

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

Supplementary Material

1. More Visualization Results

In this section, we summarize additional qualitative results from our **supplementary webpage**. These visualizations highlight the effectiveness of our approach across controlled multi-species datasets, in-the-wild videos, and cross-species retargeting scenarios, showing that our model produces high-fidelity and temporally smooth motion under a broad range of conditions.

Comparison with GenZoo. We compare our results with GenZoo, a single-image animal pose and shape estimator trained on synthetic quadruped data. Without temporal modeling, GenZoo exhibits frame-wise inconsistencies and pose fluctuations when applied to video sequences, even for quadruped inputs. In contrast, our method models motion dynamics explicitly, yielding smoother and more coherent 4D reconstructions that better follow ground-truth trajectories.

Mocap Results. The supplementary webpage provides additional mocap visualizations. From Truebones Zoo, we show examples spanning multiple animal species with diverse skeletal structures; from Objaverse, we include bipedal characters to demonstrate adaptability across different asset types. We also present in-the-wild cases such as flying birds and crocodiles to illustrate performance on real video inputs.

Arbitrary Motion Retargeting. We further include motion retargeting examples: Zoo2Zoo transfer across different animal species, Human2Zoo transfer applying human motions to animal skeletons, and Zoo2Human transfer mapping animal motions to a human skeleton. For In-the-Wild2Human results, motions from videos of animals such as eagles and leopards are retargeted to a human skeleton. These examples show that our model handles large variations in morphology, topology, and motion dynamics.

IK Visualization. We also provide IK fitting visualizations, showing recovered joint rotations and the improved temporal stability and orientation consistency achieved through geometric initialization, temporal warm-starting, and twist-regularized refinement. We additionally report an average geodesic rotation error of approximately 17° , indicating reasonable rotation accuracy after IK.

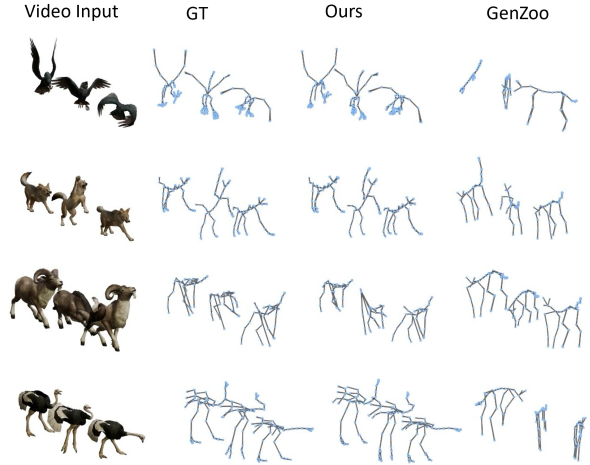


Figure 1. **Qualitative comparison with GenZoo on the Truebones Zoo dataset.** Our method produces smoother trajectories and maintains stable, anatomically plausible motions across a wide variety of skeleton types, including non-quadrupeds. In contrast, GenZoo is limited to quadruped structures and often fails to generalize to more diverse or complex skeletal configurations. Visualizations highlight our approach’s superior accuracy, robustness, and generalization ability.

2. More Experiment Results

Model	Quad	Non-Quad	All
genzoo	0.4466	0.4740	0.4580
ours	0.2354	0.2821	0.2549

Table 1. Chamfer Distance (CD) results on the Truebones Zoo dataset.

To our knowledge, GenZoo [1] is among the few works that attempt category-agnostic animal motion capture. However, it mainly supports quadruped species and struggles to generalize to more diverse skeletons. Since GenZoo does not produce joint-aligned skeletons compatible with MPJPE/MPJVE evaluation, we adopt **Chamfer Distance (CD-Skeleton)** as a structural metric for comparison. For a comprehensive comparison, we evaluate both methods on the Truebones Zoo dataset, using the CD-Skeleton metric to measure the structural accuracy of the predicted skeletal motion.

As shown in Table 1, our approach achieves significantly

lower CD-Skeleton errors than GenZoo across all categories. On the overall test set, our method reduces the average error from 0.4580 to 0.2549, indicating a substantial improvement in capturing and reconstructing diverse skeletal motions, especially for non-quadruped species where existing methods perform poorly.

Figure 1 presents representative qualitative results on the Truebones Zoo dataset. Compared to GenZoo, our predictions exhibit smoother motion trajectories, higher anatomical fidelity, and robust stability across both quadruped and non-quadruped skeletons—including bipeds, birds, reptiles, and even non-biological assets. GenZoo, while currently the most widely applicable animal motion capture method, is fundamentally constrained by its reliance on quadruped skeleton templates and struggles to generalize to broader categories.

