MotionGlot: A Multi-Embodied Motion Generation Model

Sudarshan Harithas[1], **Srinath Sridhar**[1]
[1]Brown University
Providence, RI, USA
{ sudarshan_harithas, srinath_sridhar}@brown.edu
https://ivl.cs.brown.edu/research/motionglot.html

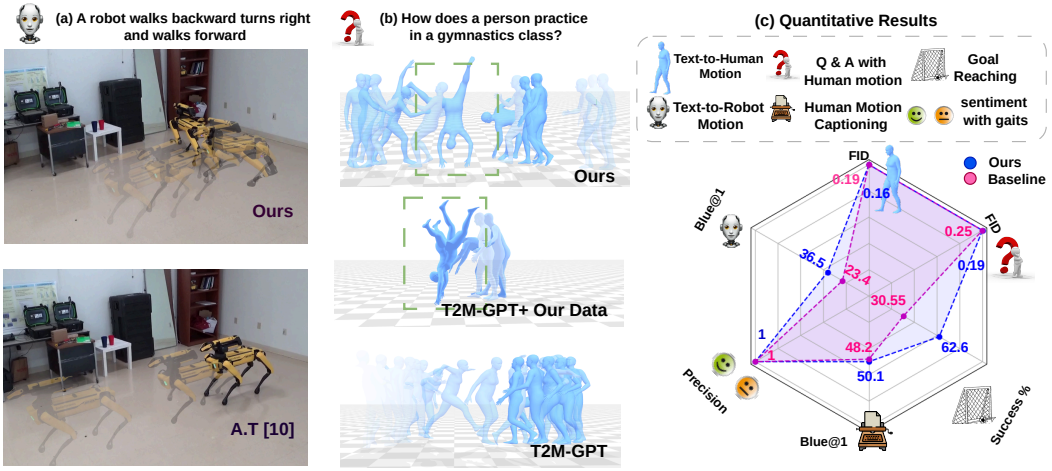

Figure 1: **Overview**: MotionGlot is a model that can generate motion trajectories that obey user instructions across multiple embodiments with different action dimensions, such as (a) quadruped robots, and (b) humans. The figures (a,b) depict the qualitative benchmark of MotionGlot against the adapted templates **(A.T)** of [1] on the text-to-robot motion (Section 4.1.1), Q&A with human motion (Section 4.3) tasks respectively. The overall quantitative performance across tasks is shown in (c). In (a,b), increasing opacity indicates forward time.

Abstract:

This paper introduces MotionGlot, a model that can generate motion across multiple embodiments with different action dimensions, such as quadruped robots and human bodies. By leveraging the well-established training procedures commonly used in large language models (LLMs), we introduce an instruction-tuning template specifically designed for motion-related tasks. Our approach demonstrates that the principles underlying LLM training can be successfully adapted to learn a wide range of motion generation tasks across multiple embodiments with different action dimensions. We demonstrate the various abilities of MotionGlot on a set of **6** tasks and report an average improvement of 35.3% across tasks. Additionally, we contribute two new datasets: (1) a dataset of expert-controlled quadruped locomotion with approximately 48,000 trajectories paired with direction-based text annotations, and (2) a dataset of over 23,000 situational text prompts for human motion generation tasks. Finally, we conduct hardware experiments to validate the capabilities of our system in real-world applications.

1 INTRODUCTION

Large Language Models (LLMs) [2, 3, 4, 5, 6, 7] have seen tremendous success recently with models that can produce text indistinguishable from human-generated text. These models have also shown to be useful in applications beyond just text generation, for example, in multi-lingual translation [5, 8], multi-task learning [5, 3, 4, 5, 6, 7], or instruction following [9].

LLMs use transformers [2] to model language as a sequence of tokens and are trained in a next-token or masked-token prediction framework. Indeed, some research has looked into modeling other forms of sequential data using the same machinery, for example, in audio [10] and weather data [11]. Unsurprisingly, recent work has also modeled motion and action as a sequential generation problem [12, 1, 13]. However, these approaches have thus far been limited to a single embodiment [14, 15, 16] or embodiments with the same number of action space dimensions [13, 1].

In this paper, we investigate the problem of building models of action that can cover multiple embodiments with different action spaces (*e.g.,* humans vs. quadrupeds). This is a hard problem because (1) motion data is not always plentifully available for all embodiments (*e.g.,* quadrupeds vs. humans), and (2) the templates used in training [13, 1] involves discretizing each action dimension into uniform bins and contacting the string to obtain the target for the transformer. Such a training template is suitable only for single embodiments and not easily extendable to create a model for motion generation across multiple embodiments.

We overcome these limitations with **MotionGlot**, a motion generation model that can span multiple embodiments with different action spaces. MotionGlot builds on top of the well-established instruction-tuning techniques from multilingual LLMs [8, 9, 17, 18] and proposes an instruction template to train a GPT [5] for motion generation.

