Two2Four: Generative Quadruped Puppeteering from Human Motion

Fatemeh Zargarbashi[1,2,*], Zehong Qiu[1,*], Dhruv Agrawal[1,2], Stelian Coros[2], Robert W. Sumner[1,2], Martin Guay[1], Jakob Buhmann[1]

[1]DisneyResearch|Studios, Switzerland [2]ETH Zürich, Switzerland [*]Equal contribution

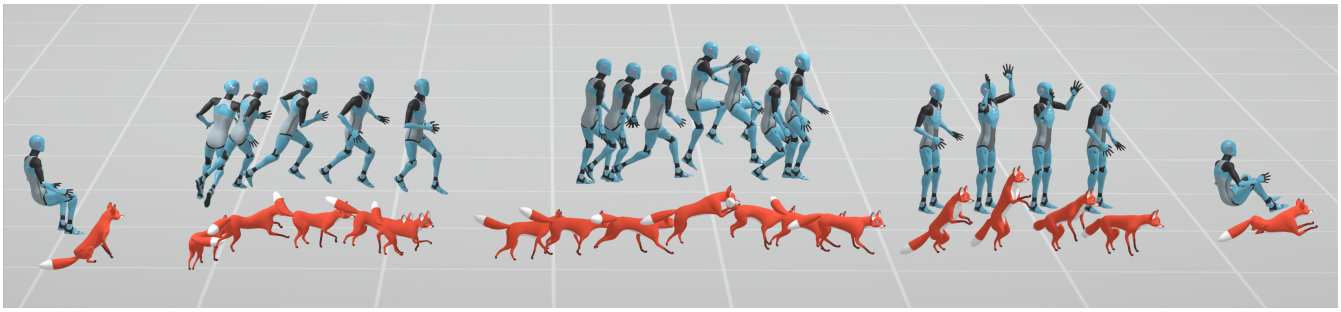

Figure 1: *Given ordinary human motion, Two2Four generates realistic and controllable quadruped motion via a two-stage diffusion model, supporting diverse behaviors and fine-grained control through structured conditioning and inpainting.*

Abstract

Realistic animal motion for virtual production is typically obtained either through motion capture of highly trained performers who accurately mimic animal behavior, or by retargeting ordinary human motion using complex control setups. Both approaches are challenging and often fail to fully reproduce the nuances of natural animal motion, motivating data-driven alternatives. We present an automatic human-to-quadruped puppeteering framework that produces plausible and controllable quadruped motions from ordinary human motion data. Our approach employs a two-stage generative diffusion model trained purely on quadruped motion data. By introducing a structured conditioning and inpainting strategy, our method supports a wide range of actions, including walking, running, jumping, sitting, and lying. Furthermore, we enable fine-grained intuitive control of the quadruped motion such as head movement control and individual limb puppeteering. Experimental results demonstrate improved motion realism and controllability compared to existing retargeting approaches, highlighting the effectiveness of our framework as a tool for animation and virtual production applications.

CCS Concepts

• ***Computing methodologies*** → ***Animation; Artificial intelligence;***

1. Introduction

Producing believable animal motion is a core challenge in animation, virtual production, and visual effects. While human motion capture has become widely accessible, leveraging it to drive animal characters is still non-trivial. In practice, it must either rely on highly trained performers who mimic animal behavior [Per14; WLSN24], or use standard human motion paired with complex hand-crafted control rigs [VFX25]. Both approaches present limitations: the former is difficult to perform consistently and requires expertise, while the latter may produce unnatural results.

In this work, our goal is to create a system that enables a human to steer quadruped motion, as required in a virtual production use case [VFX25]. The quadruped motion must spatially conform to the human motion during locomotion, while still giving the human the ability to "puppeteer" semantic actions, such as sitting, jumping, or the look-at direction of the character. Thus, the final motion should satisfy the following conditions: 1) **Naturalness**, the resulting motion lies within the distribution of natural quadruped motion, exhibiting plausible gaits, contact patterns, and coordination. 2) **Semantic consistency**, high-level or abstract motion characteris-

tics such as path, direction, speed, and behavior type are preserved from human motion.

Cross-morphology motion transfer poses substantial challenges due to severe morphological, kinematic, and proportion differences. Even at the level of overall scale and speed, it remains ambiguous what it means for a quadruped to "follow" a person while producing realistic motion. For example, a dog runs much faster than a human, while it can walk at a similar speed to human walking. Moreover, quadrupeds show a more diverse set of gait patterns compared to bipeds. Importantly, not all motion attributes are affected by this ambiguity to the same extent. We distinguish two sets of motion attributes: *precise features* that can be directly transferred across the two morphologies (e.g., travel direction or target location), and *imprecise features* (e.g., limb positions or root height) that only provide rough semantic cues rather than exact attributes to follow. These discrepancies make direct motion transfer an inherently ambiguous task.

Existing approaches typically fall into two categories. Optimization-based retargeting methods [SOL13] enforce strong alignment between source and target motion, offering controllability but often lacking realism. Conversely, data-driven approaches [ZSKS18; LJG*24; LSYS24] can produce natural-looking motions, but suffer from the absence of paired data, often lack semantic alignment, and provide limited control over how specific motion features are preserved. In this work, we bridge this gap using a quadruped motion prior that generates natural motion while enabling multiple levels of control from human motion. Our key insight is that a diffusion model trained exclusively on quadruped data learns a strong prior over natural animal motion (the naturalness property). Furthermore, diffusion models provide flexible methods of controlling the generation (the semantics property). More importantly, those control mechanisms support both direct control via network conditions and looser control with inference-based inpainting (imputation). This is particularly useful for the ill-posed task of cross-morphology motion transfer with precise and imprecise features as the control signals.

To this end, we present *Two2Four*, a generative framework for human driven quadruped puppeteering. It consists of two diffusion stages trained purely on quadruped data, and an inference-time inpainting strategy to achieve direct control over user-specified features of the generated quadruped motion. By abstracting the human movement into universal features, shared between human and animal, we can train a conditional diffusion model on quadruped data and condition it at test time with human features, such as 2D target location or head direction. In addition, we leverage an inpainting mechanism at high noise levels to control the quadruped motion generation with motion features that are imprecisely mapped between the two morphologies. For instance, a user can select to map the motion of the human's hands (or feet) to the quadruped front legs to achieve direct puppeteering or inpaint the root height to achieve semantic actions such as lying, sitting, or jumping. This flexible mapping of approximately mapped features stands in contrast to traditional optimization-based or purely data-driven methods, which often require rigid objectives or fixed training distributions. Furthermore, we show that splitting the generation process into two stages—first a trajectory stage for a subset of joints and a second stage for full-body motion—improves motion alignment and enables more effective inpainting when trained on a small dataset. This design makes our system robust, even when operating with out-of-domain control signals. Our method achieves smooth long-horizon motion generation by autoregressively rolling out the generation.

We evaluate our method on a diverse range of actions. Our experiments show successful transfer of locomotion behaviors such as walking, running, and jumping, as well as finer control such as look-at control and individual limb puppeteering. We further show examples of transferring sitting actions or mapping a human sitting to a quadruped lying down. Our method achieves more ease and flexibility in control, while resulting in more realistic motion and less foot sliding compared to baseline state-of-the-art methods. In summary, our contributions are as follows:

- A generative framework for quadruped motion synthesis with human driven attributes, trained purely on quadruped data.
- A two-stage diffusion model that enables inpainting when dealing with small datasets and improves motion quality.
- A hybrid formulation for robustness to out-of-domain data: conditioning on universally shared (sparse) features and an inpainting strategy for imprecisely mapped features.

