RECAST: REUSING HUMAN PRIORS FOR DATA-SCARCE 3D ANIMAL POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular 3D animal pose estimation is inherently ambiguous, while the scarcity of 3D annotations makes reliable probabilistic 2D-to-3D lifting difficult to learn. We introduce **RECAST**, an efficient flow-matching framework that **RE**uses human priors for data-s**CA**rce 3D animal pose e**ST**imation. RECAST employs a three-stage training strategy. First, probabilistic pretraining on abundant human poses learns a conditional 2D-to-3D flow that provides a transferable initialization for cross-species lifting. Second, low-rank feature and graph adaptation specializes the pretrained flow to animal morphology while retaining its conditional prediction structure. Third, skeletal refinement is introduced after adaptation has converged, further improving both pose accuracy and bone-length consistency. For efficient multi-hypothesis inference, we further propose **Chordal Trajectory Branching (CTB)**. Inspired by the linear probability paths used in OT-CFM, CTB constructs intermediate hypothesis states through a closed-form approximation from a reference trajectory and integrates only their remaining trajectory suffixes. This avoids repeated prefix integration and substantially reduces redundant flow evaluations. RECAST achieves 29.86 mm and 18.67 mm P-MPJPE on Animal3D and CtrlAni3D, respectively. CTB matches standard multi-hypothesis inference within 0.01 mm while reducing function evaluations by 49%.

1 INTRODUCTION

Monocular 3D pose estimation aims to recover the three-dimensional configuration of articulated joints from a single image (Sun et al., 2018) or its 2D keypoints (Martinez et al., 2017). This task is inherently ambiguous, as multiple 3D poses can produce the same 2D projection. To account for this ambiguity, probabilistic 3D human pose estimation methods model multiple plausible solutions through diffusion-based approaches (Gong et al., 2023; Shan et al., 2023; Xu et al., 2024) or conditional flow matching (Wang et al., 2026). Flow matching learns a velocity field that transports a source distribution toward a target distribution along prescribed probability paths (Lipman et al., 2022). Conditioned on 2D observations, it provides a probabilistic formulation of 3D pose lifting.

Extending probabilistic lifting to animals presents a distinct challenge. Large-scale human pose datasets, such as Human3.6M (Ionescu et al., 2013), provide hundreds of thousands of annotated frames, whereas Animal3D (Xu et al., 2023) and CtrlAni3D (Lyu et al., 2025) contain approximately 3.4K and 9.7K animal instances, respectively. Beyond this disparity in supervision, animal poses exhibit substantial variation in body proportions and articulated structure across species. Learning a conditional flow solely from limited animal data must therefore establish a probabilistic 2D-to-3D mapping while accommodating morphological differences and preserving skeletal geometry. Despite these differences, human and animal pose lifting share the underlying challenge of recovering articulated 3D structures from ambiguous 2D observations. A human-pretrained conditional flow may therefore

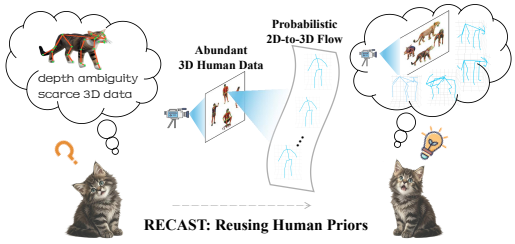


Figure 1: **Reusing human priors for data-scarce 3D animal pose estimation.** A human-pretrained conditional flow provides transferable lifting knowledge for generating plausible 3D animal poses from monocular observations.

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.

provide reusable probabilistic lifting knowledge, even though its joint dependencies and skeletal geometry cannot be transferred directly. These considerations raise a central question: which aspects of probabilistic lifting can be transferred from human poses, and which must be adapted to animal morphology?

We address this question with **RECAST**, an efficient flow-matching framework that **RE**uses human priors for data-s**CA**rce 3D animal pose e**ST**imation. As motivated in Fig. 1, our key insight is to transfer probabilistic lifting knowledge learned from large-scale human pose data to animal pose estimation. RECAST realizes this transfer through a three-stage training strategy. First, we pretrain a conditional 2D-to-3D flow on Human3.6M (Ionescu et al., 2013), a large-scale human pose benchmark, to learn a probabilistic lifting prior before introducing animal supervision. Second, low-rank adaptation (Hu et al., 2021a) specializes the feature transformations and graph operators of the pretrained model to animal-specific joint dependencies and body proportions. Finally, after the adapted model converges, skeletal refinement introduces species-specific bone-length supervision to improve pose accuracy and structural consistency. This progression separates the acquisition of general lifting knowledge from adaptation to animal morphology and subsequent skeletal refinement. It allows the pretrained flow to be reused while addressing the structural differences that prevent direct human-to-animal transfer.

At inference, multiple plausible 3D hypotheses are generated by sampling distinct source states conditioned on the same 2D observation and integrating their trajectories through the learned conditional flow. Integrating each trajectory independently, however, requires repeated velocity-field evaluations along its entire path, with computational cost growing linearly with the number of hypotheses. Existing acceleration methods primarily reduce the cost of individual trajectories through efficient numerical solvers (Lu et al., 2022) or learned few-step generation (Song & Dhariwal, 2024). They do not, however, directly address redundant computation across hypotheses generated under the same conditioning. These hypotheses share a learned velocity field but differ in their sampled source states, motivating the approximation of their intermediate states from a reference trajectory rather than independently integrating every prefix.

We address this redundancy with **Chordal Trajectory Branching (CTB)**. Motivated by the linear conditional probability paths of optimal-transport conditional flow matching (OT-CFM) (Tong et al., 2023), CTB first integrates a reference trajectory and then uses its intermediate states, together with displacements between source states, to approximate branching states for the remaining hypotheses. Each hypothesis is subsequently integrated only from its approximate branching state to the target. By replacing repeated prefix integration with this closed-form approximation, CTB reduces velocity-field evaluations without modifying the learned flow.

Our contributions are summarized as follows:

- **Progressive cross-species adaptation.** We develop a three-stage training strategy that transfers human-learned probabilistic lifting knowledge, adapts it to animal morphology, and refines skeletal geometry under limited animal supervision.
- **Efficient multi-hypothesis inference.** We introduce CTB to reuse computation across hypotheses, reducing function evaluations by 49% with only a 0.01 mm difference in P-MPJPE relative to standard multi-hypothesis inference.
- **3D animal pose estimation performance.** RECAST achieves 29.9 mm and 18.7 mm P-MPJPE on Animal3D and CtrlAni3D, respectively, outperforming the compared methods on both benchmarks.

2 Related Work

2.1 Monocular 3D Human Pose Estimation

Monocular 3D human pose estimation is commonly formulated as direct regression from RGB images (Pavlakos et al., 2017; Sun et al., 2018; Kanazawa et al., 2018) or 3D lifting from detected 2D keypoints (Newell et al., 2016; Chen et al., 2018; Sun et al., 2019), with earlier work drawing on geometric and skeletal priors to constrain pose reconstruction (Agarwal & Triggs, 2005; Andriluka et al., 2009). For 2D-to-3D lifting, methods model spatial joint dependencies through diverse representations (Martinez et al., 2017; Chen & Ramanan, 2017; Fang et al., 2018; Zhao et al., 2019; Tang

et al., 2023a) and exploit temporal context to improve prediction consistency across frames (Pavllo et al., 2019; Liu et al., 2020; Cai et al., 2019; **?**; Hu et al., 2021b; Chen et al., 2021).