While our insights and framework can be generalized and extended to multiple morphologies, we are primarily interested in two embodiments with different action spaces: **human bodies** and **quadruped robots**. MotionGlot is a single model that exhibits core capabilities such as text-conditioned motion generation for different embodiments, and captioning motion across multiple embodiments Figure 1.

To overcome the challenges due by limited data availability for quadrupeds, we propose QUAD-LOCO , a dataset of expert-controlled quadruped locomotion with direction-based text annotation Figure 2 (c). Furthermore, we propose an additional dataset of text captions for human motions, where we leverage the few-shot learning abilities of *GPT-4* [4] to create a dataset with more than 23000 situational descriptions of human actions, this data would be used for the Q&A with human motion task (Section 4.3).

QUAD-LOCO not only enables our core capability such as text-conditioned locomotion for quadruped, but also additional capabilities such as goal-conditioned motion generation for quadrupeds. Our experiments (Section 4) demonstrate that MotionGlot is a generalist method can generate motion across multiple embodiments, handle unseen user instructions, and express the multi-modal distribution in motion trajectories. MotionGlot also performs better than existing methods as shown in Figure 1 and Section 4.

Overall, our contributions are: **(1)** MotionGlot, a model that learns to generate motions across multiple embodiments with different action spaces. **(2)** an instruction tuning template that uses a single decoder-only transformer to generate motion across multiple embodiments and operate as a multi-task learner, and **(3)** The QUAD-LOCO dataset which consists of **48000** quadruped trajectories with direction-based textual descriptions for robot motion and the QUES-CAP dataset which consists of more than **23000** prompts that enable Q&A with motion Section 4.3.

2 Related Works

In this brief review, we focus on the closest work in language, robotics, motion generation, and captioning. Please see Table 1 for a summary of related works.

Language and Robotics: There has been an explosion of recent work at the intersection of language and robotic navigation or manipulation [19, 20, 21, 12] that treat language as an additional modality and have separate branches in the network to process text instructions.

Methods such as RT-2 [1] or OpenVLA [13] have attempted to unify language and action into a common vocabulary to train models

for manipulation tasks. However, their instruction tuning template is largely limited to embodiments with the same action dimension (*e.g.,* 7DoF action space of a manipulator). Driven by insights from multi-lingual instruction tuning [9, 8, 17, 22] our proposed method enables us to build a common vocabulary across embodiments with very different action spaces, specifically, human motions and quadruped motions.

Works such as Gato [23, 24] leverage autoregressive transformers to create a common controller policy for multiple embodiments. Unlike these methods, MotionGlot serves a different objective and caters towards generative tasks. While RoboCat [25] attends towards building a common model across different output dimensions, their approach is demonstrated only on manipulators, whereas MotionGlot explores diverse embodiments such as quadrupeds and human bodies. Additionally, our proposed training procedures bring the instruction-following and multi-task learning abilities of LLMs into motion motion generators.

Human and Robot Motion Generation: Motion generation for human bodies and mobile robots has been largely studied in separate communities. Human motion generation methods can be classified into two categories [26]: (1) methods that use pre-trained vision-language models like CLIP [27] for motion generation [14, 28, 29, 30], and (2) methods such as [15, 16], which jointly learn a text and motion representation. Works related to robot motion generation have largely focused on embodiments with the same action dimensions such as [31, 32, 1, 13]. MotionGlot belongs to the second category, and unlike the aforementioned models, it is a multi-embodied motion generator.

Datasets: While there exist large pools of data for manipulation [33] and navigation [34, 35, 36], there are no large data sources for quadruped locomotion paired with text. While [37] proposes to model quadruped gaits using their feet-floor contact pattern, the dataset largely ignores direction based annotation such as the captions shown in Figure 2 (c). Therefore, to expand the text-conditioned motion generation capabilities to robots, we propose QUAD-LOCO , a dataset with over **48000** (after data-augmentation) pairs of expert-controlled real-world quadruped motion trajectories with direction-based text annotation (Section 3.3).

Method	M-E	M-T	H/R M-G	H/R M-C
Adapted templates of [1, 13]	✗	✓	✗/ ✓	✗/ ✗
RoboCat [25]	✓	✓	✗/ ✓	✗/ ✗
T2MGPT [14]	✗	✗	✓/ ✓	✗/ ✗
T2MT [15]	✗	✗	✓/ ✓	✓/ ✗
MotionGPT [16]	✗	✓	✓/ ✗	✓/ ✗
MDM [30]	✗	✗	✓/ ✗	✗/ ✗
Ours	✓	✓	✓/ ✓	✓/ ✓

Table 1: Acronyms: **M-T**: Multi task ability, **H/R M-G**: Human/ Robot motion generation ability. **H/R M-C**: Human / Robot Motion captioning ability. *Robot* refers to a quadruped robot whose locomotion can be controlled with *SE2* velocity commands. **M-E** refer to the ability to perform generative tasks on multiple embodiements with different action dimensions. refer to Sec. 4.1.1 for adapted templates of [13, 1].

