AToM: Aligning Text-to-Motion Model at Event-Level with GPT-4Vision Reward

Haonan Han[1*] Xiangzuo Wu[1*] Huan Liao[1*] Zunnan Xu[1] Zhongyuan Hu[1]
Ronghui Li[1] Yachao Zhang[2†] Xiu Li[1†]
[1] Shenzhen International Graduate School, Tsinghua University [2] School of Informatics, Xiamen University

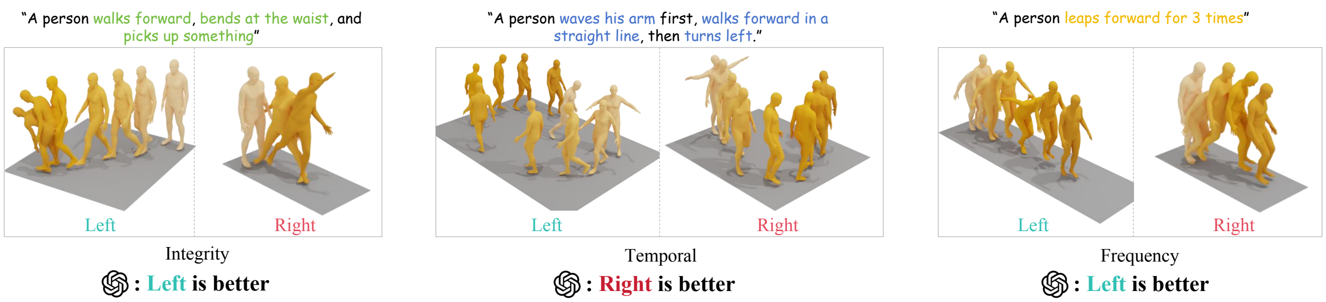

Figure 1. Showcases of motion samples for three scenarios. The two motion samples for each scenario were generated based on the prompt above the samples. Moreover, we leverage GPT-4V to compare two motion samples according to the degree of alignment between the motion samples and the input prompt.

Abstract

Recently, text-to-motion models have opened new possibilities for creating realistic human motion with greater efficiency and flexibility. However, aligning motion generation with event-level textual descriptions presents unique challenges due to the complex relationship between textual prompts and desired motion outcomes. To address this, we introduce ***AToM****, a framework that enhances the alignment between generated motion and text prompts by leveraging reward from GPT-4Vision. AToM comprises three main stages: Firstly, we construct a dataset* `MotionPrefer` *that pairs three types of event-level textual prompts with generated motions, which cover the integrity, temporal relationship and frequency of motion. Secondly, we design a paradigm that utilizes GPT-4Vision for detailed motion annotation, including visual data formatting, task-specific instructions and scoring rules for each sub-task. Finally, we fine-tune an existing text-to-motion model using reinforcement learning guided by this paradigm. Experimental results demonstrate that AToM significantly improves the event-level alignment quality of text-to-motion generation. Project page is available at https://atom-motion.github.io/.*

1. Introduction

Generating high-quality human motions from textual descriptions is a promising task and plays an important role in fields such as game production, animation, film and virtual reality. Recently, generative models leveraging autoregressive [12, 13, 16, 38] and diffusion-based approaches [9, 31, 32, 40] have shown remarkable performance on this task. These models generally perform well with short text prompts. However, due to the scarcity of text-motion pairs as well as the coarse-grained text descriptions that cover limited motion scenarios, there are challenges in mapping complex descriptions (e.g., multi-motion events or motions with temporal relationships and specified frequency) to corresponding motion sequences, limiting the model's ability to generalize effectively. As shown in the second group of Figure 1, given a prompt describing three motions in temporal order, the left example motion accurately captures the expression "waves his arm" but misinterprets "walks forward in a straight line" as "walks in a circle" and fails to account for the "turning left" action.

One potential solution to address this issue is to alleviate data scarcity by collecting additional text-motion pair data. However, unlike tasks involving language or image data, gathering motion data typically requires specialized motion capture equipment and expert annotations, which are both costly and labor-intensive. An alternative approach is to leverage the success of language models by fine-tuning pre-trained models with preference data, thereby enhancing their alignment capabilities. To the best of our knowledge, InstructMotion [33] is the first work to fine-tune a text-to-motion model using human preference data through reinforcement learning from human feedback (RLHF), leverag-

[*] Equal contribution. [†] Corresponding authors.

Table 1. Statistics of existing preference datasets for text-to-motion generative models. "Fine Grained" represents containing preference regarding multiple aspects or not.

