This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

AniMo: Species-Aware Model for Text-Driven Animal Motion Generation

Xuan Wang[1*] Kai Ruan[2*] Xing Zhang[3] Gaoang Wang[1✉]

[1] Zhejiang University [2] Renmin University of China [3] New Hope Liuhe Co., Ltd.

xuanw@zju.edu.cn, kairuan@ruc.edu.cn, zhangxing@newhope.cn, gaoangwang@intl.zju.edu.cn

[*] Equal contribution [✉] Corresponding author

https://github.com/WandererXX/AniMo

Abstract

*Text-driven motion generation has made significant strides in recent years. However, most existing works focus on human motion, largely overlooking the rich and diverse behaviors of animals. Understanding and synthesizing animal motion have important applications in wildlife conservation, animal ecology, and biomechanics. Animal motion modeling presents unique challenges due to species diversity, varied morphological structures, and different behavioral patterns in response to similar textual descriptions. To address these challenges, we propose **AniMo** for text-driven animal motion generation. AniMo consists of two stages: motion tokenization and text-to-motion generation. In the motion tokenization stage, we encode motions using a joint-aware spatiotemporal encoder with species-aware feature modulation, enabling the model to adapt to diverse skeletal structures across species. In the text-to-motion generation stage, we employ masked modeling to jointly learn the mapping from textual descriptions to motion tokens. Additionally, we introduce **AniMo4D**, a large-scale dataset containing 78,149 motion sequences and 185,435 textual descriptions across 114 animal species. Experimental results show that AniMo achieves superior performance on both the AniMo4D and AnimalML3D datasets, effectively capturing diverse morphological structures and behavioral patterns across animal species.*

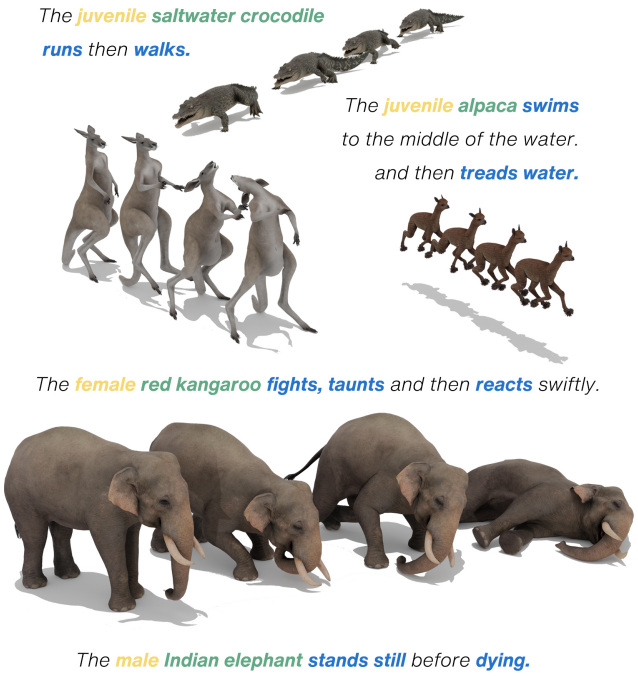


Figure 1. Examples of animal motions. Animals exhibit more diverse morphological structures and behavior patterns, making it challenging to generate motions from text-only prompts. (Yellow represents animal attributes, Green represents animal species, and Blue represents motions.)

1. Introduction

Text-to-motion generation, which transforms textual descriptions into semantically consistent and realistic 3D motions, has attracted significant attention recently. This technology has diverse applications in domains such as robotics, virtual/augmented reality (VR/AR), video games, and content creation [65]. While existing methods succeed in generating realistic motion [12, 27, 34, 47, 56, 61], they predominantly focus on human behavior, largely ignoring the rich and diverse motion patterns exhibited by animals. Expanding motion generation beyond human holds significant potential for animal management, wildlife conservation, ecological research, and biomechanics [9, 30, 39, 41].

In this work, we focus on animal motion, an area that remains relatively under-explored compared to human motion. As shown in Figure 1, modeling animal motion presents unique challenges due to the vast diversity in

species morphology and behavioral patterns. Unlike the relatively uniform human skeletal structure, animals display diverse body forms, from quadrupeds to birds, each with distinct motion patterns. Furthermore, while human behaviors tend to follow consistent patterns, animal behaviors exhibit significant variability influenced by species, morphology, and environmental factors. For example, "walking" differs fundamentally between crocodiles and giraffes due to their distinct limb structures and dynamics.

To address these challenges, we introduce **AniMo**, a model specifically designed for text-driven animal motion generation. AniMo follows a two-stage framework: motion tokenization and text-to-motion generation. During motion tokenization, we discretize motion data using an encoder-decoder architecture equipped with a quantization mechanism. Unlike existing methods [12, 15, 54], which encode motion sequences using a Temporal Convolutional Network (TCN) [21], we adopt a joint-aware spatiotemporal encoder to better capture spatial and temporal dynamics. Furthermore, to accommodate species-specific morphology, we introduce species-aware feature modulation, which adaptively adjusts skeletal structures based on the animal species and attributes. In the text-to-motion generation stage, we condition the model on textual descriptions, along with animal species and attributes, to establish a mapping from textual descriptions to motion tokens.

Additionally, we introduce **AniMo4D**, a large-scale dataset with 78,149 motions and 185,435 textual descriptions across 114 species. We conduct experiments on AniMo4D and use the AnimalML3D dataset [52] as an out-of-distribution (OOD) test set. AniMo achieves strong performance across multiple metrics, demonstrating its ability to generate diverse, realistic animal motions from text.

In summary, our key contributions are as follows:

- We propose AniMo, a novel model using a joint-aware spatiotemporal encoder to capture both spatial and temporal features of animal motion, along with species-aware feature modulation for adaptive scaling across species.
- We introduce AniMo4D, a comprehensive dataset comprising 78,149 motion sequences and 185,435 annotated textual descriptions spanning 114 animal species.
- We conduct extensive experiments on the AniMo4D and AnimalML3D datasets, demonstrating the superiority of AniMo. Our quantitative evaluations show that AniMo effectively generates species-specific morphological structures and motion sequences that align with the behavioral patterns of each species.

