MoVi: Motion-Centric 4D Dog Reconstruction from Monocular Videos

Anonymous authors
Paper under double-blind review

Abstract

We present **MoVi**, a motion-centric framework for reconstructing 4D dog motion from unconstrained monocular RGB videos for quadruped imitation learning. Existing animal motion acquisition pipelines typically depend on controlled capture setups, specialized sensing, or calibrated recording environments, which limits the scale and diversity of motion data available for downstream learning. MoVi instead targets readily available in-the-wild videos and converts them into retargeting-friendly motion representations. Our framework first estimates per-frame 3D skeletons with a flow-matching-based pose model, then refines them using reprojection constraints and temporal modeling to improve consistency under monocular ambiguity and occlusion. We further introduce a trajectory predictor that recovers global root motion from articulated motion and temporal cues, producing spatially coherent 4D motion rather than appearance-oriented reconstruction. By focusing on articulated pose and global trajectory, MoVi provides smoother and more task-relevant motion representations for quadruped retargeting and imitation learning, offering a scalable way to leverage large collections of ordinary monocular videos as training data. Code and project materials are available at `https://anonymous.4open.science/r/MoVi`.

1 Introduction

Reference-motion learning has become a powerful route to natural and agile quadruped control. Physics-based imitation methods can track complex motion clips and acquire robust locomotion skills (Peng et al., 2018; 2021; 2020). More recent systems further demonstrate that animal motion and visual demonstrations can be transferred to physical quadruped robots (Yoon et al., 2025; Zhao et al., 2025). However, the quality and diversity of the resulting behaviors remain strongly constrained by the motion sources available for training.

In contrast, monocular dog videos are abundant. Casual phone recordings and Internet videos contain a much broader range of gaits, interactions, viewpoints, and environments than laboratory motion databases. Turning such videos into useful robot-learning data is challenging because the desired signal is not only a locally accurate pose, but also a temporally coherent motion sequence with meaningful global root displacement. Monocular observations introduce depth and scale ambiguity, occlusions can cause abrupt pose inconsistencies, and moving cameras entangle apparent image motion with true animal displacement. These ambiguities make frame-wise root-relative reconstruction insufficient for downstream imitation learning.

MoVi addresses this problem from a motion-centric perspective. As illustrated in Fig. 1, our pipeline first performs 2D-to-3D Pose Estimation (FMPose3D) to generate per-frame 3D pose hypotheses from monocular observations. These predictions are then processed by a Triggered Temporal & Joint Attention Refiner, which selectively corrects unreliable joints using temporal and inter-joint context while preserving confident observations. The refined sequence is transformed through Scale-Aware Canonicalization, separating motion dynamics from subject-specific body scale. Finally, the Trajectory Prediction module recovers global root motion using a Motion-prior Anchor Branch and a Geometry-Camera Residual Branch. Their predictions are combined through Reliability Gated Fusion, allowing view-dependent geometric evidence to correct the motion anchor only when it is expected to be beneficial.

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.