Dataset	Annotator	Prompts	Pairs	Fine Grained
InstructMotion [33]	Human	3.5K	3.5K	✗
Pick-a-Move [27]	TMR [28]	–	–	✗
	Guo *et al.* [11]	–	–	✗
`MotionPrefer` (ours)	GPT-4V [25]	**5.3K**	**80K**	✓

ing human-labeled data to enhance model alignment with preferred outputs. Subsequent approaches [24, 27] have explored the use of automated preference datasets, employing pretrained models and reward functions to approximate human feedback. Such methods aim to reduce reliance on manual annotation and improve model performance across a range of alignment metrics

While human annotation reduces workload, it remains labor-intensive and challenging to scale. AI feedback-based methods reduce reliance on human annotators but rely on models trained on standard motion datasets, like HumanML3D [11], limiting their capacity to score out-of-distribution text-motion pairs. Additionally, both human and AI-based approaches often treat text and motion data as unified wholes, overlooking the need for fine-grained alignment evaluation, such as event-level correspondence.

To address these limitations, we propose leveraging advancements in rapidly evolving Vision-Language Large Models, such as GPT-4Vision. These models, through sophisticated architectures and large-scale training, have demonstrated state-of-the-art performance in tasks requiring seamless integration of textual and visual information, making them well-suited for the text-motion alignment task. By utilizing GPT-4Vision, we aim to simultaneously address challenges such as data scarcity, labor-intensive annotation, scalability issues, and the need for granular alignment evaluation—thus paving the way for more robust, scalable, and accurate alignment solutions.

To investigate GPT-4Vision's ability in aligning text and motion modalities, we propose a new framework named AToM, designed for aligning text-to-motion models using feedback from GPT-4Vision or other Vision-Language Large Models. AToM consists of three main stages: **(1)** we generate initial text prompts using GPT-4. These prompts are then fed into a motion generation model to generate several different motions for each text prompt. For ease of comparison with human-based work InstructMotion [33], we also adopt MotionGPT [16] as our motion generator. **(2)** We then render these generated motions and sample the rendered motion as a sequence of sampled frames. The input text prompt, sampled frames, along with an instruction describing the alignment rules, are then fed into the GPT-4Vision model. And the model will evaluate the text-motion alignment score at event-level based on the given frames and rules. The text prompts, generated motions, and the corresponding alignment scores jointly constitute our `MotionPrefer` dataset. In Table 1, we compare our `MotionPrefer` dataset with two other preference datasets, InstructMotion [33] and Pick-a-Move [27]. In terms of the scale, our `MotionPrefer` dataset surpasses InstructMotion [33], while the Pick-a-Move [27] does not specify its data volume. Furthermore, our dataset is the only one to provide fine-grained preference annotations across multiple aspects, including motion integrity, temporal order, and frequency. **(3)** We finetune the motion generator, MotionGPT [16], on our `MotionPrefer` dataset with LoRA[15] and IPO[6] RL strategy.

We summarize our contributions as follows:

- We created a dataset named `MotionPrefer`, consisting of 5.3K text prompts and 80K motion preference pairs. Each text-motion pair is scored based on event-level correspondences which contain three dimensions: motion integrity, temporal order, and motion frequency, thereby surpassing existing datasets in both scale and quality.
- We designed an annotation and reward paradigm leveraging GPT-4V, which encompasses key elements such as motion instructions, motion injection methods and scoring rules for each sub-task. This comprehensive paradigm is applied to evaluate the text-motion pairs collected in the `MotionPrefer` dataset, providing preference-based reward scoring.
- We used AToM to fine-tune an off-the-shelf motion generator across three sub-tasks: motion integrity, temporal order and frequency, achieving substantial performance improvements. Additionally, ablation studies explored the effects of various motion injection methods, LoRA, score filtering, and reinforcement learning strategies. The experimental results strongly support AToM's effectiveness in improving text-motion alignment across tasks.

2. Related Works

2.1. Text-to-Motion Generative Models

Like other generation tasks [14, 18, 19, 22, 37], text-to-motion (T2M) generation aims to generate human motion that corresponds to free-form natural language descriptions, serving as a fundamental task in the field of motion generation. Text2Action [2] pioneered this area by utilizing a GAN based on a SEQ2SEQ model to map short descriptions to human actions. Language2Pose [3] introduced a curriculum learning approach to develop joint-level embeddings for text and pose, while Lin et al. [21] proposed an end-to-end SEQ2SEQ model for generating more realistic animations. However, the limited availability of large-scale supervised datasets hinders generalization to novel descriptions, such as unseen combinations of motions.