2. Related Work

2.1. Text-driven Human Motion Generation

Numerous studies attempt to generate motion from text, primarily due to its user-friendliness. Existing text-to-motion generation approaches can be broadly categorized into three types: diffusion-based, autoregressive, and masked modeling methods. Diffusion methods describe motion generation as a text-conditioned denoising process [2, 4, 16, 18, 19, 25, 40, 47, 49, 55, 56, 60, 62–64]. The latter two approaches first encode motions into discrete tokens, which are then reconstructed. Autoregressive methods [3, 15, 27, 50, 54, 61] use a GPT-like [37] approach. For instance, T2M-GPT [54] encodes motions using Vector Quantized Variational Autoencoders (VQ-VAE) [48] and conditions on text to guide token generation. AttT2M [61] models individual body parts, incorporating spatial information. Masked modeling methods [12, 32–34, 58] follow a BERT-like [7] approach. MMM [34] encodes motions with VQ-VAE, constructs a large-scale codebook with a factorized code method [28], and employs a Masked Transformer to predict masked tokens. MoMask [12] utilizes Residual Vector Quantization (RVQ) [22, 53] for quantization, with a Masked Transformer for predicting base layer tokens and a Residual Transformer for predicting residual layer tokens. However, existing masked modeling methods do not explicitly incorporate spatial information in motion encoding. To address this limitation, we propose a joint-aware spatiotemporal encoder that captures skeletal spatial dependencies, ensuring the generated motion tokens effectively represent spatiotemporal dynamics.

2.2. Text-driven Animal Motion Generation

Progress in animal text-to-motion generation has lagged behind that of human motion generation due to limited data. OmniMotionGPT [52] addresses this issue through transfer learning, jointly training on human and animal motion data. However, it still relies on human motion during inference, limiting purely text-driven generation. Motion Avatar [59] uses SinMDM [36] for data augmentation to create the Zoo-300K dataset. It extends MoMask [12] to model animal motion and integrates a large language model as planner with 3D generation for avatar meshes. However, these methods often overlook essential animal-specific characteristics, particularly the morphological diversity across different species. To address this limitation, we introduce species-aware feature modulation, which accommodates the diverse morphological structures of different animal species.

2.3. Animal Motion Dataset

Several studies have constructed animal motion datasets. The RGBD-Dog dataset [17] captured skeletal data from seven dogs using motion capture devices. Similarly, Shooter *et al.* [42] introduced the 3DDogs-Wild dataset using indoor motion capture systems. The DeformingThings4D dataset [23] was synthesized using Blender scripts [5] with Mixamo characters [14] to produce diverse mo-

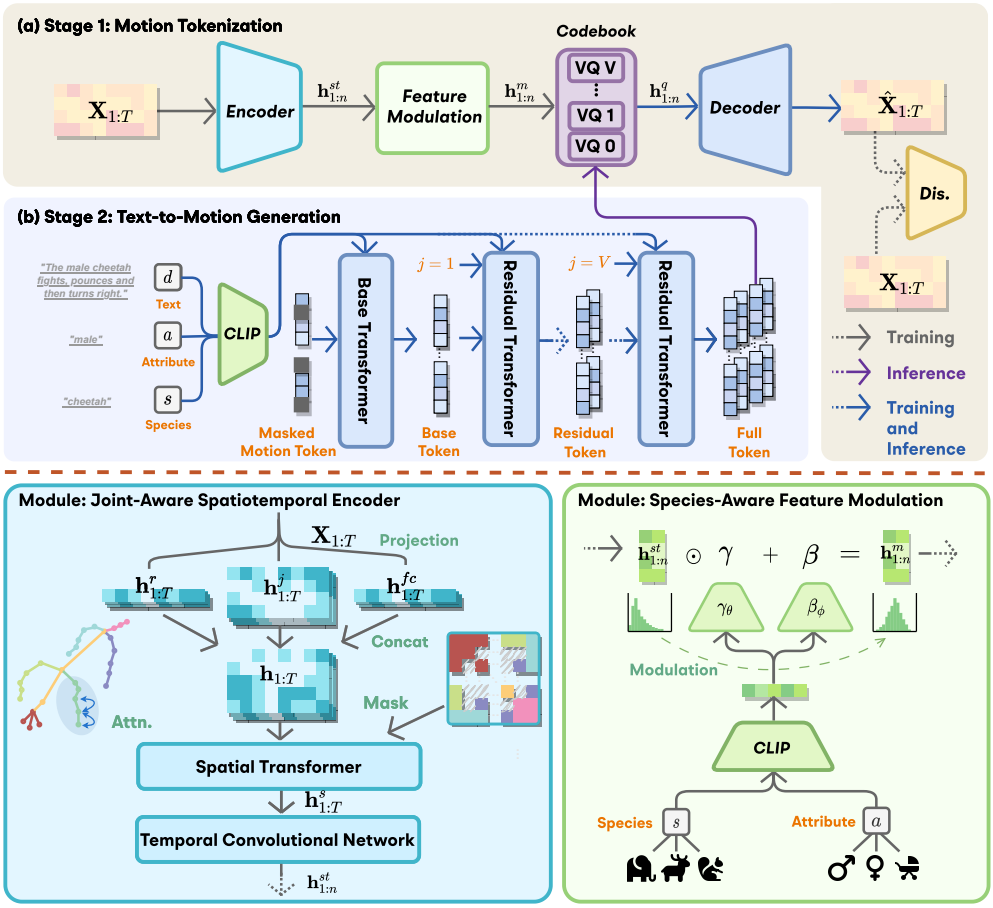


Figure 2. **The architecture of AniMo.** Our network operates in two main stages: (a) Motion Tokenization, which discretizes continuous motion representations. It adopts an encoder-decoder structure with Residual Vector Quantization. The encoder leverages a Joint-Aware Spatiotemporal Encoder to capture spatiotemporal motion features, followed by Species-Aware Feature Modulation for adaptive feature scaling. A Discriminator is employed during training to ensure realistic motion reconstruction. (b) Text-to-Motion Generation, which maps text inputs to motion tokens. A Base Transformer with masked modeling predicts the base layer tokens, while a Residual Transformer predicts the tokens for the remaining layers. (c) Inference: Given a text prompt, motion sequences that conform to animal morphology and behavior are generated sequentially using the Base Transformer, Residual Transformer, Codebook, and Decoder.