For human body motion, the *AMASS* [38] dataset, which includes text annotations from [39], has been a key resource [28, 16, 14, 15, 39]. While [39] offers a broad range of action descriptions, it often lacks the contextual details of specific situations where these actions occur. To address this, we utilized GPT-4 [] to expand the [39] descriptions into **23000** situation-based text descriptions, rephrasing them as questions (see Section 3.3). This new data enables applications like Q&A with human motion task (see Section 4.3).

Motion Captioning: Motion captioning is the task of generating a text description for the input motion. T2MT [15] uses an Encoder-Decoder transformer to caption human motion, however, such approaches are constrained to a single task of bidirectional translation between text and motion. MotionGPT [16] leverages a T5 [40] model for motion captioning and motion synthesis, however, [16] is constrained to a single embodiment. [41] performs captioning of robot actions, however, they are single-task, single embodiment models. In contrast, our model natively supports text captioning.

3 Method

We intend to build a model capable of motion generation across multiple embodiments with different action spaces. We approach this problem as a next-token prediction problem similar to LLMs.

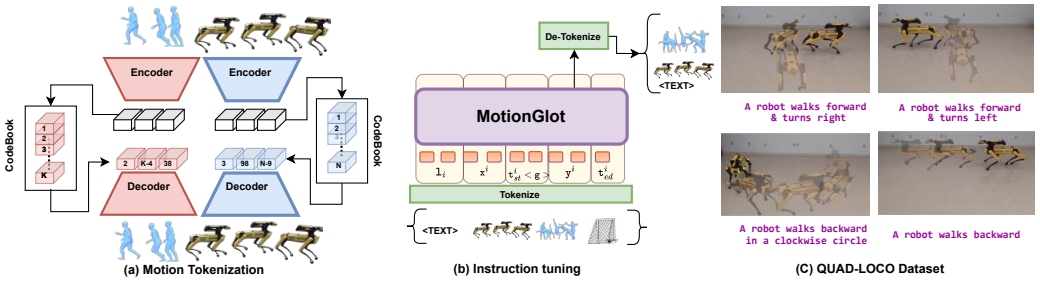

Figure 2: (a) Trajectories from different embodiments are tokenized using their associate VQ-VAE [42] (Section 3.1). (b) The proposed instruction template (Section 3.2) is used to train GPT for motion and text generation. Note that the tokenizer and de-tokenizer operate on the expanded vocabulary Section 3.2 ($\mathcal{V}$) (c) The preview of the QUAD-LOCO dataset, the captions indicate the direction-based text annotation.

Figure 2 shows an overview of our approach. Below, we describe individual components. Our training procedure involves two steps, in the first stage a VQ-VAE [42] learns a discrete latent codebook that represents a motion vocabulary per embodiment. This process, known as motion tokenization, is similar to text tokenization [43]. The motion vocabulary across embodiments are then appended to the existing vocabulary of GPT2 [3] creating a unified motion and text vocabulary. In the second step, our proposed instruction template is used to train the autoregressive GPT [2, 3, 5].

3.1 Trajectory Parameterization & Tokenization

For a given embodiment, a motion trajectory of length $\mathcal{T}$ is parameterized as $\mathbf{x}^e = [p_0^e, p_1^e, \cdots, p_{\mathcal{T}}^e]$, where p denotes motion represented as the embodiment's pose, and e denotes different embodiments – in our case either the quadruped robot (r) or human (h). The quadruped trajectory is parameterized by a sequence of 2D linear ($\dot{xz}$) and angular velocities ($\dot{r_a}$) where a pose at a discrete time t is given by $p_t^r = (\dot{xz}, \dot{r_a}) \in R^{SE2}$. Here, we assume that the y-axis is perpendicular to the ground plane (xz). The human pose is parameterized using the canonical representation from SMPL [44, 39] as $p_t^h = (\dot{r}_a, \dot{r}_{xz}, r_y, j_p, j_v, j_r, c_f) \in R^{263}$, where $\dot{r}_{xz} \in R^2$ is the root velocity along the ground plane, $\dot{r}_a \in R^1$ is the root angular velocity along the y-axis, $r_y \in R^1$ is the height of root from ground, $j_p, j_v \in R^{3k}$ and $j_r \in R^{6k}$ refer joint positions, joint velocities and joint angles represented as continuous 6D vectors, and $c_f \in R^4$ are the foot contact features, the number of joints $k = 22$ for the [39] dataset.

The goal of the tokenizer is to develop representations that allow a trajectory to be expressed as a series of discrete tokens, where each token is a unique element belonging to a finite vocabulary. We employ a *VQ-VAE* Figure 2 (a) [42] which consists of an autoencoder with a learnable codebook $\mathcal{C} \in R^{N \times d}$ with N tokens each of embedding dimension d. A separate VQ-VAE [42] is maintained for each embodiment, where the codebook represents the learned vocabulary for that embodiment.