To address these issues, Ghosh et al. [10] developed a hierarchical two-stream sequential model capable of handling long sentences that describe multiple actions. MotionCLIP [34] aligns the human motion manifold with CLIP space to endow the model with zero-shot capabilities. More recent work, such as the Transformer-based TEACH [5], generates realistic 3D human motions that follow complex, sequential action instructions, facilitating flexible temporal action composition. TEMOS [41] uses a Transformer-based VAE and an additional text encoder for multi-object 3D scene generation and editing, guided by multi-level contrastive supervision. T2M-GPT [38] combines VQ-VAE and GPT to obtain high-quality discrete representations, achieving competitive motion generation results. The diffusion-based model MotionDiffuse [39] allows for fine-grained control over body parts and supports arbitrary-length sequences through a series of denoising steps with injected variations. The classifier-free diffusion model MDM [31] predicts motion samples instead of noise, facilitating geometric loss application and setting state-of-the-art performance. MLD [9] further advances motion generation using a latent diffusion model. MotionGPT[16] develops unified large motion-language models that represent human motion via discrete vector quantization, enabling versatile performance across tasks like motion generation, captioning, and prediction. However, due to the coarse grained motion-paired text descriptions, achieving robust performance in zero-shot and multi-event scenarios remains challenging.

2.2. Aligning Models with Human/AI Feedback

Reinforcement Learning from Human Feedback (RLHF) [7, 26] has emerged as a transformative technique for model alignment, especially in applications with complex or ambiguous objectives. It has become the primary approach for aligning large language models (LLMs) [1, 4, 43] with user intent and has been effectively extended to generative models in image [17, 36] and audio generation [20]. In addition to traditional methods like PPO [30] and RLHF-PPO [26], which use explicit reward models, alternative approaches such as Direct Preference Optimization (DPO) [29] and Slic-hf [42] streamline the alignment process. These methods directly optimize model policies based on human preferences, which reduces computational overhead and enables potentially more robust optimization by working directly with preference data.

In aligning generated motion with human perception, existing methods often struggle with overfitting specific motion expressions due to limited training data, which relies heavily on expert-labeled motion. InstructMotion [33] addresses this by incorporating human preference data, where non-expert labelers compare generated motions, introducing preference learning to T2M generation and achieving performance improvements over traditional methods.

Given the high cost of obtaining quality preference labels, Reinforcement Learning from AI Feedback (RLAIF) [8] presents another promising alternative. Mao et al. [24] decompose motion descriptions into meta motions, combining them to generate novel descriptions. They use a trial-and-error approach in reinforcement learning, designing a reward model based on contrastive pre-trained text and motion encoders, enhancing semantic alignment and leveraging synthetic text-only data for better generalization. MoDiPO [27] applies DPO with AI feedback, to confine diffusion-based motion generation within realistic and text-aligned boundaries. Unlike InstructMotion, MoDiPO focuses on optimizing model alignment rather than generalizable generation, ensuring high-quality, contextually consistent outputs.

Our approach significantly diverges from previous work in two key aspects: 1) Instead of relying on labor-intensive, human-feedback-driven methods, we are the first to utilize a more efficient, LLM-based AI feedback mechanism from GPT-4V. 2) We are the first to focus on fine-grained alignment between motion and text, particularly at the event level. We emphasize integrity, temporal and frequency correspondence between the motion described in the text and the motion generated, providing a more detailed approach to motion-text alignment.

3. Method

The framework of AToM, as shown in Figure 2, consists of three stages. Firstly, we constructed a synthetic dataset of motion-text pairs using task-specific selected prompts and corresponding multiple output motion sequences from the motion generator, illustrated in Section 3.1. Furthermore, in Section 3.2, We developed a reward paradigm based on GPT-4Vision to score the alignment of visual signals rendered from motion sequences across three aspects. Finally, as elaborated in Section 3.3, we utilized the synthetic dataset collected in stage 1 and the alignment score annotated in stage 2 to fine-tune the text-to-motion model, thereby enhancing its task-specific alignment performance.

3.1. Dataset Construction

Prior studies [27, 33] have found that text-to-motion models face challenges in producing motions aligning with input textual descriptions, primarily manifest in low integrity, incorrect temporal relationships and frequency. As shown in Figure 1, the left example exhibits poor integrity because of a missed motion event and the middle example has an incorrect temporal relationship due to a wrong time order, while the right example shows the wrong frequency. Specifically, we focus on evaluating the integrity, temporal relationships, and frequency of the generated motion to assess the effectiveness of AI feedback in text-to-motion model. In stage 1, we targetedly construct prompts and generate text-motion samples related to the above three aspects to facilitate re-

