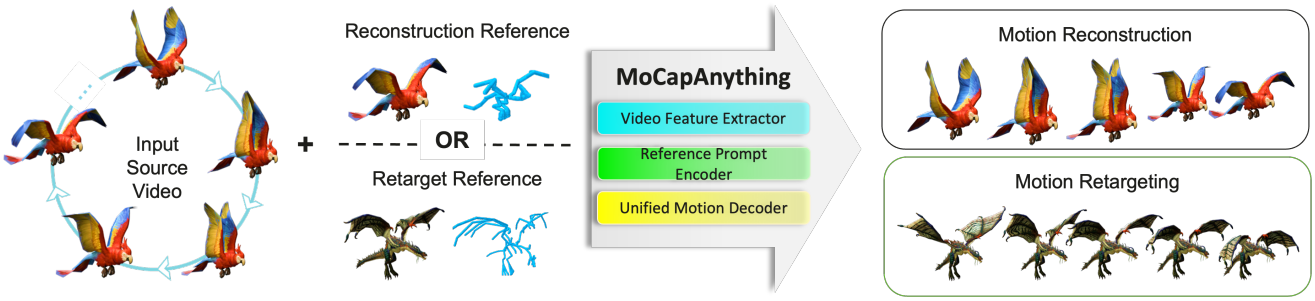


Figure 1. Overview of our MoCapAnything framework. Given a monocular video and a reference 3D asset (mesh/skeleton/appearance), our system first reconstructs a 4D mesh sequence from the video and encodes the reference asset via a multi-modal prompt encoder. A unified motion decoder then predicts joint trajectories, followed by an IK fitting stage that outputs animation in the asset's own rig convention. The framework supports both direct motion capture (reference matches video subject) and cross-asset retargeting (reference differs from video).

Abstract

*Motion capture now underpins content creation far beyond digital humans, yet most pipelines remain species- or template-specific. We formalize this gap as Category-Agnostic Motion Capture (CAMoCap): given a monocular video and an arbitrary rigged 3D asset as a prompt, the goal is to reconstruct a rotation-based animation (e.g., BVH) that directly drives the specific asset. We present MoCapAnything, a reference-guided, factorized framework that first predicts 3D joint trajectories and then recovers asset-specific rotations via constraint-aware Inverse Kinematics (IK) Fitting. MoCapAnything comprises three learnable modules and a lightweight IK stage: a **Reference Prompt Encoder** that distills per-joint queries from the asset's skeleton, mesh, and rendered image set; a **Video Feature Extractor** that computes dense visual descriptors and reconstructs a coarse 4D deforming mesh to bridge the modality gap between RGB tokens and the point-cloud–like joint space; and a **Unified Motion Decoder** that fuses these cues to produce temporally coherent trajectories. We also curate Truebones Zoo [37] with 1,038 motion clips, each providing a standardized skeleton–mesh–rendered-video triad. Experiments on in-domain benchmarks and in-the-wild videos show that MoCapAnything delivers high-quality skeletal animations and exhibits non-trivial cross-species retargeting across heterogeneous rigs, offering a scalable path toward prompt-based 3D motion capture for arbitrary assets. The codes are available on our project page: https://animotionlab.github.io/MoCapAnything/*

1. Introduction

Motion capture underpins modern content creation beyond digital humans, yet most pipelines remain tied to a single species or template. Human-centric systems typically regress SMPL-family [22, 26] parameters from monocular inputs (e.g., DeepPose [35] for 2D keypoints and HMR [13] for SMPL-based 3D recovery) and work well only within that fixed topology. For non-human subjects, category-agnostic keypoint detection (CAPE) broadens 2D landmark coverage via promptable support examples, but it stops short of producing animation-ready 3D motion [28]. On the motion side, animal mocap usually builds on SMAL [54] and is limited to a few quadruped categories, with models and rig assumptions that do not transfer to diverse assets. Consequently, existing solutions fall short in practi-

cal pipelines where creators must (i) retarget human/animal motion to non-biological rigs (robots, mechs, toys, articulated props), (ii) animate large heterogeneous asset libraries for games and crowd scenes, (iii) drive VTuber/virtual-production avatars that frequently change topology, and (iv) spin up IP-specific characters (mascots, creatures) without building a new parametric model per species.

To address the limitations of fixed-species motion capture, we recast the problem as prompt-based 3D motion capture: given a monocular video and an arbitrary rigged 3D asset, the goal is to reconstruct a rotation-based animation (e.g., BVH joint rotations) that directly drives that specific character. We refer to this setting as Category-Agnostic Motion Capture (CAMoCap). To make this concrete and reproducible, we curate the Truebones Zoo benchmark, where each motion instance provides a clean bundle comprising the rigged skeleton (with standardized joint names and hierarchy), the mesh, and an asset-aligned rendered video. The dataset contains 1,038 motion clips. We hold out 60 for the test set and use the remaining 978 for training.

CAMoCap raises three core challenges. First, *motion representation*: joint rotations are defined in asset-local frames, so direct angle regression across diverse rest poses is brittle. Second, *reference-guided estimation*: the model must inject information about the target asset into video-based 3D keypoint prediction effectively. Third, *multimodal integration*: there is a gap between dense RGB features and the point-cloud–like structure of keypoints. Bridging them naively may lead to suboptimal accuracy.

To tackles above mentioned challenges, we propose a novel framework, MoCapAnything, which factorizes motion recovery into (i) 3D keypoint trajectory prediction and (ii) per-joint rotation recovery. It uses three learnable modules followed by a lightweight IK stage. The **Reference Prompt Encoder** distills the asset's mesh, skeleton, and rendered image set into structure-aware per-joint queries. The **Video Feature Extractor** computes dense visual descriptors (e.g., DINOv2 [25]) and reconstructs a coarse 4D deforming mesh from the input video. This mesh contributes topology- and geometry-aware cues that bridge the gap between RGB tokens and the point-cloud nature of joints. The **Unified Motion Decoder** attends over reference queries and video features to produce temporally consistent 3D joint trajectories. Finally, IK Fitting converts these trajectories into asset-specific rotations while respecting hierarchy, bone lengths, joint limits, and temporal smoothness. This modular factorization naturally supports both motion capture (same skeleton) and retargeting (different skeletons) across heterogeneous rigs.