Transformer-based approaches extend spatio-temporal interaction modeling (Vaswani et al., 2017; Zheng et al., 2021; Li et al., 2022b;a; Hassanin et al., 2025; Xue et al., 2022; Zhang et al., 2022; Tang et al., 2023b), while graph, kinematic, pretraining, and state-space mechanisms provide complementary modeling strategies (Mehraban et al., 2024; Peng et al., 2024; Zhu et al., 2023; Huang et al., 2025; Lu et al., 2025). Despite advances in representation learning, monocular lifting remains inherently ambiguous: multiple 3D poses with different depths and joint configurations can produce similar 2D projections. Probabilistic approaches explicitly model this one-to-many mapping through diffusion-based sampling (Gong et al., 2023; Shan et al., 2023; Xu et al., 2024) or conditional flow matching (Wang et al., 2026), generating multiple plausible 3D pose hypotheses.

2.2 3D Animal Pose Estimation

Compared with human pose estimation, 3D animal pose estimation faces limited 3D supervision and substantial cross-species morphological variation. Animal3D (Xu et al., 2023) and CtrlAni3D (Lyu et al., 2025) expand available supervision through pseudo-3D annotations and synthetic animal imagery, respectively. Existing approaches include human mesh recovery frameworks adopted as cross-domain reconstruction baselines, such as HMR (Kanazawa et al., 2018) and HMR2.0 (Goel et al., 2023), as well as animal-oriented reconstruction methods such as VLDO (Biggs et al., 2020) and AniMer (Lyu et al., 2025). FMPose3D (Wang et al., 2026) further explores probabilistic animal pose lifting through conditional flow matching.

Nevertheless, transferring human-learned probabilistic lifting knowledge to animal pose estimation remains underexplored. RECAST addresses this gap by adapting conditional 2D-to-3D transport learned from human poses to animal morphology.

2.3 Cross-Species Transfer and Efficient Probabilistic Inference

Transfer learning leverages source-domain knowledge to reduce target-domain supervision. Parameter-efficient methods, including LoRA (Hu et al., 2021a) and AdaLoRA (Zhang et al., 2023), adapt pretrained models through low-rank updates. Cross-species pose transfer additionally requires adaptation to articulated structure, for which bone-length and kinematic priors provide complementary supervision (Chen et al., 2021; Peng et al., 2024). RECAST combines parameter-efficient adaptation with species-specific structural refinement to specialize the transferred conditional flow.

Generating multiple pose hypotheses requires repeated network evaluations (Wang et al., 2026). Existing methods reduce sampling cost through efficient numerical solvers (Lu et al., 2022), learned few-step generation (Song & Dhariwal, 2024), or low-energy transport for one-step image editing (Lu et al., 2026). Unlike these approaches, which focus on accelerating or stabilizing individual generation processes, Chordal Trajectory Branching (CTB) targets computation shared across hypotheses. It uses a reference trajectory to approximate their intermediate branching states and integrates only the remaining suffixes, thereby reducing redundant velocity-field evaluations.

3 Method

3.1 Overview

Given a monocular 2D animal pose $\mathbf{x} \in \mathbb{R}^{J\times 2}$, we aim to model the conditional distribution of its 3D pose $\mathbf{y} \in \mathbb{R}^{J\times 3}$. Learning this distribution solely from animal data is challenging because 3D annotations are scarce, while skeletal proportions and joint configurations vary across species.

As illustrated in Fig. 2, RECAST follows a three-stage training strategy. **Stage I** learns a probabilistic 2D-to-3D lifting prior from human poses. **Stage II** adapts pretrained feature transformations and graph-convolution operators to animal morphology through low-rank updates. **Stage III** introduces species-specific skeletal supervision after morphology adaptation has converged. For efficient multi-hypothesis inference, we further propose **Chordal Trajectory Branching (CTB)**, which constructs approximate intermediate hypothesis states and integrates only their remaining trajectory suffixes.

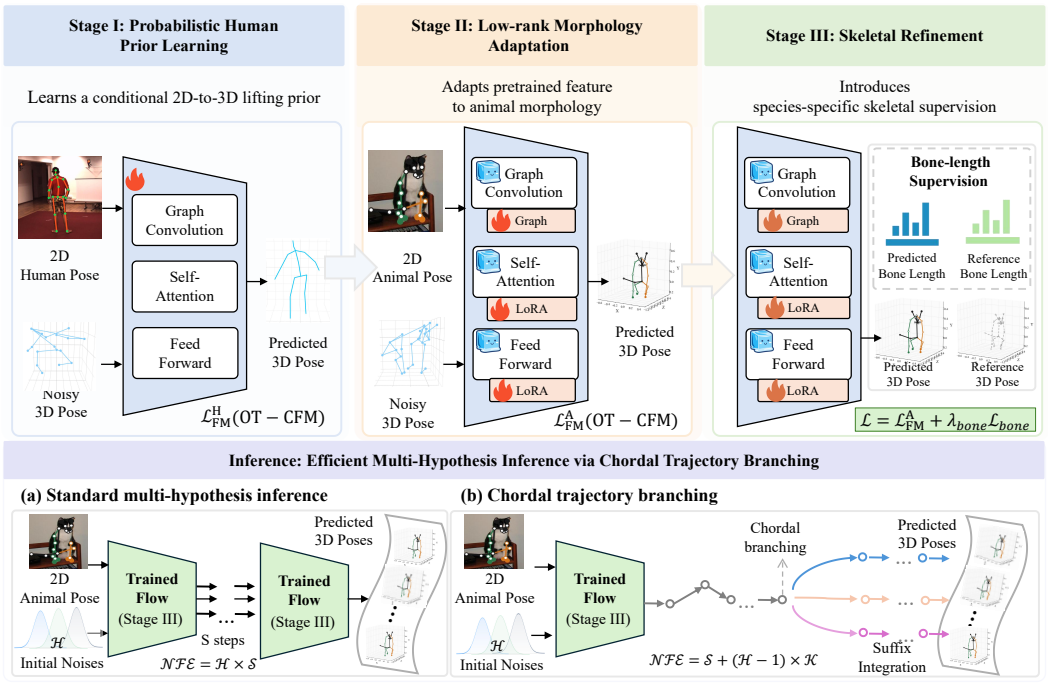


Figure 2: **Overview of RECAST. Training:** Stage I learns a conditional flow from human poses; Stage II adapts pretrained linear transformations and learns low-rank adjacency corrections on a fixed animal skeleton; Stage III introduces post-convergence skeletal supervision. **Inference:** CTB approximates intermediate hypothesis states using a reference trajectory and integrates only their remaining suffixes.

3.2 Stage I: Probabilistic Human Prior Learning

Conditional flow backbone. We adopt the Graph-Transformer backbone of FMPose3D (Wang et al., 2026). The 2D pose $\mathbf{x}$, intermediate 3D state $\mathbf{y}_t$, and flow time $t \in [0, 1]$ are encoded separately and concatenated before entering the backbone. The conditional velocity field is

$$v_{\theta_H}(\mathbf{y}_t, t \mid \mathbf{x}) = D(F_{\theta_H}(\text{Concat}[E_x(\mathbf{x}), E_y(\mathbf{y}_t), E_t(t)])), \tag{1}$$

where E_x and E_y are pose encoders, E_t is the time encoder, and D is the velocity decoder. The backbone F_{θ_H} consists of five Graph-Transformer blocks. Each block contains parallel graph-convolution and multi-head self-attention branches, denoted by $\mathcal{G}^{(l)}$ and $\mathcal{A}^{(l)}$, followed by a feed-forward network $\mathcal{F}^{(l)}$. Detailed architectural configurations are provided in Appendix B.