For further qualitative comparison, we provide side-by-side visualizations of our results and GenZoo’s on our project homepage, showcasing the advantages of our approach in both accuracy and generalization.

3. Implementation Details

A. Dataset and Training Details

Dataset Processing. Our 60-sequence benchmark composition: 27 mammals, 9 birds, 12 dinosaurs+dragons, 7 reptiles (incl. snakes), 2 aquatic, 3 arthropods. All meshes and joints are first scaled by the bounding box of their rest pose, normalizing each mesh into a unit-volume space. For sequence data, we remove the global translation of every frame, compute a sequence-level super bounding box, and uniformly scale the entire sequence into the range $[-1, 1]^3$. For in-the-wild video inputs, we assume a fixed camera position throughout the sequence.

Training details. The network consists of 12 layers for decoder, and a prompt encoder composed of 4 layers. All experiments are conducted on 8 GPUs, each equipped with 64 GB of memory. The model is trained for 60 epochs using the Adam optimizer, requiring approximately 36 hours in total. We use a learning rate of 1×10^{-4} and a batch size of 1 per GPU. Training is performed with paired supervision for motion capture (not retargeting). For each sample, we select a reference asset from the same species (one frame providing image, unordered mesh, and skeleton as prompt) and predict 3D joint positions for a 24-frame input video. The loss is defined on joint positions, followed by a lightweight inverse kinematics (IK) fitting step to recover joint rotations for deployment. We employ a sliding-window mechanism to support inference on arbitrarily long videos. During attention computation, masked joints are excluded, and both joint identity embeddings and skeletal topology are incorporated as conditioning signals. We do not explicitly train

on retargeting pairs. Nevertheless, the learned reference-conditioned motion representation enables cross-species retargeting behaviors at inference time. During training, we use ground-truth mesh sequences for efficiency. At inference, we replace them with meshes predicted by a video-to-mesh module (e.g., SWIFT4D), which provide sufficiently stable conditioning in practice, and we empirically observe negligible degradation compared to GT-mesh conditioning.

B. Inverse Kinematics Fitting

Given a predicted sequence of 3D joint locations $\{\mathbf{X}_{t,i}\}$ and a kinematic tree with rest-pose offsets \mathbf{o}_i and parent indices $p(i)$, our goal is to recover temporally stable joint rotations $\mathbf{R}_{t,i} \in SO(3)$ such that the forward kinematics (FK) matches the observed joints:

$$\mathbf{P}_{t,i} = \begin{cases} \mathbf{0}, & p(i) = -1, \\ \mathbf{P}_{t,p(i)} + \mathbf{R}_{t,p(i)} \mathbf{o}_i, & \text{otherwise.} \end{cases}$$

Because FK is not injective, position-only constraints do not fully determine local orientation, especially twist around the bone axis. We therefore combine geometric initialization, temporal warm-starting, and differentiable refinement with twist suppression.

Geometric Initialization. For each frame, we compute a closed-form IK estimate $\mathbf{R}_{t,i}^{\text{geo}}$. For single-child joints, we align rest-pose and observed bone vectors via axis-angle rotation. For multi-child joints, we solve the orthogonal Procrustes problem:

$$\mathbf{R}_{t,i}^{\text{geo}} = \arg \min_{\mathbf{R} \in SO(3)} \sum_k \|\mathbf{R} \mathbf{v}_{i,k}^{\text{rest}} - \mathbf{v}_{t,i,k}^{\text{obs}}\|^2,$$

where \mathbf{v}^{rest} are rest-space bone directions and \mathbf{v}^{obs} are normalized directions from predicted joints. This provides consistent orientations at branching structures (e.g., pelvis, shoulders).

Temporal Warm-Starting. To avoid frame-to-frame drift, optimization for frame t is initialized using the solution from the previous frame:

$$\boldsymbol{\theta}_t^{(0)} = \boldsymbol{\theta}_{t-1}^*.$$

Differentiable Refinement. Local rotations are parameterized as axis-angle vectors $\boldsymbol{\theta}_{t,i} \in \mathbb{R}^3$ and refined via the loss:

$$\mathcal{L}_t = \mathcal{L}_{\text{pos}} + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}} + \lambda_{\text{twist}} \mathcal{L}_{\text{twist}}.$$

The FK position loss is:

$$\mathcal{L}_{\text{pos}} = \frac{1}{N} \sum_i \|\mathbf{P}_{t,i}(\boldsymbol{\theta}_t) - \mathbf{X}_{t,i}\|^2.$$