2. Related Work

Early work in human-driven puppeteering of virtual creatures [Gle98] used rigid feature correspondences for motion transfer, demanding significant manual effort and limiting runtime flexibility. Recently, learning-based approaches [ZLW*24] have relied on inferring correlations across morphologies, improving motion quality but offering less direct control. In contrast, robustness of generative models [LDR*22; TRG*23] to out-of-domain conditions suggests their potential to enable human-to-quadruped motion transfer with precise actor control without complex feature engineering. Consequently, the remaining section reviews non-parametric motion transfer, neural motion retargeting, and generative approaches.

2.1. Non-parametric Motion Transfer

Early approaches to cross-morphology control relied heavily on explicit pose-level mathematical mappings. [SOL13] proposed building correspondence between human and target creature poses using manually defined optimization goals, followed by per-pose replacement during the puppetry phase. [RTK*14] similarly learned pose-mapping from a small user-assembled dataset and allowed transferring motion directly to unrigged target meshes. More recently, [GFK*23] use differentiable optimal control to transfer animal motion to different robot skeletons of varying proportions. [AI25] build a transport plan between two motion sources by aligning the pairwise distance matrix among the source frames to that in the target frames. Motion2Motion [CZY*25] extract overlapping patches from source motion and project them into the target skeleton space using sparse mapping, followed by per-patch retrieval and blending, to give the final motion. Although these methods demonstrate realistic results for their specific applications, for generalized motion they exhibit reduced naturalness due to maintaining rigid correspondences that do not generalize to new contexts.

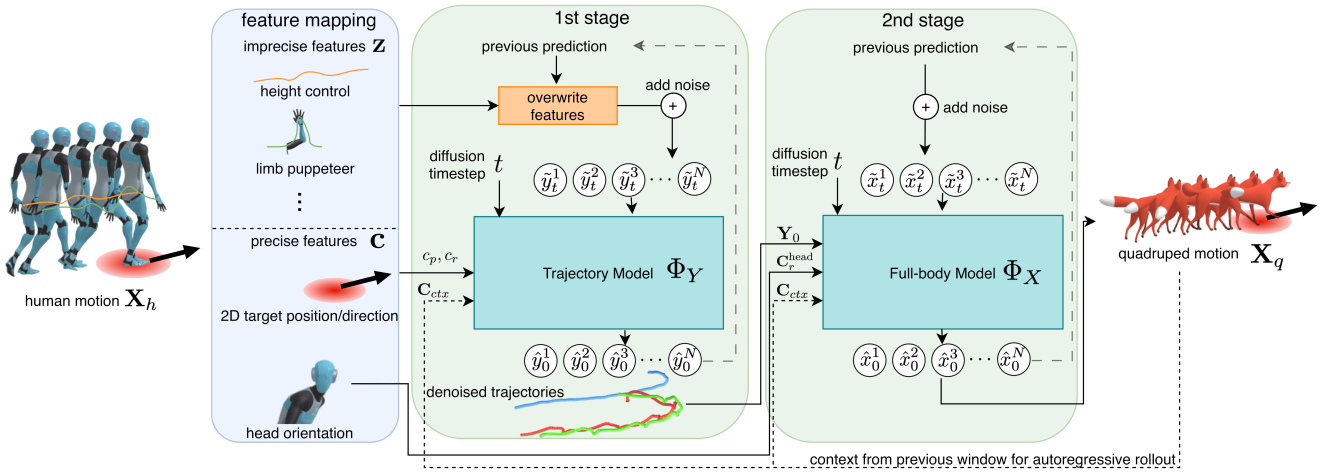

Figure 2: *Two2Four is a two-stage diffusion-based pipeline for transferring human motions to quadruped movements. The first stage denoises quadruped trajectories on a subset of joints and enables controllability via conditioning and inpainting. The second stage maps the subset of trajectories to full quadruped motion. At test time, human motion features are mapped to quadruped motion features and drive the two-stage model via directs conditions for precise features or inpainting for more roughly mapped motion features. For long horizon control, the motion is autoregressively generated by using context frames from the previous motion window.*

2.2. Neural Motion Retargeting

In the context of intra-morphology retargeting, human-to-human retargeting has flourished with the application of deep learning to large datasets with diverse skeletal structure and body sizes [LCC19; ALL*20; LKP*23]. Early work from Villegas et al. [VYCL18] used adversarial cyclic training for unsupervised motion retargeting. [ALL*20] introduced using a skeletal convolution network to unify diverse morphologies to a simplified primal skeleton, with [LKP*23] extending this to learn skeleton-agnostic motion embeddings. Conversely, little research has focused on creature-to-creature retargeting, with the notable exception of [ZLW*24]. However, this approach assumes similar motion priors across source and target creatures and requires a specific pose prior for the target skeleton. Ultimately, these methods fall short when applied to different morphologies due to the substantial domain gap between human and creature motion data.

Regarding cross-morphology motion transfer, [LWC*23] introduced adversarial losses to facilitate retargeting from humans to different morphologies. [EJS*24] leveraged shared codebooks learned from manually designed rule-based retargeting, whereas Li et al. [LSYS24] proposed VQ-PAE to learn common motion representations in a fully unsupervised manner for aligning human and quadruped data. However, their method lacks global information, resulting in significant foot sliding artifacts when employing for our use case. Moreover, in the absence of semantically grounded features to guide the alignment, it may incorrectly map different behaviors across domains. Li et al. [LJG*24] maximize mutual information between human pose and quadruped robot pose together with a cycle-consistency loss to learn a motion transfer policy, but do not generate natural animal motion.

2.3. Generative Motion

MDM [TRG*23] and EDGE [TCL23] first implemented diffusion models [HJA20] for human motion synthesis from text and music, respectively. For inference-time control, MDM and [MLS*25] explore inpainting, GMD [KPST23] leverage guidance, and DNO [KPA*24] show the use of optimization. [SAB*24; VAG*25] and [HBLK25] employ a two-stage approach with both using a reduced skeletal representation in the first stage. However, they use frame-based Neural IK and temporally-aware diffusion models for the second stage, respectively. However, none of these methods have demonstrated cross-morphology control or training on smaller datasets such as MANN [ZSKS18].

In contrast, for non-human motion generation, research remains limited due to scarce data. [LAZ*22] and [RLT*24] learn generating motion from a single animation clip using adversarial networks and diffusion models, respectively. However, such methods are limited to recomposing the training clip in contrast to generating new behavior. More recently, OmniMotionGPT [YZS*24] jointly train a motion autoencoder on human and animal datasets, aligning latent representations with text, generating diverse animal motions from text even with limited data. However, this requires text-annotated animal motion datasets. MAS [KTCB24] addresses data scarcity by learning 2D motion priors directly from video and show retargeting to different species. However, it lacks realism especially going from human to quadrupeds. [GRT*25] circumvent data sparsity by training on a diverse set of animals. While they can generalize to new morphologies, they suffer from reduced motion fidelity for any particular skeleton. [WRQ*26] similarly released a large scale motion dataset across 115 species but limited motion data for any particular non-human species. Consequently, high-fidelity generalized quadruped motion generation remains underexplored due to the difficulty in curating large animal motion capture datasets.

Hence, our work is focused on building a controllable motion generator from a small dataset that is robust to out-of-domain conditions, enabling human actor-driven puppeteering without requiring fine-tuning or complex feature engineering.

3. Method

We address the problem of transferring human motion to a quadruped such that the result looks natural and preserves the semantic intent of the source human performance. Our goal is to generate a corresponding quadruped motion that satisfies naturalness and can transfer different actions such as locomotion, jumping, and sitting. For virtual production, look-at control and direct limb puppeteering are also important.