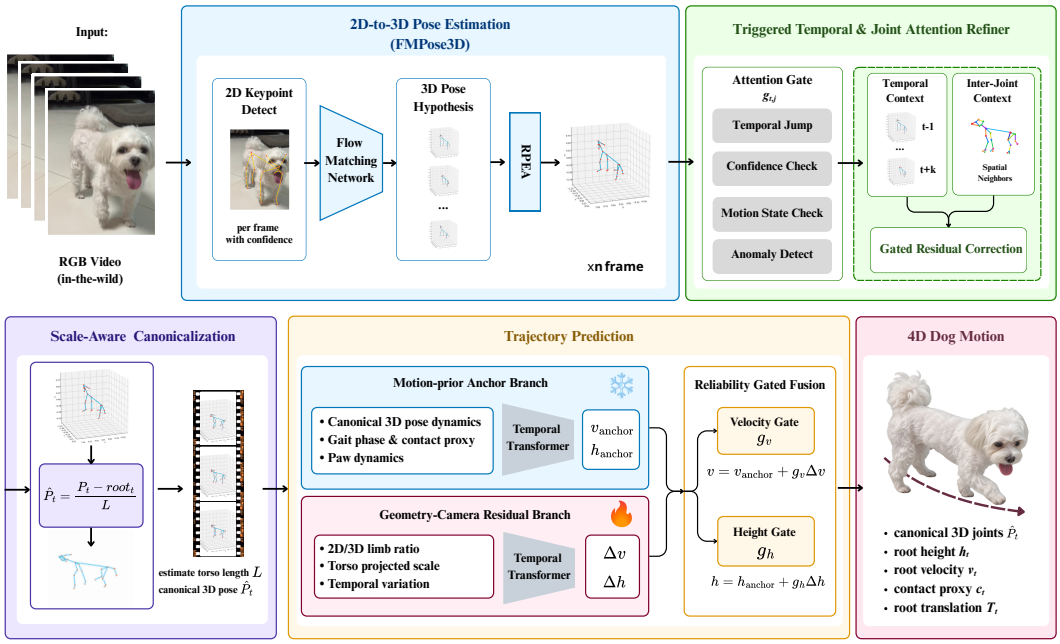

Figure 1: **Overview of MoVi.** MoVi estimates 3D dog poses from monocular video, refines unreliable joints with triggered temporal–joint attention, canonicalizes the motion, and recovers global root motion using motion-prior and geometry-camera branches with reliability-gated fusion.

The resulting representation focuses on the quantities most relevant to quadruped retargeting: canonical 3D joint motion, root height, root velocity, contact cues, and global root translation. By separating canonical articulated motion from trajectory recovery, MoVi converts unconstrained monocular videos into motion representations that are better suited for downstream imitation learning.

Our contributions are:

- We formulate monocular canine reconstruction around the motion representation required for quadruped retargeting, emphasizing temporally coherent articulated pose and global root trajectory rather than appearance reconstruction.
- We introduce a Triggered Temporal & Joint Attention Refiner that improves robustness to occlusion and temporal discontinuities while largely preserving reliable pose estimates.
- We propose a scale-aware Trajectory Prediction framework that combines a Motion-prior Anchor Branch, a Geometry-Camera Residual Branch, and Reliability Gated Fusion to recover stable global root motion under monocular viewpoint and camera-motion ambiguity.

2 Related Work

2.1 Animal Reconstruction and Motion Representation

Parametric animal models such as SMAL and subsequent learning-based methods recover animal pose, shape, and articulated appearance from images or videos (Zuffi et al., 2017; Biggs et al., 2018; Wu et al., 2023; Li et al., 2024; Sabathier et al., 2024; Lyu et al., 2025). Datasets such as Animal3D and 3DDogs further support articulated animal reconstruction (Xu et al., 2023; Shooter et al., 2024). Generative pose estimators preserve multiple plausible 3D hypotheses under monocular ambiguity (Gong et al., 2023; Shan et al., 2023); MoVi uses the flow-matching-based FMPose3D front-end (Wang et al., 2026) and further refines unreliable predictions temporally. Unlike appearance-oriented reconstruction, MoVi focuses on a compact motion representation with temporally coherent pose and global root displacement.

2.2 Global Trajectory Recovery from Monocular Video

Recovering global motion requires separating subject motion from camera-induced image motion. SfM and monocular SLAM estimate camera motion and scene geometry from multi-view correspondences (Schönberger & Frahm, 2016; Campos et al., 2021), while dynamic scenes often require additional handling of moving foregrounds (Bescós et al., 2018). Human-motion methods such as GLAMR, WHAM, and TRAM combine articulated-motion priors with camera or scene reasoning to recover global trajectories (Yuan et al., 2022; Shin et al., 2024; Wang et al., 2024). MoVi follows this direction but directly predicts animal root motion through a Motion-Prior Anchor Branch and a Geometry-Camera Residual Branch, without requiring explicit camera extrinsics.

2.3 Quadruped Imitation and Retargeting

Animal motion and visual demonstrations have been used to learn and retarget quadruped skills (Peng et al., 2020; Yoon et al., 2025; Zhao et al., 2025; Stamatopoulou et al., 2025). These works motivate scalable extraction of reference motion from ordinary videos. MoVi addresses this upstream problem by recovering articulated motion together with global root trajectories suitable for later retargeting and imitation.

3 Method

3.1 Problem Formulation and Pipeline

Given a monocular RGB sequence $I_{1:T}$, our goal is to recover a 4D skeleton motion

$$\mathcal{M} = \left\{ \tilde{\mathbf{P}}_t,\, \hat{h}_t,\, \hat{\mathbf{v}}_t,\, \mathbf{T}_t,\, \mathbf{c}_t \right\}_{t=1}^{T}, \tag{1}$$

where $\tilde{\mathbf{P}}_t \in \mathbb{R}^{J\times 3}$ denotes the canonical root-relative articulated pose, $\hat{\mathbf{v}}_t$ is the torso-normalized root velocity, $\hat{h}_t$ is normalized root height, $\mathbf{T}_t$ is the integrated relative root translation, and $\mathbf{c}_t$ denotes contact-related cues.