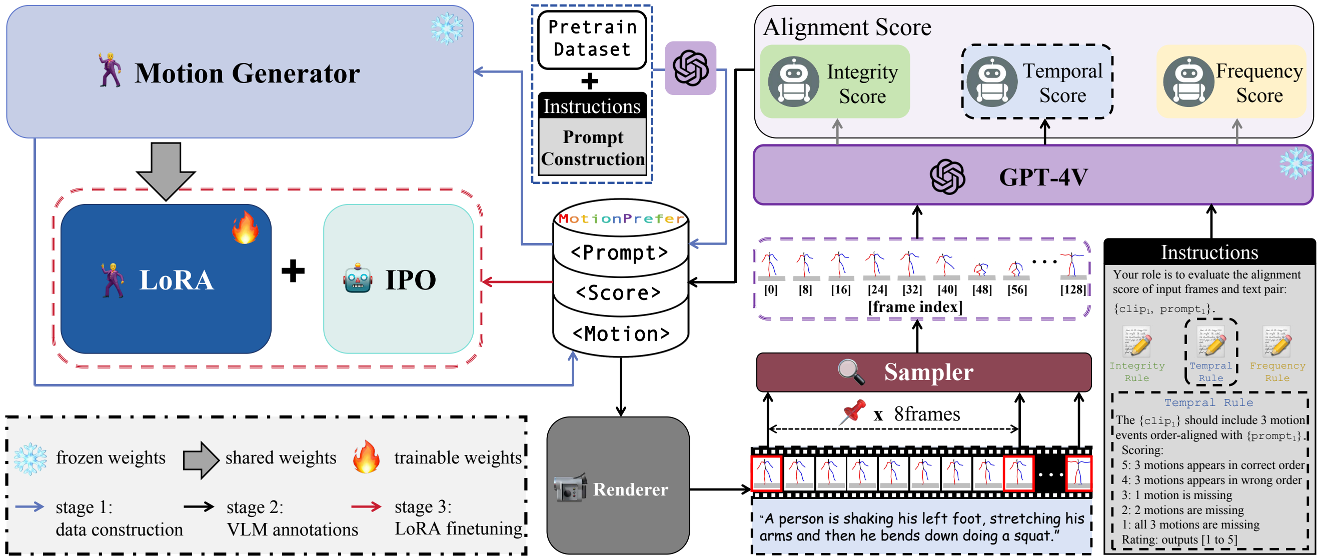

Figure 2. **The framework of AToM.** AToM encompasses three stages: (1) A motion generation process using task-specific prompts constructed by LLM; (2) Evaluation of alignment score for text-motion pairs using a predefined reward paradigm based on LVLM; (3) A fine-tuning mechanism based on LoRA and RL strategy that enhances the original motion generator using the dataset `MotionPrefer`.

> Given a label group, $\mathrm{X}_{\mathrm{task}}$, the three labels in it will be described into a motion event group:
> $\{\mathtt{Event}_1, \mathtt{Event}_2, \mathtt{Event}_3\}$.
> Then, please join the motion event group with conjunction to form a motion prompt, defined as
> "$\mathtt{Event}_1, \mathtt{Conjunction}_1, \mathtt{Event}_2, \mathtt{Conjunction}_2, \mathtt{Event}_3$".
>
> Conjunction list: $\mathrm{X}_{\mathrm{conj}}$
>
> Please ensure to randomly select conjunctions from the list and avoid relying on a single conjunction. Please try to generate complete prompts that match human language expressions as much as possible.

Table 2. GPT instruction for prompt construction in temporal task.

ward evaluation in the later stage. As the process shown by the blue arrow in Figure 2, we randomly select n=3.5K motion events from the dataset HumanML3D [11] as our meta motion labels which can be denoted as $\mathrm{X}_{\mathrm{meta}} = \{l_1, ..., l_n\}$. For different sub-tasks, we randomly pick labels from the $\mathrm{X}_{\mathrm{meta}}$ to form the task-specified label group $\mathrm{X}_{\mathrm{task}}$. Based on the label group $\mathrm{X}_{\mathrm{task}}$ and a predefined conjunction group $\mathrm{X}_{\mathrm{conj}}$, we instruct GPT-4, named $\mathrm{M}_{\mathrm{Language}}$, to generate meaningful and complete sentences matching human language with instruction $\mathrm{I}_{\mathrm{prompt}}$:

$$\mathbb{D}_{\text{prompt}} \sim \mathrm{M}_{\mathrm{Language}}(\mathrm{I}_{\mathrm{prompt}}|\mathrm{X}_{\mathrm{task}}, \mathrm{X}_{\mathrm{conj}}), \quad (1)$$

As shown in Table 2, here we give a template of the instruction $\mathrm{I}_{\mathrm{prompt}}$ for the temporal sub-task to construct prompt $\mathbb{D}_{\text{prompt}}$. The conjunction list $\mathrm{X}_{\mathrm{conj}}$ for temporal sub-task

Amount	Prompt	Motion	Pair	Motion events
Integrity	2500	25K	35.3K	2-5
Temporal	1360	13.6K	35.3K	3
Frequency	1416	8.5K	9.4K	1
Total	5276	47.1K	80K	—

Table 3. Details of amounts of `MotionPrefer` dataset.