Our model consists of two-stage diffusion modules, where the first stage captures high-level semantics of motion by only predicting the trajectories of selected joints. Then, a second diffusion stage is conditioned on the predicted trajectories and generates the full-body quadruped motion. At inference time, we map specific features from the human motion to control signals of the diffusion processes via conditions for precisely transferrable features, and via inpainting for imprecise features leaving more flexibility. This two-stage decomposition improves motion quality and reduces sliding artifacts compared to a single-stage model [HBLK25]. Additionally, our decomposition increases the effect of inpainting individual trajectories, enabling inpainting as a viable and flexible inference-based control mechanism. An overview of the full pipeline is depicted in Figure 2.

Formally, let $\mathbf{X}_h = \{\mathbf{x}_h^i\}_{i=1}^T$ denote a human motion sequence of length T, where each frame $\mathbf{x}_h^i$ represents the human pose. Our goal is to generate a corresponding quadruped motion $\mathbf{X}_q = \{\mathbf{x}_q^i\}_{i=1}^T$, satisfying naturalness and semantic consistency. Each motion frame is represented as $\mathbf{x} = \{\mathbf{p}, \mathbf{r}\}$ where $\mathbf{p} \in \mathbb{R}^{n\times 3}$ represents the positions and $\mathbf{r} \in \mathbb{R}^{n\times 6}$ denotes the 6D rotations [ZBL*19] of all joints in the skeleton and n denotes the number of joints, which differs for the human and the quadruped. In the following, we will drop the subscript of the data source $_{q/h}$ for better readability. We always assume quadruped motion unless specifically mentioned.

3.1. Trajectory Model

Our first stage aims to predict trajectories of a sparse set of joints J, capturing the global structure of quadruped movements. We represent the trajectory as a sequence $\mathbf{Y} \subset \mathbf{X}$ with $\mathbf{Y} = \{\mathbf{x}^i\}_{i=1}^T\big|_{j\in J}$, where each frame consists of the selected joints' 3D position and 6D orientation. The subset includes three joints, namely root, and the two end joints of the front limbs. These joints are selected as they capture the important semantics of the motion, yet are a compact representation, facilitating the inpainting. We model this distribution using a conditional diffusion model Φ_Y based on the U-Net model of [CTR*24] that operates in the trajectory space and predicts the clean trajectories $\hat{\mathbf{Y}}_0$. Specifically, the model learns to denoise a noisy trajectory $\tilde{\mathbf{Y}}_t$ conditioned on control signals $\mathbf{C}_Y$, parametrized as $\Phi_Y(\tilde{\mathbf{Y}}_{t-1}|\tilde{\mathbf{Y}}_t, \mathbf{C}_Y, t)$ where t is the diffusion timestep. The condition consists of two components. First, to enable locomotion control, we specify a target root position c_p and orientation c_r to be reached after T frames, which defines the global displacement of the character. Since this condition is later obtained from the human character, we only maintain the horizontal components of the target position and the yaw component of the rotation. Additionally, we rely on a sparse target signal compared to a dense trajectory condition to give the model more freedom in the generative process, leading to more realism as shown in our ablation in Section 4.5.2. Second, to ensure temporal continuity when generating long sequences, we condition on a set of preceding context frames $\mathbf{C}_{ctx} = \{\mathbf{x}^1, ..., \mathbf{x}^{n_c}\}$, where n_c denotes the number of context frames. This context enables a smooth transition between consecutive motion windows and mitigates discontinuities that would otherwise arise from independent generation. Thus, the full condition signal is given by $\mathbf{C}_Y = \{c_p^T, c_r^T, \mathbf{C}_{ctx}\}$. Both conditions are randomly dropped out during training, with the dropout mask also included in the model input.

The first stage model is trained using a combination of losses:

$$\mathcal{L} = \mathcal{L}_{recons} + \mathcal{L}_{cond}, \tag{1}$$

where $\mathcal{L}_{recons}$ is a standard diffusion reconstruction loss between $\mathbf{Y}$ and $\hat{\mathbf{Y}}_0$, and $\mathcal{L}_{cond}$ is a loss for matching the conditions at context frames and final frame.

$$\mathcal{L}_{recons} = ||\hat{\mathbf{p}}_0 - \mathbf{p}||_2^2 + ||\hat{\mathbf{r}}_0 - \mathbf{r}||_2^2, \tag{2}$$

$$\mathcal{L}_{cond} = ||\hat{\mathbf{p}}_{0,xy}^T - c_p||_2^2 + ||\hat{\mathbf{r}}_{0,xy}^T - c_r||_2^2 \tag{3}$$

$$+ ||\{\hat{\mathbf{x}}_0\}_{i=1}^{n_c}|_{j\in J} - \mathbf{C}_{ctx}|_{j\in J}||_2^2, \tag{4}$$

where $\hat{\mathbf{p}}_{0,xy}^T$ and $\hat{\mathbf{r}}_{0,xy}^T$ are the 2D position and direction of the root at time T, computed based on the predicted $\hat{\mathbf{Y}}_0$.

3.2. Full-body Model

Given the predicted joint trajectories $\hat{\mathbf{Y}}_0$, the second stage aims to generate a full body quadruped motion $\mathbf{X}$ that conforms to the trajectories. We model this mapping using a conditional diffusion model defined over the full motion space $\Phi_X(\tilde{\mathbf{X}}_{t-1}|\tilde{\mathbf{X}}_t, \mathbf{C}_X, t)$. The second stage model is conditioned on the subset of joint trajectories $\hat{\mathbf{Y}}_0$ generated by the first stage and the context frames $\mathbf{C}_{ctx}$. In addition, head orientations $\mathbf{C}_r^{\text{head}}$ are included in the conditions to enable direct control of the quadruped look-at direction. Thus the full condition of the second stage model is $\mathbf{C}_X = \{\hat{\mathbf{Y}}_0, \mathbf{C}_r^{\text{head}}, \mathbf{C}_{ctx}\}$, and it directly predicts $\hat{\mathbf{X}}_0$. The model is trained to predict positions and orientations of all joints, along with contact labels of the four limbs. The loss to train this stage consists of

$$\mathcal{L} = \mathcal{L}_{recons} + \mathcal{L}_{FK} + \mathcal{L}_{sliding} + \mathcal{L}_{cond}, \tag{5}$$

where $\mathcal{L}_{FK}$ denotes a reconstruction-style loss on the generated movements after applying forward kinematics (FK) with fixed bone length:

$$\mathcal{L}_{FK} = ||\text{FK}(\hat{\mathbf{r}}_0) - \mathbf{p}_0||_2^2. \tag{6}$$

$$\mathcal{L}_{cond} = ||\hat{\mathbf{X}}_0|_{j\in J} - \mathbf{X}|_{j\in J}||_2^2 + ||\hat{\mathbf{r}}_0|_{\text{head}} - \mathbf{C}_r^{\text{head}}||_2^2 \tag{7}$$

$$+ ||\{\hat{\mathbf{x}}_0\}_{i=1}^{n_c} - \mathbf{C}_{ctx}||_2^2. \tag{8}$$

$\mathcal{L}_{sliding}$ is a standard foot sliding loss based on the predicted contact labels as used in [TRG*23]. During training, the trajectory conditions are directly obtained from the ground truth animal dataset.

3.3. Human to Quadruped Motion Transfer

At inference time, different conditioning and inpainting strategies can be applied to influence the quadruped motion. While the trained conditions can be used to map shared features between human and quadruped motion, inpainting enables a looser control by specifying partial abstract or imprecise mappings of motion features between the two domains. Additionally, due to the inference-based nature of inpainting, this control can be flexibly switched on or off based on the user's preference.