The overall pipeline is illustrated in Fig. 1. MoVi first performs 2D-to-3D Pose Estimation using FMPose3D, followed by reprojection-based hypothesis aggregation and Triggered Temporal & Joint Attention refinement. The refined sequence is canonicalized and passed to the Trajectory Prediction module, which combines a Motion-Prior Anchor Branch and a Geometry-Camera Residual Branch through Reliability Gated Fusion.

3.2 2D-to-3D Pose Estimation

For frame t, the detector provides 2D keypoints $\mathbf{x}_t^{2D} \in \mathbb{R}^{J\times 2}$ and confidence scores $\mathbf{q}_t$. The flow-matching pose model starts from $\mathbf{z}_0 \sim \mathcal{N}(0, I)$ and learns

$$\frac{d\mathbf{z}}{d\tau} = u_\theta\left(\mathbf{z}, \tau \mid \mathbf{x}_t^{2D}, \mathbf{q}_t\right), \tag{2}$$

which transports noise samples into plausible 3D pose hypotheses.

Multiple hypotheses represent monocular depth ambiguity and are aggregated according to reprojection consistency. Confidence and reprojection statistics are retained as reliability cues for later temporal refinement and trajectory prediction.

3.3 Triggered Temporal & Joint Attention Refiner

Frame-wise pose estimates may become inconsistent under occlusion, low-confidence detection, or depth-mode switching. We therefore use a triggered refiner that only modifies unreliable joint-frame observations.

For each joint j at frame t, an attention gate $g_{t,j}^{\mathrm{attn}}$ is estimated from temporal discontinuity, confidence, motion consistency, and anomaly statistics. Temporal context is collected from nearby

reliable observations, while inter-joint context is obtained from structurally informative neighboring joints:

$$\mathbf{P}'_{t,j} = (1 - g^{\text{attn}}_{t,j})\mathbf{P}_{t,j} + g^{\text{attn}}_{t,j}\mathbf{P}^{\text{context}}_{t,j}. \tag{3}$$

A conservative motion-prior gate further suppresses implausible acceleration spikes.

3.4 Scale-Aware Canonicalization

To separate body size from motion, we normalize the root-relative pose by a robust sequence-level torso length L:

$$\tilde{\mathbf{P}}_t = \frac{\mathbf{P}_t - \mathbf{P}^{\text{root}}_t}{L}. \tag{4}$$

The canonical body axes follow the animal anatomy: x is approximately forward, y is lateral, and z is upward. The first valid frame fixes the sequence orientation gauge and the initial horizontal root position is set to zero.

The predicted root velocity is expressed in this torso-normalized animal-centered representation, $\hat{\mathbf{v}}_t = (v^{\text{forward}}_t, v^{\text{lateral}}_t, v^{\text{vertical}}_t)$, with units of torso-length/s. Root height is the root-to-ground distance normalized by the same torso length.

The canonical trajectory is obtained by

$$\mathbf{T}^{\text{canonical}}_t = \mathbf{T}_0 + \sum_{\tau \le t} \hat{\mathbf{v}}_\tau \Delta t_\tau. \tag{5}$$

A target torso scale s converts canonical motion to metric-equivalent motion:

$$\mathbf{T}^{\text{metric}}_t = s\,\mathbf{T}^{\text{canonical}}_t, \qquad h^{\text{metric}}_t = s\,\hat{h}_t. \tag{6}$$

3.5 Trajectory Prediction

Motion-Prior Anchor Branch. The Motion-Prior Anchor Branch receives canonical pose dynamics, pose differences, gait phase, contact cues, paw dynamics, and reliability statistics. A temporal Transformer produces $\mathbf{h}^m_t$ and predicts anchor velocity, root height, and contact probability:

$$\mathbf{v}^{\text{anchor}}_t = f_v(\mathbf{h}^m_t), \qquad h^{\text{anchor}}_t = f_h(\mathbf{h}^m_t), \qquad \mathbf{c}_t = \sigma(f_c(\mathbf{h}^m_t)). \tag{7}$$

This branch captures pose-dynamics–gait–translation relationships from canonical mocap motion.