Our main contributions are summarized as follows:

1. We formalize a new task, *Category-Agnostic Motion Capture* (CAMoCap), prompt-based 3D motion capture from a monocular video and an arbitrary rigged 3D asset. We also release reorganized Truebones Zoo [37] with 1,038 clips, each providing a skeleton–mesh–rendered-video triad.
2. We present the first framework for CAMoCap, MoCapAnything, to yield temporally coherent, animation-ready results across heterogeneous rigs. Specifically, we decouple motion into 3D joint trajectories followed by IK-based rotations to stablize training process and introduce mesh as an auxiliary modality to bridge RGB tokens and joint space.
3. MoCapAnything attains strong in-domain accuracy, generalizes to in-the-wild videos, and shows non-trivial cross-species mocap and retargeting.

2. Related Works

2.1. Pose Estimation

Human 2D pose estimation aims to localize anatomical keypoints in images. Classic methods are typically grouped into bottom-up and top-down paradigms: bottom-up approaches first detect all keypoints and then group them into person instances [4], while top-down pipelines detect person bounding boxes and run a single-person pose head on each crop [42]. Within the top-down family, heatmap-based networks such as Stacked Hourglass [23], CPN [3], SimpleBaseline [42], HRNet [33], Simple Pose [17], and ViTPose [45] predict per-joint likelihood maps from multi-scale or high-resolution features, whereas regression-style methods including DeepPose [36], RLE [18], and SimCC [19] directly output coordinates or 1D classifications to alleviate heatmap quantization. More recent DETR-style frameworks [2] treat poses and/or keypoints as query sets and perform end-to-end multi-person estimation without handcrafted grouping [30, 43, 50], and vision–language approaches such as LocLLM [38] encode keypoints as text descriptions to enable some zero-shot generalization to new landmarks; however, all these architectures remain tightly coupled to a predefined human skeleton [10, 11] and keypoint set.

Beyond these category-specific keypoint detectors, an emerging line of work aims to relax the dependence on fixed object categories through category-agnostic pose estimation (CAPE). CAPE formulates pose estimation as a few-shot problem, where a single model predicts keypoints for unseen categories by comparing support keypoints with query images in a shared embedding space [44]. POMNet [44] instantiates CAPE with a transformer encoder over query images and support keypoints, and regresses similarity scores from their concatenated features. CapeFormer [31] further adopts a two-stage matching framework that first proposes candidate correspondences and then refines unreliable matches to improve localization accuracy. Pose Anything [12] departs from treating keypoints as isolated en-

tities and instead models them as nodes in a graph, using graph convolutions to exploit structural relationships, break symmetries, and better handle occlusions. More recently, CapeX [27] pushes CAPE beyond purely visual correspondence by replacing annotated support images with text prompts attached to graph nodes, aligning query image features to open-vocabulary textual keypoint descriptions. While these CAPE methods significantly improve generalization across categories, they operate in 2D and focus on static keypoint localization, without modeling 3D trajectories, temporal consistency, or animation-ready joint representations, which are central to our monocular motion capture setting.

2.2. Motion Capture

Monocular human motion capture is typically formulated as recovering pose and shape parameters of parametric whole-body models such as SMPL [22] and SMPL-X [26]. With whole-body models, expressive human pose and shape (EHPS) estimation from a single RGB image or video—jointly modeling body, hands, and face—has attracted much attention. Early optimization-based methods (e.g., SMPLify-X [26]) fit SMPL-X to detected 2D keypoints but are slow and brittle. One-stage frameworks such as OSX [21], AiOS [34], and MultiHMR [1] instead use ViT-style backbones to jointly localize and regress full SMPL-X parameters in a single forward pass, alleviating error accumulation and improving robustness. Beyond image-aligned meshes, recent work distinguishes between camera-space and world-grounded human motion recovery. Most HMR and video-based approaches follow the camera-space formulation, regressing SMPL parameters directly from images or clips with CNN, RNN, or transformer encoders (e.g., HMR [13]/HMR2.0 [9], VIBE [15], TCMR [5]), which yields accurate pose but entangles motion with camera movements. To obtain physically meaningful trajectories, multi-camera studios and IMU-based systems rely on calibration or inertial sensors, while recent monocular methods integrate SLAM or visual odometry with learned motion priors (e.g., SLAHMR [53], PACE [16], TRAM [39], WHAC [52], WHAM [32]) to estimate global motion. However, these pipelines remain tied to a single human template and are difficult to extend to more general, non-human skeletons.

Beyond humans, 3D animal reconstruction has been explored under two main paradigms: model-free and model-based. Model-free methods make minimal assumptions about anatomy and directly recover a deformable surface, e.g., CMR [14] deforms a spherical template to reconstruct birds, while LASSIE [51], MagicPony [40], and 3D-Fauna [20] learn articulated 3D shape from image collections; ViSER [47], LASR [46], BANMo [48], and PPR [49] extend this idea to videos. In contrast, model-based approaches assume a species-specific or parametric 3D template is given or retrievable [41], enabling pose- and shape-aware analysis over time. SMAL [54], an articulated quadruped model learned from toy scans, has been widely used [24]. However, these pipelines remain species- and template-specific, and do not generalize to the diverse, non-animal skeletons required by arbitrary animatable assets.