A geometric prior encourages solutions close to the closed-form initialization:

$$\mathcal{L}_{\text{prior}} = \frac{1}{N} \sum_i \|\theta_{t,i} - \theta_{t,i}^{\text{geo}}\|^2.$$

Twist Suppression. Since bone-axis twist is under-constrained, we penalize rotation components parallel to the bone direction $\mathbf{u}_i = \mathbf{o}_i / \|\mathbf{o}_i\|$. Let $\theta_{t,i} = \alpha_{t,i} \hat{\mathbf{a}}_{t,i}$. The twist magnitude is:

$$\alpha_{t,i}^{\text{twist}} = \alpha_{t,i} (\hat{\mathbf{a}}_{t,i} \cdot \mathbf{u}_i).$$

We minimize:

$$\mathcal{L}_{\text{twist}} = \frac{1}{N} \sum_i (\alpha_{t,i}^{\text{twist}})^2.$$

This term suppresses candy-wrapper artifacts while preserving natural motion around long chains such as tails.

Summary. The combination of geometric IK, temporal warm-starting, and twist-regularized refinement yields stable and anatomically consistent joint rotations, significantly improving reconstruction quality. Further implementation details are provided in the code release.

4. Evaluation Metrics

This section describes the computation of the proposed metric (CD-Skeleton) that evaluates the alignment between two articulated skeletons. Each skeleton is represented by a set of 3D joint positions and a kinematic hierarchy defined by a parent array.

Notation

Let Skeleton A and Skeleton B be defined as:

- Joint positions:

$$\begin{aligned} \mathbf{X}^A &= \{\mathbf{x}_i^A \in \mathbb{R}^3 \mid i = 1, \dots, N\}, \\ \mathbf{X}^B &= \{\mathbf{x}_i^B \in \mathbb{R}^3 \mid i = 1, \dots, N\}. \end{aligned}$$

where N is the number of joints.

- Kinematic hierarchy, defined by a parent array:

$$\mathbf{p}^A, \mathbf{p}^B \in \{-1, 1, \dots, N\}^N,$$

where $p_i^A = -1$ (or $p_i^B = -1$) indicates a root joint.

Although the parent arrays may differ, the metric assumes a known correspondence of joint indices between the two skeletons.

Distance From Joint to the Other Skeleton

For each joint of Skeleton A, we compute its distance to the closest point on the bone segments of Skeleton B. Skeleton B consists of line segments defined by its kinematic tree:

$$\mathcal{S}^B = \{(\mathbf{x}_i^B, \mathbf{x}_{p_i^B}^B) \mid p_i^B \neq -1\}.$$

For a joint \mathbf{x}_i^A , its distance to Skeleton B is defined as:

$$d(\mathbf{x}_i^A, \mathcal{S}^B) = \min_{(\mathbf{b}_1, \mathbf{b}_2) \in \mathcal{S}^B} \|\mathbf{x}_i^A - \Pi_{\mathbf{b}_1, \mathbf{b}_2}(\mathbf{x}_i^A)\|,$$

where $\Pi_{\mathbf{b}_1, \mathbf{b}_2}(\mathbf{v})$ denotes the orthogonal projection of point \mathbf{v} onto the line segment connecting \mathbf{b}_1 and \mathbf{b}_2 . This projection is computed as:

$$\Pi_{\mathbf{b}_1, \mathbf{b}_2}(\mathbf{v}) = \mathbf{b}_1 + \text{clip}\left(\frac{(\mathbf{v} - \mathbf{b}_1) \cdot (\mathbf{b}_2 - \mathbf{b}_1)}{\|\mathbf{b}_2 - \mathbf{b}_1\|^2}, 0, 1\right) (\mathbf{b}_2 - \mathbf{b}_1),$$

where $\text{clip}(t, 0, 1) = \max(0, \min(t, 1))$ ensures the projected point lies on the segment.

Similarly, we can compute the distance from joints of Skeleton B to Skeleton A.

Skeleton-to-Skeleton Distance

The asymmetric distance from Skeleton A to Skeleton B is:

$$D(A \rightarrow B) = \frac{1}{N} \sum_{i=1}^N d(\mathbf{x}_i^A, \mathcal{S}^B).$$

The symmetric distance is defined as:

$$D_{\text{sym}}(A, B) = \frac{1}{2} (D(A \rightarrow B) + D(B \rightarrow A)).$$

Interpretation

This metric evaluates how closely each joint of one skeleton lies to the structure of the other skeleton, capturing differences in global pose, limb orientation, and proportions. The symmetric version provides a balanced measure when neither skeleton should be considered the reference.

References

- [1] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J. Black, Silvia Zuffi, and Peter Kulits. Generative zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1