CAST: REAL-TIME MOTION CAPTURE FOR ANY SKELETON TOPOLOGY

Anonymous authors
Paper under double-blind review

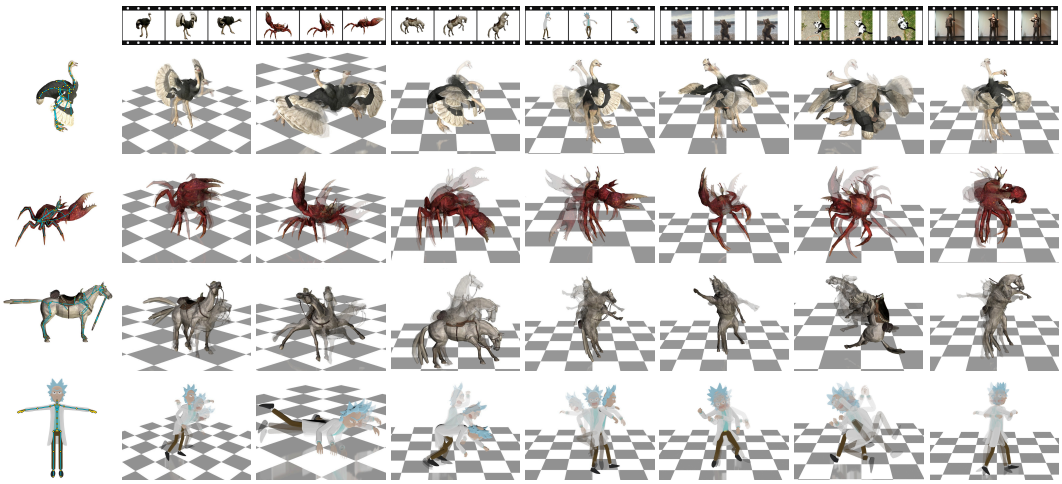


Figure 1: CAST animates four target skeletons (left) from each of seven input clips (top).

ABSTRACT

This paper studies the problem of recovering motion from monocular video, even when the target skeleton topology differs from the subject observed in the video. Although previous methods have made efforts to resolve this problem, they remain vulnerable to joint-rotation errors that are coupled across the skeleton and can propagate along its kinematic hierarchy. To mitigate these errors, we present CAST, a topology-aware motion capture framework that combines learned local rotations with geometric correction guided by predicted global bone directions. Our observability analysis characterizes which rotational components can be constrained by outgoing bone directions. Guided by this analysis, GRACE uses reliability-gated alignment to correct composed orientations in parent-to-child order, leaving unconstrained components to the learned rotation predictor. The motion predictor and correction module are trained jointly with motion-derived supervision, using a frozen image encoder. Experiments under the TopoCap evaluation protocol show that the proposed framework improves motion recovery across different target topologies, including unseen structures. Reference-based cross-skeleton evaluation on a curated Mixamo benchmark and a user preference study further assess transfer to different target structures. A causal CAST-B variant achieves over 51 FPS on a single NVIDIA A100 GPU, with timing covering image encoding and motion prediction. Code is available at `https://cast-mocap.github.io/`.

1 INTRODUCTION

Motion capture animates digital characters in games, films, virtual avatars, and interactive applications (Zhang et al., 2024; Yang et al., 2025). AI-generated videos broaden this task to creatures and stylized characters whose implicit articulated structures may differ from the rigs used for animation (Hu, 2024; Men et al., 2025; Wang et al., 2026). This calls for recovering motion on a supplied target skeleton, even when its topology differs from that of the video subject. However, most

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

video-to-motion systems assume fixed human or animal templates, known joint correspondences, and predefined rotation conventions (Zhu et al., 2023; Shin et al., 2024; Shen et al., 2024), limiting their applicability to targets with different joint counts, branching structures, or rest geometries.

Video-driven 4D-mesh methods (Wu et al., 2026) recover vertex trajectories rather than rotations for an existing target rig. Estimating source-skeleton motion and then retargeting it typically requires source–target bone correspondences or target motion examples (Aberman et al., 2020; Chen et al., 2025). Direct target-skeleton methods avoid this explicit retargeting step. MoCapAnything (Gong et al., 2026a) fits rotations to predicted joint trajectories using inverse kinematics (IK), while TopoCap (Pu et al., 2026) derives them from predicted global rotations. MoCapAnything V2 (Gong et al., 2026b) replaces IK with a learned, reference-conditioned pose-to-rotation decoder.

However, arbitrary-skeleton motion capture presents two related challenges, as joint rotations are inherently ambiguous in monocular video and the target skeleton's kinematic hierarchy couples errors across joints. Pose-based inverse kinematics fits rotations to predicted joint positions, but noisy trajectories can introduce fitting residuals, while joint positions alone do not fully constrain axial twist (Gong et al., 2026a). Learned local-rotation prediction faces a different challenge because an error in a parent's rotation alters the global orientations of its descendants and can propagate along the kinematic chain (Pavllo et al., 2018; Pan et al., 2024). Predicting global orientations avoids recursive local-rotation composition during estimation, but recovering accurate parent-relative rotations still requires consistent orientation estimates across neighboring joints (Pu et al., 2026). The impact of these errors on downstream poses and end-effector positions depends on the target skeleton's chain depth, branching structure, and bone lengths.

The target skeleton determines not only error propagation but also which rotational components outgoing bone directions can constrain. We therefore introduce CAST, a motion Capture system for Arbitrary Skeleton Topologies that combines learned local rotations with topology-dependent geometric correction. CAST predicts initial local rotations and global bone directions with associated reliabilities; the latter provide geometric constraints without recursive local-rotation composition. Our Gated Rotation Alignment and CorrEction (GRACE) module uses reliability-weighted alignment and a learned gate to correct composed global rotations while leaving unconstrained components to the rotation predictor. Corrected parent orientations are then used to compose child rotations, allowing correction to intervene during hierarchical propagation.