3.3.1. Precise Feature Mapping with Conditions

Given the human motion $\mathbf{X}_h$, the target 2D root position and direction of a future frame T are obtained and provided to the first stage model as a condition. A scaling factor γ can be applied on the target 2D position to obtain different action mappings according to the user preference. For instance, scaling up a human walk leads to a running quadruped, while scaling down will keep the quadruped walking. The resulting joint trajectories from the first stage, along with human head orientation are passed as conditions to the second stage model. For controlling the head direction C_r^{head}, the human head orientations $C_{r,h}^{\text{head}}$ are mapped to the orientation of quadruped head via a fixed retargeting offset.

3.3.2. Imprecise Feature Mapping with Inpainting

To enable additional control, such as limb puppeteering or the height of the quadruped for sitting and jumping, we leverage an inpainting strategy. This overwrites selective components $\mathbf{z}$ of the motion in the first diffusion stage in order to steer the generative process towards these target movements. At each diffusion timestep t, the corresponding components of the noisy sample are overwritten by the noised reference signal.

$$\tilde{\mathbf{X}}_t|_z = \begin{cases} \sqrt{\bar{\alpha}_t}\mathbf{z} + \sqrt{1-\bar{\alpha}_t}\epsilon, & \text{if } t \geq \tau \\ \sqrt{\bar{\alpha}_t}\hat{\mathbf{X}}_0|_z + \sqrt{1-\bar{\alpha}_t}\epsilon, & \text{if } t < \tau \end{cases} \quad \text{where } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{9}$$

where $|_z$ denotes the subset of features to be inpainted. This enforces consistency with the reference signal while allowing the remaining degrees of freedom to be generated by the model. To balance control and naturalness, the inpainting mask is only active during early diffusion steps and lifted in later diffusion steps $< \tau$. Thus, rough structural features, which are generated early on in the diffusion process, can be controlled while later diffusion steps maintain realism by matching to the quadruped motion distribution. In practice, we find that inpainting the first 80% denoising steps strikes a good balance between control and naturalness.

In the following, we describe how to transfer different human actions and movements to the desired quadruped movements within our inpainting framework.

Sit, lie, and jump: Sitting, lying, and jumping behaviors require information about root height. To transfer these motions, we inpaint the height of the root trajectory during the diffusion process of the trajectory model. Due to different scales, the target root height is obtained by applying a scale and offset to the human root height,

$$z_q^{root} = \frac{M_q - m_q}{M_h - m_h}(z_h^{root} - m_h) + m_q. \tag{10}$$

In the above equation, m and M denote the walking and sitting root height z^{root} in the global frame, respectively. For the quadruped, m_q and M_q are computed from dataset statistic using the median and minimum height, respectively. For human, m_h and M_h can be similarly computed from statistics of the dataset, or set manually.

In practice, different offset values are required depending on the type of source motion. For example, human sitting motions may occur either on elevated surfaces (e.g., chairs) or directly on the ground, leading to different relative height offsets. We leverage this distinction to map different human sitting styles to different quadruped behaviors: sitting on a chair is mapped to a quadruped sitting posture, whereas sitting on the ground is mapped to a lying-down behavior.

For jumping, we replace M to correspond to the maximum heights of human or quadruped jumping. In practice, we observed that inpainting only the root height provides a weak constraint, as it represents a small subset of the trajectory features and can be ignored by the diffusion model. We observe a similar issue when trying to inpaint one feature in a single-stage diffusion, as will be discussed in Section 4.5. Thus, to influence the generation stronger, we additionally inpaint the heights of the remaining joints z_q^e in the trajectory model by propagating the root height offset, effectively dragging other joints with the root as it starts to jump,

$$z_q^e = \begin{cases} \hat{z}_q^e + (z_q^{\text{root}} - \hat{z}_q^{\text{root}}) & \text{if } z_q^{\text{root}} > m_q + \varepsilon \\ \hat{z}_q^e & \text{if } z_q^{\text{root}} < m_q + \varepsilon \end{cases}, \tag{11}$$

where $\hat{z}$ denotes the predicted height from the previous diffusion step and ε is a tolerance threshold. To accommodate the discrepancy in jumping characteristics, quadrupeds typically cover greater horizontal distances compared to their size, we scale the 2D target position by a constant factor $\gamma > 1$ during jump inpainting.

Limb puppeteer: To enable user-guided manipulation of the quadruped's limbs using human hand motion, we inpaint the heights of the corresponding limb joints. The inpainting process operates on the first-stage joint trajectories, where the vertical positions of the quadruped's limbs are adjusted based on the human hand positions. Specifically, the human hand heights are scaled and shifted to match the quadruped's limb range, producing target heights,

$$z_q^e = s z_h^e + b. \tag{12}$$

By inpainting only the height dimension, the diffusion model maintains natural joint trajectories while allowing the user to control vertical motion, effectively puppeteering the quadruped's limbs in a manner consistent with the human demonstration.

3.4. Implementation Details

All positions and orientations are expressed in the coordinate frame of the root projected on the ground of the first context frame. Furthermore, all channels are normalized with a per-joint normalization factor based on the mean and variance of that channel for the specific joint. We find this to improve inpainting compared to normalizing all joints with a single factor.

Our pipeline works in an autoregressive manner to generate long

motion sequences. At the start of a motion, if no previous frames exist, context frames are selected from a default pose (i.e., standing, sitting). Otherwise the last n_c frames of the previously generated window are used as the context. After generation, the overlapping context frames are linearly blended between the two consecutive windows. Due to the shared context, no complex blending during the diffusion process [TCL23] is required.

Both diffusion stages are trained purely on quadruped motion capture data from [ZSKS18], containing various actions (walk, run, jump, sit, lie down). This includes about 40 minutes of quadruped motion, which is small compared to datasets typically used for generative models, which range from tens to hundreds of hours [GZZ*22; RPY*26]. The model architecture is a temporal U-Net as in [CTR*24] that operates on 32-frame windows with 2-frame context conditioning for temporal coherence and we keep the remaining architecture unchanged. Choosing a long window size may result in ambiguous behaviors as the human may turn left or right and come to the same target point, whereas a very small window makes the generation more restrictive due to context frames.

We use 1000 diffusion steps during training and DDIM sampling [SME21] with 25 steps during inference. The combined inference time of both diffusion stages per 32-frame window is 203 ms on a single RTX 4090. In an offline setting, this allows a 10-second sequence to be generated in approximately 4 seconds. For online applications, the system can generate motions in real-time with a latency of 0.7 s, consisting of 0.5 s of future-target context and 0.2 s of computation time. The models are optimized with AdamW optimizer [LH19] with adaptive learning rate and batch size of 128. The training takes roughly 15 and 21 hours for the first and second stage models on a single Nvidia GeForce RTX 3090 GPU.

The mapping parameters for inpainting are summarized in Table 1. Since these parameters are applied only at inference time, tuning and iterating on their values is fast in practice. Once a suitable set of parameters has been selected, it can be fixed and reused, enabling users to simply choose a behavior mode (e.g., sitting, jumping, or puppeteering) and directly apply the corresponding preset without further adjustment.

Table 1: *Mapping parameters for different motion categories.*

Category	Parameter	Value	Parameter	Value
General	m_q	0.47	m_h	0.92
Jump	M_q	1.24	M_h	1.32
Sit & Lie	M_q	0.05	M_h	0.46
Puppeteer	s	0.77	b	-0.68

4. Results

In this section, we present several experiments using our proposed approach for transferring various human motion to quadruped motions. The human motions are obtained either from the Humanloco dataset [SZKS19], or from our captured motion videos which are converted to skeleton format using a markerless motion capture system [BMBG25]. We refer readers to the supplementary video for a more comprehensive demonstration of the results.