encompasses terms such as "and", "then", "followed by", and so forth. Similarly, but distinctively, for the integrity sub-task, we construct prompts by selecting 2-5 motion events from $\mathrm{X}_{\mathrm{task}}$ combined with conjunctions to cover most cases; for the frequency sub-task, prompts are typically formed by selecting a single motion event paired with a frequency-descriptive conjunction (refer to the Appendix for detailed instructions regarding the other two sub-tasks). Based on the prompt data $\mathbb{D}_{\text{prompt}}$ constructed earlier, we use the widely adopted text-to-motion model, MotionGPT [16], named $\mathrm{M}_{\mathrm{Motion}}$, as our motion generator to produce motion samples, ranging from 6 to 10 per prompt:

$$\mathbb{D}_{\text{motion}} \sim \mathrm{M}_{\mathrm{Motion}}(\mathbb{D}_{\text{prompt}}). \quad (2)$$

The details of `MotionPrefer` are provided in Table 3. In total, the `MotionPrefer` dataset consists of 5,276 prompts and 47.1k motion samples.

3.2. Reward Paradigm Design

The motion sequences generated by MotionGPT [16] were rendered into video using the renderer. This rendered video sequence is denoted as $\mathbb{D}_{\text{motion-video}}$:

$$\mathbb{D}_{\text{motion-video}} \sim \mathrm{R}_{\mathrm{motion}}(\mathbb{D}_{\text{motion}}), \quad (3)$$

Table 4. Scoring rules for sub-tasks.

Task	Example	Scoring Criteria
Integrity	A man walks forward, walks backward, and squats.	5: All described motions appear in the frames. 0: Some motions are missing or incomplete in frames.
Temporal	A person walks forwards doing ballet, then raising one leg, then skipping and raising the other leg.	5: Three motions appear in correct order. 4: Three motions appear in wrong order. 3: One motion is missing. 2: Two motions are missing. 1: All three motions are missing.
Frequency	A person jumps forward one time.	3: The motion is correct and the frequency is accurate. 2: The motion is present but the frequency is incorrect. 1: The motion is incorrect, regardless of the frequency.

where $\mathbb{D}_{\text{motion}}$ represents the original motion data and $\mathrm{R}_{\text{motion}}$ denotes the renderer for obtaining the motion video.

The video is then sampled at 8-frame intervals, extracting a sequence of frame F, using the sampling function $\mathrm{Sampler}$ as follows:

$$F = \{f_0, f_8, \dots\} = \mathrm{Sampler}(\mathbb{D}_{\text{motion-video}}, 8), \quad (4)$$

where $\mathrm{Sampler}$ selects frames at 8-frame intervals from the rendered video. These sampled motion frames F, along with the corresponding text description $\mathbb{D}_{\text{prompt}}$, are sequentially injected into GPT-4V for evaluation.

We leverage GPT-4V, named M_{VL}, to assess the alignment between generated motion sequences and textual descriptions. With the specific instructions $\mathrm{I}_{\text{score}}$ (refer to the Appendix for the detailed scoring instructions for the three tasks), GPT-4V evaluates the alignment score across three tasks, based on the scoring criteria $\mathrm{C}_{\text{score}}$ detailed in Table 4. This scoring process can be represented as:

$$\mathbb{D}_{\text{reward}} = \mathrm{M}_{\mathrm{VL}}\left((F, \mathbb{D}_{\text{prompt}}), \mathrm{I}_{\text{score}}\right), \quad (5)$$

where the motion frames and text description are input into GPT-4V, along with instructions $\mathrm{I}_{\text{score}}$, to compute the alignment score.

For each task, GPT-4V assigns a score reflecting how well the motion sequence aligns with the text description, considering factors such as the integrity, temporal order, and frequency of motions. Through this process, we obtain a set of text-motion pairs with corresponding scores, denoted as $\mathbb{D}_{\text{motionprefer}}$, which can be represented as:

$$\mathbb{D}_{\text{motionprefer}} = \{\mathbb{D}_{\text{motion}}, \mathbb{D}_{\text{prompt}}, \mathbb{D}_{\text{reward}}\}, \quad (6)$$

where $\mathbb{D}_{\text{motion}}$ represents the generated motion sequence, $\mathbb{D}_{\text{prompt}}$ represents the text description, and the $\mathbb{D}_{\text{score}}$ is the text-motion alignment score annotated by GPT-4V. This annotation framework enables an evaluation of the generated motions across multiple dimensions, serving as a foundation for the next-step finetuning of the pretrain model.

3.3. Text-to-Motion Model Fine-tuning

To fine-tune our motion generator with GPT-4V feedback, we apply algorithm 1 as a method to construct our training data $\mathbb{D}$ from $\mathbb{D}_{\text{motionprefer}}$, where $\mathbb{D}_{\text{motionprefer}}$ contains pairs in the form (m, p, r), with m, p and r derived from the sets $\mathbb{D}_{\text{motion}}$, $\mathbb{D}_{\text{prompt}}$ and $\mathbb{D}_{\text{reward}}$, respectively. This algorithm firstly filter out motion samples relevant to a specific subtask. It then groups the samples by identical prompts and ranks each group by their quality scores in descending order. For each prompt group, the algorithm iterates over all possible pairs of motion samples, selecting pairs where the difference in quality scores exceeds a predefined threshold δ. These selected pairs, consisting of a high-quality sample m_w, a low-quality sample m_l, and their associated prompt p, are then added to the training dataset $\mathbb{D}$.