3. Method

3.1. Task Formulation

In this work, we propose a new task, **Category-Agnostic Motion Capture** (CAMoCap), which aims to reconstruct motion for arbitrary 3D assets with diverse skeletal topologies from monocular videos. This formulation transcends traditional paradigms centered on human or category-specific mocap, enabling both motion capture and retargeting for assets of any type or skeletal structure, and thus brings broader applicability and flexibility to animation, virtual production, and creative content creation.

Formally, given a monocular RGB video $V = \{I_t\}_{t=1}^{T}$ depicting a moving character or creature, and a rigged 3D asset $A = (\mathcal{M}, \mathcal{S}, \mathcal{I}_A)$ with arbitrary skeletal structure, the goal is to predict a sequence of joint rotations $\{\mathbf{R}_t\}_{t=1}^{T}$ that animates A in accordance with the motion in V:

$$(V, A) \longrightarrow \{\mathbf{R}_t\}_{t=1}^{T},\ \mathbf{R}_t = \{R_{t,j}\}_{j\in\mathcal{J}},\ R_{t,j} \in \mathrm{SO}(3). \tag{1}$$

Here, $\mathcal{M}$ denotes the mesh, and the skeleton is $\mathcal{S} = (\mathcal{J}, \mathcal{E}, \mathbf{o})$, where $\mathcal{J}$ is the joint set and $\mathcal{E} \subseteq \mathcal{J} \times \mathcal{J}$ denotes the directed parent→child edges. For each edge $(i \to j) \in \mathcal{E}$, $\mathbf{o}_{i\to j} \in \mathbb{R}^3$ is the offset of joint j relative to its parent i. The rest rig also specifies canonical joint labels via a naming function $\ell : \mathcal{J} \to \mathcal{N}$. The optional appearance is provided as a reference image set $\mathcal{I}_A = \{I_A^{(k)}\}_{k=1}^{K}$ (e.g., renders or photos of A). This general formulation covers both *motion capture* (when A matches the subject in V) and *motion retargeting* (when A differs from the subject).

3.2. Overview

To tackle the CAMoCap task, we employ three dedicated, learnable branches to extract features from the reference prompt and the input monocular video, fuse them, and estimate motion sequences. A naive alternative is to regress joint rotations directly after feature fusion, but in monocular settings this is brittle due to: (i) parameterization and rig–frame ambiguities that make angles asset-dependent, (ii) under-constrained evidence where depth and camera motion entangle local rotations, and (iii) poor temporal continuity from per-frame angle regression. We therefore decompose the problem into 3D keypoint trajectory estimation followed by rotation recovery via inverse kinematics (IK).

Accordingly, our approach, MoCapAnything, comprises four components (see Fig. 2):

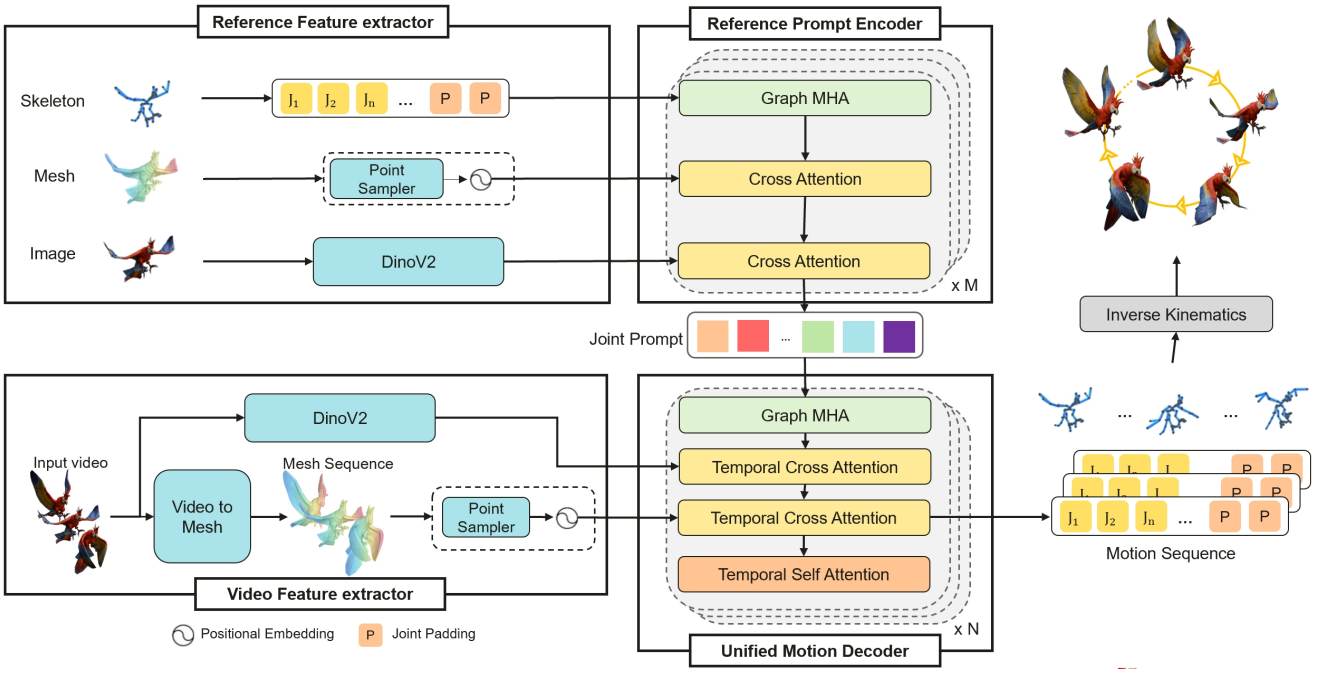


Figure 2. Detailed architecture of our method. A multi-modal Reference Prompt Encoder fuses mesh, skeleton, and appearance of the target asset into per-joint queries. A monocular video is converted into a 4D mesh sequence, and both mesh and video features are extracted. The Unified Motion Decoder fuses these signals via multi-branch attention to predict 3D keypoints, which are converted to asset-specific joint rotations via an optimization-based IK layer.