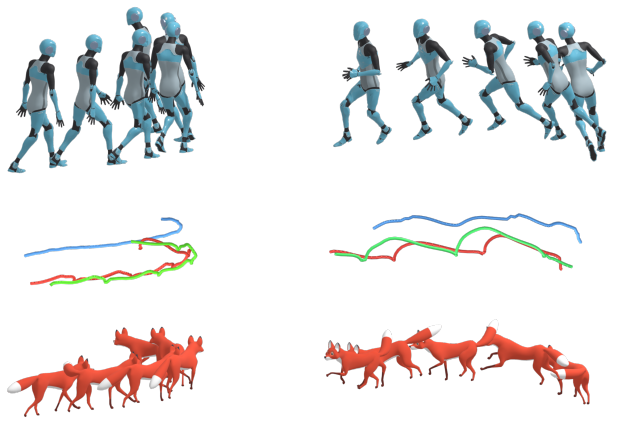

Figure 3: *Retargeting Locomotion. Top: Human source motion. Middle: Key trajectories generated by trajectory model. Bottom: Quadruped motion generated by full-body model.*

4.1. Locomotion and Head Control

Our two-stage model can map human locomotion to quadruped motions by direct conditioning. Figure 3 shows two examples where human walking and running motions are transferred to quadruped walking and running, respectively. A scaling factor of $\gamma = 0.8$ is used to adjust the 2D target location from human data. By varying this scaling factor, one can achieve different behaviors for the quadruped (e.g., different gait patterns), or make the motion compatible for various quadruped sizes as shown in our supplementary video.

Look-at control is particularly important in movie production [VFX25] when acting out the interactions between characters or the scene. As shown in the supplementary video, in the absence of explicit conditioning, the quadruped tends to look toward the ground, reflecting the dominant look-at direction in the training data. Despite the discrepancy between human and quadruped motion domains, the model successfully adapts these signals, producing motions that follow the intended look-at direction, as illustrated in Figure 4.

4.2. Jumping, Sitting, and Lying

Our method leverages inpainting-based control to transfer high-level human actions to quadruped motion. Using the jump inpainting strategy along with $\gamma = 1.5$, our method enables converting human jumping motions to corresponding quadruped jumps, as depicted in Figure 5. For sitting and lying experiments, we employ the scaling values detailed in Table 1. This maps human sitting on a chair to quadruped sitting, and human sitting on the ground to quadruped lying, as shown in Figure 6. By modifying this scale, one can also map human sitting on ground to quadruped sitting if desired. We note that jumping or sitting inpainting strategies can remain active during the locomotion phase without disrupting the motion.

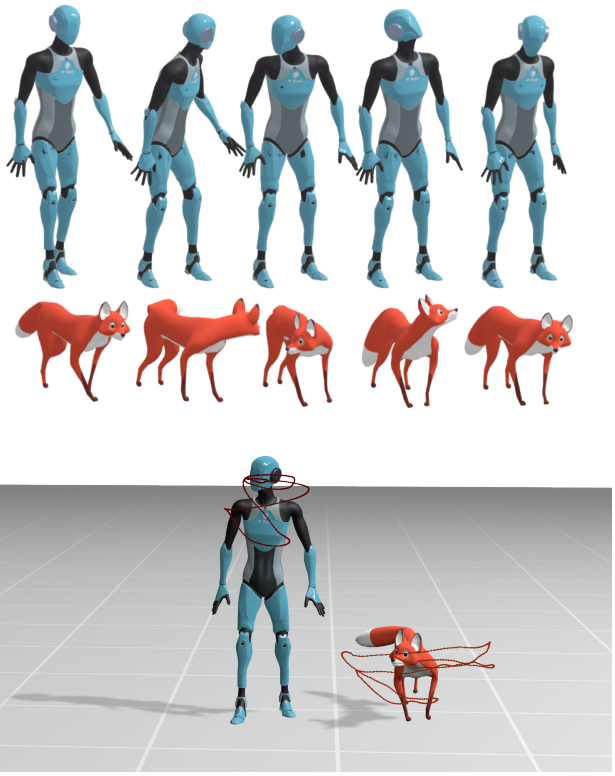

Figure 4: *Our second stage full-body model provides direct control of the quadruped's head movements, given the actor's head orientation. Top images show a few snapshots at specific times, bottom shows a full trajectory of the head joint.*

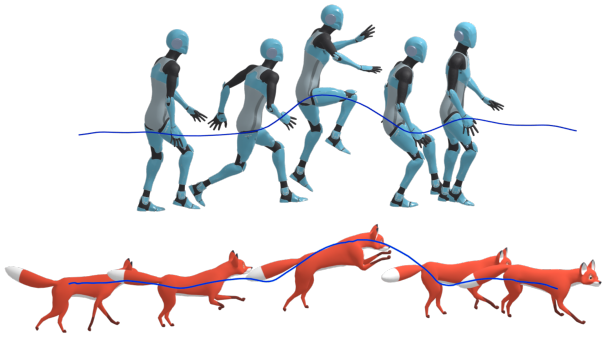

Figure 5: *An example jumping motion being produced via inpainting in the trajectory model.*

4.3. Limb Puppeteering

As shown in Figure 7 and our supplementary video, the limb puppeteering inpainting can be used to faithfully control the front legs of the quadruped. This allows for an intuitive fine-grained control while maintaining realism. Notably, when the human raises a single arm, the quadruped responds by "giving a paw", naturally transitioning into a sitting posture for stability. When both arms are raised, the quadruped briefly rears up, demonstrating coherent and context-aware whole-body adaptation. As additionally shown in the supplementary video, the mapping can also be done from another joint source, such as the legs.

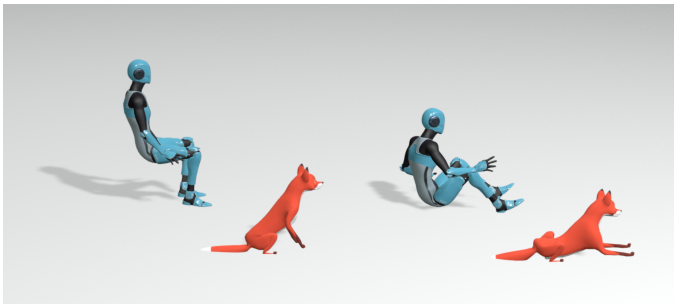

Figure 6: *An example of our sitting inpainting strategy: "sitting on a chair" gets mapped to quadruped sit (left) and "sitting on the ground" is mapped to lying down (right).*

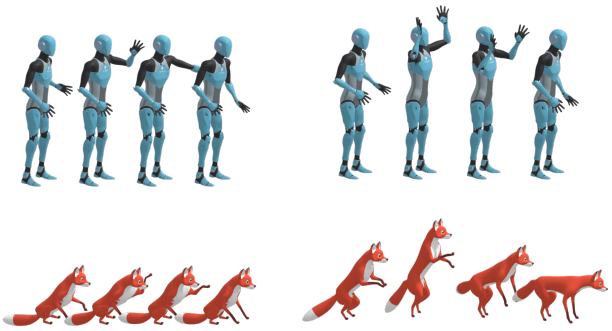

Figure 7: *Examples of limb puppeteering via inpainting.*

4.4. Baseline Comparison

In this section, we compare our method with state-of-the-art baselines most relevant to human-to-quadruped motion transfer.