Human flow pretraining. We pretrain the conditional flow on Human3.6M (Ionescu et al., 2013). Let $\mathbf{y}_1$ denote a target 3D pose and $\mathbf{y}_0 \sim p_0$ a source sample. Following optimal-transport conditional flow matching (OT-CFM) (Tong et al., 2023), we construct the linear conditional path

$$\mathbf{y}_t = (1-t)\mathbf{y}_0 + t\mathbf{y}_1, \qquad t \sim \mathcal{U}(0,1), \tag{2}$$

with target velocity

$$\mathbf{u}_t = \mathbf{y}_1 - \mathbf{y}_0. \tag{3}$$

The human flow is trained by minimizing

$$\mathcal{L}_{\text{FM}}^{H} = \mathbb{E}_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_0,t}\left[\|v_{\theta_H}(\mathbf{y}_t, t \mid \mathbf{x}) - \mathbf{u}_t\|_2^2\right]. \tag{4}$$

The resulting model provides a human-learned probabilistic lifting prior that can be efficiently reused for subsequent animal adaptation.

3.3 Stage II: Low-Rank Morphology Adaptation

Human pretraining provides reusable feature and channel transformations, but the human adjacency matrix cannot be directly transferred to an animal skeleton with a different joint configuration. We therefore adopt a predefined animal skeleton graph while retaining the pretrained channel transformations. GraphLoRA adapts the animal adjacency structure within the graph branches $\mathcal{G}^{(l)}$, while feature-level LoRA adapts linear transformations in the attention branches $\mathcal{A}^{(l)}$, feed-forward networks $\mathcal{F}^{(l)}$, and pose encoders.

Graph-level adaptation. We introduce **GraphLoRA** to adapt the animal adjacency structure while preserving the predefined animal skeleton as a fixed structural prior. Let $\mathbf{A}_{\text{animal}} \in \mathbb{R}^{K_g \times J \times J}$ denote the animal adjacency matrix, where $K_g = 4$ is the number of graph partitions and $J = 26$ is the number of animal joints. For each graph-convolution block l, the adapted adjacency matrix is

$$\mathbf{A}^{(l)}_{\text{adapt}} = \mathbf{A}_{\text{animal}} + \frac{1}{r_g}\mathbf{B}^{(l)}_g \mathbf{A}^{(l)}_g, \tag{5}$$

where $\mathbf{A}^{(l)}_g \in \mathbb{R}^{r_g \times J}$ and $\mathbf{B}^{(l)}_g \in \mathbb{R}^{K_g \times J \times r_g}$ are trainable low-rank factors, with $r_g = 4$. The matrix product is computed independently for each graph partition.

The resulting graph convolution can be written as

$$\mathbf{X}'_{b,j} = \sum_{k=1}^{K_g} \sum_{i=1}^{J} \mathbf{A}^{(l)}_{\text{adapt}}[k, i, j]\, \mathbf{W}^{(l)}_k \mathbf{X}_{b,i}, \tag{6}$$

where $\mathbf{W}^{(l)}_k$ denotes the human-pretrained graph-convolution channel transformation. We apply GraphLoRA to all five graph-convolution blocks. The animal adjacency matrix and pretrained channel weights remain frozen; only the low-rank adjacency corrections are optimized. Thus, GraphLoRA learns low-rank corrections to the fixed animal adjacency matrix while preserving the human-pretrained graph-convolution channel weights.

Attention and feed-forward adaptation. To adapt the pretrained feature transformations, we apply LoRA (Hu et al., 2021a) to linear layers in the attention branches $\mathcal{A}^{(l)}$ and feed-forward networks $\mathcal{F}^{(l)}$. For a pretrained linear weight $\mathbf{W}_H \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the adapted weight is

$$\mathbf{W}_A = \mathbf{W}_H + \Delta\mathbf{W}, \qquad \Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{B}\mathbf{A}, \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ and $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ are trainable low-rank factors. We apply LoRA to 28 linear transformations with $r = 128$ and $\alpha = 128$. These updates adapt pretrained feature transformations to animal poses.

Animal flow adaptation. The resulting animal velocity field v_{θ_A} is trained on animal pose data using the same flow-matching objective:

$$\mathcal{L}^{A}_{\text{FM}} = \mathbb{E}_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_0,t}\left[\left\|v_{\theta_A}(\mathbf{y}_t, t \mid \mathbf{x}) - \mathbf{u}_t\right\|_2^2\right]. \tag{8}$$

We train Stage II to convergence and denote the resulting parameters by θ^*_A.

3.4 Stage III: Post-Convergence Skeletal Refinement

Stage II adapts the conditional velocity field to animal poses, but flow matching does not explicitly constrain skeletal proportions. After morphology adaptation has converged, we introduce species-specific bone-length supervision through a terminal-pose estimate.

For a training sample, we construct $\mathbf{y}_t$ using the conditional path in Eq. 2. At the sampled time t, the predicted velocity yields

$$\hat{\mathbf{y}}_1 = \mathbf{y}_t + (1-t)\, v_{\theta_A}(\mathbf{y}_t, t \mid \mathbf{x}). \tag{9}$$

This one-step estimate extrapolates from the current state without requiring ODE integration during training. Let $\mathcal{E}$ denote the set of skeletal edges, and let $\bar{\ell}_{ij}^{(s)}$ be the reference length of bone (i,j) for species s, estimated from training-set 3D poses. The skeletal loss is defined as

$$\mathcal{L}_{\text{bone}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left(\|\hat{\mathbf{y}}_{1,i} - \hat{\mathbf{y}}_{1,j}\|_2 - \bar{\ell}_{ij}^{(s)} \right)^2. \tag{10}$$

Starting from the converged Stage II parameters θ_A^*, Stage III minimizes

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{FM}}^{A} + \lambda_{\text{bone}} \mathcal{L}_{\text{bone}}, \tag{11}$$

where $\lambda_{\text{bone}} = 0.1$. The skeletal loss backpropagates through the velocity-induced terminal-pose estimate to refine the conditional flow while retaining the flow-matching objective. Full ODE integration is used for hypothesis generation at inference.

3.5 Efficient Multi-Hypothesis Inference via Chordal Trajectory Branching

Given the trained animal velocity field v_{θ_A}, a 3D pose hypothesis is generated by solving

$$\frac{d\mathbf{y}_t}{dt} = v_{\theta_A}(\mathbf{y}_t, t \mid \mathbf{x}), \qquad \mathbf{y}_0 \sim p_0. \tag{12}$$

For H source samples, standard multi-hypothesis inference independently integrates

$$\hat{\mathbf{y}}^{(h)} = \Phi_{0\to 1}\left(\mathbf{y}_0^{(h)}; \mathbf{x}\right), \qquad h = 0, \ldots, H-1, \tag{13}$$

where $\Phi_{s\to t}$ denotes the flow map. With S integration steps per hypothesis, the total number of function evaluations is

$$\mathcal{C}_{\text{full}} = HS. \tag{14}$$

Chordal trajectory branching. Independently integrating multiple hypotheses repeats velocity-field evaluations along their trajectories. CTB reduces this cost by constructing approximate intermediate states from a fully integrated reference trajectory. The linear conditional path in Eq. 2 scales the source-state contribution by $1-t$. Motivated by this property, CTB approximates the displacement between inference trajectories at a branching time t_b.

We first fully integrate a reference trajectory initialized at $\mathbf{y}_0^{(0)}$. For each additional source state $\mathbf{y}_0^{(h)}$, we define

$$\Delta\mathbf{y}_0^{(h)} = \mathbf{y}_0^{(h)} - \mathbf{y}_0^{(0)}. \tag{15}$$

Using the reference state $\mathbf{y}_{t_b}^{(0)}$, we construct the approximate branching state

$$\tilde{\mathbf{y}}_{t_b}^{(h)} = \mathbf{y}_{t_b}^{(0)} + (1 - t_b)\Delta\mathbf{y}_0^{(h)}, \qquad h = 1, \ldots, H-1. \tag{16}$$

This construction requires no additional velocity-field evaluations. It approximates intermediate states of the learned flow rather than assuming that its inference trajectories are exactly linear.