Geometry-Camera Residual Branch. The Geometry-Camera Residual Branch encodes 2D/3D limb ratios, projected torso scale, foreshortening, limb orientation, confidence, and their temporal variation. A second temporal Transformer produces a geometry representation $\mathbf{h}^g_t$ and camera/view latent $\mathbf{z}^{\text{cam}}_t$, which are fused with the motion feature:

$$\mathbf{h}^f_t = E_{\text{fusion}}\left([\mathbf{h}^m_t, \mathbf{h}^g_t]\right). \tag{8}$$

The fused feature predicts residual proposals $\Delta\mathbf{v}^{\text{proposal}}_t$ and $\Delta h^{\text{proposal}}_t$.

Reliability Gated Fusion. A Reliability Module determines whether these residuals should be applied. Separate gates are predicted for velocity and height:

$$g^v_t = \sigma(f^v_{\text{rel}}(\cdot)), \qquad g^h_t = \sigma(f^h_{\text{rel}}(\cdot)). \tag{9}$$

The final outputs are

$$\hat{\mathbf{v}}_t = \mathbf{v}^{\text{anchor}}_t + g^v_t \Delta\mathbf{v}^{\text{proposal}}_t, \qquad \hat{h}_t = h^{\text{anchor}}_t + g^h_t \Delta h^{\text{proposal}}_t. \tag{10}$$

This preserves a stable motion anchor while allowing view-dependent geometry to correct trajectory estimates when needed.

3.6 Training Strategy

Training consists of two stages, as shown in Fig. 2.

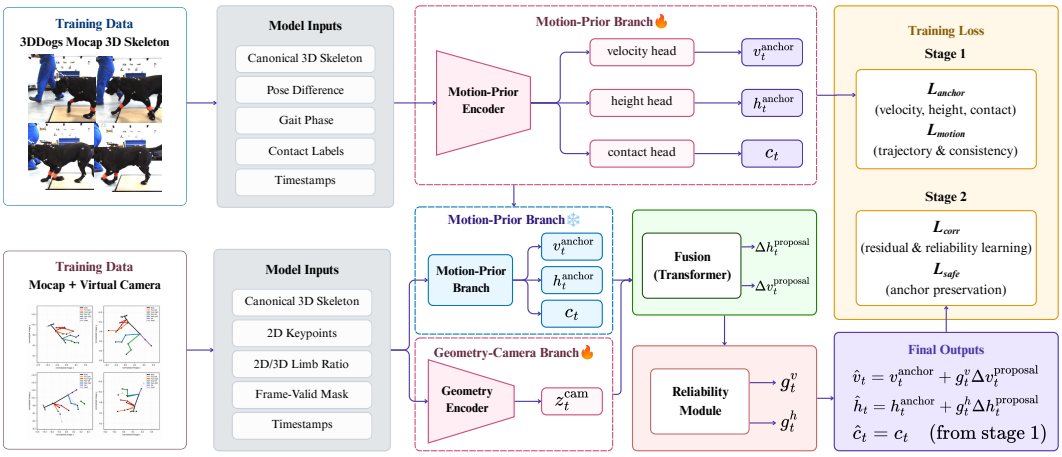

Figure 2: **Training strategy of MoVi.** Stage 1 trains the Motion-Prior Branch and anchor heads using canonicalized 3DDogs mocap. Stage 2 freezes the motion prior and trains the Geometry-Camera Branch, Fusion module, and Reliability Module with mocap and virtual-camera augmentation.

Stage 1: Motion-Anchor Learning. We train the Motion-Prior Encoder and the velocity, height, and contact heads using canonicalized 3DDogs mocap. The objective is

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{anchor}} + \lambda_m \mathcal{L}_{\text{motion}}, \tag{11}$$

where $\mathcal{L}_{\text{anchor}}$ supervises root velocity, height, and contact, while $\mathcal{L}_{\text{motion}}$ constrains integrated trajectory and temporal consistency.

Stage 2: Geometry and Reliability Learning. The Motion-Prior Branch is then frozen. We train the Geometry-Camera Branch, Fusion module, and Reliability Module using canonical mocap and virtual-camera projections under fixed, follow, orbit, and handheld camera motion.