The motion trajectories ($\mathbf{x}_e$) are first passed through the encoder that applies 1D convolutions to create a latent code $z \in R^{d \times T/l}$, where l is the temporal down-sampling from the encoder. The quantization process substitutes each entry of the latent space $z_i \in R^d$ with the closest element in the codebook $\hat{z}_i \in R^d$ given by Equation (1). The quantized embeddings $\hat{z}_i$, are then fed into the decoder to reconstruct the input signal $\hat{x} \in \mathbb{R}^{d_e \times T}$ as

$$\hat{z}_i = \underset{c_k \in \mathcal{C}}{\arg\min} ||z_i - c_k||_2. \tag{1}$$

The tokenizer is trained using three loss functions [42, 14]: $L = L_r + L_e + L_c$, where L_r represents the reconstruction loss, L_e is the embedding loss, and L_c denotes the commitment loss. Following the approach outlined in [14], all loss functions are L_1 loss with smoothing, velocity regularization, and EMA with codebook reset techniques [42] are included.

Note, that in contrast to discrete binning-based tokenization used in [13, 1] where N tokens are used to represent a single pose of $N - DOF$ output space, using the *VQ-VAE* based tokenization one token would return l poses. Leading to a total compression of the order of $\mathcal{O}(lN)$, thereby, improving the use of the finite context window of the transformer [2, 3, 6].

3.2 Instruction Tuning

To enable multi-embodiment motion synthesis we leverage insights from instruction tuning for multi-lingual models [9, 17, 8]. The process involves two steps, first, we merge the motion and text vocabularies to create a unified vocabulary suitable for generating motion and text. In the second step, we propose an instruction template for motion synthesis is proposed. We first define various vocabularies and their objectives.

Vocabulary Definition: We choose GPT-2 [3] as the backbone model for training, its vocabulary ($\mathcal{V}_l$) size of 50,257 primarily consists of tokens from the English language. The VQ-VAE [42] results in a motion vocabulary denoted as $\mathcal{V}_r, \mathcal{V}_h$ for the robot and human motion respectively. Additionally, the ground plane is divided into uniform cells and each cell is treated as a token, the complete set of these cells forms the vocabulary $\mathcal{V}_g$. Furthermore, a vocabulary of gait tokens $\mathcal{V}_{gait}$ are defined that indicate the choice of gait the quadruped must choose while executing the trajectory, the gait tokens are associated with an RL-controller trained using proximal policy optimization (*PPO*) [45], which execute the trajectory with the chosen gait. Following works from machine translation [8], task-specific special tokens are included that indicate the start and end of the response, the vocabulary of special task identification tokens is given by $\mathcal{V}_s$.

Vocabulary Expansion: Following insights from instruction tuning strategies from multi-lingual LLMs [46, 9, 17, 18], we merge all the vocabularies, to create a single vocabulary given as $\mathcal{V} = \{\mathcal{V}_l, \mathcal{V}_r, \mathcal{V}_h, \mathcal{V}_s, \mathcal{V}_g, \mathcal{V}_{gait}\}$. Performing next-token prediction on such a unified vocabulary ($\mathcal{V}$), across text, human, robot trajectories, and $2D$ ground plane enables the generation of motion across embodiments with different action dimensions in the same way text is generated.

Training Template: Given a corpus $\mathcal{M}$ of input-output $(\mathtt{x}^i, \mathtt{y}^i)$ pairs, a prefix (l) and the corresponding task-specific start (t^i_{st}) and end (t^i_{ed}) special tokens, the dataset is represented as $\mathcal{M} = \{(t^i_{st}, t^i_{ed}, \mathtt{x}^i, \mathtt{y}^i, l_i)\}$. For a given sample $p_i \in \mathcal{M}$, we leverage a template $\hat{\mathcal{T}}$ to create a task instruction d^i, i.e. $d^i = \hat{\mathcal{T}}(p_i)$. The template $\hat{\mathcal{T}}$ is defined in Eq. 2, where $< \mathtt{g} >$ is an optional field for the gait indicator token, which would only be active for robot trajectory generation. This stage is depicted in Figure 2 (b).

$$\hat{\mathcal{T}} := l_i : \mathtt{x}^i \, t^i_{st} \; < \mathtt{g} > \; \mathtt{y}^i \, t^i_{ed} \tag{2}$$

Note that unlike the training strategies used [1, 13] our template is not restricted to a single embodiment. The standard next-token prediction objective from [3, 2] on the vocabulary $\mathcal{V}$ is used to train the GPT. The task-specific substitution for $l_i, \mathtt{x}_i, \mathtt{y}^i$ are detailed in Sec. 4.

3.3 Dataset Creation

3.3.1 QUAD-LOCO Dataset

Motion generation has largely been limited to single human embodiments due to the lack of data beyond human bodies [38, 39, 47]. Therefore, we propose the QUAD-LOCO dataset with around 48000 pairs (with data augmentation) of trajectories and direction-based text annotation. A preview of the QUAD-LOCO dataset is displayed in Fig. 2 (c). Here, an expert operator remotely controls a spot quadruped robot to follow direction-based text-based instructions. The resulting movements of the robot are recorded, creating a dataset with quadruped motion and textual command correspondences. More than 1000 trajectories have been recorded over 2.5 hours from the expert teleoperator. Furthermore, the mirroring strategies from [39] are used to augment, furthermore we time-scale these trajectories as an additional augmentation strategy. The QUAD-LOCO dataset has been crucial in enabling text-to-robot motion (Sec. 4.1.1) and goal-conditioned motion generation (Sec. 4.2).