Each additional hypothesis is then integrated only from its branching state to the terminal time:

$$\hat{\mathbf{y}}^{(h)} = \Phi_{t_b \to 1}\left(\tilde{\mathbf{y}}_{t_b}^{(h)}; \mathbf{x}\right), \qquad h = 1, \ldots, H-1. \tag{17}$$

The reference hypothesis is obtained from its fully integrated trajectory.

Inference complexity. Let K denote the number of integration steps after t_b. CTB requires S evaluations for the reference trajectory and K for each additional hypothesis:

$$\mathcal{C}_{\mathrm{CTB}} = S + (H-1)K. \tag{18}$$

With $H = 50$, $S = 20$, and $K = 10$, CTB reduces function evaluations from 1000 to 510, a 49% reduction. The final 3D pose is obtained by averaging the generated hypotheses.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation metrics. We evaluate RECAST on Animal3D (Xu et al., 2023) and CtrlAni3D (Lyu et al., 2025), which contain approximately 3.4K and 9.7K animal instances, respectively. We use 3,065 training and 320 test instances from Animal3D, and 8,277 training and 1,434 test instances from CtrlAni3D. Both datasets use a unified 26-joint pose representation. Following prior work (Wang et al., 2026), we report Procrustes-aligned mean per-joint position error (P-MPJPE, mm) as the primary accuracy metric. We additionally evaluate skeletal consistency using absolute bone-length error and inference cost using the number of function evaluations (NFE).

Implementation details. Stage I pretrains the conditional flow on Human3.6M (Ionescu et al., 2013). Stage II adapts it to animal poses using Linear LoRA ($r = 128$) and GraphLoRA ($r_g = 4$). Stage III resumes from the converged Stage II checkpoint with $\lambda_{\mathrm{bone}} = 0.1$. By default, inference uses $H = 50$ hypotheses and $S = 20$ Euler steps. The final 3D pose is obtained by averaging the hypotheses. CTB uses $t_b = 0.5$.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares RECAST with representative regression-based and probabilistic methods on Animal3D and CtrlAni3D. The probabilistic FMPose3D (Wang et al., 2026) outperforms the compared regression-based methods (Kanazawa et al., 2018; Biggs et al., 2020; Goel et al., 2023; Lyu et al., 2025), achieving 61.5 mm and 44.0 mm P-MPJPE, respectively. This result highlights the potential of probabilistic lifting for 3D animal pose estimation. However, FMPose3D does not leverage human-pretrained lifting knowledge. By reusing human priors through progressive cross-species adaptation, RECAST further reduces P-MPJPE to 29.9 mm on Animal3D and 18.7 mm on CtrlAni3D. The consistent improvements across both benchmarks show the performance gains achieved by RECAST under limited 3D supervision across diverse animal species.

Table 1: **Comparison with state-of-the-art methods.** P-MPJPE is reported in millimeters; lower is better.

Method	Animal3D ↓	CtrlAni3D ↓
HMR (Kanazawa et al., 2018)	123.5	123.5
VLDO (Biggs et al., 2020)	112.3	71.5
HMR2.0 (Goel et al., 2023)	94.1	60.9
AniMer (Lyu et al., 2025)	80.4	44.1
FMPose3D (Wang et al., 2026)	61.5	44.0
RECAST (Ours)	**29.9**	**18.7**

Table 2: **Component and inference ablation.** Rows compare training stages and configurations; columns vary the hypothesis count H or integration-step count S. All results use CTB with $t_b = 0.5$. Sample-weighted P-MPJPE (mm) is evaluated over the combined Animal3D and CtrlAni3D test sets; lower is better. The default setting is $H = 50$, $S = 20$.

Stage	Configuration	Hypothesis sweep ($S = 20$)				Step sweep ($H = 50$)		**Default**
		$H = 1$	$H = 5$	$H = 10$	$H = 20$	$S = 4$	$S = 10$	
NFE		20	60	110	210	102	255	510
Baseline	Animal training from scratch	47.12	37.62	36.20	35.63	37.15	35.50	35.18
	Human initialization	612.86	612.42	612.24	612.20	611.25	611.94	612.11
Stage II	LoRA + GraphLoRA (400 epochs)	30.24	25.36	24.65	24.37	25.41	24.32	24.13
	LoRA + GraphLoRA (800 epochs)	28.83	24.20	23.52	23.23	24.40	23.23	23.00
Stage III	**RECAST**	**24.26**	**21.41**	**21.04**	**20.86**	**21.96**	**20.99**	**20.71**

Table 3: **Training-stage progression and skeletal refinement.** Both Stage III runs use cosine restarts (3×10^{-4}, 400-epoch cycle) and differ in bone-length supervision. P-MPJPE and bone-length error are reported in mm; lower is better.

Stage	Configuration	P-MPJPE ↓			Bone-length error ↓		
		Animal3D	CtrlAni3D	Weighted	Animal3D	CtrlAni3D	Weighted
Stage II	400 epochs	31.95	22.38	24.13	18.70	13.80	14.69
	800 epochs	31.53	21.09	23.00	18.16	12.61	13.62
Stage III	w/o bone loss	30.15	19.59	21.51	17.19	11.39	12.45
	+ bone loss	**29.86**	**18.67**	**20.71**	**16.89**	**10.65**	**11.79**

4.3 Ablation Studies

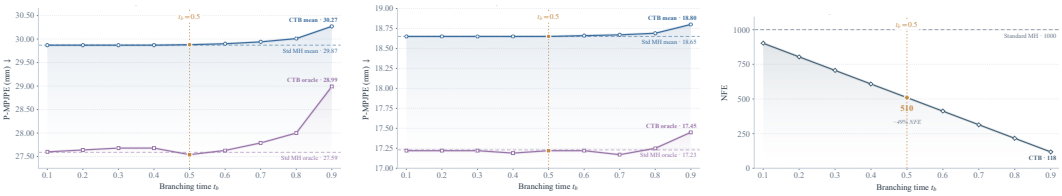
We conduct a series of ablation studies to examine the effectiveness of cross-species adaptation and post-convergence skeletal refinement. We further investigate how the hypothesis count, integration steps, and CTB branching time affect pose accuracy and computational cost.

Training components and inference settings. Table 2 evaluates training configurations and inference settings using CTB with $t_b = 0.5$. At the default setting $(H, S) = (50, 20)$, training from scratch on animal data yields 35.18 mm P-MPJPE, whereas directly applying the human-pretrained model yields 612.11 mm, indicating that human priors require cross-species adaptation. With Linear LoRA and GraphLoRA, Stage II reduces the error to 24.13 mm after 400 epochs and 23.00 mm after 800 epochs. Stage III further improves P-MPJPE to 20.71 mm.

We next examine the inference settings of the final model. With $S = 20$, increasing the hypothesis count from $H = 1$ to $H = 50$ reduces P-MPJPE from 24.26 mm to 20.71 mm, with the largest gain occurring between $H = 1$ and $H = 5$. With $H = 50$, increasing the integration steps from $S = 4$ to $S = 20$ reduces the error from 21.96 mm to 20.71 mm, while NFE increases from 102 to 510.

Skeletal refinement. Table 3 compares different training configurations and evaluates the contribution of species-specific bone-length supervision. Extending Stage II to 800 epochs with a single cosine schedule reduces weighted P-MPJPE from 24.13 mm to 23.00 mm. Resuming training from the converged Stage II checkpoint with cosine restarts achieves 21.51 mm without bone loss. Under the same resumed-training configuration, introducing the bone-length objective further reduces P-MPJPE to 20.71 mm and weighted bone-length error from 12.45 mm to 11.79 mm. These matched results demonstrate that skeletal supervision improves both pose accuracy and bone-length consistency beyond the gains obtained from additional training alone.