tion data. However, these datasets lack textual descriptions, limiting their applicability in text-to-motion tasks. Recent efforts have incorporated text annotations into animal motion datasets. Yang *et al.* [52] extracted skeletal data from DeformingThings4D, and Zhang *et al.* [59] expanded animal motions in the Zoo-300K dataset using SinMDM [36]. However, these datasets contain a limited number of original animal motion sequences and exhibit restricted species representation. To address these gaps, we introduce the AniMo4D dataset with 78,149 motion sequences and 185,435 textual descriptions across 114 species.

3. Method

Our goal is to generate animal motions based on textual descriptions while ensuring consistency with animal morphology and behavioral patterns. As shown in Figure 2, we propose AniMo, which comprises two training stages: motion tokenization (Section 3.1) and text-to-motion generation (Section 3.2). In the motion tokenization stage, we utilize a species-aware encoder-decoder architecture with a quantization mechanism to discretize continuous motion representations into motion tokens. In the text-to-motion generation stage, we employ a masked motion modeling mechanism to establish a mapping from text to motion tokens. During inference, textual descriptions are first mapped to discrete motion tokens, and then transformed into continuous motion sequences through the pre-trained decoder.

3.1. Stage 1: Motion Tokenization

Given a motion sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$, where $\mathbf{x}_t \in \mathbb{R}^d$, T represents the number of frames and d denotes

the motion dimension, our goal is to train an autoencoder and a learnable codebook to discretize motion embeddings while ensuring accurate reconstruction of the motion sequence $\hat{\mathbf{X}}$. Specifically, we employ Residual Vector Quantization (RVQ) [12, 53] to learn fine-grained multi-layer discrete codebooks. To enrich feature learning and capture complex motion patterns from diverse species, we introduce a joint-aware spatiotemporal encoder enhanced with species-aware feature modulation for motion encoding. Additionally, following the concept of GANs [10, 20, 57], we incorporate a discriminator to differentiate between $\mathbf{X}$ and $\hat{\mathbf{X}}$, thereby enhancing the reconstruction quality.

Joint-Aware Spatiotemporal Encoder. In widely used motion representations [11], $\mathbf{x}_t$ is represented as a feature vector in a one-dimensional format. Most existing methods process $\mathbf{x}_t$ directly through the encoder [15, 54] without explicitly modeling its spatial feature. To address this limitation, we propose a joint-aware spatiotemporal encoder designed to explicitly capture both spatial and temporal dependencies in motion data. For spatial motion encoding, we reorganize the motion data with n_j joints into a joint-aware representation: $\left[\mathbf{x}_t^r; \mathbf{x}_t^j; \mathbf{x}_t^f\right]$, where $\mathbf{x}_t^r$, $\mathbf{x}_t^j$, and $\mathbf{x}_t^f$ denote the root joint, non-root joints, and foot-ground contact features, respectively. These components are then projected into the feature space via linear transformations:

$$\mathbf{h}_t = \left[\mathbf{h}_t^r; \mathbf{h}_t^j; \mathbf{h}_t^f\right] = \left[\mathbf{W}_r\mathbf{x}_t^r; \mathbf{W}_j\mathbf{x}_t^j; \mathbf{W}_f\mathbf{x}_t^f\right], \quad (1)$$

where $\mathbf{W}_r$, $\mathbf{W}_j$, $\mathbf{W}_f$ are learnable weight parameters.

Next, the joint-aware features $\mathbf{h}_t$ are processed by a Spatial Transformer (S-Trans). To incorporate the morphological structure of animal motion, we impose constraints on the S-Trans's attention mask. Specifically, we represent the skeleton as a tree rooted at the pelvis joint. For any joint pair (A, B), the attention mask $\mathcal{M}_{A,B}$ is defined as:

$$\mathcal{M}_{A,B} = \begin{cases} 0, & \text{if A, B are on the same limb} \\ -\infty, & \text{otherwise} \end{cases} \quad (2)$$

The S-Trans processes the motion embeddings $\mathbf{h}_{1:T}$ using self-attention with the attention mask $\mathcal{M}$ as follows:

$$\mathbf{Q} = \mathbf{h}_{1:T}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{h}_{1:T}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{h}_{1:T}\mathbf{W}_V, \quad (3)$$

$$\mathbf{h}_{1:T}^s = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathcal{M}\right)\mathbf{V}, \quad (4)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are the learnable parameters of the S-Trans. The resulting motion embeddings $\mathbf{h}_{1:T}^s$ are enhanced with spatial structural information.

Finally, the spatially enhanced features $\mathbf{h}_{1:T}^s$ processed by a Temporal Convolutional Network (TCN) [21] to capture temporal dynamics, producing the final spatiotemporal features $\mathbf{h}_{1:n}^{st}$, where $n = \lceil T/r \rceil$ and r is the downsampling rate. Without loss of generality, we denote $\mathbf{h}^{st}$ without the subscript t for simplicity.

Species-Aware Feature Modulation. Different species possess diverse morphological structures. As shown in Figure 3(c), the relative spine lengths across species can vary by orders of magnitude. While the feature representation $\mathbf{h}^{st}$ primarily encodes motion information, it also implicitly captures certain morphological characteristics. To better account for species-specific morphological variations, we propose Species-Aware Feature Modulation (SAFM). Specifically, we employ a feature-wise linear modulation strategy [31] to adaptively scale $\mathbf{h}^{st}$:

$$\mathbf{h}^m = \boldsymbol{\gamma} \odot \mathbf{h}^{st} + \boldsymbol{\beta}, \quad (5)$$

where $\odot$ denotes element-wise multiplication, and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are scaling parameters generated by combining the CLIP [38] text encoder with a Multi-Layer Perceptron (MLP). These parameters are learned by encoding animal species s along with specific animal attributes a (male, female, or juvenile):

$$\boldsymbol{\gamma}, \boldsymbol{\beta} = \text{MLP}(\text{CLIP}_{\text{text}}([a, s])). \quad (6)$$

As a result, the latent motion feature $\mathbf{h}^m$ is dynamically modulated to accommodate different species.

