QuadForecaster: Diffusion-Based Quadruped Pose Prediction for Animal Communication Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Animal communication relies on subtle temporal patterns in movement that current pose estimation systems cannot anticipate, thus limiting their utility. Existing frameworks excel at detecting present configurations but fail to predict future poses, forcing interaction systems to remain reactive rather than proactive. We introduce QuadForecaster, the first diffusion-based model specifically designed for quadrupedal pose prediction, enabling automated systems to decode animal communication through movement forecasting. Our temporally cascaded diffusion architecture treats pose prediction as structured denoising, iteratively refining uncertain future poses while providing essential uncertainty quantification for safe deployment. Evaluated on the cheetah and cow datasets, QuadForecaster achieves 0.116m MPJPE for cheetah behaviors and 0.86m MPJPE for complex cow social interactions, successfully capturing rapid behavioral transitions and multi-modal dynamics. QuadForecaster paves the way for robust animal motion and communication analysis, enabling proactive cross-species interaction across conservation, agriculture, and research applications.

1 Introduction

Understanding animal movement patterns unlocks the secret language of non-human communication. Animals communicate through subtle posture shifts, precise gait transitions, and complex multi-modal signals that current technology struggles to decode [1, 23]. Predicting future poses allows robots and automated systems to anticipate animal intentions, thus fostering safer interspecies interactions [7]. This capability benefits wildlife conservation, preventing human-animal conflicts, and agricultural monitoring, where early detection of lameness protects welfare.

The landscape of animal pose estimation has evolved from basic 2D tracking to sophisticated multi-modal systems for behavioral analysis [8]. DeepLabCut and LEAP pioneered markerless tracking with minimal annotation, while SLEAP extended capabilities to multi-animal scenarios essential for social behavior studies [21, 24, 25]. Beyond 2D, DANNCE [16] and OpenMonkeyStudio [5] enable 3D multi-camera tracking, and SuperAnimal demonstrates cross-species generalization across 45+ species [30]. However, current frameworks remain fundamentally reactive, detecting poses after they occur, which severely limits communication analysis [26]. In human-robot interaction, forecasting human poses enables fluid communication [28], a principle that applies with greater urgency to animal interactions.

Quadrupedal motion presents unique forecasting challenges. Anatomical diversity introduces species-specific kinematic constraints, quadrupeds with flexible spines exhibit dynamics distinct from bipedal humans [22]. Rapid gait transitions, social interactions, and defensive behaviors create highly variable motion distributions [6]. Data scarcity, environmental noise, vegetation occlusions, lighting variations, camera motion, and multi-agent interactions further complicate prediction [14, 11, 17].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Early human motion prediction relied on recurrent architectures [10, 20] and later on graph convolutional networks [19] or spatio-temporal transformers [2, 18]. Stochastic generative models captured multi-modality [4], and diffusion-based methods such as DePOSit framed prediction as iterative denoising [27]. However, these methods incorporate human-specific priors, limiting applicability to quadrupeds. Diffusion models are particularly promising for animal pose forecasting due to their robustness to noise and missing data [12, 9].

We introduce the first diffusion-based approach to animal pose forecasting, engineered to decode the temporal language of quadruped communication. Motion prediction is treated as structured denoising, where missing or uncertain future poses emerge through iterative refinement, mirroring how animals anticipate each other. To capture multi-scale temporal dynamics, we propose a *temporally cascaded diffusion architecture* for both short- and long-term forecasts: short-term predictions capture immediate intention signals, while long-term forecasts reveal behavioral patterns crucial for interaction planning.

In summary, our contributions are: (i) the first diffusion-based approach to animal pose forecasting robust to noisy and incomplete observations; (ii) a temporally cascaded architecture capturing immediate signals and extended behavioral patterns; and (iii) robust prediction of out-of-distribution wireframes.

2 Methodology

QuadForecaster pioneers animal-specific motion prediction by recognizing that quadrupedal communication operates through fundamentally different kinematic principles than human movement. While human pose prediction relies on bipedal constraints and cyclical walking patterns, quadrupeds communicate through complex gait transitions involving 2-3 ground contact points, dynamic spine articulation, and rapid behavioral state changes that traditional models cannot capture [6].

We address this challenge through a novel diffusion-based architecture that treats pose forecasting as structured communication decoding. The model processes temporal sequences encompassing both observed poses and masked future frames, using iterative denoising to reveal plausible motion continuations. This approach mirrors how animals themselves process and predict each other's movements, through gradual refinement of uncertain sensory information into actionable behavioral predictions.