Algorithm 1 Paired Data Construction for a Subtask

Require: Dataset $\mathbb{D}_{\text{motionprefer}}$
Ensure: Filtered Training Set $\mathbb{D}$
  Extract subset $\mathbb{D}_{\text{subtask}} = \{(m, p, r) \in \mathbb{D}_{\text{motionprefer}} \mid \text{issubtask}(p)\}$
  Group $\mathbb{D}_{\text{subtask}}$ by p to form $\mathbb{D}_{\text{grouped}}$
  Initialize an empty set $\mathbb{D}$
  for each group g in $\mathbb{D}_{\text{grouped}}$ **do**
    Sort g in descending order by r
    for each pair (i, j) in g where $i < j$ **do**
      if $r_i - r_j > \delta$ **then**
        Add (m_i, m_j, p) to $\mathbb{D}$
      end if
    end for
  end for
  return $\mathbb{D}$

We then follow the definition in IPO[6] to define IPO loss $h_\pi(m_w, m_l, p)$:

$$h_\pi(m_w, m_l, p) = \log\left(\frac{\pi(m_w|p)\pi_{ref}(m_l|p)}{\pi(m_l|p)\pi_{ref}(m_w|p)}\right), \quad (7)$$

where π represents our training policy, and π_{ref} is the ref-

Table 5. **Comparison of AToM with baselines in different tasks**. AToM♠ represents the process of mixing preference data from three tasks and randomly selecting a subset of preference data (approximately 3.5K pairs) that matches the size of the RLHF framework InstructMotion, ensuring fair comparison with the baseline model.

Task	Method	MM Dist↓	Top-1↑	Top-2↑	Top-3↑	FID↓	Diversity↑	MModality↑
Temporal	MotionGPT [16]	$5.652^{\pm.048}$	$0.193^{\pm.004}$	$0.312^{\pm.007}$	$0.403^{\pm.009}$	$0.655^{\pm.049}$	$\mathbf{8.936^{\pm.048}}$	$\mathbf{3.846^{\pm.048}}$
	AToM (ours)	$\mathbf{5.576^{\pm.027}}$	$\mathbf{0.199^{\pm.005}}$	$\mathbf{0.322^{\pm.006}}$	$\mathbf{0.412^{\pm.008}}$	$\mathbf{0.613^{\pm.044}}$	$8.926^{\pm.086}$	$3.495^{\pm.147}$
Frequency	MotionGPT [16]	$5.328^{\pm.063}$	$0.182^{\pm.011}$	$\mathbf{0.315^{\pm.012}}$	$0.406^{\pm.012}$	$1.409^{\pm.148}$	$8.857^{\pm.163}$	$\mathbf{3.591^{\pm.132}}$
	AToM (ours)	$\mathbf{5.259^{\pm.045}}$	$\mathbf{0.183^{\pm.009}}$	$0.314^{\pm.012}$	$\mathbf{0.407^{\pm.009}}$	$\mathbf{1.165^{\pm.124}}$	$\mathbf{8.861^{\pm.163}}$	$3.261^{\pm.115}$
Integrity	MotionGPT [16]	$5.897^{\pm.036}$	$0.160^{\pm.007}$	$0.271^{\pm.007}$	$0.350^{\pm.006}$	$\mathbf{0.340^{\pm.023}}$	$8.815^{\pm.130}$	$\mathbf{3.752^{\pm.162}}$
	AToM (ours)	$\mathbf{5.871^{\pm.034}}$	$\mathbf{0.160^{\pm.007}}$	$\mathbf{0.274^{\pm.008}}$	$\mathbf{0.360^{\pm.008}}$	$0.400^{\pm.040}$	$\mathbf{8.821^{\pm.180}}$	$3.511^{\pm.172}$
General	MotionGPT [16]	$3.998^{\pm.016}$	$0.407^{\pm.002}$	$0.571^{\pm.003}$	$0.661^{\pm.002}$	$0.188^{\pm.006}$	$\mathbf{9.452^{\pm.080}}$	$\mathbf{3.421^{\pm.170}}$
	InstructMotion [33]	$3.947^{\pm.013}$	$0.402^{\pm.003}$	$0.570^{\pm.003}$	$0.663^{\pm.003}$	$0.260^{\pm.009}$	$9.411^{\pm.120}$	$2.534^{\pm.152}$
	AToM♠ (ours)	$\mathbf{3.943^{\pm.012}}$	$\mathbf{0.410^{\pm.008}}$	$\mathbf{0.575^{\pm.003}}$	$\mathbf{0.669^{\pm.002}}$	$\mathbf{0.177^{\pm.008}}$	$9.401^{\pm.107}$	$3.025^{\pm.174}$