CTB branching time. Figure 3 compares CTB with standard multi-hypothesis inference at $H = 50$ and $S = 20$. The oracle curves assess the generated hypothesis sets using ground-truth poses, whereas the mean curves evaluate the averaged final prediction. As branching time increases, NFE decreases, while prediction accuracy remains nearly unchanged for $t_b \in [0.1, 0.5]$. Within this



(a) Animal3D. (b) CtrlAni3D. (c) Computational cost.

Figure 3: **Effect of branching time** t_b ($H = 50$, $S = 20$). (a) and (b) Mean and oracle P-MPJPE on Animal3D and CtrlAni3D; solid curves denote CTB and dashed lines denote standard multi-hypothesis inference. (c) Number of function evaluations (NFE). Orange markers indicate $t_b = 0.5$.

range, CTB matches standard inference at the reported precision on both benchmarks, with NFE decreasing from 902 to 510. We adopt $t_b = 0.5$ as the default to balance accuracy and computational cost, achieving 20.71 mm weighted P-MPJPE with 510 NFE.

4.4 Qualitative Analysis

Flow Trajectory Visualization. Fig. 4(a)–(c) visualizes the coordinate evolution of joint 3 (the left front paw). Unlike standard inference, CTB integrates one reference trajectory and approximates branching states for the remaining hypotheses at step 10, requiring only suffix integration thereafter.

Attention Visualization. Fig. 4(d) shows non-uniform joint-to-joint attention weights in the final block, illustrating variation across joint interactions. Additional examples in the Appendix exhibit distinct attention patterns across samples.

Pose estimation. Fig. 5 presents representative 3D pose estimation results together with the corresponding ground truth, providing a direct qualitative comparison of the estimated skeletal configurations across different animal instances. Additional qualitative results are provided in the Appendix.

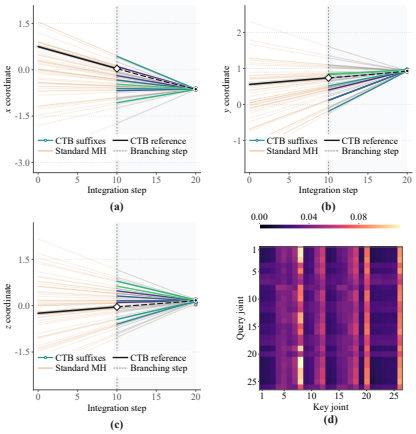


Figure 4: **Visualization of flow trajectories and attention.** (a)–(c) Flow trajectories under standard inference and CTB ($t_b = 0.5$); (d) final-block joint attention.

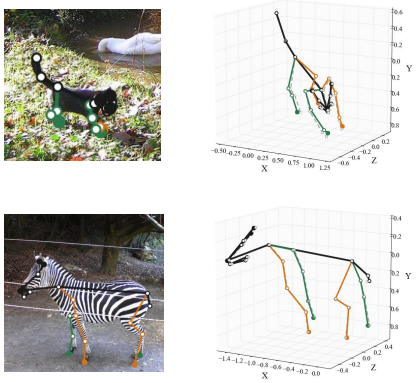


Figure 5: **Qualitative 3D pose estimation results.** Left: input 2D poses. Right: estimated 3D poses (colored solid lines) and ground truth (gray dashed lines).

5 Conclusion

We presented RECAST, an efficient flow-matching framework for data-scarce 3D animal pose estimation that reuses human-learned probabilistic lifting knowledge instead of learning solely from limited animal supervision. Progressive cross-species adaptation and post-convergence skeletal refinement improve pose accuracy and bone-length consistency. For efficient multi-hypothesis inference, Chordal Trajectory Branching (CTB) reuses a reference trajectory to construct approximate branching states and integrates only the remaining hypothesis suffixes. RECAST achieves 29.86 mm and 18.67 mm P-MPJPE on Animal3D and CtrlAni3D, respectively, while CTB reduces function evaluations by 49% with only a 0.01 mm difference in P-MPJPE relative to standard inference. These results demonstrate the effectiveness of human-prior reuse and trajectory reuse for probabilistic 3D animal pose estimation.

AI USE STATEMENT

DeepSeek V4 was used to assist with language polishing and improve the clarity and readability of the manuscript. The tool was not used to generate experimental results or scientific claims. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work does not involve the collection of new human-subject data or the release of new datasets. All experiments are conducted on existing human and animal pose estimation benchmarks. The study focuses on cross-species knowledge transfer, conditional flow modeling, and 3D animal pose estimation. The authors have reviewed the experimental design and reporting to ensure that the results and conclusions are presented accurately and responsibly.

REPRODUCIBILITY STATEMENT

We provide the implementation and experimental details required to reproduce RECAST in the main paper and Appendix. These include the FMPose3D backbone, low-rank adaptation configuration, three-stage training protocol, skeletal refinement objective, and CTB inference procedure. The Appendix further provides dataset and skeletal statistics, theoretical and empirical analyses of CTB, additional Linear LoRA and GraphLoRA rank ablations, and qualitative visualizations. These details are intended to support independent reproduction of the reported results.

REFERENCES

Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2005.

Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2020.

Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *ICCV*, 2019.

Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017.

Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):198–209, 2021.

Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018.

Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 2018.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14737–14748. IEEE, 2023.

Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, 2023.

Mohammed Hassanin, Abdelwahed Khamis, Mohammed Bennamoun, Farid Boussaid, and Ibrahim Radwan. Crossformer3d: cross spatio-temporal transformer for 3d human pose estimation. *Signal, Image and Video Processing*, 19(8):618, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021a.

Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. Conditional directed graph convolution for 3d human pose estimation. In *ACM MM*, 2021b.

Yunlong Huang, Junshuo Liu, Ke Xian, and Robert Caiming Qiu. Posemamba: Monocular 3d human pose estimation with bidirectional global-local spatio-temporal state space model. In *AAAI*, 2025.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 2022a.

Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, 2022b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *CVPR*, 2020.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Liangsi Lu, Xuhang Chen, Minzhe Guo, Shichu Li, Jingchao Wang, and Yang Shi. Chordedit: One-step low-energy transport for image editing. *arXiv preprint arXiv:2602.19083*, 2026.

Ye Lu, Jie Wang, Jianjun Gao, Rui Gong, Chen Cai, and Kim-Hui Yap. A structure-aware and motion-adaptive framework for 3d human pose estimation with mamba. In *ICCV*, pp. 7958–7968, 2025.

Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, and Liang An. Animer: Animal pose and shape estimation using family aware transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17486–17496. IEEE, 2025.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6920–6930, 2024.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

Jihua Peng, Yanghong Zhou, and PY Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *CVPR*, 2024.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14715–14725. IEEE, 2023.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018.

Zhenhua Tang, Jia Li, Yanbin Hao, and Richang Hong. Mlp-jcg: Multi-layer perceptron with joint-coordinate gating for efficient 3d human pose estimation. *IEEE Transactions on Multimedia*, 2023a.

Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *CVPR*, 2023b.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Ti Wang, Xiaohang Yu, and Mackenzie Weygandt Mathis. Fmpose3d: monocular 3d pose estimation via flow matching. *arXiv preprint arXiv:2602.05755*, 2026.

Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9065–9075. IEEE, 2023.

Jinglin Xu, Yijie Guo, and Yuxin Peng. Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 561–570. IEEE, 2024.

Youze Xue, Jiansheng Chen, Xiangming Gu, Huimin Ma, and Hongbing Ma. Boosting monocular 3d human pose estimation with part aware attention. *IEEE Transactions on Image Processing*, 31, 2022.

Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *CVPR*, 2022.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, 2019.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, 2021.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15085–15099, 2023.

Appendix–RECAST: REUSING HUMAN PRIORS FOR DATA-SCARCE 3D ANIMAL POSE ESTIMATION

This appendix includes additional details and analyses of RECAST. Specifically, we provide: (1) dataset statistics and skeletal distributions; (2) the architecture and cross-species adaptation of FMPose3D; (3) theoretical and empirical analyses of CTB; (4) ablations on Linear LoRA and GraphLoRA ranks; and (5) additional qualitative visualizations.