Our contributions are as follows:

- We introduce CAST, a unified motion capture framework for diverse skeleton topologies. CAST-L reduces MPJPE by 27.6–62.5% on matched-rig benchmarks and by 54.7% on the cross-skeleton benchmark relative to our reimplementation of MoCapAnything V2.
- We establish the relationship between bone-direction constraints and rotation recoverability and use this analysis to develop GRACE. The full GRACE configuration reduces MPJPE by 8.6–15.3% across different visual backbones and dataset splits.
- We extend CAST to real-time streaming motion capture and motion completion from sparse observations, with the streaming variant exceeding 51 FPS on a single A100 GPU. We also demonstrate sketch-guided in-betweening and provide a Blender add-on for real-time motion capture.

2 Related Work

Video-based motion capture for arbitrary skeletons. Early video-based motion capture methods typically rely on fixed human or animal body models. MotionBERT (Zhu et al., 2023) learns general human-motion representations, while WHAM and GVHMR (Shin et al., 2024; Shen et al., 2024) recover world-grounded human motion with body-specific priors. Recent work extends video-based capture to arbitrary rigged assets. MoCapAnything (Gong et al., 2026a) predicts target-joint trajectories and recovers asset-specific rotations with inverse kinematics. MoCapAnything V2 (Gong et al., 2026b) replaces this analytical stage with a learnable, reference-conditioned pose-to-rotation module and jointly trains the video-to-pose and pose-to-rotation paths. These methods differ in how they represent the intermediate motion and how the target skeleton enters the recovery process, ranging from joint positions and inverse kinematics to directly learned local rotations.

Cross-skeleton motion transfer and topology-agnostic representations. Motion retargeting methods learn to transfer motion between characters with different structures using skeleton-aware operators or shared motion representations. Neural kinematic networks and skeleton-aware networks learn motion mappings across character structures (Villegas et al., 2018; Aberman et al., 2020), while HuMoT uses a topology-agnostic transformer representation for character animation retargeting (Mourot et al., 2023). More recent approaches use topology-conditioned generation or shared structural representations, including AnyTop (Gat et al., 2025), TopoCap (Pu et al., 2026), and SAMoR (Zhang et al., 2026). Motion2Motion (Chen et al., 2025) transfers motion across topologies using sparse source–target correspondences and target motion examples. Other video-guided asset animation methods, such as Puppeteer (Song et al., 2026) and R-DMesh (Wu et al., 2026), use rigging or dynamic-mesh representations to connect visual input with animation. These approaches differ in their input modality, use of source–target correspondence, and output representation. CAST focuses on RGB-conditioned motion capture with an existing target skeleton and predicts local rotations for that skeleton without requiring an explicit source–target joint map.

3 Problem Setup

Given an RGB video $\mathcal{I}_{1:T}$ and a target skeleton $\mathcal{S} = (\mathcal{V}, \mathcal{E}, p, \mathbf{o}, \mathbf{L}^{\text{rest}})$, we predict per-frame local joint rotations $\mathbf{L}_{t,j} \in \mathtt{SO}(3)$ that drive the target rig through the observed motion. The skeleton is a rooted tree with joints $\mathcal{V} = \{0, \dots, J-1\}$ and parent map $p(\cdot)$, where $p(0) = -1$. Its rest pose is specified by parent-local offsets $\mathbf{o}_j \in \mathbb{R}^3$ and local joint rotations $\mathbf{L}_j^{\text{rest}} \in \mathtt{SO}(3)$. In this paper, we consider two instances of this problem. For *matched-rig capture*, the video depicts the supplied target rig. For *cross-skeleton animation*, the video depicts a different source rig from the supplied target skeleton, and no source–target joint correspondence is assumed.

4 Method

Figure 2(a) presents CAST, our framework for motion capture with topology-dependent rotation correction. Its central component, GRACE, uses predicted global bone directions to correct rotation errors propagated through the skeleton hierarchy. The geometry of a joint's outgoing bones determines the rotational constraints available for correction, while directional reliability and a learned gate control the updates. GRACE applies these updates in parent-to-child order, incorporating each corrected orientation into the subsequent rotation composition. To provide the initial local rotations and features for direction prediction, we adapt the prediction backbone of MoCapAnything V2 (Gong et al., 2026b) and train it jointly with GRACE. The backbone architecture is detailed in Appendix B.1.

4.1 Topology-dependent observability and error correction

Rotation composition. As shown in Figure 2(a), the method extends from the backbone of MoCapAnything V2 (Gong et al., 2026b), predicting the initial local prediction $\widehat{\mathbf{L}}^0_{t,i}$ of joint i. For joint i with parent $p(i)$, composing its local rotation with the parent's global rotation gives $\widehat{\mathbf{G}}_{t,i}$, which transforms its outgoing rest-pose bone directions into global directions $\widehat{\mathbf{d}}_{t,ij}$,

$$\widehat{\mathbf{G}}_{t,i} = \begin{cases} \widehat{\mathbf{L}}^0_{t,0}, & i = 0, \\ \widehat{\mathbf{G}}_{t,p(i)} \widehat{\mathbf{L}}^0_{t,i}, & i > 0, \end{cases} \qquad \widehat{\mathbf{d}}_{t,ij} = \widehat{\mathbf{G}}_{t,i} \mathbf{r}_{ij}. \tag{1}$$

Here $\mathbf{r}_{ij} = \mathbf{o}_j / \|\mathbf{o}_j\|_2$ is the unit rest-pose direction of bone $i \to j$ in joint i's local axes, assuming $\|\mathbf{o}_j\|_2 > 0$ in the directional analysis, and $\widehat{\mathbf{d}}_{t,ij} \in \mathbb{S}^2$ is the composed bone direction before correcting joint i. The hat marks quantities before correction: $\widehat{\mathbf{L}}^0$ for the initial local rotations, and $\widehat{\mathbf{G}}$ for global rotations and forward-kinematic (FK) directions.