Residual Vector Quantization. To encode motion features, we employ a quantization mechanism $Q(\cdot)$, where a continuous vector is replaced by its nearest code entry (referred to as motion tokens) from a codebook $\mathcal{C}$. To obtain more fine-grained motion tokens, we perform iterative quantization to the latent sequence $\mathbf{h}^m$ using $V + 1$ quantization layers [12, 22, 53], where each layer successively refines the residual from the previous step.

Formally, let $v = 0, 1, \ldots, V$ denote the index of the quantization layer. We initialize the residual representation as $\mathbf{r}^0 = \mathbf{h}^m$. The quantization process is then formulated as:

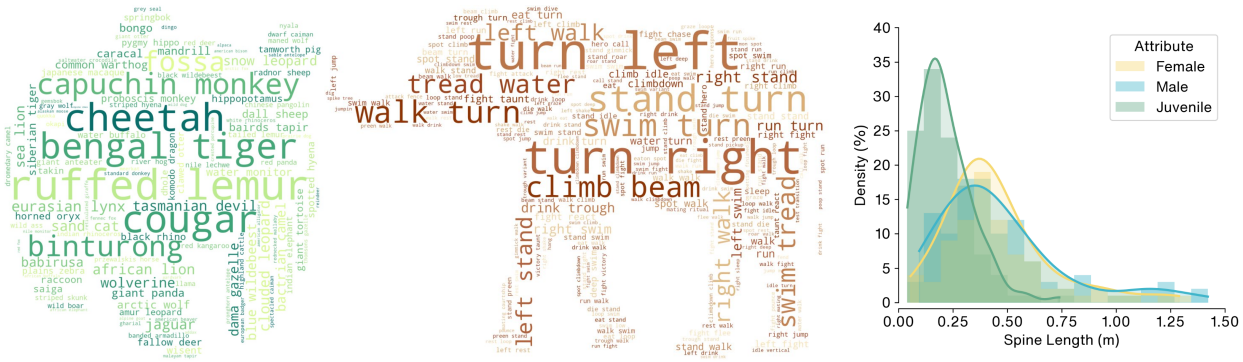
$$\mathbf{h}^v = Q(\mathbf{r}^v), \quad \mathbf{r}^{v+1} = \mathbf{r}^v - \mathbf{h}^v. \quad (7)$$

The final quantized motion representation is obtained by summing the quantized sequences across all layers:

$$\mathbf{h}^q = \sum_{v=0}^{V} \mathbf{h}^v. \quad (8)$$

During training, the codebook $\mathcal{C}$ is updated using an Exponential Moving Average (EMA) and a codebook reset strategy [51, 54], facilitating stable and high-quality representation learning.

The quantized representation $\mathbf{h}^q$ is then passed through a decoder to reconstruct the motion sequence, resulting in



(a) Word cloud of the species. (b) Word cloud of the motion labels. (c) Distribution of animal spine length.

Figure 3. Visualization results on the AniMo4D dataset.

the output $\hat{\mathbf{X}}$. The overall training objective combines a reconstruction loss and a latent embedding loss at each quantization layer:

$$\mathcal{L}_{\text{rvq}} = \|\mathbf{X} - \hat{\mathbf{X}}\|_1 + \alpha \sum_{v=1}^{V} \|\mathbf{r}^v - \text{sg}[\mathbf{h}^v]\|_2^2, \tag{9}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation, and α is a weighting factor for the embedding constraint.

To ensure the realism of the reconstructed motion, we introduce a discriminator $D(\cdot)$ that differentiates between the generated motion $\hat{\mathbf{X}}$ and the real motion $\mathbf{X}$. The adversarial loss is formulated as:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}[\log D(\mathbf{X})] + \mathbb{E}_{\hat{\mathbf{X}}\sim p_{\text{gen}}}\Big[\log(1 - D(\hat{\mathbf{X}}))\Big]. \tag{10}$$

The overall training objective combines the RVQ loss and the adversarial loss, and is expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rvq}} + \lambda \mathcal{L}_{\text{adv}}, \tag{11}$$

where λ is a hyperparameter that controls the balance between the RVQ loss and the adversarial loss.

After training, each motion sequence $\mathbf{X}$ can be represented as $V+1$ discrete motion token sequences $M = [m^v_{1:n}]^V_{v=0}$, where motion token $m^j_i \in \{1, 2, ..., |\mathcal{C}|\}$ is the index of codebook embedding.

3.2. Stage 2: Text-to-Motion Generation

In this stage, our goal is to learn the mapping from textual descriptions to pre-trained motion tokens. Since the first layer of our multi-layer codebook contains the most essential information, we use a Base Transformer (B-Trans) to predict the base layer motion tokens $m^0_{1:n} \in \mathcal{R}^n$. To condition the model, we input the animal attributes a, species s, and textual description d into the CLIP text encoder, resulting in a condition vector $\mathbf{c} = \text{CLIP}_{\text{text}}([a, s, d])$.

Table 1. Statistics of the text-motion datasets.

Dataset	#Motions	#Texts	Duration (h)	#Species
HumanML3D [11]	14,616	44,970	28.59	1
KIT-ML [35]	3,911	6,248	10.33	1
AnimalML3D [52]	1,240	3,720	0.82	36
Zoo-300K [59]	1,153	**300,000***	-	65
AniMo4D (Ours)	**78,149**	185,435	**84.29**	**114**

Note: * utilized SinMDM [36] to augment motion data to 300K, and generated 300K corresponding texts.