Our training methodology operates on two distinct datasets that capture diverse quadrupedal communication patterns. AcinoSet [15] features cheetahs with 20-joint skeletons representing high-speed predator dynamics, while MBE-ARI [23] captures cow behaviors with 39-joint configurations encompassing complex social and feeding interactions. The model processes approximately 12,000 cheetah frames and 7,000 cow frames, learning species-specific movement vocabularies essential for accurate communication analysis.

Given past pose sequences, QuadForecaster predicts future motion over specified horizons through a sophisticated spatio-temporal encoding scheme. Temporal dynamics receive 128-dimensional embeddings that capture multi-scale motion patterns from immediate micro-movements to extended behavioral sequences. Joint identity encoding uses 16-dimensional representations that preserve anatomical relationships while enabling cross-species generalization. These unified representations condition a diffusion process that iteratively refines noisy future poses into physically consistent, behaviorally plausible trajectories.

QuadForecaster employs a Conditional Score-based Diffusion for Imputation (CSDI) [29] as its backbone architecture, which we adapt for quadrupedal motion prediction. During training, the model learns to predict noise on masked future frames using the standard diffusion objective. This teaches the system to recover clean future poses from noisy intermediate representations by minimizing mean-squared error between predicted and ground-truth trajectories. Critically, we evaluate using Mean Per-Joint Position Error (MPJPE) [13] rather than training on it directly, ensuring the model learns robust generative capabilities rather than overfitting to specific error metrics.

For cow skeletons, we incorporate bone-length regularization to preserve anatomical constraints during prediction. This additional loss term maintains realistic skeletal proportions essential for accurate behavioral interpretation. The 20-joint cheetah skeleton uses a simplified topology where this regularization becomes inactive, allowing the model to capture rapid and dynamic characteristics of

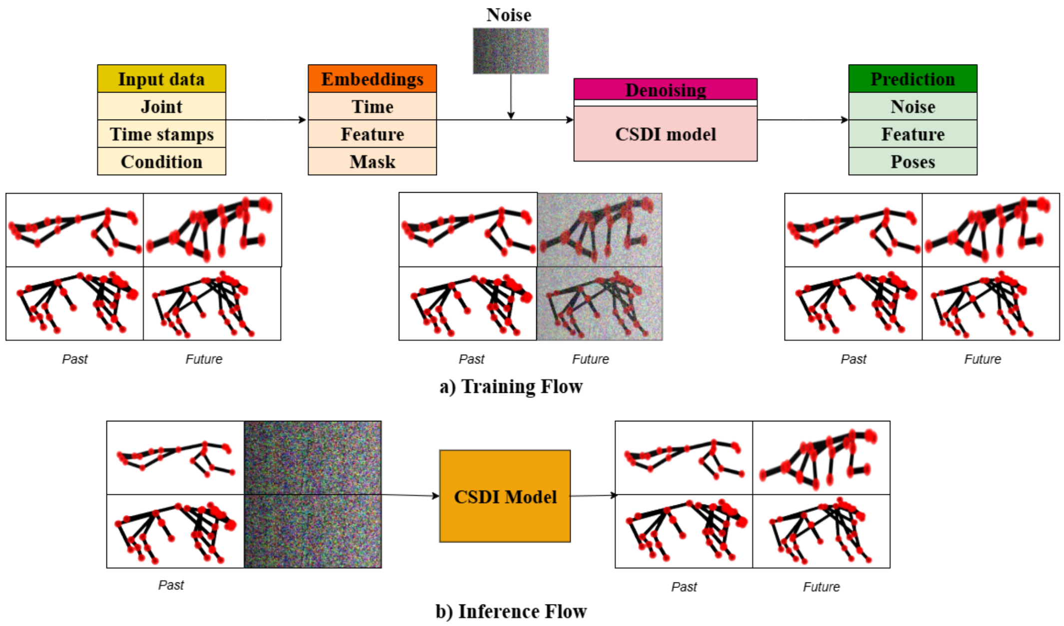

Figure 1: Overview of QuadForecaster Architecture

high-speed predator behavior. These species-specific adaptations enable QuadForecaster to generate physically consistent forecasts that respect each animal's unique biomechanical constraints.