A Dataset Statistics and Skeletal Distributions

We provide additional statistics for Animal3D and CtrlAni3D to characterize the data distribution and skeletal variation involved in cross-species adaptation.

Dataset composition. Animal3D contains 3,065 training and 320 test instances, while CtrlAni3D contains 8,277 training and 1,434 test instances. Both datasets use a unified 26-joint pose representation. Figure 6 shows the species-wise sample distribution of Animal3D, highlighting the limited and unevenly distributed animal annotations. This data scarcity motivates reusing human-learned lifting knowledge for cross-species adaptation.

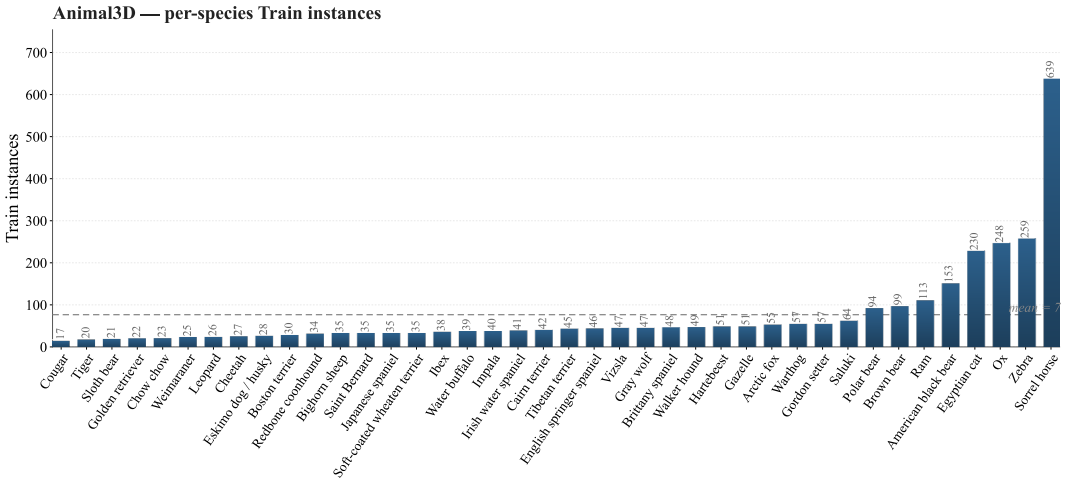


Figure 6: **Species distribution in Animal3D.** Number of samples per species.

Skeletal distributions. Figure 7 illustrates the variation in mean bone length across Animal3D species. For species s, we compute the average length over all bones and ground-truth 3D poses:

$$\bar{\ell}^{(s)} = \frac{1}{N_s|\mathcal{B}|} \sum_{n=1}^{N_s} \sum_{(i,j)\in\mathcal{B}} \left\| \mathbf{y}_{n,i}^{(s)} - \mathbf{y}_{n,j}^{(s)} \right\|_2, \tag{19}$$

where N_s is the number of instances of species s and $\mathcal{B}$ denotes the bone set. The species-dependent bone lengths show differences in skeletal scale across species despite the shared joint representation.

B FMPose3D Architecture

Backbone architecture. RECAST builds upon the conditional flow-matching backbone of FMPose3D (Wang et al., 2026). As illustrated in Fig. 8, the network takes an intermediate 3D skeleton $\mathbf{y}_t$, a conditioning 2D pose $\mathbf{x}$, and flow time t as inputs. Their embeddings are fused and processed by a stack of blocks combining graph convolution and attention. The resulting features are passed through a regression head to predict the conditional velocity field.

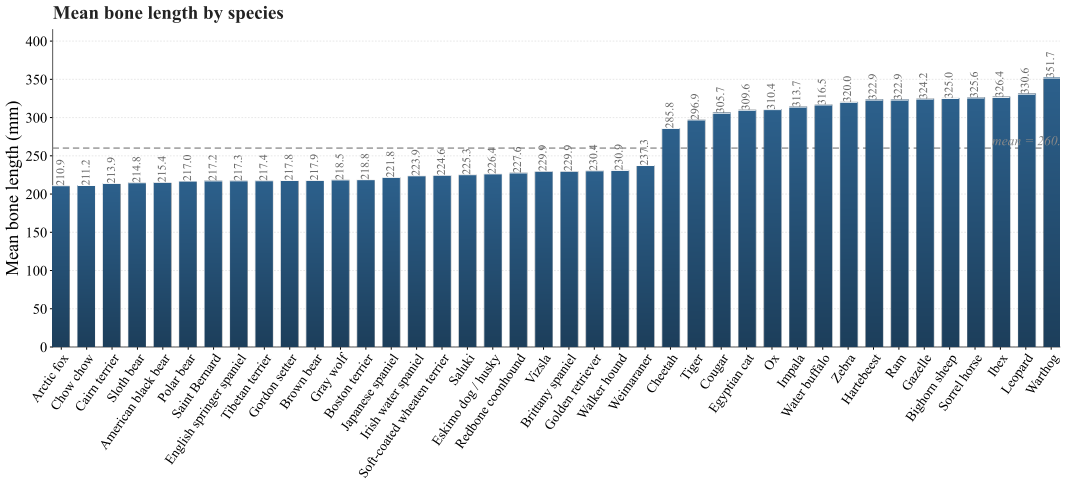


Figure 7: **Skeletal distribution in Animal3D.** Mean bone length per species.

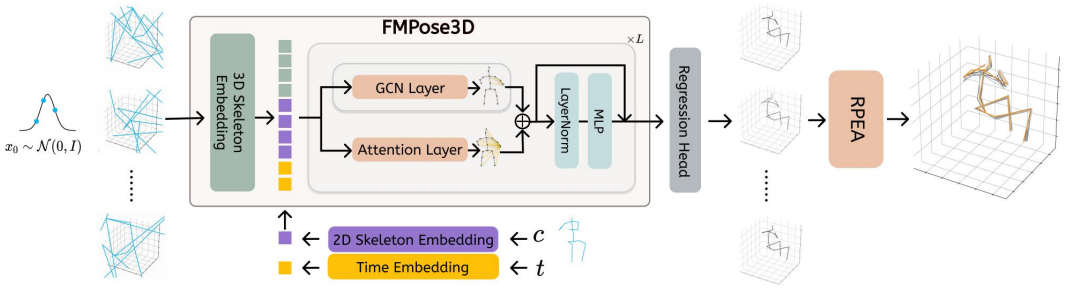


Figure 8: **FMPose3D backbone.** The network combines 3D skeleton, 2D pose, and time embeddings to predict the conditional velocity field.

Low-rank adaptation. During Stage II, RECAST freezes the pretrained human backbone and optimizes animal-specific adaptation parameters. Linear LoRA is applied to 28 linear transformations in the attention and feed-forward components, with rank $r = 128$ and scaling $\alpha = 128$. GraphLoRA is applied to five graph-convolution blocks, with rank $r_g = 4$, to adapt joint dependencies to animal morphology.

The backbone contains approximately 16.3M parameters, while 4.4M parameters are optimized during animal adaptation. Parameters with compatible dimensions are initialized from the human checkpoint; components affected by the change in joint configuration are initialized for the animal domain.

Training protocol. Stage I pretrains the conditional flow on Human3.6M using the flow-matching objective. Stage II adapts the pretrained model to animal poses using Linear LoRA and GraphLoRA.

Stage III resumes training from the converged Stage II checkpoint and introduces species-specific bone-length supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}}^{A} + \lambda_{\mathrm{bone}} \mathcal{L}_{\mathrm{bone}}, \tag{20}$$

where $\lambda_{\mathrm{bone}} = 0.1$. The resumed training uses cosine restarts with an initial learning rate of 3×10^{-4} and a 400-epoch cycle.