The Stage 2 objective is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{corr}} + \lambda_s \mathcal{L}_{\text{safe}}, \tag{12}$$

where $\mathcal{L}_{\text{corr}}$ supervises residual prediction and reliability gating, and $\mathcal{L}_{\text{safe}}$ discourages corrections that degrade an accurate anchor. The residual target is defined relative to the frozen anchor, so the Geometry-Camera Branch learns to correct rather than replace the motion prior.

4 Experiments

4.1 Evaluation Protocol

We evaluate MoVi from three complementary perspectives. First, we evaluate 3D motion reconstruction and robustness to occlusion on the Animal4D dog subset. Second, we ablate the major components of the temporal refinement and trajectory prediction modules. Third, we evaluate whether recovering global displacement produces more useful reference motions for downstream Go2 retargeting than root-relative motion without trajectory recovery.

For Animal4D, we compare MoVi with FMPose3D, AniMer, and SMALify using PCK@0.1, PCK@0.05, MPJVE, and mean error. Since the public Animal4D dog subset does not provide reliable metric root-trajectory ground truth, metric trajectory evaluation is conducted separately using mocap-derived clips with controlled virtual cameras.

To evaluate robustness under imperfect observations, we additionally construct an occluded version of the Animal4D dog subset by inserting synthetic visual occluders while keeping the underlying motion unchanged.

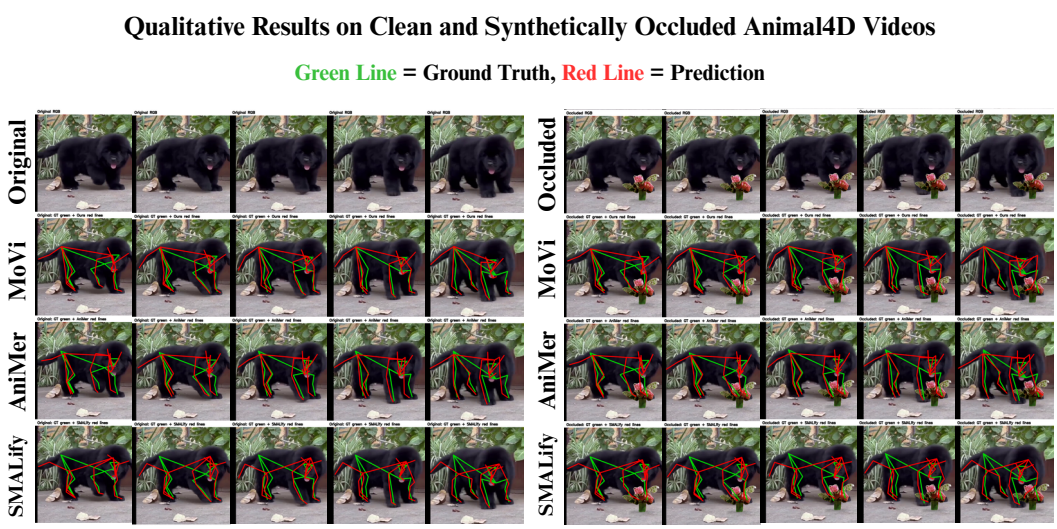

Figure 3: **Qualitative results on clean and synthetically occluded Animal4D videos.** We compare MoVi with AniMer and SMALify across multiple frames. Green lines denote ground-truth skeletons and red lines denote predictions.

Table 1: Comparison on the Animal4D dog subset under clean and synthetically occluded settings. Higher is better for PCK, while lower is better for MPJVE and mean error.

Method	Clean				Occluded			
	PCK@0.1↑	PCK@0.05↑	MPJVE↓	Mean Error↓	PCK@0.1↑	PCK@0.05↑	MPJVE↓	Mean Error↓
Ours: MoVi	**0.762**	**0.392**	14.64	19.87	**0.555**	0.300	**36.92**	**51.15**
FMPose3D	0.761	0.391	14.70	**19.83**	0.550	**0.305**	50.62	58.12
AniMer	0.732	0.329	**12.28**	20.65	0.259	0.085	45.78	64.76
SMALify	0.471	0.197	19.76	39.95	0.354	0.173	75.89	91.55

4.2 Animal4D Dog Subset

Table 1 compares MoVi with existing animal pose and reconstruction methods under clean and synthetically occluded inputs. On clean videos, MoVi retains the strong spatial accuracy of its pose front-end, achieving the highest PCK@0.1 and PCK@0.05 while remaining comparable in motion error. This is consistent with the design of the Triggered Temporal & Joint Attention Refiner, which preserves reliable observations rather than uniformly smoothing the sequence.