The inference pipeline demonstrates the model's practical utility for real-time communication analysis. The system processes formatted motion sequences, automatically parsing temporal trajectories and splitting them into observed (past) and target (future) sequences. During prediction, the model receives observed sequences concatenated with masked future frames, along with temporal indices and visibility masks indicating known timesteps. The spatio-temporal diffusion process then iteratively denoises masked portions through 50 refinement steps, generating future motion predictions suitable for immediate use in animal-robot interaction systems. This iterative denoising framework provides natural uncertainty quantification essential for safe animal interaction. Rather than producing single deterministic predictions, the model generates probability distributions over future poses, enabling downstream systems to assess prediction confidence and adjust interaction strategies accordingly. Such uncertainty awareness is crucial when deploying autonomous systems in animal environments, where prediction failures can have serious welfare consequences

3 Results

Our evaluation protocol rigorously assesses QuadForecaster's ability to decode animal communication through pose prediction across diverse behavioral contexts. We employ a 90/10 train-test split for both datasets, ensuring robust evaluation on unseen behavioral patterns. For cows, this yields 6,452 training frames and 717 test frames; for cheetahs, 11,834 training and 2,548 test frames. This evaluation strategy tests the model's generalization to novel behavioral sequences critical for real-world deployment.

We assess prediction accuracy using sliding window evaluation that simulates real-time communication analysis scenarios. Each evaluation window comprises 20 input frames followed by 10 target frames. This configuration captures both immediate intention signals (1-5 frames) and medium-term behavioral patterns (6-10 frames) essential for interaction planning. We report Mean Per-Joint Position Error (MPJPE), ADE/FDE (Average and Final Displacement Error) [3], all measured in absolute meters, to facilitate practical deployment considerations

Our framework demonstrates excellent predictive performance for the Cheetah test cases across diverse behavioral states. Evaluated over 2,225 test windows, the model achieves 0.116m MPJPE with corresponding ADE and FDE values of 0.116m and 0.205m, respectively. Per-sequence analysis reveals behavioral specificity in prediction accuracy: walking behaviors achieve 0.104m MPJPE, while more complex pacing/lunging sequences reach 0.146m MPJPE. These results indicate that QuadForecaster successfully captures both routine locomotion and dynamic behavioral transitions crucial for understanding cheetah communication patterns.

Cow behavioral prediction presents greater challenges due to anatomical complexity and social interaction patterns. Across 717 test windows, the model achieves 0.86m MPJPE with ADE and FDE values of 0.86m and 0.80m, respectively (Table 1). Behavioral analysis shows walking sequences achieve 0.949m MPJPE, while turning behaviors (0.801m) and observing states (0.840m) demonstrate more accurate predictions. This pattern suggests that stationary and slow-motion behaviors enable more precise prediction than rapid locomotion, consistent with the biomechanical complexity of 39-joint cow skeletons.

Table 1: Per-Sequence Motion Prediction Evaluation

Animal	**Action**	**MPJPE (m)**	**ADE (m)**	**FDE (m)**
Cheetah	Walking	0.104	0.104	0.189
Cheetah	Pacing / Galloping	0.104	0.104	0.164
Cheetah	Pacing / Lunging	0.146	0.146	0.242
Cheetah	Pacing (grouped avg.)	0.122	0.122	0.222
Cheetah	Lunging / Pacing	0.084	0.084	0.141
Cheetah	Galloping	0.137	0.137	0.254
Cow	Walking	0.949	0.949	1.079
Cow	Turning	0.801	0.801	0.652
Cow	Observing	0.840	0.840	0.841

Temporal consistency metrics reveal motion quality. Mean Velocity Error (MVE), equivalent to the Mean Per-Joint Velocity Error (MPJVE) used in prior work [31], measures prediction smoothness essential for natural animal interaction. We also report cosine similarity between predicted and ground-truth joint velocities, which quantifies directional alignment. Cheetah predictions achieve MVE of 0.0367 ± 0.0185 m/frame with cosine similarity of 0.762 ± 0.302, indicating excellent preservation of motion direction and magnitude. Cow predictions show higher variance with MVE of 1.309 ± 0.393 and cosine similarity of 0.216 ± 0.111, reflecting the greater complexity in predicting multi-joint social behaviors.

Comparative analysis against human motion prediction benchmarks provides context for our achievements. While direct comparison proves impossible due to scale and dataset differences, our framework with the configurations we used for the Cheetah datasets achieves competitive performance with ADE/FDE of 0.116m/0.205m compared to DePOSit's long-term human predictions of 0.356m/0.396m [27]. However, DePOSit's short-term performance (9.9mm FDE at 80ms) highlights opportunities for improving our temporal resolution, particularly for rapid events.

4 Conclusion

In this work, building on the intuition of using motion as a guide for communication and interaction in the animal world, we introduce QuadForecaster. This is the first diffusion-based quadrupedal pose prediction model specifically designed for animal communication analysis, achieving 0.116m MPJPE on cheetah behaviors and 0.86m MPJPE on complex cow interactions. We demonstrate that temporally cascaded architectures capture both immediate intention signals and extended behavioral patterns essential for cross-species interactions. Unlike previous reactive systems, QuadForecaster enables proactive interaction by predicting animal intentions through iterative denoising that mirrors biological perception processes, while providing uncertainty quantification essential for safe deployment. Our results prove that species-specific diffusion models can decode the temporal language of quadrupedal movement, establishing the foundation for next-generation animal-robot communication systems across conservation, agriculture, and research applications.