C Analysis of CTB

We analyze Chordal Trajectory Branching (CTB) relative to standard Euler inference under the same learned velocity field, source states, and time discretization. In standard inference, each hypothesis

is integrated from its source state over the complete trajectory. For a pose tensor, $\|\cdot\|_2$ denotes the Euclidean norm after vectorizing all 26×3 coordinates.

Branching approximation. Let the Euler step size be $\Delta t = 1/S$, and let the branching time lie on the integration grid:

$$t_b = \frac{b}{S}, \qquad K = S - b, \tag{21}$$

where b is the number of prefix steps and K is the number of remaining suffix steps. Let $\mathbf{y}_{t_b}^{(h)}$ denote the state obtained by standard Euler integration of hypothesis h from $\mathbf{y}_0^{(h)}$ to t_b. Following Eq. 16, CTB constructs

$$\tilde{\mathbf{y}}_{t_b}^{(h)} = \mathbf{y}_{t_b}^{(0)} + (1 - t_b)\left(\mathbf{y}_0^{(h)} - \mathbf{y}_0^{(0)}\right). \tag{22}$$

For the linear probability path used in flow-matching training, $\mathbf{y}_t = (1-t)\mathbf{y}_0 + t\mathbf{y}_1$, the contribution of the source state is scaled by $1 - t$. Equation 22 applies the same scaling to the source-state displacement relative to the reference trajectory. For the learned velocity field, the resulting state is an approximation to the state obtained by standard Euler integration. We define the branching error as

$$\epsilon_b^{(h)} = \left\|\tilde{\mathbf{y}}_{t_b}^{(h)} - \mathbf{y}_{t_b}^{(h)}\right\|_2. \tag{23}$$

Error propagation. Assume that the learned velocity field v_θ is L-Lipschitz with respect to the state over the region traversed by the considered inference trajectories:

$$\|v_\theta(\mathbf{y}, t \mid \mathbf{x}) - v_\theta(\mathbf{z}, t \mid \mathbf{x})\|_2 \le L\,\|\mathbf{y} - \mathbf{z}\|_2. \tag{24}$$

Let $\mathbf{y}_k^{(h)}$ and $\tilde{\mathbf{y}}_k^{(h)}$ denote the standard and CTB states, respectively, after k suffix steps from t_b. Since both use the same Euler update after branching,

$$\begin{aligned}\left\|\tilde{\mathbf{y}}_{k+1}^{(h)} - \mathbf{y}_{k+1}^{(h)}\right\|_2 &\le \left\|\tilde{\mathbf{y}}_k^{(h)} - \mathbf{y}_k^{(h)}\right\|_2 \\ &\quad + \Delta t \left\|v_\theta(\tilde{\mathbf{y}}_k^{(h)}, t_b + k\Delta t \mid \mathbf{x}) - v_\theta(\mathbf{y}_k^{(h)}, t_b + k\Delta t \mid \mathbf{x})\right\|_2 \\ &\le (1 + L\Delta t)\left\|\tilde{\mathbf{y}}_k^{(h)} - \mathbf{y}_k^{(h)}\right\|_2.\end{aligned} \tag{25}$$

Applying this relation recursively over the remaining K steps gives

$$\left\|\tilde{\mathbf{y}}_1^{(h)} - \mathbf{y}_1^{(h)}\right\|_2 \le (1 + L\Delta t)^K \epsilon_b^{(h)} \le \exp(LK\Delta t)\epsilon_b^{(h)}. \tag{26}$$

Here, $\mathbf{y}_1^{(h)}$ is the output of standard Euler inference rather than the exact solution of the continuous ODE. Equation 26 therefore bounds the additional terminal discrepancy of CTB relative to standard Euler inference.

Averaged prediction. The final pose is obtained by averaging the generated hypotheses. Let

$$\bar{\mathbf{y}}_1 = \frac{1}{H}\sum_{h=0}^{H-1} \mathbf{y}_1^{(h)}, \qquad \tilde{\bar{\mathbf{y}}}_1 = \frac{1}{H}\sum_{h=0}^{H-1} \tilde{\mathbf{y}}_1^{(h)} \tag{27}$$

denote the standard and CTB outputs, respectively. Since the reference hypothesis is fully integrated in both cases, $\tilde{\mathbf{y}}_1^{(0)} = \mathbf{y}_1^{(0)}$. Using Eq. 26 and the triangle inequality gives

$$\left\|\tilde{\bar{\mathbf{y}}}_1 - \bar{\mathbf{y}}_1\right\|_2 \le \frac{(1 + L\Delta t)^K}{H}\sum_{h=1}^{H-1} \epsilon_b^{(h)}. \tag{28}$$

Thus, the discrepancy between the two averaged outputs is bounded in terms of the branching errors and the subsequent Euler updates.

Computational complexity. Standard multi-hypothesis inference requires

$$\mathcal{C}_{\text{std}} = HS. \tag{29}$$

CTB fully integrates one reference trajectory and performs only the remaining K steps for each of the other $H-1$ hypotheses:

$$\mathcal{C}_{\text{CTB}} = S + (H-1)K. \tag{30}$$

The relative reduction is

$$\eta = 1 - \frac{\mathcal{C}_{\text{CTB}}}{\mathcal{C}_{\text{std}}} = \frac{(H-1)(S-K)}{HS}. \tag{31}$$

For $H = 50$, $S = 20$, and $b = 10$, we have $t_b = 0.5$ and $K = 10$. The number of velocity-field evaluations is therefore reduced from 1000 to 510, corresponding to a 49% reduction.

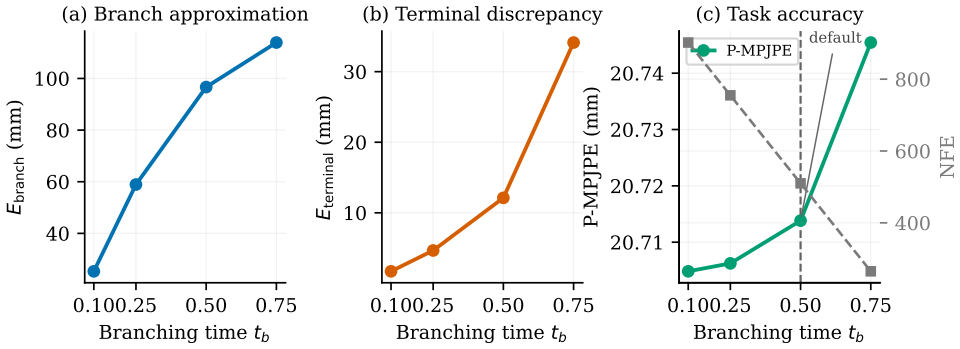


Figure 9: **Approximation–accuracy–efficiency analysis of CTB.** (a) Branching-state discrepancy relative to standard Euler inference. (b) Terminal discrepancy after suffix integration. (c) Weighted P-MPJPE and NFE at different branching times. All results use $H = 50$ and $S = 20$; the dashed line marks the default $t_b = 0.5$.

Empirical analysis. Figure 9 measures the approximation introduced by CTB relative to standard Euler inference. For N test samples, we report

$$E_{\text{branch}}(t_b) = \frac{1}{N(H-1)} \sum_{n=1}^{N} \sum_{h=1}^{H-1} \left\| \tilde{\mathbf{y}}^{(h)}_{t_b,n} - \mathbf{y}^{(h)}_{t_b,n} \right\|_2, \tag{32}$$

and define $E_{\text{terminal}}(t_b)$ analogously using the corresponding terminal states. These discrepancies are measured before the deterministic root-joint normalization used for evaluation. P-MPJPE is computed after averaging all hypotheses and applying the standard evaluation protocol.