Ideal geometric correction. Let $\mathcal{C}(i) = \{j \mid p(j) = i\}$ denote the children of i. As there is a gap between the composed direction $\widehat{\mathbf{d}}_{t,ij}$ and ground-truth direction $\mathbf{d}^{\text{gt}}_{t,ij}$, we introduce a rotation factor $\mathbf{C} \in \mathtt{SO}(3)$ to correct it. We define the ideal correction as

$$\mathbf{C}^{\text{ideal}}_{t,i} = \operatorname*{arg\,min}_{\mathbf{C} \in \mathtt{SO}(3)} \sum_{j \in \mathcal{C}(i)} \|\mathbf{C} \widehat{\mathbf{d}}_{t,ij} - \mathbf{d}^{\text{gt}}_{t,ij}\|_2^2. \tag{2}$$

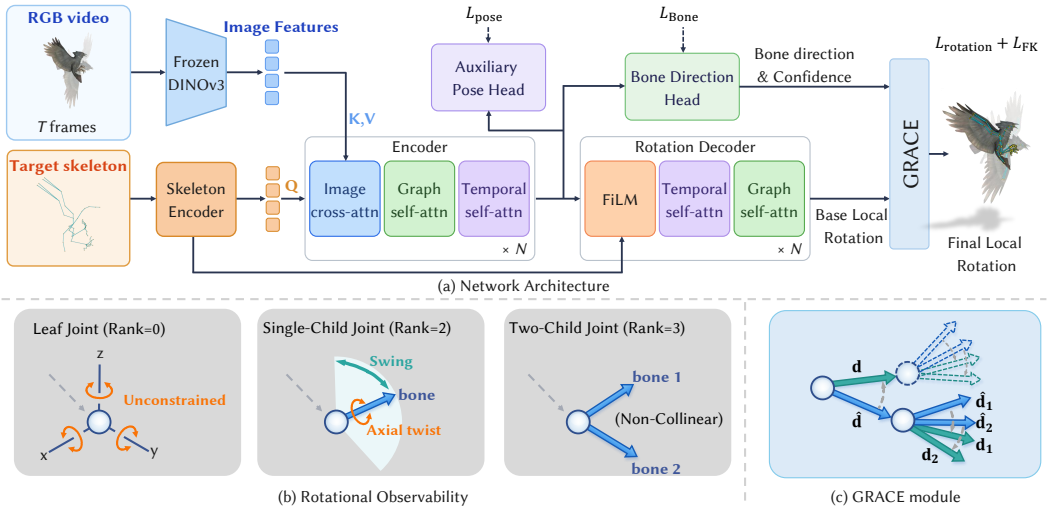


Figure 2: CAST overview. (a) CAST predicts initial local rotations and global bone directions from an RGB video and a target skeleton, then refines the rotations with GRACE. (b) Rotational observability: a leaf provides no directional constraint, a single child constrains swing but not axial twist, and two non-collinear child bones constrain the full rotation. (c) GRACE aligns FK directions $\hat{\mathbf{d}}$ (blue) with predicted directions $\mathbf{d}$ (teal) through minimum-angle swing for a single direction or joint alignment of multiple directions.

Rotational observability. The outgoing bone geometry determines which rotational degrees of freedom can be constrained in Eq. 2. For an infinitesimal rotation $\delta\boldsymbol{\omega} \in \mathbb{R}^3$ applied on the left of $\widehat{\mathbf{G}}_{t,i}$, the first-order change in each bone direction is

$$\delta\widehat{\mathbf{d}}_{t,ij} = \delta\boldsymbol{\omega} \times \widehat{\mathbf{d}}_{t,ij}. \tag{3}$$

Thus, each bone constrains the two rotation components perpendicular to its axis, represented by the projector $\mathbf{I} - \widehat{\mathbf{d}}_{t,ij}\widehat{\mathbf{d}}_{t,ij}^{\mathsf{T}}$, where $\mathbf{I}$ is the 3×3 identity matrix.

For nonempty $\mathcal{C}(i)$, averaging these per-bone projectors gives the normalized Gauss–Newton matrix

$$\mathbf{A}_{t,i}^{\text{glob}} = \frac{1}{|\mathcal{C}(i)|} \sum_{j \in \mathcal{C}(i)} \left(\mathbf{I} - \widehat{\mathbf{d}}_{t,ij}\widehat{\mathbf{d}}_{t,ij}^{\mathsf{T}}\right) = \widehat{\mathbf{G}}_{t,i} \underbrace{\left[\frac{1}{|\mathcal{C}(i)|} \sum_{j \in \mathcal{C}(i)} \left(\mathbf{I} - \mathbf{r}_{ij}\mathbf{r}_{ij}^{\mathsf{T}}\right)\right]}_{\mathbf{A}_i^{\text{loc}}} \widehat{\mathbf{G}}_{t,i}^{\mathsf{T}}. \tag{4}$$

Definition 1 (Rotational observability)**.** The locally observable rotational subspace of joint i, expressed in its rest-pose local axes, is defined as the range of $\mathbf{A}_i^{\text{loc}}$. Its dimension, $\text{rank}(\mathbf{A}_i^{\text{loc}})$, is the number of observable rotational degrees of freedom constrained by its outgoing bone directions.