The base layer tokens are predicted using a masked motion modeling approach. Specifically, a portion of the tokens in the original motion sequence m^0 is replaced with [MASK] tokens according to a masking probability p_{mask}. This modified sequence $\tilde{m}^0_i$ is subsequently fed into the B-Trans, which predicts the actual tokens at the [MASK] positions. The model is trained by minimizing the negative log-likelihood of the target token predictions:

$$\mathcal{L}_{\text{base}} = \sum_{\tilde{m}^0_i = [\text{MASK}]} -\log p_\theta(m^0_i | \tilde{m}^0, \mathbf{c}), \tag{12}$$

To predict the motion tokens for the remaining V layers, we employ a Residual Transformer (R-Trans). In addition to the text condition $\mathbf{c}$, we also condition the model on the residual layer ID $j \in (1, V)$. The R-Trans predicts the tokens for the j-th quantization layer, with the following training objective:

$$\mathcal{L}_{\text{res}} = \sum_{j=1}^{V} \sum_{i=1}^{n} -\log p_\phi(m^j_i | m^{1:j-1}_i, \mathbf{c}, j), \tag{13}$$

where m^j_i denotes the token from the j-th quantization layer for sequence index i, and $m^{1:j-1}_i$ represents the embedded tokens from the preceding layers.

Inference. During inference, given a text condition $\mathbf{c}$, we first initialize all base-layer tokens as [MASK], *i.e.*, $m^0 = \text{[MASK]}$. The B-Trans predicts the base layer tokens as $\hat{m}^0 = \text{B-Trans}(m^0, c)$. For each subsequent quantization layer $j \in \{1, 2, \ldots, V\}$, the R-Trans predicts the tokens $\hat{m}^j$ for the j-th layer based on the previously predicted tokens from layers 1 to $j-1$ and the condition $\mathbf{c}$, *i.e.*, $\hat{m}^j = \text{R-Trans}(\hat{m}^{1:j-1}, \mathbf{c}, j)$. After predicting all layers, the full quantized motion token sequence is obtained as $\hat{m}^{1:V} = \left[\hat{m}^0, \hat{m}^1, \ldots, \hat{m}^V\right]$. The final motion embedding $\hat{\mathbf{h}}^q$ is retrieved from the pre-trained codebook, and the motion sequence $\hat{\mathbf{X}}$ is reconstructed using the pre-trained decoder: $\hat{\mathbf{X}} = \text{Decoder}(\hat{\mathbf{h}}^q)$.

4. Experiments

Datasets. To validate the effectiveness of our model, we conducted experiments using the AniMo4D and AnimalML3D [52] datasets. Building on prior success with synthetic data techniques demonstrated in previous research [17, 29, 43], we curated AniMo4D from *Planet Zoo* [6], a simulation game known for its high-quality animal models and motions. Specifically, we utilize `cobra-tools`[1] to extract the animal models and animations. We then conduct exploratory data analysis and define a unified skeletal structure with $n_j = 30$ joints, selecting the most frequently occurring and cross-species-compatible joints, balancing joint coverage with species inclusiveness (with some incompatible species, such as penguins, discarded). Furthermore, we annotate each motion with textual descriptions generated by Llama [8], Qwen [1], and Deepseek [24], followed by manual cleaning to ensure high quality. AniMo4D contains 78,149 motion sequences and 185,435 textual descriptions across 114 species. Figure 3(a) and Figure 3(b) illustrate the distribution of animal species and motion labels. The motion sequences range from 20 to 300 frames. Table 1 provides a comparison with existing text-to-motion datasets, highlighting AniMo4D's compatibility. We split the dataset into training, validation, and testing sets with ratios of 0.80, 0.05, and 0.15, respectively.

For the AnimalML3D dataset, used as an out-of-distribution (OOD) test set, we selected 1,150 motion samples with sequence lengths ranging from 20 to 300 frames. To ensure consistency, we unified the number of joints to 30 across all datasets and preprocessed the joint data in the T2M format [11], where the data has dimensionality (T, d), with T representing the sequence length and $d = 359$ denoting the motion dimension.

Evaluation metrics. Following T2M [11], we trained text-motion matching components on the AniMo4D and AnimalML3D datasets to encode motions and text for evaluation. We evaluate performance using the following metrics: (1) Frechet Inception Distance (FID): This metric calculates the distance between feature distributions of generated and ground truth motions, assessing quality and realism. (2) R-Precision and Multimodal Distance (MM-Dist): R-Precision measures semantic alignment between text and generated motions, reporting Top-1, Top-2, and Top-3 accuracy. MM-Dist quantifies the distance between motion and text embeddings. (3) Diversity: This metric computes the mean pairwise Euclidean distance among generated motions for different descriptions, reflecting variability. (4) Multimodality: This evaluates the diversity of motions generated from a single text prompt. (5) Mean Per Joint Position Error (MPJPE): MPJPE measures the average Euclidean distance between predicted and ground truth joint positions, directly assessing skeletal motion accuracy.

Baselines. To the best of our knowledge, no publicly available models are specifically designed for pure text-driven animal motion generation. While OmniMotionGPT [52] can generate animal motions, it requires additional human motions as references. Motion Avatar [59] focuses on generating mesh-based animal avatars rather than motion sequences. Therefore, we selected several state-of-the-art text-driven human motion generation models to serve as our baselines, including: T2M [11], MDM [47], T2M-GPT [54], AttT2M [61], MMM [34] and MoMask [12]. To adapt these models for animal motion, we modified their skeletal definitions to align with those used in AniMo4D. Each baseline model was retrained using its default parameters, and we evaluated their performance on both the AniMo4D and AnimalML3D datasets.

Implementation Details. In line with the configuration in Guo *et al.* [12], we configure RVQ with 6 quantization layers in the first stage, each with a codebook size of 512×512 and a downsampling rate of $r = 4$. We use the AdamW optimizer [26] with $[\beta_1, \beta_2] = [0.9, 0.99]$. The weighting factor α is set to 0.02, and the adversarial loss weight coefficient λ is set to 0.002. In the second stage, both the B-Trans and R-Trans consist of 6 transformer layers, each with 6 attention heads and a latent dimension of 384. The learning rate is set to 2×10^{-4}, with linear warm-up periods of 80k and 16k iterations for the first and second stages, respectively. We use batch sizes of 128 and 64 for these stages. For text encoding, we employ the CLIP ViT-B/32 model [38]. All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs.