1. **Reference Prompt Encoder**: Extracts per-joint features from the reference asset, including skeletal, mesh, and appearance image-set cues.
2. **Video Feature Extractor**: Uses off-the-shelf models to obtain visual descriptors (e.g., DINOv2 [25]) and reconstruct a coarse 4D deforming mesh from the video. The mesh supplies topology- and geometry-aware signals that bridge the modality gap between dense visual tokens and the point-cloud–like joint representation, stabilizing and improving trajectory estimation.
3. **Unified Motion Decoder**: Fuses reference, geometric, and visual information to predict temporally coherent 3D joint trajectories for the target asset.
4. **IK Fitting Process**: Converts predicted joint trajectories into asset-specific joint rotations via an optimization-based IK procedure that respects hierarchy, bone lengths, joint limits, and temporal smoothness.

This modular pipeline flexibly supports both motion capture (same skeleton) and retargeting (different skeletons) for arbitrary 3D assets and rig topologies.

3.3. Architecture Design

3.3.1. Reference Prompt Encoder

Let the reference asset be $A = (\mathcal{M}, \mathcal{S}, \mathcal{I}_A)$ with mesh $\mathcal{M}$, skeleton $\mathcal{S} = (\mathcal{J}, \mathcal{E}, \mathbf{o})$, and an image set $\mathcal{I}_A = \{I_A^{(k)}\}_{k=1}^{K}$. The encoder outputs per-joint queries $Q = \{\mathbf{q}_j \in \mathbb{R}^d\}_{j\in\mathcal{J}}$. For each joint j with coordinate $\mathbf{x}_j \in \mathbb{R}^3$, we apply a sinusoidal positional encoding $\mathrm{pe}(\mathbf{x}_j)$ and a linear projection to obtain an initial embedding $\mathbf{q}_j^{(0)} = \mathbf{W}_p[\mathrm{pe}(\mathbf{x}_j);\ \mathbf{e}_{\text{name}}(\ell(j))] + \mathbf{b}_p$ (where $\mathbf{e}_{\text{name}}$ is optional). Variable joint counts are handled by a binary mask $\mathbf{m} \in \{0,1\}^{|\mathcal{J}_{\max}|}$ that zeroes padded joints in all attention operations.

We then apply L stacked fusion blocks, each with three submodules in a row:

1. **Self-Attention with Skeleton Topology.** We use a graph multi-head attention (Graph-MHA) on $\{\mathbf{q}_j^{(\ell)}\}$ with an attention bias $\mathbf{B}_{ij}$ computed from skeleton topology (adjacency in $\mathcal{E}$ and geodesic/kinematic distances), following the AnyTop [8] design:

$$\mathrm{Attn}(\mathbf{q}_i, \mathbf{q}_j) \propto \frac{\langle \mathbf{W}_Q\mathbf{q}_i,\ \mathbf{W}_K\mathbf{q}_j \rangle}{\sqrt{d}} + \mathbf{B}_{ij}, \qquad \mathbf{B}_{ij} = f_{\text{topo}}(\mathcal{E}, i, j). \tag{2}$$

   This encourages structure-aware message passing along the kinematic tree. Details will be illustrated in the supplementary material.
2. **Cross-Attention to Mesh Geometry.** We sample surface points from $\mathcal{M}$ to form $\mathcal{P} = \{(\mathbf{p}_u, \mathbf{n}_u)\}_{u=1}^{U}$ (positions and normals). Mesh tokens are $\mathbf{g}_u = \mathbf{W}_m[\mathrm{pe}(\mathbf{p}_u);\ \mathbf{n}_u]$. Joints attend to $\{\mathbf{g}_u\}$ to learn implicit

skinning-like relations between joints and local surface geometry.

3. **Cross-Attention to Appearance.** Images in $\mathcal{I}_A$ are encoded by a frozen image encoder ϕ_{img} (e.g., DINOv2) to obtain appearance tokens $\phi_{\text{img}}(\mathcal{I}_A)$. We inject appearance cues that disambiguate symmetric or visually similar parts via cross-attention mechanism.

Across layers, masked attention ensures invariance to the absolute joint count, and residual/FFN updates refine $\mathbf{q}_j^{(\ell)} \to \mathbf{q}_j^{(\ell+1)}$ by progressively integrating *structural* ($\mathcal{S}$), *geometric* ($\mathcal{M}$), and *visual* ($\mathcal{I}_A$) evidence. The final per-joint queries $Q = \{\mathbf{q}_j^{(L)}\}$ serve as asset-specific prompts for the Unified Motion Decoder and enable robust generalization across diverse characters and skeleton topologies.

3.3.2. Video Feature Extractor

Given a monocular video $V = \{I_t\}_{t=1}^{T}$, we build two complementary streams.

Visual stream. Each frame is encoded by a frozen DINOv2 image encoder ϕ_{img}, yielding per-frame dense tokens $\mathbf{A}_t = \phi_{\text{img}}(I_t)$ (and an optional global token). These serve as appearance/texture cues.

Geometry stream. We apply a pretrained image-to-3D reconstructor to obtain a coarse deforming surface sequence $\widehat{\mathcal{M}} = \{\widehat{\mathcal{M}}_t\}_{t=1}^{T}$. For each t, we randomly downsample the surface to U=1024 points $\mathcal{P}_t = \{(\mathbf{p}_{t,u}, \mathbf{n}_{t,u})\}_{u=1}^{U}$. Points are embedded as

$$\mathbf{g}_{t,u} = \mathbf{W}_m \left[\text{pe}(\mathbf{p}_{t,u});\, \mathbf{n}_{t,u};\, \text{pe}(t)\right],$$

producing geometry-aware tokens $\mathbf{G}_t = \{\mathbf{g}_{t,u}\}_{u=1}^{U}$ analogous to the mesh features used in the Reference Prompt Encoder.