References

[1] Naqash Afzal et al. "The Convergence of AI and animal-inspired robots for ecological conservation". In: *Ecological Informatics* 85 (2025), p. 102950.

[2] Emre Aksan et al. “A spatio-temporal transformer for 3D human motion prediction”. In: *International Conference on 3D Vision (3DV)*. 2021.

[3] Alexandre Alahi et al. “Social LSTM: Human trajectory prediction in crowded spaces”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 961–971.

[4] Sadegh Aliakbarian et al. “Contextually plausible and diverse 3D human motion prediction”. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[5] Prathap Bala et al. “OpenMonkeyStudio: Automated markerless pose estimation in freely moving macaques”. In: *Nature Communications* 11.1 (2020), p. 4560.

[6] James P Bohnslav et al. “DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels”. In: *eLife* 10 (2021), e63377.

[7] Frank Bonnet et al. “Robots mediating interactions between animals for interspecies collective behaviors”. In: *Science Robotics* 4.28 (2019), eaau7897.

[8] Qianyi Deng et al. “Towards Multi-Modal Animal Pose Estimation: A Survey and In-Depth Analysis”. In: *arXiv preprint arXiv:2410.09312* (2024).

[9] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat GANs on image synthesis”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[10] Katerina Fragkiadaki et al. “Recurrent network models for human dynamics”. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2015, pp. 4346–4354.

[11] Jacob M Graving et al. “DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning”. In: *eLife* 8 (2019), e47994.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[13] Catalin Ionescu et al. “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339.

[14] Catalin Ionescu et al. “Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339.

[15] Daniel Joska et al. “AcinoSet: A 3D Pose Estimation Dataset and Baseline Models for Cheetahs in the Wild”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2021, pp. 13901–13908.

[16] Phoebe Karashchuk et al. “DANNCE: multi-animal 3D pose estimation from video”. In: *Nature Methods* 18.5 (2021), pp. 586–595.

[17] Jessy Lauer et al. “Multi-animal pose estimation, identification and tracking with DeepLabCut”. In: *Nature Methods* 19.4 (2022), pp. 496–504.

[18] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. “History repeats itself: Human motion prediction via motion attention”. In: *European Conference on Computer Vision (ECCV)*. 2020.

[19] Wei Mao et al. “Learning trajectory dependencies for human motion prediction”. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.

[20] Julieta Martinez, Michael J Black, and Javier Romero. “On human motion prediction using recurrent neural networks”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2891–2900.

[21] Alexander Mathis et al. “DeepLabCut: markerless pose estimation of user-defined body parts with deep learning”. In: *Nature Neuroscience* 21.9 (2018), pp. 1281–1289.

[22] Tanmay Nath et al. “Using DeepLabCut for 3D markerless pose estimation across species and behaviors”. In: *Nature Neuroscience* 22.9 (2019), pp. 1321–1329.

[23] Ian Noronha et al. *MBE-ARI: A Multimodal Dataset Mapping Bi-directional Engagement in Animal-Robot Interaction*. 2025. arXiv: `2504.08646 [cs.CV]`.

[24] Talmo D Pereira et al. “Fast animal pose estimation using deep neural networks”. In: *Nature Methods* 16.1 (2019), pp. 117–125.

[25] Talmo D Pereira et al. “SLEAP: Multi-animal pose tracking”. In: *Nature Methods* 19 (2022), pp. 486–495.

[26] Donato Romano et al. “A review on animal–robot interaction: from bio-hybrid organisms to mixed societies”. In: *Biological cybernetics* 113.3 (2019), pp. 201–225.

[27] Saeed Saadatnejad et al. “A generic diffusion-based approach for 3D human pose prediction in the wild”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[28] Alessandro Simoni et al. “3D Pose Nowcasting: Forecast the future to improve the present”. In: *Computer Vision and Image Understanding* 251 (2025), p. 104233.

[29] Yusuke Tashiro et al. “CSDI: Conditional score-based diffusion models for probabilistic time series imputation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[30] Shaokai Ye et al. “SuperAnimal pretrained pose estimation models for behavioral analysis”. In: *Nature communications* 15.1 (2024), p. 5165.

[31] Xiaozheng Zheng et al. “Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 14678–14688.