Qualitatively, we compare our method with *Dog Code* [EJS*24], *Walk-the-Dog* [LSYS24] and *Motion2Motion* [CZY*25], as summarized in Table 4.4. We asked two professional artists to rate the realism of the resulting motion for all approaches. **Realism (Local)** refers to motion quality when viewed in local space (zero root movement). **Realism (Global)** refers to motion quality when considering the root movement in global space. **Semantic Alignment** denotes whether high-level motion intent is preserved from the source motion, e.g., jumping mapped to jumping and walking mapped to walking. **Direct control** refers to the ability to control specific features of the dog motion from human motion, such as head orientation or direct limb puppeteering.

Dog Code shows example of intuitive human to quadruped motion transfer with direct control, however, as seen in their supple-

mentary video (2:10-2:45), the generated dog motion shows issues regarding realism and foot sliding during locomotion.

Motion2Motion lacks mechanisms to include global information, leading to foot sliding artifacts. Furthermore, its evaluation is limited to short 2-second motion clips. When applied to human-to-quadruped motion transfer on a larger and more diverse motion dataset, it fails to preserve semantic consistency.

Walk-the-Dog produces realistic motions when viewed in a local frame. However, since it is not designed to incorporate global root trajectory information during motion alignment, the resulting motion exhibits noticeable foot sliding when driven with the human root, rendering it unusable for global steering. Furthermore, because *Walk-the-Dog* relies on completely unsupervised semantic alignment between human and quadruped motions, it can yield misaligned outputs for a given latent encoding. For instance, matching a dog jumping motion to a human run, as shown in our supplementary video.

In contrast, *Two2Four* explicitly models global displacement while maintaining naturalness due to the inductive bias of the diffusion model, resulting in overall better realism. It also offers inference-time inpainting techniques that enable direct control of the generated quadruped motion.

Table 2: *Qualitative comparison of SOTA human-to-quadruped motion transfer methods.*

Approach	Realism (local / global)	Semantic alignment	Direct control
Walk-the-Dog	✓/ ✗	✗	✗
Motion2Motion	✓/ ✗	✗	✗
Dog Code	✗/ ✗	✓	✓
Ours	✓/ ✓	✓	✓

In Table 3, we quantitatively compare our results with *Walk-the-Dog* on a subset of *HumanLoco* [SZKS19] dataset. *Walk-the-Dog* copies the 2D root position and yaw angle from human trajectory, while getting the height, roll, pitch and other joint rotations from the motion matching results. For a fair comparison, we use the scaling factor $\gamma = 1$ for this evaluation. We evaluate the motion quality using Fréchet Inception Distance (FID) [HRU*17] computed from a variational auto-encoder (VAE) trained on the same dataset. Our method shows better motion quality as measured by FID and significantly less foot sliding. When driving our framework with dog motions (reconstruction), FID score and foot sliding values remain close to ground truth data, as reported in Table 3. The large gap in FID values between retarget (human-to-quadruped) and reconstruct (quadruped-to-quadruped) is due to the different distributions of the datasets used for deriving the motion.

4.5. Ablations

In the following, we ablate the key design choices of our approach, such as the two-stage nature, the sparsity of the locomotion target, and the inpainting strategy.

Table 3: *Quantitative analysis of the motion quality. Our method shows better FID and foot sliding compared to Walk-the-Dog.*

Model	FID ↓	Foot sliding ↓
GT validation	0.6838	0.055
Ours (Reconstruct)	0.9668	0.059
Walk-the-dog (Retarget)	14.734	0.626
Ours (Retarget)	**8.185**	**0.076**

4.5.1. One- vs. Two-stage Diffusion

We compare our method with a single stage diffusion model (1SD) trained to predict full quadruped motion directly from 2D target position and direction. We observe that in challenging locomotion scenarios, the two-stage model shows better alignment with human motion. An example of the alignment issue is depicted in Figure 8. When the human performs sharp turns, our model can consistently exhibit similar turn behavior, while the single stage model fails to keep the original turn direction.

More importantly, we observe that inpainting individual joints is less effective for 1SD, as shown in Figure 9. A similar effect is observed when the trajectory model is trained to predict five joints instead of three. We attribute this to diminishing signal-to-noise ratio when the diffusion model has to denoise more data.

4.5.2. Sparse Goal vs. Full Trajectory

In the following, we ablate choosing a sparse root target at frame T against having the full root trajectory as condition to the trajectory model. For simple motion sequences, both approaches perform comparably when transferring human motion to quadrupeds. However, for more challenging cases, such as sharp in-place turns, the full conditioned model struggles to follow the motion and produces noticeable artifacts, as shown in Figure 10. In contrast, the sparse-target model remains stable and follows the motion even in such challenging cases. We attribute this to the condition signal getting too far out-of-distribution for the model to recover, a problem that is more prominent for models trained on small datasets such as ours.

4.5.3. Inpainting Schedule

To account for the imprecise mapping of some motion features between human and quadruped, a key component in our inpainting control is to apply it only until a fixed percentage of diffusion time steps, in our case 80%. As shown in the supplementary video, varying this threshold for inpainting leads to different results. Fewer inpainting steps give more flexibility to the diffusion model, but may weaken the adherence to the intended control signal. Conversely, inpainting all diffusion steps can introduce artifacts and reduce naturalness. For instance, in jumping motions, full inpainting leads to excessive acceleration and missing the proper steps for the jump. In contrast, in our setting the quadruped takes an extra step to prepare for the jump and the jump is smoother.

5. Discussion

Compared to optimization-based frameworks with rigid mappings, the diffusion process enables more flexible mapping to the target

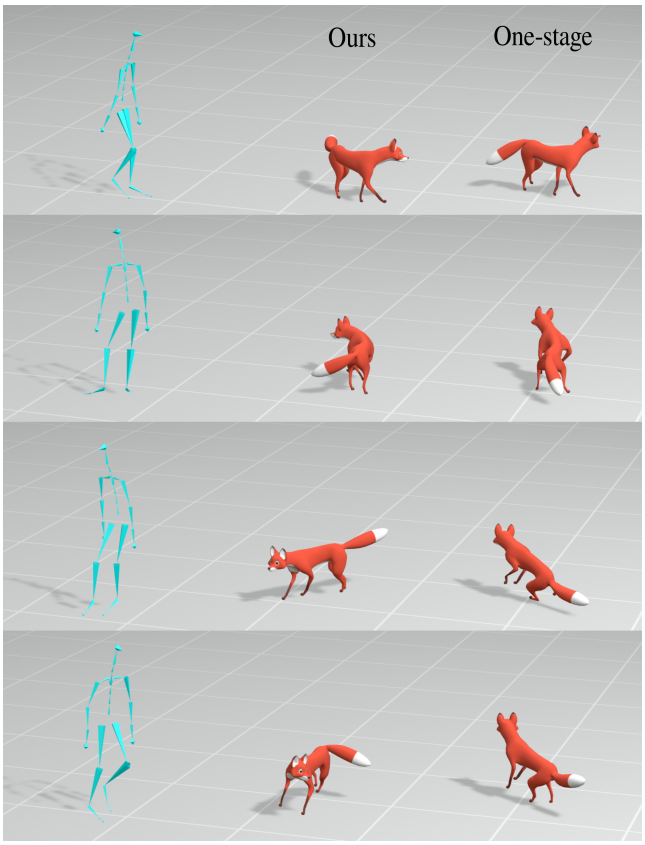

Figure 8: *Our result follows the root direction of the human better while the single stage model fails to turn in a similar manner.*