As t_b increases from 0.1 to 0.75, E_{branch} increases from 25.31 mm to 113.91 mm, while E_{terminal} increases from 1.71 mm to 34.13 mm. At the default $t_b = 0.5$, the branching and terminal discrepancies are 96.67 mm and 12.12 mm, respectively. Thus, under the evaluated model and inference settings, the discrepancy measured at the terminal states is smaller than that at the branching states.

Although $t_b = 0.1$ yields a marginally lower weighted P-MPJPE of 20.7048 mm, it reduces NFE by only 9.8% relative to standard inference. At $t_b = 0.5$, CTB achieves a comparable 20.71 mm while reducing NFE by 49%. Increasing t_b to 0.75 further reduces NFE to 265, while the terminal discrepancy increases to 34.13 mm and weighted P-MPJPE to 20.74 mm. We therefore use $t_b = 0.5$ as the default setting.

D Additional Ablation Studies

Low-rank adaptation. We investigate the effects of Linear LoRA and GraphLoRA ranks on cross-species adaptation. For each sweep, we vary one rank while fixing the other at its default value. All results are evaluated using the Stage II model after 400 epochs.

Table 4: **Effect of low-rank adaptation ranks.** We vary the Linear LoRA rank r or GraphLoRA rank r_g while fixing the other at its default. P-MPJPE is reported in mm; lower is better.

Sweep	r	r_g	Animal3D	CtrlAni3D	Weighted
Linear LoRA	8	4	40.06	29.08	31.09
	32	4	33.28	23.69	25.44
	64	4	32.33	22.32	24.15
	128	4	31.95	22.38	**24.13**
GraphLoRA	128	1	32.75	23.01	24.78
	128	2	33.12	23.09	24.92
	128	4	31.95	22.38	**24.13**

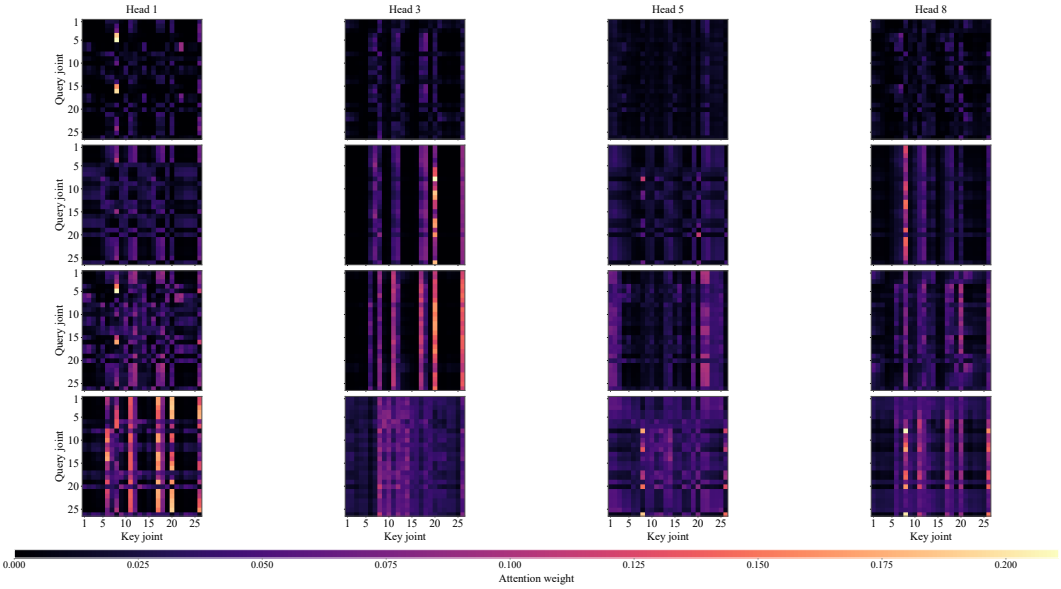


Figure 10: **Attention maps.** Joint-to-joint attention weights in the final block.

Increasing the Linear LoRA rank from 8 to 128 reduces weighted P-MPJPE from 31.09 mm to 24.13 mm, with diminishing improvements beyond $r = 64$. For GraphLoRA, $r_g = 4$ achieves the lowest weighted P-MPJPE of 24.13 mm among the evaluated ranks. We therefore adopt $r = 128$ and $r_g = 4$ as the default configuration.

E Additional Visualizations

We provide additional attention visualizations and 3D pose estimation results to complement the qualitative analysis in the main paper.

Attention visualization. Fig. 10 presents additional joint-to-joint attention maps from the final block. These examples illustrate the variation in attention weights across joint interactions.

3D pose estimation. Fig. 11 presents additional 3D pose estimation results together with the corresponding ground truth. The examples cover different animal poses and articulation patterns.

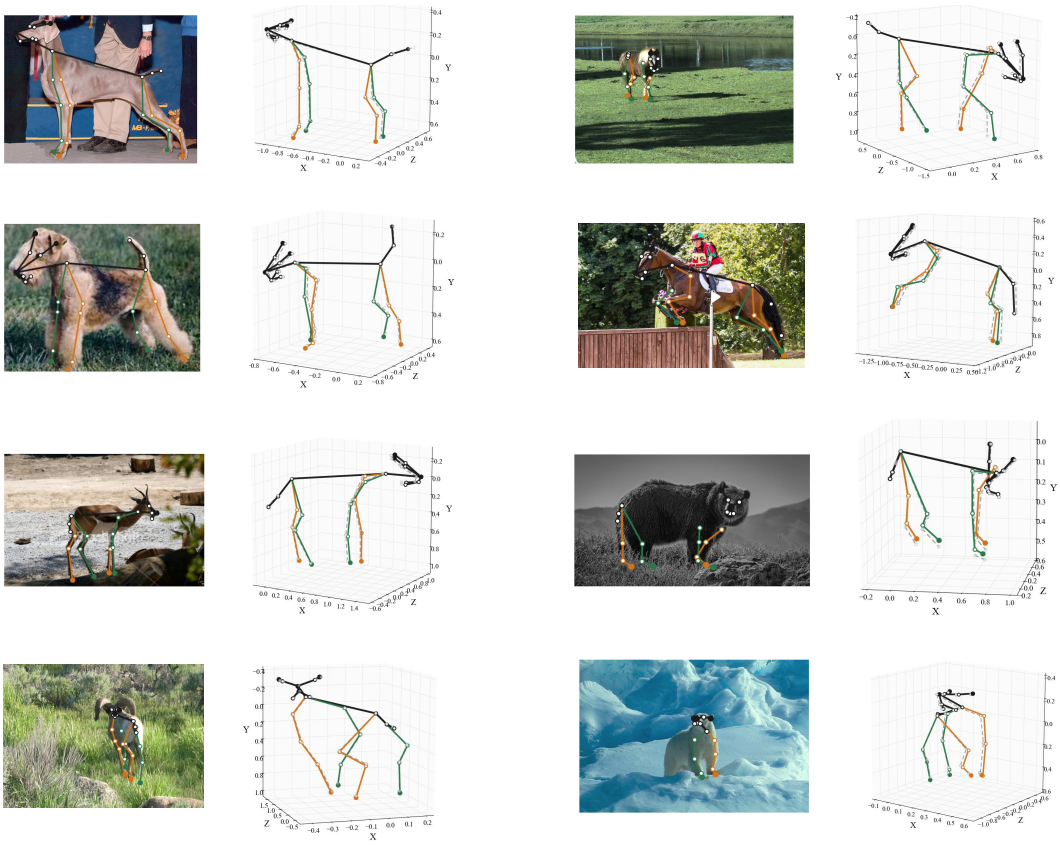


Figure 11: **Additional 3D pose results.** Predicted 3D skeletons are shown alongside ground-truth poses.