Equation 4 rotates this representation into the current pose without changing its rank, so the number of observable rotational degrees of freedom is independent of the joint's orientation. Figure 2(b) illustrates three cases: for leaves, we define both matrices as zero, giving rank zero; a single direction or collinear directions give rank two, constraining swing but leaving axial twist unobservable; two or more non-collinear directions give rank three and constrain full rotation. Derivation in Appendix A.2.

Lemma 1 (Uniqueness of the ideal correction)**.** *Suppose all ground-truth outgoing bone directions at joint i are obtained by applying the same rotation $\mathbf{G}_{t,i}^{\text{gt}}$ to their corresponding rest-pose directions. Then Eq. 2 has a unique solution if and only if $\text{rank}(\mathbf{A}_i^{\text{loc}}) = 3$, in which case $\mathbf{C}_{t,i}^{\text{ideal}} = \mathbf{G}_{t,i}^{\text{gt}}\widehat{\mathbf{G}}_{t,i}^{\mathsf{T}}$ (proof in Appendix A.3).*

In the fully observable case, form the cross-covariance matrix and its SVD:

$$\mathbf{H}_{t,i}^{\text{ideal}} = \sum\nolimits_{j \in \mathcal{C}(i)} \mathbf{d}_{t,ij}^{\text{gt}}\widehat{\mathbf{d}}_{t,ij}^{\mathsf{T}} = \mathbf{U}_{t,i}\boldsymbol{\Sigma}_{t,i}\mathbf{V}_{t,i}^{\mathsf{T}}, \tag{5}$$

where $\mathbf{U}_{t,i}$ and $\mathbf{V}_{t,i}$ are orthogonal and $\boldsymbol{\Sigma}_{t,i}$ contains the singular values in non-increasing order. The Wahba/Kabsch solution (Wahba, 1965; Kabsch, 1976) is

$$\mathbf{C}_{t,i}^{\text{ideal}} = \mathbf{U}_{t,i}\,\mathrm{diag}(1, 1, \det(\mathbf{U}_{t,i}\mathbf{V}_{t,i}^{\mathsf{T}}))\mathbf{V}_{t,i}^{\mathsf{T}}. \tag{6}$$

Applying this solution gives the corrected global rotation $\mathbf{G}_{t,i} = \mathbf{C}_{t,i}^{\text{ideal}}\widehat{\mathbf{G}}_{t,i}$. Under the conditions of Lemma 1, $\mathbf{G}_{t,i} = \mathbf{G}_{t,i}^{\text{gt}}$, so the error in the composed global rotation is removed.

With compatible directions but rank two, we select a minimum-angle swing that aligns the bone directions; axial twist remains undetermined (Appendix A.4). For rank zero, no bone direction constrains the correction, and we set $\mathbf{C}_{t,i}^{\text{ideal}} = \mathbf{I}$.

4.2 GRACE: RELIABILITY-GATED ROTATION CORRECTION

Confidence-weighted alignment. In practice, GRACE replaces the ground-truth directions $\mathbf{d}_{t,ij}^{\text{gt}}$ assumed in Section 4.1 with predicted directions $\mathbf{d}_{t,ij}$. The bone direction head uses a 7-layer MLP to fuse parent–child features and bone geometry, predicting global unit directions and reliabilities $c_{t,ij} \in [0, 1]$. We align the FK-derived directions $\widehat{\mathbf{d}}_{t,ij} = \widehat{\mathbf{G}}_{t,i}\mathbf{r}_{ij}$ with these predictions, weighting each bone by its reliability,

$$\mathbf{C}_{t,i}^{*} = \underset{\mathbf{C}\in\mathtt{SO}(3)}{\arg\min} \sum\nolimits_{j\in\mathcal{C}(i)} c_{t,ij}\|\mathbf{C}\widehat{\mathbf{d}}_{t,ij} - \mathbf{d}_{t,ij}\|_2^2. \tag{7}$$

We compute $\mathbf{C}_{t,i}^{*}$ numerically using stabilized Wahba/Kabsch alignment (Wahba, 1965; Kabsch, 1976) for well-conditioned multi-direction inputs. Single-direction inputs use a minimum-angle swing. Figure 2(c) illustrates the alignment, with solver details provided in Appendices A.4 and B.3.

Gated correction. To prevent inaccurate direction predictions from disrupting early training, we use a learned gate initialized to favor small corrections. Joint features, directional reliability and disagreement, and target-skeleton geometry determine its strength $g_{t,i} \in [0, 1]$, which scales the alignment proposal in the Lie algebra,

$$\widetilde{\mathbf{C}}_{t,i} = \mathtt{exp}\big(g_{t,i}\mathtt{log}(\mathbf{C}_{t,i}^{*})\big)\,. \tag{8}$$

Here $\widetilde{\mathbf{C}}_{t,i}$ is the applied correction, and `log` and `exp` denote the rotation logarithm and matrix exponential. Gate and logarithm details are given in Appendix B.3.

GRACE corrects rotations level by level from the root to the leaves. The root starts from $\widehat{\mathbf{G}}_{t,0} = \widehat{\mathbf{L}}_{t,0}^{0}$. Each non-root joint composes its initial local rotation with its parent's corrected global rotation, giving $\widehat{\mathbf{G}}_{t,i} = \mathbf{G}_{t,p(i)}\widehat{\mathbf{L}}_{t,i}^{0}$. At each level, we recompute the outgoing FK directions, solve the alignment in Eq. 7, and left-multiply each provisional rotation by its gated correction $\widetilde{\mathbf{C}}_{t,i}$. The corrected global rotations then serve as the parent rotations for the next level. Finally, we recover local rotations as $\mathbf{L}_{t,i} = \mathbf{G}_{t,p(i)}^{\mathsf{T}}\mathbf{G}_{t,i}$ for $i > 0$, with $\mathbf{L}_{t,0} = \mathbf{G}_{t,0}$.