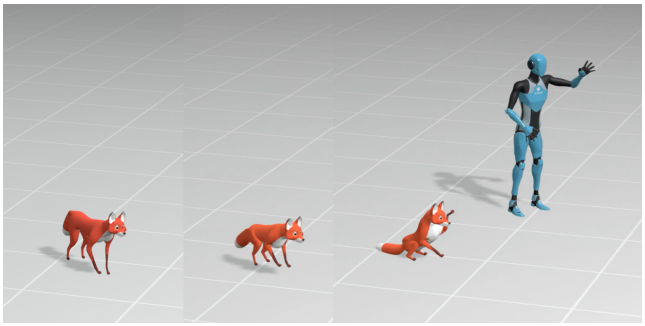

Figure 9: *Inpainting front limb trajectories is effective when diffusion model is predicting fewer signals. Single stage diffusion 1SD (left) and two-stage model predicting five joints (middle) fail to follow human hand motion, while our two-stage model predicting three joints (right) can successfully be directed.*

distribution, as precise features can be concretely mapped as conditions and imprecisely mapped motion features can be weakly transferred via inpainting. Crucially, we condition the generator on a sparse goal position rather than a full body trajectory. Because quadrupeds cannot turn in place like humans, over-constraining the trajectory leads to unnatural movement; sparse conditioning avoids

Figure 10: *A model trained on full trajectory as condition (right) leads to artifacts compared two our sparse target condition (left).*

this, allowing for realistic quadrupedal motion to the human target location. In contrast, the head orientation trajectory can be seamlessly transferred without artifacts. Hence, we condition our second stage on ground-truth quadruped head orientation during training and use the corresponding human look-at signal during inference.

Moreover, at early diffusion steps, the out-of-domain inpainted features, such as root height, are sufficiently distorted by the large added noise to resemble the in-distribution features. Hence, the generative model synthesizes motion conforming to these imprecise features while maintaining realism as long as the inpainting is dropped at lower noise levels. Besides inpainting, other forms of inference-based influence, such as guidance [KPST23] and noise optimization [KPA*24], have been explored for motion generation. However, compared to our inpainting strategy, they can come with a larger computational cost which is more challenging for a real-time use case, such as live puppeteering in virtual production.

When training on small datasets such as MANN [ZSKS18], we observe that increasing the dimensionality of the features, i.e. the number of trajectories generated, in the first stage diminishes inpainting effectiveness. We hypothesize that as more joint trajectories are introduced, the inter-feature correlations increase. Hence, rather than utilizing the inpainted signal, the model opts to regenerate the trajectory based on the remaining highly correlated features. Consequently, we observe that the first-stage model generating three trajectories (hip and front feet) outperforms one generating five (hip and all feet), which in turn outperforms a single-stage model generating all joint trajectories.

Although our evaluation focuses on dogs, the proposed framework is largely species-agnostic and does not incorporate any dog-specific modeling assumptions. Consequently, given sufficient motion data, Two2Four can be adapted to other quadruped species with minimal changes to the overall pipeline. In practice, however, the availability of high-quality quadruped motion datasets remains limited, with canine datasets being the most accessible public resources.

6. Conclusion

We introduced Two2Four, a two-stage diffusion model along with inference-based inpainting strategies to generate quadruped motion driven by human motion. We showed that a two-stage diffusion approach results in higher-quality movements than a single-stage model, particularly when driven by out-of-distribution human data.

Our method shows high quality motions and can imitate different actions including locomotion, jumping, sitting, lying, as well as individual limb movements. However, due to the lack of backward quadruped motion in our dataset, our model struggles when the human is walking backwards for more than a few seconds. In this case, the quadruped sits down and shows sliding artifacts. Nevertheless, due to the sparse goal condition mechanism and our autoregressive rollout scheme, the quadruped can recover and catch up to the human source after a while.

In this work, we focus on enabling multiple control modes for human-to-quadruped motion transfer. Currently, these modes are determined by the user. Designing an automated system that switches between these modes (for example by setting some thresholds), or finding a unified representation for these modes remains a design challenge to be explored in the future. Additional future directions could also explore Bézier motion models as in [VAG*25] to reduce the data dimensionality in order to include more joints in the first stage, thus enabling more user control of the motion, or to allow for artist-friendly editing of the quadruped motion.

Acknowledgment

The authors would like to thank Dominik Borer for his technical support and Violaine Fayolle and Doriano van Essen for their artistic support during this project.

References

[AI25] AGATA, NAOKI and IGARASHI, TAKEO. "Motion Control via Metric-Aligning Motion Matching". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 2025, 1–12 2.

[ALL*20] ABERMAN, KFIR, LI, PEIZHUO, LISCHINSKI, DANI, et al. "Skeleton-Aware Networks for Deep Motion Retargeting". *ACM Transactions on Graphics* 39.4 (2020), 62:1–62:14 3.

[BMBG25] BUHMANN, JAKOB, MOORE, DOUGLAS L., BORER, DOMINIK, and GUAY, MARTIN. "Synth2Track Editor for Efficient Match-Animation". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks*. SIGGRAPH Talks '25. Association for Computing Machinery, 2025 6.

[CTR*24] COHAN, SETAREH, TEVET, GUY, REDA, DANIELE, et al. "Flexible motion in-betweening with diffusion models". *ACM SIGGRAPH 2024 conference papers*. 2024, 1–9 4, 6.

[CZY*25] CHEN, LING-HAO, ZHANG, YUHONG, YIN, ZIXIN, et al. "Motion2motion: Cross-topology motion transfer with sparse correspondence". *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 2025, 1–11 2, 7.

[EJS*24] EGAN, DONAL, JOVANE, ALBERTO, SZKARADEK, JAN, et al. "Dog Code: Human to Quadruped Embodiment Using Shared Codebooks". *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games*. 2024, 1–11 3, 7.

[GFK*23] GRANDIA, RUBEN, FARSHIDIAN, FARBOD, KNOOP, ESPEN, et al. "DOC: Differentiable Optimal Control for Retargeting Motions onto Legged Robots". *ACM Trans. Graph.* 42.4 (July 2023). ISSN: 0730-0301. DOI: 10.1145/3592454 2.

[Gle98] GLEICHER, MICHAEL. "Retargetting Motion to New Characters". *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. ACM, 1998, 33–42 2.

[GRT*25] GAT, INBAR, RAAB, SIGAL, TEVET, GUY, et al. "Anytop: Character animation diffusion with any topology". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 2025, 1–10 3.

[GZZ*22] GUO, CHUAN, ZOU, SHIHAO, ZUO, XINXIN, et al. "Generating Diverse and Natural 3D Human Motions From Text". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, 5152–5161 6.

[HBLK25] HWANG, INWOO, BAE, JINSEOK, LIM, DONGGEUN, and KIM, YOUNG MIN. "Motion Synthesis with Sparse and Flexible Keyjoint Control". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025 3, 4.

[HJA20] HO, JONATHAN, JAIN, AJAY, and ABBEEL, PIETER. "Denoising Diffusion Probabilistic Models". *Advances in Neural Information Processing Systems*. Vol. 33. 2020, 6840–6851 3.

[HRU*17] HEUSEL, MARTIN, RAMSAUER, HUBERT, UNTERTHINER, THOMAS, et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 8.

[KPA*24] KARUNRATANAKUL, KORRAWE, PREECHAKUL, KONPAT, AKSAN, EMRE, et al. "Optimizing Diffusion Noise Can Serve as Universal Motion riors". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 1334–1345 3, 9.