The benefit becomes clearer under occlusion. MoVi reduces MPJVE from 50.62 for FMPose3D to 36.92 and mean error from 58.12 to 51.15, while maintaining competitive PCK. Fig. 3 further shows that the refined sequence preserves more temporally coherent articulated motion when local observations are corrupted. Together, these results indicate that MoVi preserves the accuracy of the underlying pose reconstruction while improving temporal robustness under occlusion.

4.3 Ablation Study

We evaluate the contribution of the major components in Fig. 1 and Fig. 2. As shown in Table 2, removing either temporal or inter-joint context degrades trajectory quality, with temporal context having the larger impact. The Motion-Prior Anchor Branch and Geometry-Camera Residual Branch are also complementary: removing either one substantially increases Root RMSE and trajectory error.

Reliability Gated Fusion further improves stability by preventing unnecessary residual corrections, while the explicit 2D/3D limb-ratio cue provides useful view-dependent geometry. The largest degradation occurs without virtual-camera augmentation, confirming that controlled camera variation is important for robust trajectory recovery.

Table 2: Ablation study of the major components in MoVi.

Variant	MPJPE↓	Root RMSE↓	Traj. RMSE↓	Endpoint↓	Direction Err. (°)↓
Full MoVi	0.5438 ± 0.0293	$\mathbf{0.3797}$	$\mathbf{0.1620 \pm 0.0494}$	$\mathbf{0.3803 \pm 0.1036}$	$\mathbf{2.82 \pm 1.29}$
w/o Temporal Context	0.5123 ± 0.0116	0.9669	0.7658 ± 0.0674	1.3609 ± 0.1812	22.97 ± 2.40
w/o Inter-Joint Context	$\mathbf{0.4593 \pm 0.0009}$	0.4852	0.3055 ± 0.1586	0.6420 ± 0.3847	6.20 ± 3.03
w/o Motion-Prior Anchor Branch	0.5228 ± 0.0057	0.6721	0.5014 ± 0.2427	0.7066 ± 0.2388	11.48 ± 6.11
w/o Geometry-Camera Residual Branch	0.5726 ± 0.0480	0.7000	0.5142 ± 0.3469	0.8991 ± 0.4260	11.86 ± 10.06
w/o 2D/3D Limb Ratio	0.5300 ± 0.0331	0.7355	0.5831 ± 0.0119	1.0455 ± 0.0946	14.04 ± 2.74
w/o Residual Correction	0.5231 ± 0.0315	0.8277	0.6458 ± 0.0373	0.9955 ± 0.2724	16.61 ± 1.34
w/o Reliability Gated Fusion	0.5652 ± 0.0392	0.4913	0.3290 ± 0.0711	0.5158 ± 0.0815	6.78 ± 2.37
w/o Virtual-Camera Augmentation	0.6888 ± 0.0199	1.2343	1.0284 ± 0.0345	1.9546 ± 0.0258	31.85 ± 2.77

Table 3: Trajectory recovery and Go2 execution over 30 mocap-derived virtual-camera clips. The upper block measures recovered displacement. The lower block reports execution dynamics; Static-root is trivially easier on these metrics because its commanded global motion is nearly zero, so comparison to GT is more informative.

Metric	GT	Ours: MoVi	Static-root
Trajectory and displacement			
Trajectory RMSE (m) ↓	0.1684 ± 0.1172	0.3057 ± 0.2397	0.4208 ± 0.2071
Speed RMSE (m/s) ↓	0.3437 ± 0.1305	0.5080 ± 0.3236	0.6316 ± 0.3013
FDE (m) ↓	0.2317 ± 0.1953	0.4949 ± 0.4116	0.7029 ± 0.3437
Path-length Error (m) ↓	0.2167 ± 0.1665	0.4266 ± 0.3400	0.6716 ± 0.3373
Realized Displacement (m)	0.4847 ± 0.2626	0.2742 ± 0.1946	0.0262 ± 0.0114
Mean Command Speed (m/s)	0.6140 ± 0.3071	0.3545 ± 0.2170	0.0000
Execution dynamics			
Velocity-command Tracking RMSE (m/s) ↓	0.2819 ± 0.1130	0.1781 ± 0.1024	0.0526 ± 0.0185
Root Acceleration RMS (m/s^2) ↓	4.5001 ± 0.5889	3.9240 ± 0.5169	3.5117 ± 0.2657
Root Jerk RMS (m/s^3) ↓	198.32 ± 44.76	163.58 ± 22.27	147.51 ± 10.92
Joint Acceleration RMS (rad/s^2) ↓	112.20 ± 30.04	89.11 ± 22.48	62.32 ± 6.40
Action-rate RMS (/s) ↓	20.29 ± 4.53	16.48 ± 3.57	11.49 ± 2.42
Contact Foot Sliding RMS (m/s) ↓	1.0206 ± 0.4212	0.6511 ± 0.3403	0.1710 ± 0.1066