We form the video feature set $\mathcal{V} = \{\mathbf{A}_t, \mathbf{G}_t\}_{t=1}^{T}$ (keys/values for the decoder). The 4D mesh tokens provide topology/geometry signals that bridge dense RGB features and the point-cloud–like joint space, stabilizing subsequent 3D keypoint estimation.

3.3.3. Unified Motion Decoder

Given the per-joint prompts $Q = \{\mathbf{q}_j\}_{j\in\mathcal{J}}$, the skeleton $\mathcal{S} = (\mathcal{J}, \mathcal{E})$, and video features $\mathcal{V} = \{\mathbf{A}_t, \mathbf{G}_t\}_{t=1}^{T}$ (DINOv2-based visual tokens $\mathbf{A}_t$ and 4D-mesh point tokens $\mathbf{G}_t$), we tile Q across time, add a temporal encoding to obtain $\{\mathbf{h}_{t,j}^{(0)}\}$, and apply a binary joint mask to accommodate variable-size skeletons. Each decoder layer refines these tokens through the following four stages:

1. **Graph-based self-attention (intra-frame).** Within each frame, joint tokens are updated using an attention layer with an explicit topology bias derived from $\mathcal{E}$ (same as AnyTop [8]), ensuring that updates respect the kinematic tree and local limb couplings.
2. **Temporal video cross-attention.** For each joint at time t, a sliding window over neighboring frames provides visual tokens that supply short-range appearance cues. Attending to this window improves continuity, fills in details under occlusion or motion blur, and stabilizes rapid movements.
3. **Temporal point-cloud cross-attention.** Joint tokens then aggregate geometry-aware evidence from the corresponding 4D mesh window. These point tokens inject topology/shape signals that bridge dense RGB features and the point-cloud–like joint space, disambiguating depth and self-occlusion and capturing non-rigid deformations.
4. **Temporal self-attention (per joint).** Finally, a windowed self-attention along the time axis mixes each joint's past and future states to enforce longer-range consistency and reduce jitter, while better modeling higher-order dynamics.

Residual connections, normalization, and feed-forward updates follow each block, and stacking L layers progressively integrates *structural* ($\mathcal{S}$), *visual* ($\mathbf{A}_t$), and *geometric* ($\mathbf{G}_t$) cues. A lightweight MLP head then predicts per-frame joint positions $\{\widehat{\mathbf{x}}_{t,j} \in \mathbb{R}^3\}$, yielding trajectories for the subsequent IK stage.

3.4. Training Objective

We supervise the decoder with a masked position regression loss consistent with our notation above. Let $\widehat{\mathbf{x}}_{t,j} \in \mathbb{R}^3$ be the predicted 3D position of joint $j \in \mathcal{J}$ at time $t \in \{1, \ldots, T\}$, and let $\mathbf{x}_{t,j}$ be the ground-truth position. Since assets have different skeleton sizes, we pad all sequences to $|\mathcal{J}_{\max}|$ joints and use a binary joint-validity mask $\mathbf{m} \in \{0,1\}^{|\mathcal{J}_{\max}|}$ (with $m_j = 1$ iff $j \in \mathcal{J}$ for this asset).

$$\mathcal{L}_{\text{pos}} = \frac{1}{\sum_{t=1}^{T}\sum_j m_j} \sum_{t=1}^{T}\sum_j m_j \left\|\widehat{\mathbf{x}}_{t,j} - \mathbf{x}_{t,j}\right\|_1.$$

We do not apply rotation space or explicit temporal losses during training: the network predicts joint positions, and rotations are obtained afterwards by the IK stage.

3.5. IK Fitting Process

We recover joint rotations from the predicted 3D joint trajectories using a lightweight two-stage IK procedure. First, we compute a per-frame geometric IK initialization by aligning rest-pose bone directions with the observed joint positions along each kinematic chain. This closed-form step provides a stable rotation estimate that respects the skeleton hierarchy. Then, we refine the rotations with a small differentiable IK optimization that minimizes the discrepancy between FK-reconstructed joints and the predicted 3D

positions, while regularizing the solution toward the geometric initialization. The optimization is warm-started from the previous frame to ensure temporal stability and suppress unnecessary twist. This hybrid strategy produces accurate and smooth joint rotations at minimal computational cost. Additional implementation details are provided in the supplementary material.

4. Experiments

4.1. Dataset and Evaluation Protocol

We evaluate our approach on the **Truebones Zoo** [37] dataset, which contains 1,038 animal motion sequences (totaling 104,715 frames) spanning a broad range of species and kinematic structures and 1000 random samples from objaverse [6, 7]. For testing, we curate a set of 60 sequences with enough diversity, stratified into three groups: **Seen** (species with abundant training data), **Rare** (species with limited training data), and **Unseen** (species never seen during training). This protocol enables a thorough assessment of model generalization.

4.2. Evaluation Metrics

To disentangle the contributions of the two stages, we evaluate them separately: (i) 3D joint positions $\widehat{\mathbf{x}_{t,j}}$ and (ii) joint rotations $\widehat{R_{t,j}}$. In the main paper we focus on quantitative results for 3D keypoint prediction, while rotation-level evaluation (after IK) is deferred to the supplementary material. As for 3D keypoints, we report the following metrics:

- **MPJPE** (Mean Per Joint Position Error): the mean Euclidean distance between predicted and ground-truth joint positions (lower is better).
- **MPJVE** (Mean Per Joint Velocity Error): the average velocity difference per joint, capturing temporal consistency and motion plausibility.

To account for large inter-species scale variations, we normalize all samples to a $[-1, 1]^3$ cube for training and rescale both predictions and ground truth to a unified $1\mathrm{m}^3$ cube for evaluation. All metrics are reported in centimeters (cm). For quantitative evaluation, we use ground-truth meshes to compute joint positions for both training and testing data to avoid interference from predicted mesh quality, while all visualizations are based on predicted meshes.