[KPST23] KARUNRATANAKUL, KORRAWE, PREECHAKUL, KONPAT, SUWAJANAKORN, SUPASORN, and TANG, SIYU. *GMD: Controllable Human Motion Synthesis via Guided Diffusion Models*. 2023 3, 9.

[KTCB24] KAPON, ROY, TEVET, GUY, COHEN-OR, DANIEL, and BERMANO, AMIT H. "MAS: Multi-view Ancestral Sampling for 3D Motion Generation Using 2D Diffusion". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 1965–1974 3.

[LAZ*22] LI, PEIZHUO, ABERMAN, KFIR, ZHANG, ZIHAN, et al. "GANimator: Neural Motion Synthesis from a Single Sequence". *ACM Transactions on Graphics (TOG)* 41.4 (2022), 1–12 3.

[LCC19] LIM, JONGIN, CHANG, HYUNG JIN, and CHOI, JIN YOUNG. "PMnet: Learning of Disentangled Pose and Movement for Unsupervised Motion Retargeting". *Proceedings of the British Machine Vision Conference (BMVC)*. Vol. 2. 2019, 7 3.

[LDR*22] LUGMAYR, ANDREAS, DANELLJAN, MARTIN, ROMERO, ANDRES, et al. "Repaint: Inpainting using denoising diffusion probabilistic models". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 11461–11471 2.

[LH19] LOSHCHILOV, ILYA and HUTTER, FRANK. "Decoupled Weight Decay Regularization". *International Conference on Learning Representations (ICLR)* (2019) 6.

[LJG*24] LI, TIANYU, JUNG, HYUNYOUNG, GOMBOLAY, MATTHEW, et al. "CrossLoco: Human Motion Driven Control of Legged Robots via Guided Unsupervised Reinforcement Learning". *The Twelfth International Conference on Learning Representations*. 2024 2, 3.

[LKP*23] LEE, SUNMIN, KANG, TAEHO, PARK, JUNGNAM, et al. "SAME: Skeleton-Agnostic Motion Embedding for Character Animation". *SIGGRAPH Asia 2023 Conference Papers*. 2023, 1–11 3.

[LSYS24] LI, PEIZHUO, STARKE, SEBASTIAN, YE, YUTING, and SORKINE-HORNUNG, OLGA. "WalktheDog: Cross-Morphology Motion Alignment via Phase Manifolds". *ACM SIGGRAPH 2024 Conference Papers*. 2024, 1–10 2, 3, 7.

[LWC*23] LI, TIANYU, WON, JUNGDAM, CLEGG, ALEXANDER, et al. "ACE: Adversarial Correspondence Embedding for Cross Morphology Motion Retargeting from Human to Nonhuman Characters". *SIGGRAPH Asia 2023 Conference Papers*. 2023, 1–11 3.

[MLS*25] MU, YUXUAN, LING, HUNG YU, SHI, YI, et al. "StableMotion: Training Motion Cleanup Models with Unpaired Corrupted Data". *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 2025, 1–12 3.

[Per14] PERRY, TEKLA S. "Motion Capture Technology Goes Into the Wild for Dawn of the Planet of the Apes". *IEEE Spectrum* (2014). Accessed: 2026-04-15. URL: https://spectrum.ieee.org/motion-capture-technology-goes-into-the-wild-for-dawn-of-the-planet-of-the-apes 1.

[RLT*24] RAAB, SIGAL, LEIBOVITCH, INBAL, TEVET, GUY, et al. "Single Motion Diffusion". *The Twelfth International Conference on Learning Representations*. 2024 3.

[RPY*26] REMPE, DAVIS, PETROVICH, MATHIS, YUAN, YE, et al. "Kimodo: Scaling Controllable Human Motion Generation". *arXiv preprint arXiv:2603.15546* (2026) 6.

[RTK*14] RHODIN, HELGE, TOMPKIN, JAMES, KIM, KWANG IN, et al. "Interactive Motion Mapping for Real-time Character Control". *Computer Graphics Forum* 33.2 (2014), 273–282 2.

[SAB*24] STUDER, JUSTIN, AGRAWAL, DHRUV, BORER, DOMINIK, et al. "Factorized Motion Diffusion for Precise and Character-Agnostic Motion In-Betweening". *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games*. 2024, 1–10 3.

[SME21] SONG, JIAMING, MENG, CHENLIN, and ERMON, STEFANO. "Denoising Diffusion Implicit Models". *International Conference on Learning Representations*. 2021 6.

[SOL13] SEOL, YEONGHO, O'SULLIVAN, CAROL, and LEE, JEHEE. "Creature Features: Online Motion Pppetry for Non-Human Characters". *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 2013, 213–221 2.

[SZKS19] STARKE, SEBASTIAN, ZHANG, HE, KOMURA, TAKU, and SAITO, JUN. "Neural State Machine for Character-Scene Interactions". *ACM Transactions on Graphics* 38.6 (2019), 209:1–209:14 6, 8.

[TCL23] TSENG, JONATHAN, CASTELLON, RODRIGO, and LIU, KAREN. "EDGE: Editable Dance Generation from Music". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 448–458 3, 6.

[TRG*23] TEVET, GUY, RAAB, SIGAL, GORDON, BRIAN, et al. "Human Motion Diffusion Model". *The Eleventh International Conference on Learning Representations*. 2023 2–4.

[VAG*25] VÖGELI, LUCA, AGRAWAL, DHRUV, GUAY, MARTIN, et al. "Implicit Bézier Motion Model for Precise Spatial and Temporal Control". *Proceedings of the 2025 18th ACM SIGGRAPH Conference on Motion, Interaction, and Games*. 2025, 1–10 3, 10.

[VFX25] VFX EXPRESS. *Mufasa: The Lion King – MPC's Character Lab Brings Taka to Life*. `https://www.youtube.com/watch?v=XCeJu8XeIJ4`. YouTube video, accessed April 10, 2026. 2025 1, 6.

[VYCL18] VILLEGAS, RUBEN, YANG, JIMEI, CEYLAN, DUYGU, and LEE, HONGLAK. "Neural Kinematic Networks for Unsupervised Motion Retargetting". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, 8639–8648 3.

[WLSN24] WINQUIST, ERIK, LEONHARDT, PHILLIP, STORY, PAUL, and NOWOTNY, ALEX. "A New Kingdom: Weta FX Returns to The Planet of The Apes". *ACM SIGGRAPH 2024 Talks*. 2024, 1–2 1.

[WRQ*26] WANG, XUAN, RUAN, KAI, QIAN, LIYANG, et al. "X-MoGen: Unified Motion Generation Across Humans and Animals". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. 12. 2026, 10234–10242 3.

[YZS*24] YANG, ZHANGSIHAO, ZHOU, MINGYUAN, SHAN, MENGYI, et al. "OmniMotionGPT: Animal Motion Generation with Limited Data". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 1249–1259 3.

[ZBL*19] ZHOU, YI, BARNES, CONNELLY, LU, JINGWAN, et al. "On the Continuity of Rotation Representations in Neural Networks". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, 5745–5753 4.

[ZLW*24] ZHAO, QINGQING, LI, PEIZHUO, WANG, YIFAN, et al. "Pose-to-Motion: Cross-Domain Motion Retargeting with Pose Prior". *Computer Graphics Forum*. Vol. 43. Wiley Online Library, 2024, e15170 2, 3.

[ZSKS18] ZHANG, HE, STARKE, SEBASTIAN, KOMURA, TAKU, and SAITO, JUN. "Mode-Adaptive Neural Networks for Quadruped Motion Control". *ACM Transactions on Graphics (ToG)* 37.4 (2018), 1–11 2, 3, 6, 9.