4.4 Trajectory Recovery and Go2 Retargeting

Finally, we study whether recovering global displacement produces more useful robot reference motion than retaining articulation alone. We construct 30 mocap-derived clips rendered with fixed, follow, and orbit virtual cameras. All references are processed through the same Unitree Go2 retargeting pipeline and evaluated with the same frozen controller. We compare three settings: *GT*, which uses the original mocap motion and trajectory; *Ours*, which reconstructs trajectory from the visual observations; and *Static-root*, which uses the GT articulated mocap motion but sets root velocity and displacement to zero. Static-root therefore isolates the effect of removing global trajectory while preserving high-quality articulation. The qualitative difference is illustrated in Fig. 4, where MoVi preserves spatial displacement while Static-root executes the same articulated motion largely in place.

As shown in Table 3, recovering root displacement substantially improves agreement with the mocap trajectory. Ours reduces trajectory RMSE, FDE, speed error, and path-length error relative to the Static-root baseline, while also recovering substantially more of the reference displacement and command speed. This shows that articulated pose alone is insufficient to reproduce the spatial motion of the original animal.

The execution-dynamics metrics require a different interpretation. Static-root naturally obtains low velocity-tracking error, acceleration, jerk, and foot-sliding values because it commands almost no global movement. It therefore represents an easier but less faithful robot reference. For these metrics, similarity to the dynamic GT reference is more meaningful than the raw minimum. Ours preserves non-trivial displacement while producing execution dynamics substantially closer to GT than a purely static-root motion. Overall, the experiment shows that MoVi adds task-relevant global motion to the reconstructed articulation, yielding Go2 references that better preserve the spatial and dynamic characteristics of mocap motion.

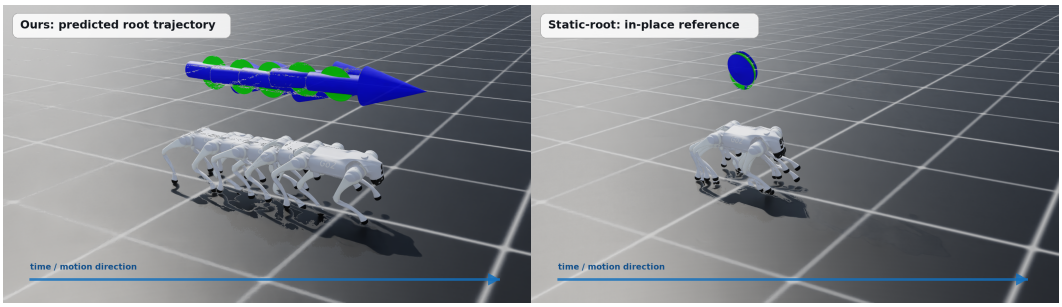

Figure 4: **Qualitative comparison of Go2 retargeting with and without recovered root trajectory.** MoVi produces a moving reference with recovered global displacement, whereas the Static-root baseline preserves articulation but removes root velocity and translation, resulting in in-place motion.

5 Limitations and Discussion

Our current motion supervision is dominated by common canine locomotion, particularly walking and trotting, which limits the behavioral diversity covered during training. As a result, highly dynamic, non-periodic, or less frequent motions are not yet sufficiently represented. Future work will focus on incorporating broader and more diverse motion data to extend MoVi toward a wider range of canine behaviors and downstream quadruped skills.

6 Conclusion

We presented MoVi, a motion-centric framework that converts unconstrained monocular dog videos into retargeting-ready 4D motion. MoVi combines flow-matching-based 3D pose estimation, triggered temporal and joint refinement, scale-aware canonicalization, and reliability-gated trajectory prediction to recover articulated motion together with global root displacement.