erence policy. The loss IPO optimizes upon is given by:

$$\mathbb{E}_{(m_w,m_l,p)\sim D}(h_\pi(m_w,m_l,p)-\frac{1}{2\beta})^2. \tag{8}$$

To enhance the adaptability and efficiency of our motion generator during fine-tuning, we employed LoRA[15], which enables us to adjust the model's parameters with significantly fewer computational resources compared to traditional fine-tuning methods. This fine-tuning approach generalizes well across various sub-tasks. As a result, the motion generator showed marked improvements in output quality, achieving stronger alignment between desired text prompts and the generated motions.

4. Experiments

4.1. Implementation Details

Dataset To evaluate the effectiveness of our AI feedback-driven fine-tuning framework, we initialized from the pre-trained MotionGPT checkpoint[1] and fine-tuned it on three subsets of our preference dataset, MotionPrefer, each including 35k (for temporal), 35k (for integrity) and 9.4k (for frequency) motion preference pairs.

Implementation Specifics The optimal hyperparameter configuration included a learning rate of 1e-3, batch size of 32, and 20 epochs, using AdamW as the optimizer. A cosine learning rate scheduler was applied, and PEFT[23] parameters were tuned with LoRA[15] (R = 8, α = 16, dropout = 0.05), allowing the model to leverage the parameter-efficient fine-tuning structure. This setup required approximately 12 GB of memory on a single RTX 4090 GPU, ensuring efficient fine-tuning throughout the process.

Evaluation Metrics For evaluation, we filtered the HumanML3D [11] test set, obtaining 418, 506, and 234 text-motion pairs for integrity, temporal, and frequency tasks. Consistent with prior research [33, 35, 38], we focus on motion quality and text-motion alignment. Multi-modal Distance (MM-Dist) calculates the average Euclidean distance between text and motion features, while R-Precision measures motion-to-text retrieval accuracy with Top-1, Top-2, and Top-3 scores based on the model's ability to rank ground-truth descriptions correctly. Motion quality is assessed using FID to compute distribution distances between generated and real motion features. Diversity measures the average Euclidean distance between randomly sampled motion pairs, and MModality evaluates the variation among motions generated from the same text by averaging distances across 20 sequences per description. For subjective analysis, 50 participants evaluated text-motion alignment on temporal, frequency, and integrity aspects, selecting their preferred case or indicating similarity for each comparison.

[1] https://huggingface.co/OpenMotionLab/MotionGPT-base

4.2. Main Results

Quantitative Experiment In our quantitative experiments across the temporal, frequency, and integrity tasks, AToM consistently outperforms MotionGPT [16] and InstructMotion [33] across most evaluation metrics, demonstrating superior text-motion alignment, motion quality, and generative realism. As shown in Table 5, AToM outperforms baselines in text-motion alignment, achieving lower MM Dist scores, higher top-1 and top-3 retrieval accuracy, and comparable top-2 accuracy, demonstrating its ability to generate semantically aligned motions. In terms of motion quality, AToM achieves a lower FID (0.613 vs. 0.655, a 6.4% improvement), indicating more realistic motions. AToM balances alignment and diversity, with slight reductions in diversity and MModality attributed to fine-tuning focused on aligning with high-quality samples. In the general task, AToM excels with the lowest MM Dist (3.943), superior retrieval accuracy, and the best FID (0.177), showcasing its ability to generate realistic, well-aligned motions while maintaining consistency and robust generalization across sub-tasks and datasets. These improvements stem from our fine-tuning approach, which integrates AI feedback to better capture motion nuances valued by humans. By combining human-like feedback with efficient AI strategies, AToM sur-