4.3 TRAINING OBJECTIVE

We directly supervise the bone direction head with a direction loss $\mathcal{L}_{\text{dir}}$ and a reliability loss $\mathcal{L}_{\text{rel}}$:

$$\mathcal{L}_{\text{dir}} = \mathbb{E}\big[1 - \mathbf{d}_{t,ij}^{\mathsf{T}}\mathbf{d}_{t,ij}^{\text{gt}}\big]\,, \tag{9}$$

$$\mathcal{L}_{\text{rel}} = \mathbb{E}[\rho(c_{t,ij} - \mathrm{sg}[\exp(-\alpha_{t,ij}/\tau)])]\,, \tag{10}$$

$$\alpha_{t,ij} = \arccos(\mathbf{d}_{t,ij}^{\mathsf{T}}\mathbf{d}_{t,ij}^{\text{gt}}), \tag{11}$$

where $\alpha_{t,ij}$ is the angular error in radians, ρ is the Smooth L1 penalty, $\tau > 0$ controls the target's decay, and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. We apply $\mathcal{L}_{\text{rotation}}$ and $\mathcal{L}_{\text{FK}}$ to the corrected local rotations $\mathbf{L}_{t,i}$ and their FK positions, respectively, with no separate loss on the initial predictions. The auxiliary pose loss $\mathcal{L}_{\text{pose}}$ supervises the pose head's joint positions and adjacent-frame displacements. Combining these terms with $\mathcal{L}_{\text{bone}} = \mathcal{L}_{\text{dir}} + 0.1\mathcal{L}_{\text{rel}}$ gives

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{rotation}} + \lambda_{\text{FK}}\mathcal{L}_{\text{FK}} + \lambda_{\text{bone}}\mathcal{L}_{\text{bone}}. \tag{12}$$

The component losses, weights, and reductions are detailed in Appendix B.4.

Table 1: Combined matched-rig and cross-skeleton results. Lower is better for all metrics, and a dash denotes an unreported result. MPJPE and PA-MPJPE are reported in millimeters, MPJVE in millimeters per frame, and CD in normalized coordinates.

Task	Dataset	Model	MPJPE	PA-MPJPE	MPJVE	CD
Matched-rig	Zoo	Puppeteer	454.70	236.38	19.74	0.245
		TopoCap	–	–	–	–
		MCA V2*	105.52	67.04	10.35	0.077
		CAST-B	42.68	33.23	5.48	0.036
		CAST-L	**39.59**	**31.05**	**5.14**	**0.034**
	MObj Seen	Puppeteer	205.85	103.89	22.09	0.136
		TopoCap	157.51	74.04	34.86	0.096
		MCA V2*	94.38	58.62	16.71	0.071
		CAST-B	55.39	32.92	9.73	0.043
		CAST-L	**52.19**	**30.91**	**9.10**	**0.041**
	MObj Unseen	Puppeteer	181.63	93.23	21.10	0.119
		TopoCap	191.77	82.11	38.93	0.109
		MCA V2*	103.04	61.98	16.44	0.074
		CAST-B	79.21	46.52	10.73	0.057
		CAST-L	**74.63**	**44.21**	**10.19**	**0.054**
Cross-skeleton	Mixamo	TopoCap	363.59	154.72	57.20	0.174
		MCA V2*	277.35	127.38	33.64	0.135
		CAST-B	149.50	85.79	19.60	0.087
		CAST-L	**125.62**	**71.16**	**15.74**	**0.074**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We train CAST jointly on Truebones Zoo (Zoo) (Truebones, 2022) and MObjaverse (MObj), using a 95%/5% train/test split for Zoo and TopoCap's Train/Seen/Unseen splits for MObj (Pu et al., 2026; Deitke et al., 2023). Cross-skeleton evaluation uses 60 held-out Mixamo animations (Adobe, 2026), ten per character skeleton across six skeletons. Pairing each source sequence with the other five skeletons yields 300 ordered source–target pairs. Mixamo-retargeted animations on the target characters serve as references used only for evaluation. Data and evaluation details are given in Appendices D.1 and D.3.

Implementation details. We evaluate two model sizes. CAST-B uses 8 blocks of width 256 in both the encoder and rotation decoder, while CAST-L uses 16 blocks of width 384. Both use frozen DINOv3 (Siméoni et al., 2025) image features by default. Ablation and task-specific experiments use CAST-B unless noted otherwise. We train for 6,000 steps with AdamW on 4 NVIDIA-A100 GPUs, using a batch size of 256 and 81-frame clips. The learning-rate schedule and loss weights are given in Appendix B.4. For baselines, we use the TopoCap (Pu et al., 2026) official checkpoint and reimplement MoCapAnything V2 (Gong et al., 2026b), denoted by MCA V2*, training it on the same Zoo and MObj data as CAST. We also evaluate Puppeteer (Song et al., 2026) for matched-rig motion capture using its released video-guided optimizer and pretrained tracking, depth, and optical-flow models on the supplied rigged assets.