A key property of MoVi is that trajectory recovery is introduced without sacrificing the quality of the underlying articulated motion. The resulting representation preserves accurate and temporally coherent canine motion while adding root velocity, height, contact cues, and global displacement that are directly useful for quadruped retargeting. By reducing the dependence on carefully captured mocap, depth, or calibrated-camera data, MoVi provides a path toward using ordinary monocular RGB videos as a broader and more scalable source of reference motion for quadruped imitation learning.

AI Use Statement

Generative AI tools were used to assist with language polishing and manuscript editing, to support retrieval and discovery of related literature, and to generate synthetic occlusion examples used in robustness evaluation. All research design, model development, implementation, experimental analysis, and scientific conclusions were conducted and verified by the authors.

References

Bescós, B., Fácil, J. M., Civera, J., et al. DynaSLAM: Tracking, Mapping and Inpainting in Dynamic Scenes. *IEEE Robotics and Automation Letters*, 2018.

Biggs, B., Roddick, T., Fitzgibbon, A., et al. Creatures Great and SMAL: Recovering the Shape and Motion of Animals from Video. *ACCV*, 2018.

Campos, C., Elvira, R., Rodríguez, J. J. G., et al. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*, 2021.

Gong, J., Foo, L. G., Fan, Z., et al. DiffPose: Toward More Reliable 3D Pose Estimation. *CVPR*, 2023.

Li, Z., Litvak, D., Li, R., et al. Learning the 3D Fauna of the Web. *CVPR*, 2024.

Lyu, J., Zhu, T., Gu, Y., et al. AniMer: Animal Pose and Shape Estimation Using Family Aware Transformer. *CVPR*, 2025.

Peng, X. B., Abbeel, P., Levine, S., et al. DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills. *ACM Transactions on Graphics*, 2018.

Peng, X. B., Coumans, E., Zhang, T., et al. Learning Agile Robotic Locomotion Skills by Imitating Animals. *Robotics: Science and Systems (RSS)*, 2020.

Peng, X. B., Ma, Z., Abbeel, P., et al. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. *ACM Transactions on Graphics*, 2021.

Sabathier, R., Novotny, D., Mitra, N. J. Animal Avatars: Reconstructing Animatable 3D Animals from Casual Videos. *ECCV*, 2024.

Schönberger, J. L. and Frahm, J.-M. Structure-from-Motion Revisited. *CVPR*, 2016.

Shan, W., Liu, Z., Zhang, X., et al. Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation. *ICCV*, 2023.

Shin, S., Kim, J., Halilaj, E., et al. WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion. *CVPR*, 2024.

Shooter, M., Malleson, C., Hilton, A. Benchmarking Monocular 3D Dog Pose Estimation Using In-The-Wild Motion Capture Data. *CVPR Workshops*, 2024.

Stamatopoulou, M., Li, J., Kanoulas, D. SDS – See it, Do it, Sorted: Quadruped Skill Synthesis from Single Video Demonstration. *Conference on Robot Learning (CoRL)*, 2025.

Wang, Y., Wang, Z., Liu, L., et al. TRAM: Global Trajectory and Motion of 3D Humans from In-the-Wild Videos. *ECCV*, 2024.

Wang, T., Yu, X., Mathis, M. W. FMPose3D: Monocular 3D Pose Estimation via Flow Matching. *CVPR*, 2026.

Wu, S., Li, R., Jakab, T., et al. MagicPony: Learning Articulated 3D Animals in the Wild. *CVPR*, 2023.

Xu, J., Zhang, Y., Peng, J., et al. Animal3D: A Comprehensive Dataset of 3D Animal Pose and Shape. *ICCV*, 2023.

Yoon, T., Kang, D., Kim, S., et al. Spatio-Temporal Motion Retargeting for Quadruped Robots. *IEEE Transactions on Robotics*, 2025.

Yuan, Y., Iqbal, U., Molchanov, P., et al. GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras. *CVPR*, 2022.

Zhao, B. N., Wu, J., Wu, S. Web-Scale Collection of Video Data for 4D Animal Reconstruction. *NeurIPS*, 2025.

Zhao, L., Luo, Z., Han, Y., et al. Learning Aggressive Animal Locomotion Skills for Quadrupedal Robots Solely from Monocular Videos. *npj Robotics*, 2025.

Zuffi, S., Kanazawa, A., Jacobs, D. W., et al. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. *CVPR*, 2017.