4.3. Compare with Baseline

We compare with adapted baselines from established pose and motion capture pipelines, including ViTPose [45], HRNet [33], VIBE [15], and GLoT [29]. All baselines are retrained on the same training set with unified skeleton representations and input modalities. For methods originally designed for human pose estimation, we adapt their output layers to match the target skeleton topology. As shown in Table 1, our method consistently outperforms all baselines across seen, rare, and unseen splits, demonstrating the effectiveness of our category-agnostic design.

	Seen		Rare		Unseen	
Method	**MPJPE ↓**	**MPJVE ↓**	**MPJPE ↓**	**MPJVE ↓**	**MPJPE ↓**	**MPJVE ↓**
HRNet	9.77	1.41	10.86	1.55	23.53	1.84
ViTPose	9.19	1.73	9.33	1.40	23.37	2.15
VIBE	4.46	0.83	4.03	0.68	8.72	0.95
GLoT	3.98	1.37	3.58	0.84	7.42	2.18
Ours	**1.06**	**0.44**	**1.28**	**0.37**	**1.76**	**0.36**

Table 1. Comparison with baseline methods on the Truebones Zoo-test set. Lower is better (↓). Results are reported across three generalization levels: seen, rare, and unseen species.

	Seen		Rare		Unseen	
Method	**MPJPE ↓**	**MPJVE ↓**	**MPJPE ↓**	**MPJVE ↓**	**MPJPE ↓**	**MPJVE ↓**
Ours w/o image	1.34	0.68	1.56	0.44	2.85	0.60
Ours w/o mesh	1.88	0.63	2.25	0.39	3.16	0.44
Ours w/o GMHA	1.08	0.55	1.49	0.38	1.82	0.37
Ours	**1.06**	**0.44**	**1.28**	**0.37**	**1.76**	**0.36**

Table 2. Ablation study on input modalities and modules. Lower is better (↓). Results are reported on the Truebones Zoo-test set across three generalization levels: seen, rare, and unseen species.

Method	**Enc/Dec**	Seen		Rare		Unseen	
		MPJPE ↓	**MPJVE ↓**	**MPJPE ↓**	**MPJVE ↓**	**MPJPE ↓**	**MPJVE ↓**
Variant 1	1 / 12	1.07	0.44	1.90	0.39	2.18	0.38
Variant 2	2 / 12	1.11	0.51	1.37	0.38	2.00	0.37
Variant 3	4 / 16	**1.05**	**0.42**	1.46	0.39	1.85	0.38
Ours	**4 / 12**	1.06	0.44	**1.28**	**0.37**	**1.76**	**0.36**

Table 3. Ablation results with encoder/decoder layer configurations. Metrics are reported on the Truebones Zoo-test set under seen, rare, and unseen generalization.

4.4. Ablation Study

We first analyze the impact of input modalities and key modules. Specifically, we consider variants that remove the reference image-set encoder and its cross-attention modules (w/o image), exclude mesh features from both reference and video streams (w/o mesh), and disable the graph multi-head attention over the skeleton (w/o GMHA), in comparison to the full model. As shown in Table 2, removing any modality or module leads to clear performance drops, especially in the rare and unseen splits. The mesh and graph-attention branches are crucial for robust transfer to new species, highlighting the importance of explicit topology and geometry modeling.

Table 3 further examines the impact of encoder and decoder layer configurations. Increasing encoder and decoder depth generally improves performance, especially on rare and unseen species, indicating the importance of sufficient model capacity for handling diverse motion patterns. Although Variant 3 achieves slightly better performance on the seen split, it introduces higher model complexity with

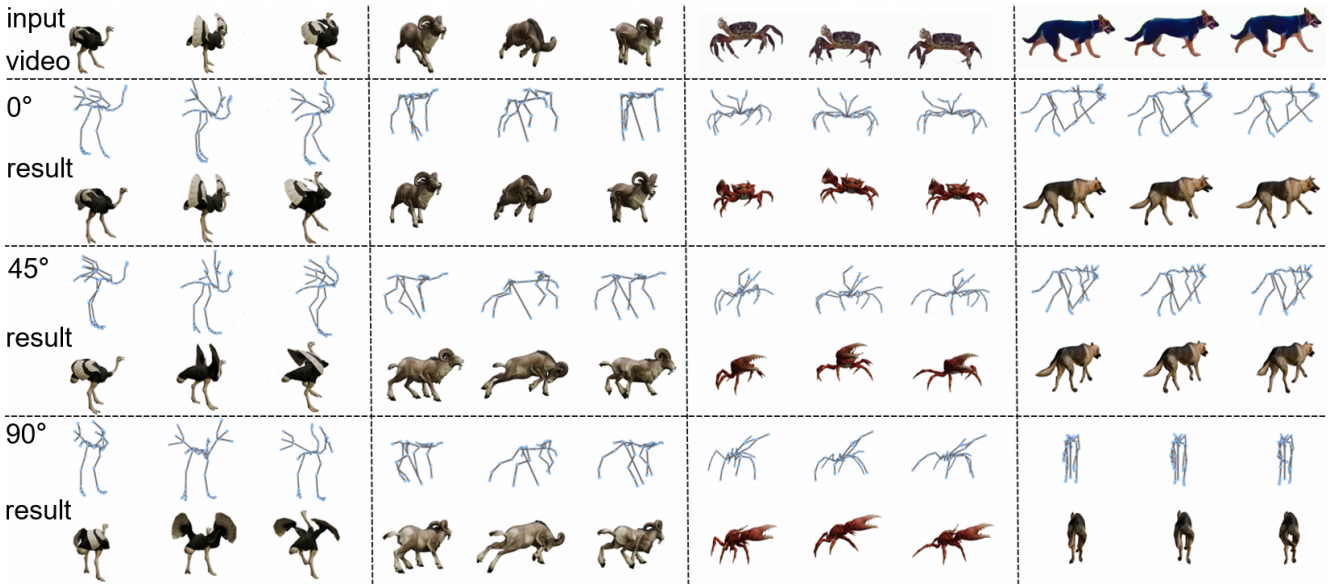


Figure 3. **IK-Driven Asset Animation** by ostrich (biped), a goat (quadruped), and in-the-wild cases such as a crab and a dog. We visualize the predicted poses and corresponding IK-driven animations from multiple viewpoints.

limited gains on rare and unseen cases. We therefore adopt a balanced configuration (4 encoder layers, 12 decoder layers), which achieves consistently strong performance across all splits.