3.3.2 QUES-CAP Dataset

Datasets like [39, 47] have advanced human motion generation, however, the captions typically lack the situational context in which the action can be performed.To enable human motion generators to synthesize motion based on situational queries, we propose the QUES-CAP dataset. We leverage

GPT-4's [4] few-shot learning [48] capabilities to generate situational questions based on everyday scenarios and rewrite the provided text descriptions from [39] to serve as potential answers. For example, for a description like *'a person is boxing; they throw an uppercut, then dodge, and throw a few right jabs'*, a corresponding situational question might be *'What sequence of movements describes a beginner learning basic boxing techniques?'*. Similarly, for a description like *'a man raises his right arm, wiggles it, and then brings it back down'*, a relevant situational question could be *'How would someone look if they were trying to get someone's attention from across a noisy room using only their arm?'*. With similar examples we prompt `gpt-4-turbo` to rewrite **23000** prompts from [39] as questions. This dataset has been used in the *Q & A* with human motion task (Sec. 4.3).

4 Experiments

We conduct experiments to specifically answer the following questions related to the generative abilities of MotionGlot: **Q1:** Can the same machinery that is used to generate text be used to generate diverse motion across embodiments? **Q2:** Can MotionGlot generalize to unseen user instructions? **Q3:**Can MotionGlot express multi-modal action distribution? Experiments in Sec. 4.1, 4.3, 6.1.2 address **Q1** they are motion equivalent tasks of classical language problems. Sec. 4.1.1, and Sec. 4.2 answers **Q2**, **Q3** respectively.

Implementation Details: We choose GPT-2 (small) [3] as our base model, the codebook size of the human motion tokenizer and robot motion tokenizer are $R^{512\times512}$ and $R^{128\times512}$ respectively. For the goal-reaching task, we divide the $14m \times 14m$ ground plane into cells with a uniform resolution of $0.5 \times 0.5m$. The downsampling rate (l) of the VQ-VAE [42] is set to 4 (($l = 4$). Our model is trained on eight $NVIDIA - A5000$, for about $20k$ steps with a per-device batch size of 16 and 4 steps of gradient accumulation. Adam optimizer [49] with an initial learning rate of 5×10^{-4}, that decays with a cosine schedule has been used during training.

Evaluation Metrics: Protocols and procedures from [39] have been used, global text and motion features are extracted to compute the metrics below. Pre-trained models ($\mathcal{M}_h$) and ($\mathcal{M}_r$) are motion feature extractors for human and robot motion, respectively. ($\mathcal{M}_h$) is pre-trained model from [39] and similarly we train another feature extractor ($\mathcal{M}_r$) which produces close features for matched text and robot-motion pairs, and vice versa. Furthermore, 95% confidence is reported similar to [39].

(1) Diversity (Div): N pairs are randomly sampled from a set of global-motion features and the average distance between them is computed. **(2) Multimodality(MMod)**: For a given query 20 motion samples are generated forming 10 pairs of motion and the average distance between them is computed. **(3) FID:** is the distribution distance between the features of generated and real motion [50]. **(4) Translation Metrics:** BERT-score [51] (BS), Rouge [52], Cider [53], Bleu@N [54] (B@N) measure similarity between the ground truth and the generated text. **(5) Success %:** 40 trajectories are sampled per goal cell and a trajectory is successful if it terminates within the target cell. **(6) R-precision (RP):** For every generated output $\hat{\mathrm{y}}$, 32 input conditions (either text or motion) are sampled $\{\tilde{\mathrm{x}}\}_{i=1}^{32}$ (1 ground truth and 31 randomly sampled from dataset). The Euclidian distance between the features of $\hat{\mathrm{y}}$ and $\{\mathrm{x}\}_{i=1}^{32}$ are ranked to measure the retrieval accuracy.

4.1 Translation

4.1.1 Text-to-Robot Motion

This experiment evaluates the ability of MotionGlot to follow unseen user instructions, the task is to generate trajectories that semantically follow the input direction-based text description from the test QUAD-LOCO dataset. While [1, 13] are primarily meant for manipulation tasks, here we adapt their instruction template to perform text-to-robot motion generation. Here we briefly detail the performed modifications to [1]. Following [13] the data has been cleaned from outliers by selecting samples between 1^{st} and 99 quantiles. Each of the continuous dimensions has been uniformly discretized into 256 bins, where each bin represents an action token. The target for the *LLM* is obtained by concatenating the action tokens for each dimension with a space character as given in Eq. 3. The string is given below where $\Delta x, \Delta y, \Delta\psi$ represent the $2D$ linear and angular velocities. [13, 1] further requires observation as the input, here we project the global $SE(2)$ position through a linear layer to serve as the observation.