4.1. Quantitative Results

Following [11, 47], each experiment was conducted 10 times, and the metrics are reported as mean values with

[1] https://github.com/OpenNaja/cobra-tools

Table 2. Comparison of text-driven motion generation on the AniMo4D test set. The right arrow (→) indicates that values closer to the ground truth motion are better. The best results for each metric are shown in bold.

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top-1 ↑	Top-2 ↑	Top-3 ↑				
Real	$0.854^{\pm.002}$	$0.947^{\pm.001}$	$0.974^{\pm.001}$	$0.002^{\pm.000}$	$0.897^{\pm.001}$	$4.751^{\pm.078}$	–
T2M [11]	$0.604^{\pm.002}$	$0.722^{\pm.002}$	$0.807^{\pm.001}$	$0.107^{\pm.000}$	$1.391^{\pm.002}$	$4.362^{\pm.053}$	$0.314^{\pm.015}$
MDM [47]	$0.303^{\pm.002}$	$0.444^{\pm.001}$	$0.534^{\pm.002}$	$1.017^{\pm.005}$	$2.525^{\pm.002}$	$3.953^{\pm.051}$	$0.264^{\pm.016}$
T2M-GPT [54]	$0.581^{\pm.002}$	$0.732^{\pm.002}$	$0.807^{\pm.002}$	$0.218^{\pm.003}$	$1.505^{\pm.003}$	$4.281^{\pm.050}$	$\mathbf{0.519}^{\pm.018}$
AttT2M [61]	$0.573^{\pm.002}$	$0.734^{\pm.002}$	$0.816^{\pm.001}$	$0.265^{\pm.002}$	$1.497^{\pm.002}$	$4.360^{\pm.042}$	$0.362^{\pm.012}$
MMM [34]	$0.628^{\pm.002}$	$0.770^{\pm.002}$	$0.838^{\pm.002}$	$0.268^{\pm.001}$	$1.411^{\pm.002}$	$4.318^{\pm.053}$	$0.178^{\pm.008}$
MoMask [12]	$0.708^{\pm.002}$	$0.870^{\pm.002}$	$0.932^{\pm.001}$	$0.143^{\pm.000}$	$1.223^{\pm.001}$	$4.441^{\pm.053}$	$0.415^{\pm.013}$
AniMo (Ours)	$\mathbf{0.774}^{\pm.002}$	$\mathbf{0.908}^{\pm.002}$	$\mathbf{0.953}^{\pm.001}$	$\mathbf{0.029}^{\pm.000}$	$\mathbf{1.063}^{\pm.002}$	$\mathbf{4.664}^{\pm.052}$	$0.368^{\pm.015}$

Table 3. Comparison of text-driven motion generation in out-of-distribution (OOD) experiments on the AnimalML3D dataset. The right arrow (→) indicates that values closer to the ground truth motion are better. The best results for each metric are shown in bold.

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top-1 ↑	Top-2 ↑	Top-3 ↑				
Real	$0.696^{\pm.002}$	$0.881^{\pm.004}$	$0.939^{\pm.005}$	$0.011^{\pm.001}$	$0.972^{\pm.015}$	$4.738^{\pm.037}$	–
T2M [11]	$0.102^{\pm.017}$	$0.226^{\pm.017}$	$0.304^{\pm.016}$	$2.906^{\pm.153}$	$3.972^{\pm.012}$	$3.396^{\pm.056}$	$0.803^{\pm.043}$
MDM [47]	$0.074^{\pm.042}$	$0.135^{\pm.041}$	$0.187^{\pm.050}$	$3.134^{\pm.149}$	$4.276^{\pm.015}$	$3.286^{\pm.052}$	$0.737^{\pm.044}$
T2M-GPT [54]	$0.135^{\pm.005}$	$0.233^{\pm.005}$	$0.311^{\pm.006}$	$2.843^{\pm.085}$	$3.920^{\pm.014}$	$\mathbf{3.990}^{\pm.048}$	$0.947^{\pm.038}$
AttT2M [61]	$0.122^{\pm.003}$	$0.219^{\pm.003}$	$0.290^{\pm.005}$	$2.893^{\pm.045}$	$4.043^{\pm.021}$	$3.463^{\pm.033}$	$0.887^{\pm.072}$
MMM [34]	$0.115^{\pm.003}$	$0.191^{\pm.004}$	$0.251^{\pm.004}$	$2.943^{\pm.094}$	$4.019^{\pm.008}$	$3.324^{\pm.034}$	$0.843^{\pm.069}$
MoMask [12]	$0.133^{\pm.003}$	$0.234^{\pm.004}$	$0.321^{\pm.004}$	$2.887^{\pm.032}$	$3.970^{\pm.016}$	$3.320^{\pm.047}$	$\mathbf{1.062}^{\pm.058}$
AniMo (Ours)	$\mathbf{0.144}^{\pm.004}$	$\mathbf{0.254}^{\pm.004}$	$\mathbf{0.336}^{\pm.005}$	$\mathbf{2.710}^{\pm.086}$	$\mathbf{3.867}^{\pm.017}$	$3.377^{\pm.054}$	$0.947^{\pm.042}$

95% confidence intervals. As shown in Table 2, AniMo outperforms other methods across key evaluation metrics on the AniMo4D dataset. It achieves superior R-Precision, reflecting stronger alignment between text prompts and motions. Notably, AniMo achieves a significantly lower FID score, indicating more realistic and higher-quality motions. Moreover, Its lower MM-Dist further demonstrates strong semantic consistency between text and motion. Although MModality is slightly lower than that of some other methods, AniMo's high R-Precision and low MM-Dist emphasize text adherence over extreme diversity while still maintaining strong diversity.