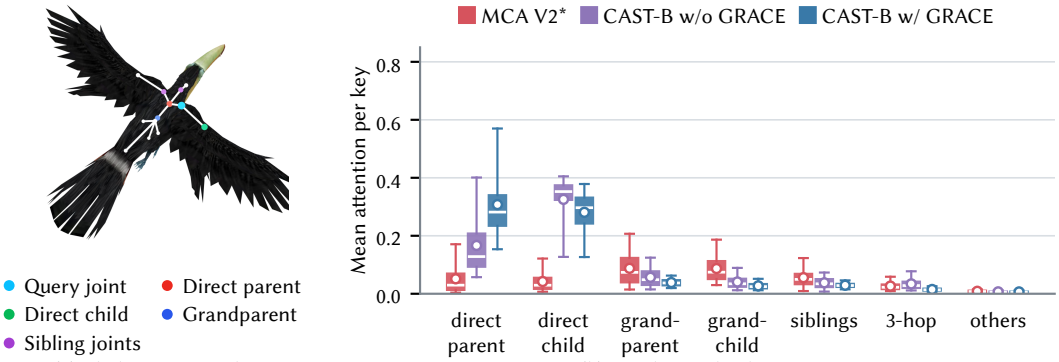
Metrics. For both tasks, we report MPJPE, PA-MPJPE, MPJVE, and Chamfer distance (CD) on root-centered predictions and references over complete animations. Following TopoCap's unit-bounding-box normalization and reporting convention (Pu et al., 2026), MPJPE and PA-MPJPE are reported in millimeters, MPJVE in millimeters per frame, and CD in normalized coordinates. Metric definitions and temporal-window handling are detailed in Appendices D.2 and D.3.

5.2 MOTION CAPTURE AND CROSS-SKELETON ANIMATION

Matched-rig motion capture. CAST recovers the character's articulation more faithfully across both animal and humanoid rigs. In Figure 3 (left), CAST-L preserves the folded leg configuration of the multi-limbed character, whereas the baselines spread or raise limbs that are folded in the input.



Figure 3: Qualitative comparison with baselines. **Left:** matched-rig motion capture. **Right:** cross-skeleton animation, with retargeted target-character references in the final column (Retarget). Each example shows an input frame and the three methods' outputs on the same target character.



(a) Skeleton example (b) Final encoder-layer attention

Figure 4: Graph attention by kinematic relation on Zoo. (a) An illustrative query joint and its topological relations. (b) Final encoder-layer attention distributions pooled over valid frames and query joints across the evaluated Zoo sequences.

For the humanoid, it also avoids the outstretched-arm pose produced by MCA V2*. These differences concern the relative arrangement of body parts, which an overall orientation adjustment alone cannot resolve. The quantitative results support this observation: CAST-L gives the lowest reported errors on all four metrics across Zoo, MObj Seen, and MObj Unseen (Table 1), with MPJPE reductions of 27.6–62.5% relative to MCA V2*. The improvements persist after Procrustes alignment, indicating better articulated pose recovery beyond global alignment, while lower MPJVE shows that joint displacements also follow the reference motion more closely.

This behavior is consistent with CAST's use of target-skeleton queries to organize visual evidence and global bone directions to constrain the orientations produced by local-rotation composition. The contribution of the complete GRACE configuration is examined separately in Section 5.4. Generalization to unseen structures remains more difficult: CAST-L's MPJPE rises from 52.19 on MObj Seen to 74.63 on MObj Unseen, even though it retains an advantage over every evaluated baseline. Thus, accommodating a variable rig does not remove the difficulty of interpreting unfamiliar articulations from images. Additional matched-rig examples are provided in Appendix G.

Cross-skeleton animation. Cross-skeleton animation requires preserving the observed action while expressing it through a different body structure. Figure 3 (right) shows this distinction: CAST-L retains the bent arms and forward-leaning torso of the running example, and the hand-together gesture of the compact source character. In the latter example, the retargeted reference places the arms inside the bulkier target's torso, whereas CAST-L keeps the hands in front of the body. The action is recognizable despite a change in arm configuration, illustrating the value of conditioning motion on the target rig. This also explains why distance to a retargeted reference captures only part of transfer quality: retargeting across different body proportions can introduce interpenetration (Villegas et al., 2021). Qualitative examples of the four transfer directions are in Appendix G.4.

Both CAST variants nevertheless improve all four reference-based metrics over the evaluated baselines (Table 1). CAST-L reduces MPJPE by 54.7% relative to MCA V2*, and the accompanying reductions in PA-MPJPE and MPJVE indicate gains in articulation and frame-to-frame motion as well as overall pose agreement. Increasing capacity from CAST-B to CAST-L brings a larger MPJPE reduction here (16.0%) than in the matched-rig evaluations (5.8–7.2%). This suggests that mapping

Table 3: GRACE and visual-encoder ablations on CAST-B. Bold indicates the best result.

Model	Zoo		MObj Seen		MObj Unseen	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
CAST-B w/o GRACE (DINOv2)	**49.58**	**38.70**	64.56	37.13	93.32	52.12
CAST-B w/o GRACE (DINOv3)	50.41	38.78	**63.71**	**36.87**	**86.64**	**50.71**
CAST-B (DINOv2)	**42.41**	**33.03**	57.34	33.27	83.15	48.19
CAST-B (DINOv3)	42.68	33.23	**55.39**	**32.92**	**79.21**	**46.52**

visual action cues to a different target structure benefits particularly from additional model capacity in the evaluated setting. The qualitative body adaptation and quantitative reference agreement provide complementary evidence, although neither establishes collision-free motion in general.

Local and global modeling in CAST. For local-rotation predictors such as MCA V2*, nonlocal joint context can help maintain global consistency as ancestor errors propagate through rotation composition. GRACE reduces this burden by correcting composed orientations with predicted global bone directions, allowing the predictor to emphasize parent-relative articulation. On Zoo, attention in the final encoder layer concentrates on direct parents and children. Adding GRACE increases parent attention while reducing attention to distant relations (Figure 4), consistent with partial decoupling of local articulation estimation from global orientation correction.

5.3 User study

18 participants separately ranked 20 anonymized result triplets across both tasks by overall animation quality (protocol in Appendix D.4). CAST-L obtains 63.3% of first-choice votes across 360 rankings and achieves the best mean rank of 1.531 (Table 2), supporting the perceptual quality of its predictions.

Table 2: Overall user-study preferences.