4.5. Qualitative Results on Truebones Zoo

Figure 4 presents representative Truebones Zoo-test results. **Row 1** shows input video Jugar. **Row 2** displays the same-species reference and predicted mocap outputs. **Rows 3–5** show results when retargeting to skeletons of three different species. Our approach generalizes robustly across species and maintains temporally consistent, anatomically plausible 3D motion even with significant appearance and shape variation.

4.6. Qualitative Results on Objaverse

In addition to animal skeletons, our framework also supports human-like rigs, enabling both human motion capture and cross-domain retargeting between humans and animals. Our model can transfer motion from humans to animals and vice versa, demonstrating strong versatility across different skeleton types. Representative qualitative results: including humanoid mocap, human to animal, and animal to human retargeting, are provided on our project homepage.

4.7. In-the-Wild Generalization

To further assess robustness, we apply our trained model to a variety of in-the-wild animal videos collected from the Internet, including birds (chickens, eagles, seagulls), quadrupeds (tigers, leopards, elephants, cats, dogs), and other animals such as crabs, fish, and snakes. As shown in Figure 5, our method successfully reconstructs plausible 3D skeletal motion for both mocap and retargeting (Jaguar pretend to fly), demonstrating strong generalization.

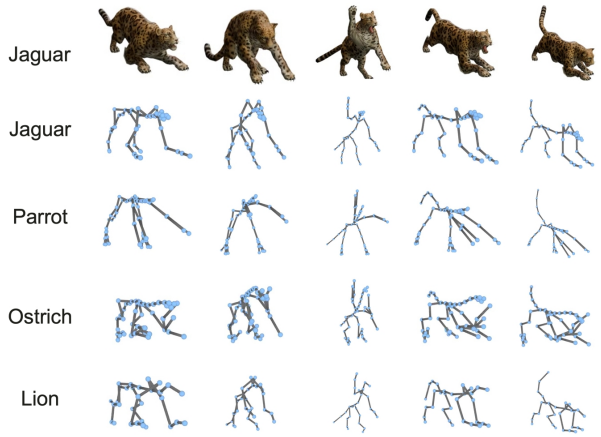


Figure 4. **Truebones Zoo mocap and retargeting results.** Each row visualizes one evaluation sequence. **Row 1**: Input video frames. **Row 2**: Same-species reference skeleton and predicted mocap results. **Rows 3–5**: Reference skeletons from three different species and retargeted motions by our method. Our method generalizes across species and produces stable, anatomically plausible 3D motion.

4.8. Visualization of IK-Driven Asset Animation

We further visualize the IK-driven animation results to assess the full pipeline from predicted skeletons to rigged assets. For each example, we show multi-view renderings of



Figure 5. **Real-world (Wild) results.** Similar layout as Figure 4. **Row 1**: Input wild video frames. **Row 2**: Same-species reference skeleton and predicted mocap outputs. **Rows 3–4**: Cross-species reference skeletons and our retargeted motion predictions. Despite real-world challenges, our method maintains robustness and stability.

the predicted 3D skeletons together with the corresponding IK-driven asset animations. As illustrated (see Figures 3) in our examples—including an ostrich (biped), a goat (quadruped), and in-the-wild cases such as a crab and a dog. These results demonstrate that our approach generalizes well across different species and real-world scenarios, enabling reliable motion capture and animation beyond controlled settings.

4.9. Arbitrary Cross-Species Retargeting

A unique feature of our approach is **prompt-based retargeting** across arbitrary asset types: even for reference skeletons entirely unrelated to the subject in the input video. Although not explicitly trained for cross-species transfer, our model leverages structural, visual, and geometric cues to synthesize plausible retargeted motion.

We observe a wide range of creative results: bird videos drive quadrupeds to perform flapping-like actions or animate pterosaurs; fish swimming is transferred to crocodiles or snakes; dog running animates bipedal birds; crocodile tail-whipping is retargeted to leopards or parrots. Such unconstrained retargeting enables new workflows for animation (see Figures 6 and 7 for more results).

Given the lack of directly comparable baselines, we focus on extensive qualitative analysis and ablation, providing thorough visualization of our results and highlighting the practical versatility of our approach.

5. Conclusion

In this work, we reformulate the motion capture problem as Category-Agnostic Motion Capture (CAMoCap), a novel paradigm in which a monocular video and an arbitrary rigged asset function as input prompts to generate rotation-based animations tailored to the target character. We further propose MoCapAnything, a reference-guided factorized architecture that initially estimates 3D joint trajectories and subsequently reconstructs asset-specific rotations through constraint-aware inverse kinematics, while mitigating cross-modal discrepancies between RGB and joint representations via an intermediate coarse 4D mesh. Using our proposed reorganized the **Truebones Zoo** benchmark, comprising 1,038 annotated clips with 60 test sequences and providing standardized skeleton-mesh-rendered video triples, MoCapAnything consistently produces temporally stable, animation-ready outputs across diverse rigging systems, demonstrating notable in-domain precision, robust generalization to in-the-wild scenarios, and semantically meaningful cross-species motion retargeting capabilities.