Furthermore, we performed out-of-distribution (OOD) evaluations on the AnimalML3D dataset. As shown in Table 3, due to the distributional shift between datasets, the performance of all models degraded. Nonetheless, AniMo continues to achieve superior performance across multiple metrics despite these challenging conditions.

4.2. Qualitative Results

Figure 4(a) idemonstrates the model's ability to generate motions for different species given similar textual descriptions, highlighting its capacity to capture the unique morphological characteristics of each species. For example, the okapi and the proboscis monkey exhibit substantial differences in skeletal structure and size, with the okapi being larger. The model also learns species-specific behaviors: the proboscis monkey appears more agile in swimming, using its upper limbs more, while the snow leopard uses all four limbs in a coordinated way. Additionally, Figure D2 (in Supp.) presents the qualitative results of AniMo compared to baseline models. These qualitative results further substantiate the model's ability to generate diverse and realistic animal motions.

4.3. Ablation Study

To assess the effectiveness of our model components, we performed an ablation study, evaluating both the first-stage reconstruction and the full generation pipeline, as shown in Table 4. The results indicate that the removal of any component leads to a decline in performance across all key metrics. Among the components, the Spatial Transformer (S-Trans) proves to be the most critical, as its removal results in the largest increase in FID, underscoring its importance in capturing spatiotemporal features for high-quality motion synthesis. Moreover, removing the discriminator $D(\cdot)$ yields the highest MPJPE, emphasizing its crucial role in ensuring realistic skeletal motion. Eliminating the Species-Aware Feature Modulation (SAFM) also degrades performance. These findings confirm the contributions of each component to the overall architecture and further demonstrate the robustness of our design.

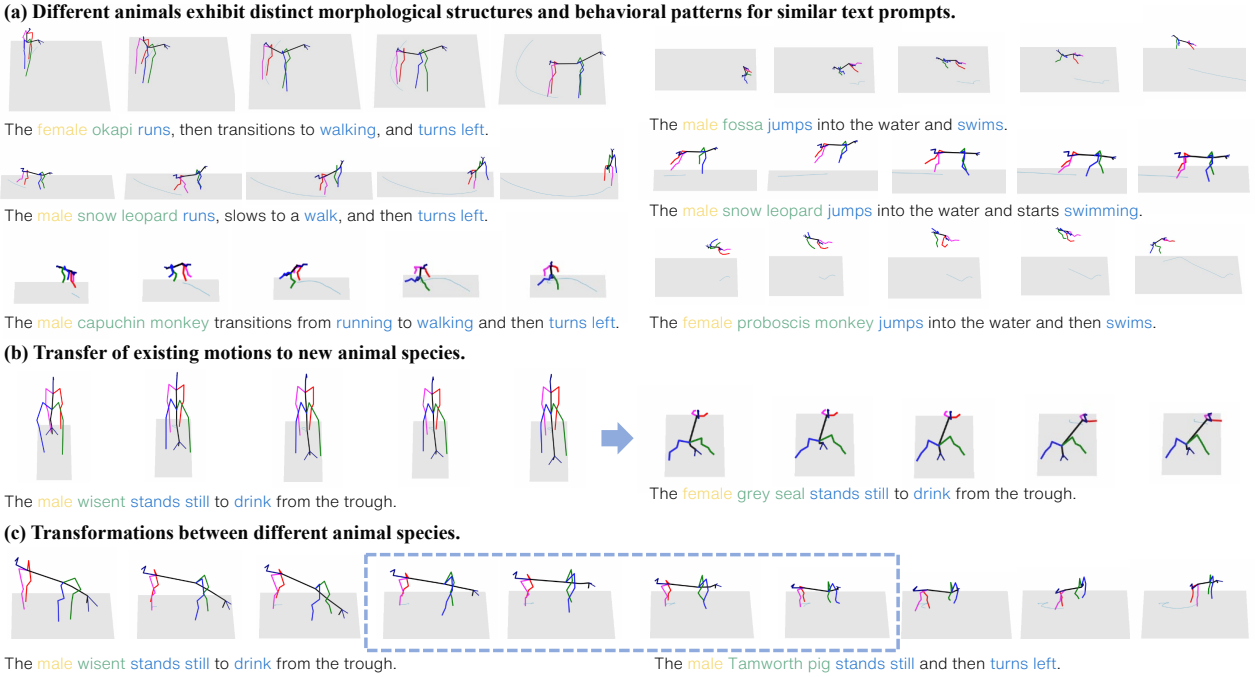


Figure 4. Motions generated by AniMo. The results demonstrate that AniMo has learned different animal species and behavior patterns, and can transfer motions to novel animals as well as perform cross-species transformations.

4.4. Application

Motion Transfer. AniMo demonstrates a strong understanding of intrinsic animal behaviors, facilitating effective motion transfer across species. This capability is evidenced in Figure 4(b), where the model successfully transfers a drinking motion from a wisent to a grey seal, even though such examples are absent in the AniMo4D training set. The generated motion naturally aligns with the grey seal's morphological structure, enabling a wide range of applications, including animation, wildlife simulation, behavioral studies, and creative design, such as character development and digital art creation [44, 45]. Furthermore, as both the text input and animal attributes are encoded using CLIP, the model leverages CLIP's semantic representations to map species into the latent space, enabling seamless motion transfer to unseen species.

Cross-species Motion Inpainting. AniMo achieves cross-species motion inpainting through a mask-based inference approach, leveraging its species-aware architecture and comprehensive understanding of animal behavior patterns.The process consists of tokenizing the original motion sequence, replacing target segments with [MASK] tokens, and feeding the modified sequence along with textual descriptions into the B-Trans and R-Trans. Figure 4(c) illustrates this capability, showing smooth transitions between concatenated motions from different species, with the inpainted process highlighted in blue. This enables seamless motion editing and completion, particularly valuable for animation production where continuous motion sequences need to be generated from fragmented recordings or different species [13, 46].

Table 4. Ablation study on the AniMo4D dataset.