$$terminate \Delta x \Delta y \Delta \psi \tag{3}$$

The performance results are summarized in Table. 2. To quantitatively evaluate the performance in the text-to-robot motion task, we translate the input text instruction to a robot motion and back-translate the resulting motion tokens to get the text caption (refer to Sec. 6.1.1 for the evaluation of the robot motion captioning ability), the metrics **B@4, B@1 and BS** are then used to measure the cycle consistency between the user text instruction and back-translation. Text and motion feature vectors from $\mathcal{M}_r$, are used in the measurement of **RP**. A higher value of these metrics indicates greater consistency and adherence to the input text instruction. **Div** and **MMod** are used to evaluate the generative abilities of the model

For this task "`give robot motion: `" is substituted as the prefix l_i in Eq. 2, similarly, $\mathbf{x}_i$ is the sequence of text tokens and $\mathbf{y}^i$ is the sequence of robot motion tokens. MotionGlot outperforms competitors by 31.2% on average across all back translation metrics. The qualitative results are shown in Figure 1 (a), it can be observed while MotionGlot follows the user instructions, the adapted version of [13, 1] only execute the backward motion and does not turn right and walk forward.

Method	B@4 ↑	B@1 ↑	BS ↑	RP@1/2/3 ↑	Div →	MMod ↑
Real	-	-	-	$0.26/0.47/0.579^{\pm.001}$	$4.10^{\pm.003}$	-
Ours	$\mathbf{36.5}^{\pm.002}$	$\mathbf{64.7}^{\pm.002}$	$\mathbf{57.5}^{\pm.003}$	$\mathbf{0.18/0.35/0.48}^{\pm.005}$	$\mathbf{3.74}^{\pm.011}$	$2.35^{\pm.022}$
[1]$^{\mathbf{A.T}}$	$23.4^{\pm.003}$	$51.1^{\pm.002}$	$35.9^{\pm.003}$	$0.045/0.095/0.156^{\pm.002}$	$3.35^{\pm.012}$	$\mathbf{3.18}^{\pm.015}$

Table 2: Results on the QUAD-LOCO test set. **A.T**: Adapted templates. ↑, ↓ indicate higher, lower the better respectively and → indicates closer to the real value the better. **Bold** indicates the best method, ± indicates the 95% confidence interval as [39] defines.

4.1.2 Text-to-Human Motion

We evaluate the model's ability to generate motion across various embodiments with different action dimensions by conducting text-to-human motion on the test set of [39]. text-to-human motion generation literature falls into two main categories. The first category (`Cat I`) includes methods such as [14, 28, 29, 30], which use *CLIP* [27] embeddings for motion generation. Techniques [29, 28], also use privileged information, such as ground-truth trajectory length, during evaluation.

The second category (`Cat II`) consists of methods like MotionGlot and [15, 16] which don't use privilege information like *CLIP* or ground-truth trajectory length, instead jointly learn both the text and motion representations. While `Cat I` are better than `Cat II` on metrics like *FID, R-Precision,* and *MMDist*, they are single-task specialized models. Conversely, `Cat II` methods offer greater versatility but trade-offs some performance in favor of their multi-tasking capabilities.

For this task, "`give human motion: `" is the task-specific prefix l_i in Eq. 2, x_i and y_i are seuqnce of text and human motion tokens respectively. Tab. 3 summarizes the performance in text-human motion task. Where we compare to methods within `Cat II`, as they are directly comparable when privileged information is not used, however, Tab 3 mentions `Cat I` for completeness. MotionGlot demonstrates a competitive performance against competing SOTA baselines.

Txt.Rep	Methods	RPrecision↑			FID↓	MMDist↓	Diversity→	MMod↑
		Top1	Top2	Top3				
	Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
`Cat I`	MDM [29]$^{\Delta}$	$0.32^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$
	T2M-GPT [14]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
	MO-MASK [28]$^{\Delta}$	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	-	$1.241^{\pm.040}$
`Cat II`	T2MT [15]	$\mathbf{0.424}^{\pm.003}$	$\mathbf{0.618}^{\pm.003}$	$\mathbf{0.729}^{\pm.002}$	$1.501^{\pm.017}$	$\mathbf{3.467}^{\pm.011}$	$8.589^{\pm.076}$	$2.424^{\pm.093}$
	MotionGPT [16]$^{\delta_1}$	$0.402^{\pm.003}$	$0.567^{\pm.002}$	$0.649^{\pm.002}$	$\underline{0.19}^{\pm.0056}$	$4.18^{\pm.001}$	$\mathbf{9.33}^{\pm.008}$	$\underline{3.43}^{\pm.11}$
	Ours	$\underline{0.406}^{\pm.005}$	$\underline{0.571}^{\pm.007}$	$\underline{0.652}^{\pm.007}$	$\mathbf{0.1618}^{\pm.005}$	$\underline{3.969}^{\pm.008}$	$\underline{9.724}^{\pm.065}$	$\mathbf{3.48}^{\pm.098}$

Table 3: Text to Human Motion Benchmark on the HumanML3D [39] dataset. Δ indicates results evaluated with ground truth motion length. All values for the baselines are extracted from the paper, apart from δ_1 which is from the pre-trained open source model. underline is the second best method. Real data is deterministic therefore MMod is "-", and the Diversity value of [28] is not available.