Method	First choice (%)	Mean rank
TopoCap	19.7	2.261
MCA V2*	16.9	2.208
CAST-L	**63.3**	**1.531**

5.4 Ablation study

Table 3 evaluates GRACE and the choice of frozen visual encoder in CAST-B under the same training schedule. With either DINOv2 (Oquab et al., 2023) or DINOv3, GRACE reduces MPJPE by 8.6–15.3% across all splits. The accompanying PA-MPJPE gains indicate improved articulation beyond global alignment. DINOv3 provides its largest improvement over DINOv2 on MObj Unseen, lowering MPJPE by 4.7% with GRACE, while performance on Zoo remains similar. These results support DINOv3 for transfer to unfamiliar structures and show that the complete GRACE configuration provides consistent gains under both visual encoders.

6 Application

We adapt CAST-B to key-frame motion completion and causal inference by independently finetuning each variant for 3,000 steps. The completion model uses randomly masked visual inputs, while the three causal models use chunk sizes of 1, 2, and 4, respectively.

Key-frame motion completion. The completion variant enables sketch-guided animation inbetweening on a target skeleton (examples in Appendix G.5). Using only one quarter of the input frames (k=4), the model's MPJPE is 5.7–8.5% higher than with dense input (Table 4). Errors increase with longer gaps between observations.

Causal inference. Causal temporal attention allows motion capture from incoming frames. The model trained with single-frame chunks incurs a modest accuracy loss relative to bidirectional CAST-B on Zoo and MObj Unseen (Table 4). The model trained with 4-frame chunks achieves comparable MPJPE and PA-MPJPE to bidirectional CAST-B across all three splits, at the cost of waiting for a complete chunk before prediction.

For live capture, we deploy the causal variant through a Blender add-on that applies the received skeletons to a rigged character as they arrive, allowing a performance to be inspected during recording. Figure 5 and Appendix C show the add-on interface and a live capture session.

Table 4: Causal inference and motion completion with CAST-B. The Full sequence setting denotes the bidirectional model with dense input, and k denotes the observation interval for key-frame completion. Each Real-time model is trained separately for different chunk sizes. Errors are reported in millimeters, with lower values indicating better performance. Bold and underline mark the best and second-best results.

	Zoo		MObj Seen		MObj Unseen	
Task setting	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
Full sequence	**42.68**	**33.23**	55.39	<u>32.92</u>	**79.21**	**46.52**
Key frame (k=4)	45.61	35.15	58.52	35.84	85.93	49.65
Key frame (k=8)	48.79	37.35	67.98	41.37	92.01	53.90
Key frame (k=16)	58.53	43.68	82.73	50.15	104.45	61.69
Key frame (k=32)	75.96	52.77	99.39	59.09	118.54	68.47
Real-time (chunk 1)	44.09	34.00	**52.46**	**32.32**	80.79	47.23
Real-time (chunk 2)	43.54	33.67	54.36	33.32	80.34	47.11
Real-time (chunk 4)	<u>42.88</u>	<u>33.25</u>	<u>53.98</u>	32.97	<u>79.62</u>	<u>46.99</u>

Table 5: Inference efficiency on a single NVIDIA A100. Params (millions) and Time (ms per forward pass) include the frozen image backbone. Frames denotes the input length per pass, and FPS measures streaming inference throughput.

			30 joints		52 joints		89 joints	
Model	Params	Frames	Time	FPS	Time	FPS	Time	FPS
TopoCap (10 steps)	2913.8	8	321	–	342	–	321	–
TopoCap (300 steps)	2913.8	8	8550	–	9080	–	8510	–
MCA V2	452.6	32	192	–	191	–	192	–
CAST-B	352.6	81	142	–	145	–	154	–
CAST-L	488.2	81	186	–	189	–	201	–
CAST-B (causal)	352.6	1	19.5	51.3	18.5	53.9	18.9	52.9

6.1 Inference Efficiency

Table 5 compares model size and inference efficiency on a single NVIDIA A100. CAST-B uses fewer parameters than the evaluated baselines and processes longer temporal windows in less time, providing higher offline throughput under the reported configurations. Scaling to CAST-L further improves motion accuracy with a moderate increase in inference cost, while retaining a throughput advantage over the baselines. The causal variant further enables real-time online prediction across all tested skeletons without requiring future observations. These results support both efficient offline motion capture and online use across different target skeletons. Timings include the frozen image backbone and motion prediction, excluding video acquisition and mesh rendering. CAST uses bf16 precision and `torch.compile`, with implementation optimizations detailed in Appendix B.5.

7 Conclusion

We presented CAST, a framework for recovering motion from monocular video on a supplied target skeleton, including when its topology differs from the subject in the video. Our observability analysis characterizes which rotational components outgoing bone directions can constrain and derives closed-form corrections under exact, compatible directional evidence. Building on this analysis, GRACE aligns composed rotations with predicted global bone directions through reliability-gated updates, propagating corrected orientations from parents to children. CAST improves matched-rig capture on Zoo and MObj and cross-skeleton animation on the curated Mixamo benchmark over the evaluated baselines. Ablations support the contribution of GRACE under both visual encoders, and the user study provides complementary evidence of animation quality. Fine-tuned variants support key-frame motion completion and causal inference, with single-frame CAST-B exceeding 51 FPS in model inference on an NVIDIA A100. Remaining limitations include rotational ambiguities unconstrained by bone directions and the inability to represent time-varying parent-local bone offsets (Appendix F). Together, these results demonstrate the value of combining learned local rotations with topology-dependent geometric correction for motion capture across skeletons.