Limitations and future work. Our performance depends on the quality of the pretrained image-to-3D reconstructor and assumes access to a rig with known joint structure; it also operates primarily in camera space without explicit physics or contact reasoning. Future directions include end-to-end, contact- and physics-aware IK, world-grounded trajectory recovery, reducing reliance on 4D reconstruction (e.g., video-only geometry priors), text-only or multimodal prompts beyond rendered images, and extensions to multi-character interaction.

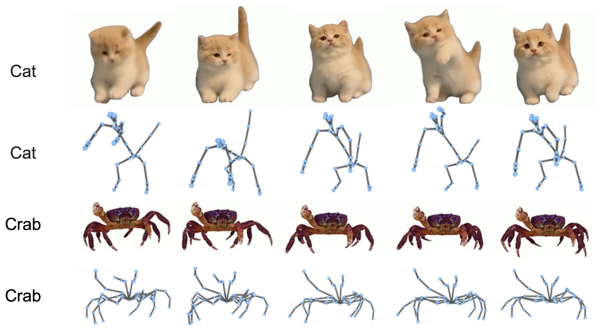


Figure 6. **More in-the-wild mocap results.** Our method generalizes to a diverse range of species and scenarios.

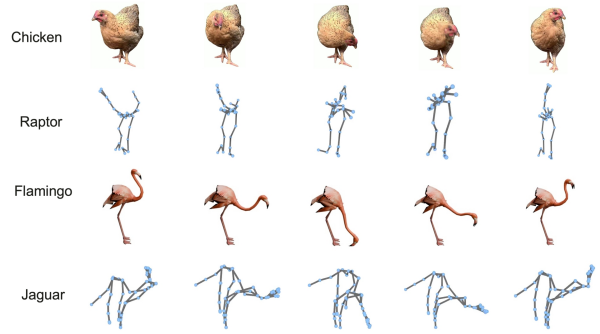


Figure 7. **Unconstrained cross-species retargeting.** Examples of using our model to retarget motion from one species to another, yielding diverse, creative, and physically plausible animations. 1st row: from chicken to Raptor, 2nd row: Flamingo to Jaguar.

References

[1] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 2

[4] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 2

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021. 3

[6] 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. *arXiv preprint arXiv:2212.08051*, 2022. 6

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 6

[8] Inbar Gat, Sigal Raab, Guy Tevet, Yuval Reshef, Amit H Bermano, and Daniel Cohen-Or. Anytop: Character animation diffusion with any topology. *arXiv preprint arXiv:2502.17327*, 2025. 4, 5

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 3

[10] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8575–8584, 2021. 2

[11] Kehong Gong, Bingbing Li, Jianfeng Zhang, Tao Wang, Jing Huang, Michael Bi Mi, Jiashi Feng, and Xinchao Wang. Posetriplet: Co-evolving 3d human pose estimation, imitation, and hallucination under self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11017–11027, 2022. 2

[12] Or Hirschorn and Shai Avidan. A graph-based approach for category-agnostic pose estimation. *arXiv preprint arXiv:2311.17891*, 2023. 2

[13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 3

[14] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European conference on computer vision (ECCV)*, pages 371–386, 2018. 3

[15] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3, 6

[16] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. In *2024 International Conference on 3D Vision (3DV)*, pages 397–408. IEEE, 2024. 3

[17] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11354–11361, 2020. 2

[18] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11025–11034, 2021. 2

[19] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[20] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9752–9762, 2024. 3

[21] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 3

[22] 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):1–16, 2015. 1, 3

[23] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[24] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J. Black, Sil-

via Zuffi, and Peter Kulits. Generative zoo. *CoRR*, abs/2412.08101, 2024. 3

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. 2, 4

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1, 3

[27] Matan Rusanovsky, Or Hirschorn, and Shai Avidan. Capex: Category-agnostic pose estimation from textual point explanation. *arXiv preprint arXiv:2406.00384*, 2024. 3

[28] Matan Rusanovsky, Or Hirschorn, and Shai Avidan. Capex: Category-agnostic pose estimation from textual point explanation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8887–8896, 2023. 6

[30] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11069–11078, 2022. 2

[31] Min Shi, Zihao Huang, Xianzheng Ma, Xiaowei Hu, and Zhiguo Cao. Matching is not enough: A two-stage framework for category-agnostic pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7308–7317, 2023. 2

[32] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 3

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2, 6

[34] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1834–1843, 2024. 3

[35] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[36] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 2

[37] Truebones. Truebones motion capture – mocap files, n.d. Accessed: 2025-05-22. 1, 2, 6

[38] Dongkai Wang, Shiyu Xuan, and Shiliang Zhang. Locllm: Exploiting generalizable human keypoint localization via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 614–623, 2024. 2

[39] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024. 3

[40] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023. 3

[41] Yuefan Wu*, Zeyuan Chen*, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. CASA: Category-agnostic skeletal animal reconstruction. In *NeurIPS*, 2022. 3

[42] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 2

[43] Yabo Xiao, Kai Su, Xiaojuan Wang, Dongdong Yu, Lei Jin, Mingshu He, and Zehuan Yuan. Querypose: Sparse multi-person pose regression via spatial-aware part-level query. *Advances in Neural Information Processing Systems*, 35: 12464–12477, 2022. 2

[44] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *European conference on computer vision*, pages 398–416. Springer, 2022. 2

[45] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 2, 6

[46] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15980–15989, 2021. 3

[47] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. *Advances in Neural Information Processing Systems*, 34:19326–19338, 2021. 3

[48] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 3

[49] Gengshan Yang, Shuo Yang, John Z. Zhang, Zachary Manchester, and Deva Ramanan. Physically plausible reconstruction from monocular videos. In *ICCV*, 2023. 3

[50] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. *arXiv preprint arXiv:2302.01593*, 2023. 2

[51] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022. 3

[52] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. 3

[53] 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/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 3

[54] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3