Embodiment	Methods	RPrecision↑		MMDist↓	Length$_{avg}$ ↑	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	BertScore↑
		Top1	Top3							
Human	Real	0.523	**0.828**	2.901	12.75	-	-	-	-	-
	TM2T [15]	0.516	0.823	2.935	10.67	48.9	7.00	38.1	16.8	0.32
	MotionGPT [16]	**0.543**	0.827	2.821	13.04	48.2	12.47	37.4	29.2	0.324
	Ours	0.508	0.805	**2.78**	**14.42**	**50.1**	**13.5**	**41.8**	**33.6**	**0.339**

Table 4: Motion Captioning Benchmark on HumanML3D [39] dataset.

4.1.3 Motion Captioning

This task involves generating a text description for the input motion trajectory, the experiment further demonstrates the multi-task learning ability of MotionGlot, the results are given in Table 4. The task-specific prefix l_i in Eq. 2 is "`give text description:` , $\mathtt{x}_i$ is the sequence of human motion tokens and $\mathtt{y}_i$ is the sequence of text tokens. We evaluate the performance of MotionGlot against the current SOTA human motion captioning techniques, our method delivers an average improvement of 6.5 % on the motion captioning tasks across *Bleu [54], Cider [53] and BertScore [51]* . The results indicate that the captions generated by MotionGlot are semantically similar relative to the ground-truth captions and accurately capture the input motion trajectory.

4.2 Goal conditioned Motion Generation

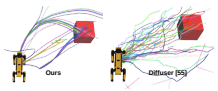

Figure 3: Qualitative results of the goal reaching task: note that our method expresses the multi-modal nature of the trajectory distribution, while [55] generates path towards the goal, its success of convergence at goal is lower.

This experiment evaluates the model's ability to express multi-modal action distributions, generating diverse trajectories that approach the goal. The task-specific prefix in Eq. 2 is l_i is "`reach goal:` ", the input token $\mathtt{x}_i$ is the goal cell token from $\mathcal{V}_g$ an the output $\mathtt{y}_i$ is the robot motion tokens. The qualitative results are shown in Fig. 3, and the quantitative results are summarized in Tab. 5. A trajectory is successful if its terminal position is within the goal cell. Diffusion with classifier guidance [55], is a promising generative approach to capture multiple-behavioural modes within the trajectory distribution, therefore, its trained on the QUAD-LOCO dataset as a baseline. It can be observed that MotionGlot achieves a significant improvement over [55] on success % and generative metrics.

Method	Success ↑ %	Diversity→	FID↓	MMod→
Real	100	$2.85^{\pm 0.031}$	$0.039^{\pm 0.00}$	$1.38^{\pm 0.0067}$
Ours	$\mathbf{62.0}^{\pm 0.061}$	$\mathbf{3.24}^{\pm 0.16}$	$\mathbf{0.33}^{\pm 0.014}$	$\mathbf{1.56}^{\pm 0.01}$
Diffusion [55]	$30.55^{\pm 0.074}$	$3.51^{\pm 0.0106}$	$0.95^{\pm 0.022}$	$2.91^{\pm 0.009}$

Table 5: Goal reaching Task.

Methods	RP@3↑	FID↓	Div→	MMod↑
Real	$0.364^{\pm .002}$	$0.002^{\pm .000}$	$9.503^{\pm .065}$	-
T2M-GPT	$\mathbf{0.38}^{\pm .003}$	$3.5^{\pm .008}$	$8.58^{\pm .078}$	$2.89^{\pm .042}$
T2m-GPT*	$0.33^{\pm .006}$	$0.25^{\pm .005}$	$9.26^{\pm .071}$	$2.44^{\pm .053}$
Ours	$0.36^{\pm .003}$	$\mathbf{0.19}^{\pm .006}$	$\mathbf{9.69}^{\pm .08}$	$\mathbf{3.06}^{\pm .042}$

Table 6: Q& A with Motion. T2M-GPT* indicates [14] trained with [39] and QUES-CAP datasets.

4.3 Q&A with Human Motion

This task presents a motion equivalent for text-based zero-shot Q&A, where motion is generated in response to user questions. Qualitative and quantitative results are shown in Fig. 1 (b) and Tab. 6, respectively. As seen in Fig. 1 (b), for a given question, the motion response form [14] generates a generic walking motion that does not relate well to a gymnastics practice question. After training on the QUES-CAP dataset, however, the response improves, producing a headstand action. Motion generated through MotionGlot is more expressive, performing a complete cartwheel relevant to the user query. These findings show that QUES-CAP dataset can train models for *Q & A* with motion. The overall performance improvement is summarized in Tab. 6, and the entries of Eq. 2 are the same as in Sec. 4.1.2.