AI USE STATEMENT

In this work, we used generative AI tools for one task with required disclosure: implementing methods, specifically to write and edit parts of the model, training, and evaluation code. We have not used generative AI tools to generate synthetic data sets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in the writing of proofs, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, to clean or reformat data sets, to support qualitative or thematic data analysis, to interpret results, or to assist with translation. Additionally, we used generative AI tools to edit parts of the manuscript to improve readability. We have reviewed all AI-assisted work: we executed and validated the AI-assisted code against our experiment records, reproduced every reported number from our own runs, revised all text ourselves, and checked every cited reference against the original publication. We created all figures and tables ourselves without any AI-generated content: the schematics are our own drawings, the rendered examples are outputs of our implementation, and the quantitative figures and tables are computed from our experiment records. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The experiments use rendered articulated assets. Use and redistribution of source assets remain subject to their respective licenses. The method predicts character motion and does not infer personal identity or sensitive attributes. The anonymous user study required explicit consent, did not request names or email addresses, and used randomly generated participant IDs. Submitted response records contained no IP addresses. Appendix D.4 describes the questionnaire and the fields recorded by the study application.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the problem formulation, observability analysis, correction rules, and training objective, and Section 5 the training setting, baselines, and evaluation benchmark. Appendix D details the metric definitions, cross-skeleton pairing, windowed inference, and user-study configuration; Appendix B.2 specifies the coordinate conventions. Appendices D.1, B.1, B.4, and B.3 give the data and split sizes, the configured architecture, the training penalties, and the numerical details, including the optimizer, schedule, seed, and precision. Appendix B.5 describes the inference implementation and optimization used for the reported latency, and Appendix E.2 clarifies the scope of the reported GRACE comparisons and their relation to the ideal analysis. Code, model configurations, checkpoint identifiers, and split manifests are available at `https://cast-mocap.github.io/`, subject to source-data licenses.

REFERENCES

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (ToG)*, 39(4):62:1–62:14, 2020.

Adobe. Mixamo. `https://www.mixamo.com/`, 2026. Accessed: 2026-09-22.

Ling-Hao Chen, Yuhong Zhang, Zixin Yin, Zhiyang Dou, Xin Chen, Jingbo Wang, Taku Komura, and Lei Zhang. Motion2Motion: Cross-topology motion transfer with sparse correspondence. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-XL: A universe of 10m+ 3D objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

Inbar Gat, Sigal Raab, Guy Tevet, Yuval Reshef, Amit Haim Bermano, and Daniel Cohen-Or. AnyTop: Character animation diffusion with any topology. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025.

Kehong Gong, Zhengyu Wen, Weixia He, Mingxi Xu, Qi Wang, Ning Zhang, Zhengyu Li, Dongze Lian, Wei Zhao, Xiaoyu He, et al. MoCapAnything: Unified 3D motion capture for arbitrary skeletons from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7089–7099, 2026a.

Kehong Gong, Zhengyu Wen, Dao Thien Phong, Mingxi Xu, Weixia He, Qi Wang, Ning Zhang, Zhengyu Li, Guanli Hou, Dongze Lian, et al. MoCapAnything V2: End-to-end motion capture for arbitrary skeletons. *arXiv preprint arXiv:2604.28130*, 2026b.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. MIMO: Controllable character video synthesis with spatial decomposed modeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21181–21191. IEEE, 2025.

Lucas Mourot, Ludovic Hoyet, François Le Clerc, and Pierre Hellier. HuMoT: Human motion representation using topology-agnostic transformers for character animation retargeting. *arXiv preprint arXiv:2305.18897*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xiaoyu Pan, Bowen Zheng, Xinwei Jiang, Zijiao Zeng, Qilong Kou, He Wang, and Xiaogang Jin. RoMo: A robust solver for full-body unlabeled optical motion capture. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Dario Pavllo, David Grangier, and Michael Auli. QuaterNet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Cheng-Feng Pu, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. TopoCap: Learning topology-agnostic motion priors for monocular video-to-animation. *arXiv preprint arXiv:2606.12153*, 2026.

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 *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–2080. IEEE, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Chaoyue Song, Xiu Li, Fan Yang, Zhongcong Xu, Jiacheng Wei, Fayao Liu, Jiashi Feng, Guosheng Lin, and Jianfeng Zhang. Puppeteer: Rig and animate your 3D models. *Advances in Neural Information Processing Systems*, 38:72152–72184, 2026.

Truebones. Free FBX/BVH zoo: Over 75 animals with animations and textures. `https://truebones.gumroad.com/l/skZMC`, 2022. Accessed: 2026-09-22.

Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8639–8648. IEEE, 2018.

Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9700–9709. IEEE, 2021.

Grace Wahba. A least squares estimate of satellite attitude. *SIAM Review*, 7(3):409, 1965.

Ruiyan Wang, Teng Hu, Kaihui Huang, Zihan Su, Ran Yi, and Lizhuang Ma. PoseAnything: General pose-guided video generation with part-aware temporal coherence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23193–23203, 2026.

Zijie Wu, Lixin Xu, Puhua Jiang, Sicong Liu, Chunchao Guo, and Xiang Bai. R-DMesh: Video-guided 3D animation via rectified dynamic mesh flow. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '26, pp. 1–11. ACM, 2026. doi: 10.1145/3799902.3811135.

Xiaohang Yang, Qing Wang, Jiahao Yang, Gregory Slabaugh, and Shanxin Yuan. STaR: Seamless spatial-temporal aware motion retargeting with penetration and consistency constraints. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12947–12955. IEEE, 2025.

Haodong Zhang, Zhike Chen, Haocheng Xu, Lei Hao, Xiaofei Wu, Songcen Xu, Zhensong Zhang, Yue Wang, and Rong Xiong. Semantics-aware motion retargeting with vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2155–2164. IEEE, 2024.

Yuhao Zhang, Gerard Pons-Moll, and Tolga Birdal. SAMoR: Motion modelling for articulated objects of any skeleton and topology. *arXiv preprint arXiv:2607.02148*, 2026.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746. IEEE, 2019.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15039–15053. IEEE, 2023.