Methods	Reconstruction		Generation	
	FID↓	MPJPE↓	FID↓	MM-Dist↓
AniMo	**0.035**	**0.206**	**0.037**	**1.063**
w/o SAFM	0.044	0.249	0.056	1.301
w/o S-Trans	0.063	0.214	0.071	1.241
w/o $D(\cdot)$	0.055	0.325	0.091	1.197

5. Conclusion

In this paper, we introduced AniMo to tackle the challenges of text-driven animal motion generation. By leveraging a joint-aware spatiotemporal encoder and species-aware feature modulation, AniMo effectively captures the unique morphological and behavioral traits of various species. Comprehensive evaluations on the AniMo4D and AnimalML3D datasets demonstrate that our approach consistently outperforms existing methods across multiple quantitative metrics, generating motions that conform to animal morphological structures and behavioral patterns.

Acknowledgments

This work was supported by the Zhejiang Provincial Natural Science Foundation of China (LZ24F030005), the Scientific Research Foundation of Sichuan Provincial Department of Science and Technology, China (2024YFHZ0001), and the Research Fund for International Scientists of National Natural Science Foundation of China (72350710798). Finally, we would like to acknowledge the Key Laboratory of Digital Intelligent Breeding Technology Innovation for Swine and Poultry, Ministry of Agriculture and Rural Affairs, and New Hope Liuhe Co., Ltd., China, for the support.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 6, 1

[2] Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms. *arXiv preprint arXiv:2410.18977*, 2024. 2

[3] Wenshuo Chen, Hongru Xiao, Erhang Zhang, Lijie Hu, Lei Wang, Mengyuan Liu, and Chen Chen. Sato: Stable text-to-motion framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6989–6997, 2024. 2

[4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2

[6] Frontier Developments. Planet zoo. https://www.planetzoogame.com. Accessed: 2024-3-12. 6

[7] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6, 1

[9] Magnus Enquist and Stefano Ghirlanda. Neural networks and animal behavior. 2013. 1

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[11] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 4, 5, 6, 7, 1, 2

[12] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1, 2, 4, 6, 7

[13] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 8

[14] Adobe Systems Inc. Mixamo. https://www.mixamo.com. Accessed: 2024-11-12. 2

[15] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2, 4

[16] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Runyi Yu, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Local action-guided motion diffusion model for text-to-motion generation. In *European Conference on Computer Vision*, pages 392–409. Springer, 2025. 2

[17] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. Rgbd-dog: Predicting canine pose from rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8336–8345, 2020. 2, 6

[18] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 2

[19] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816, 2023. 2

[20] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[21] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 2, 4

[22] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 2, 4

[23] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12706–12716, 2021. 2

[24] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. 6, 1

[25] Chang Liu, Mengyi Zhao, Bin Ren, Mengyuan Liu, Nicu Sebe, et al. Spatio-temporal graph diffusion for text-driven human motion generation. In *BMVC*, pages 722–729, 2023. 2

[26] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 1, 2

[28] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 2

[29] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12395, 2020. 6

[30] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19023–19034, 2022. 1

[31] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[32] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Controlmm: Controllable masked motion generation. *arXiv preprint arXiv:2410.10780*, 2024. 2

[33] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. *arXiv preprint arXiv:2403.19435*, 2024.

[34] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 1, 2, 6, 7

[35] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 5

[36] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023. 2, 3, 5

[37] Alec Radford. Improving language understanding by generative pre-training. 2018. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 6

[39] Ali Rohan, Muhammad Saad Rafaq, Md Junayed Hasan, Furqan Asghar, Ali Kashif Bashir, and Tania Dottorini. Application of deep learning for livestock behaviour recognition: A systematic literature review. *Computers and Electronics in Agriculture*, 224:109115, 2024. 1

[40] Alessio Sampieri, Alessio Palma, Indro Spinelli, and Fabio Galasso. Length-aware motion synthesis via latent diffusion. *arXiv preprint arXiv:2407.11532*, 2024. 2

[41] Lyes Saad Saoud, Atif Sultan, Mahmoud Elmezain, Mohamed Heshmat, Lakmal Seneviratne, and Irfan Hussain. Beyond observation: Deep learning for animal behavior and ecological conservation. *Ecological Informatics*, page 102893, 2024. 1

[42] Moira Shooter, Charles Malleson, and Adrian Hilton. Benchmarking monocular 3d dog pose estimation using in-the-wild motion capture data. *arXiv preprint arXiv:2406.14412*, 2024. 2

[43] Moira Shooter, Charles Malleson, and Adrian Hilton. Digidogs: Single-view 3d pose estimation of dogs using synthetic training data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 80–89, 2024. 6

[44] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 8

[45] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3637–3646, 2022. 8

[46] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *arXiv preprint arXiv:2409.14393*, 2024. 8

[47] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. 1, 2, 6, 7

[48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[49] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 2

[50] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024. 2

[51] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020. 4

[52] Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and Yalin Wang. Omnimotiongpt: Animal motion generation with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1249–1259, 2024. 2, 3, 5, 6

[53] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 2, 4

[54] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2, 4, 6, 7, 1

[55] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[56] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 1, 2

[57] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023. 4

[58] Zeyu Zhang, Akide Liu, Qi Chen, Feng Chen, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Infinimotion: Mamba boosts memory in transformer for arbitrary long motion generation. *arXiv preprint arXiv:2407.10061*, 2024. 2

[59] Zeyu Zhang, Yiran Wang, Biao Wu, Shuo Chen, Zhiyuan Zhang, Shiya Huang, Wenbo Zhang, Meng Fang, Ling Chen, and Yang Zhao. Motion avatar: Generate human and animal avatars with arbitrary motion. *arXiv preprint arXiv:2405.11286*, 2024. 2, 3, 5, 6

[60] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Learning uncoupled-modulation cvae for 3d action-conditioned human motion synthesis. In *European Conference on Computer Vision*, pages 716–732. Springer, 2022. 2

[61] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 509–519, 2023. 1, 2, 6, 7

[62] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, pages 405–421. Springer, 2025. 2

[63] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *arXiv preprint arXiv:2312.02256*, 2, 2023.

[64] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2025. 2

[65] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1