5 Conclusion

We introduce MotionGlot, a motion generator for multiple embodiments with various action dimensions. Our findings show that MotionGlot can follow unseen user instructions, represent multi-modal action distributions, and function as a multi-task learner for motion and text data. In the future, we aim to enhance MotionGlot to include motion planning capabilities.

Acknowledgments

This research was supported by the Office of Naval Research (ONR) grant N00014-22-1-259.

References

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[3] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] P. Lin, S. Ji, J. Tiedemann, A. F. Martins, and H. Schütze. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*, 2024.

[9] J. Li, H. Zhou, S. Huang, S. Cheng, and J. Chen. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592, 2024.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[11] S. Talukder, Y. Yue, and G. Gkioxari. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*, 2024.

[12] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[13] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[14] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023.

[15] C. Guo, X. Zuo, S. Wang, and L. Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022.

[16] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Anoop Kunchukuttan. Extending english large language models to new languages: A survey. URL https://anoopkunchukuttan.gitlab.io/publications/presentations/extend_en_llms_apr2024.pdf.

[18] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

[19] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022.

[20] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023.

[21] C. Huang, O. Mees, A. Zeng, and W. Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.

[22] O. Shliazhko, A. Fenogenova, M. Tikhonova, V. Mikhailov, A. Kozlova, and T. Shavrina. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*, 2022.

[23] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[24] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024.

[25] K. Bousmalis, G. Vezzani, D. Rao, C. M. Devin, A. X. Lee, M. B. Villalonga, T. Davchev, Y. Zhou, A. Gupta, A. Raju, et al. Robocat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2023.

[26] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[28] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng. Momask: Generative masked modeling of 3d human motions. *arXiv preprint arXiv:2312.00063*, 2023.

[29] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[30] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[31] H.-T. L. Chiang, Z. Xu, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah, et al. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775*, 2024.

[32] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[33] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Mart'in-Mart'in, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[34] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine. GNM: A General Navigation Model to Drive Any Robot. In *International Conference on Robotics and Automation (ICRA)*, 2023. URL https://arxiv.org/abs/2210.03370.

[35] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine. ViNT: A foundation model for visual navigation. In *7th Annual Conference on Robot Learning*, 2023. URL https://arxiv.org/abs/2306.14846.

[36] A. Sridhar, D. Shah, C. Glossop, and S. Levine. NoMaD: Goal Masked Diffusion Policies for Navigation and Exploration. *arXiv pre-print*, 2023. URL https://arxiv.org/abs/2310.xxxx.

[37] Y. Tang, W. Yu, J. Tan, H. Zen, A. Faust, and T. Harada. Saytap: Language to quadrupedal locomotion. *arXiv preprint arXiv:2306.07580*, 2023.

[38] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019.

[39] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022.

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[41] T. Yamada, H. Matsunaga, and T. Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters*, 3(4):3441–3448, 2018.

[42] A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[43] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*, 2020.

[44] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[46] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[47] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[48] T. B. Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[49] P. K. Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014.

[50] H. Martin, R. Hubert, U. Thomas, N. Bernhard, and H. Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30:6626–6637, 2017.

[51] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[52] L.-Y. ROUGE. A packageforautomaticevaluationof summaries. *ProcofPost-ConferenceWorkshoponText SummarizationBranchesOutofthe42nd AnnualMeeting onAssociationforComputationalLinguistics*, pages 74–81, 2004.

[53] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[54] K. Papineni, S. Roukos, T. Ward, and W. Zhu. A method for automatic evaluation of machine translation". *the Proceedings of ACL-2002, ACL, Philadelphia, PA, July 2002*, 2001.

[55] M. Janner, Y. Du, J. Tenenbaum, and S. Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022.

6 Appendix

6.1 Ablation Studies

6.1.1 Robot Motion Captioning

This ablation aims to generate direction-based text captions for robot trajectories. In Eq. 2, `'give text description for robot :'` is the substituted prefix l_i, $\mathtt{x}_i$ and $\mathtt{y}_i$ is the sequence of robot motion and text tokens respectively. The performance analysis is given in Tab, 7, the high value of translation metrics indicates that MotionGlot is a reliable motion-to-text translator.

Methods	RP@3↑	MDist↓	$\mathbf{L}_{avg}$ ↑	B@1↑	B@4↑	[52]↑	[53]↑	[51]↑
Real	0.581	3.9	9.26	-	-	-	-	-
Ours	0.2635	3.09	8.58	64.7	41.1	74.5	29.6	0.6165

Table 7: Motion Captioning ablation on QUAD-LOCO dataset.

6.1.2 Sentiment Classification with Gaits

Saytap [37] demonstrates that each sentiment class can be associated with a gait for robot locomotion. For example, the *bounding* and *trott* gait can be used to indicate happy and neutral sentiments. With MotionGlot the gait field in Eq. 2 indicates the sentiment. 100 samples from the QUAD-LOCO dataset was used to benchmark against GPT-4 [4] in a few shot setting. Both techniques perform this task with an average precision of 100%.