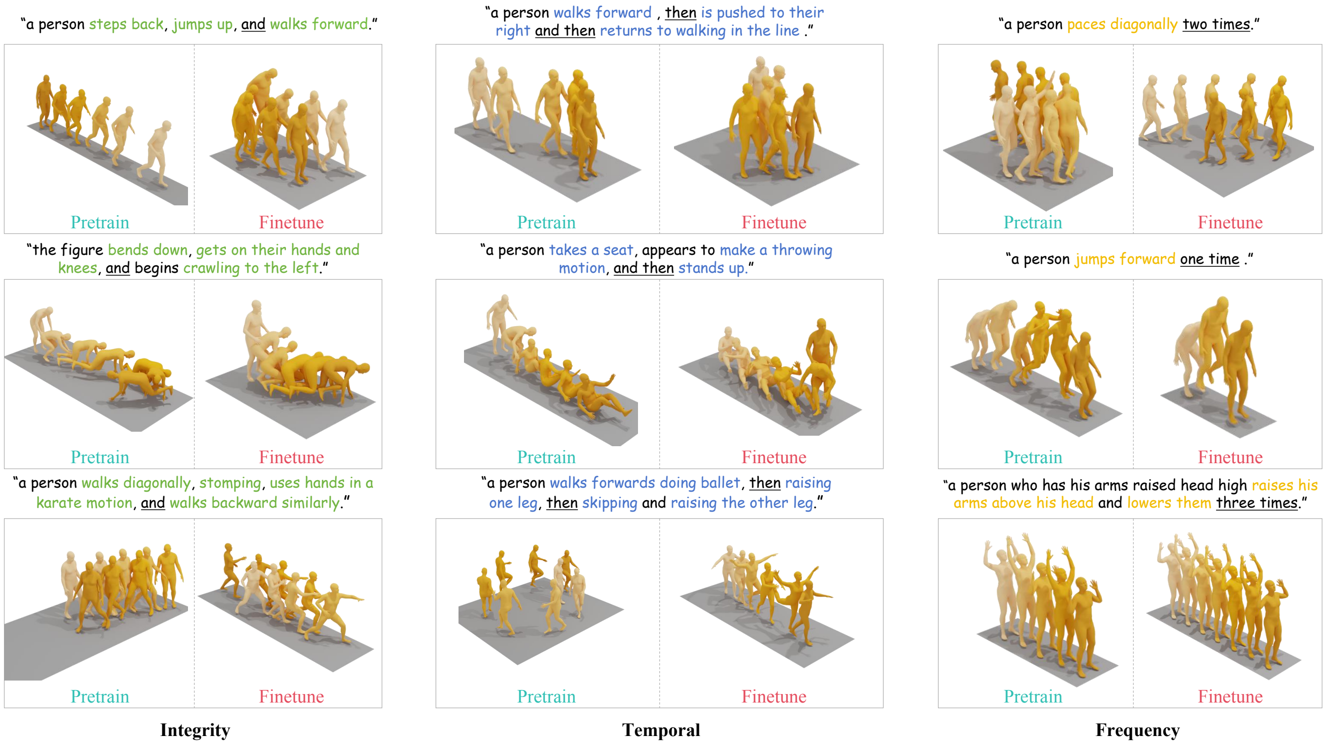

Figure 3. Generated qualitative samples comparison of pretrained model MotionGPT and finetuned model AToM.

passes traditional human-feedback-based models, offering a more effective and scalable solution.

Qualitative Experiment Figure 3 compares the MotionGPT [16] and AToM in terms of generation faithfulness for integrity, temporal and frequency tasks. In our qualitative analysis, discernible discrepancies are noted in the motion generation by the original model. For integrity task, in response to the three-event description "*a person steps back**,** jumps up,* ***and*** *walks forward.*", the generated motion included only two of the elements, but omitting specific motion event "***jumps up***". Similarly, for temporal task, the original model misrepresented the sequence of events or omitted action events. An example is the prompt "*a person walks forward,* ***then*** *is pushed to their right* ***and then*** *returns to walking in the line.*", where the generated motion incorrectly rendered the "***is pushed to their right***". For frequency task, the pretrained model generated motion "*a person jumps forward two times"*, which is inconsistent with the prompt "***one time***". In contrast, AToM demonstrates enhanced performance, achieving more faithful generation on diverse motion events, complex temporal order and specific frequency shown in prompts.

User Study We conducted a human evaluation to compare the performance of AToM with the MotionGPT baseline. For each sub-task, motions were generated from randomly selected prompts in the filtered HumanML3D test set [11] mentioned in sec 4.1. Each of the 50 participants evaluated five randomly selected pairs of generated motions each across three criteria: frequency, integrity, and temporal alignment. For each pair, the participant chose the better motion or marked a tie if both were comparable. Figure 4 presents the average win and tie rates for AToM compared to MotionGPT. As shown, AToM outperforms MotionGPT across all sub-tasks, with win rates of 74.4% for temporal, 70.0% for frequency, and 84.4% for integrity. This result reflects human evaluators' recognition of AToM's superior alignment and highlights its robustness across tasks.

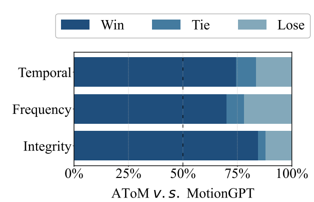

Figure 4. Win rates of AToM fine-tuned compared to MotionGPT by human judgments in three tasks.

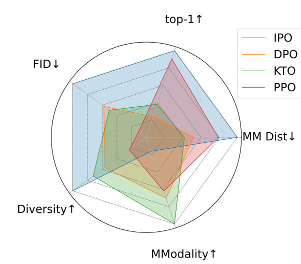

Figure 5. Performance distribution of different reinforcement learning strategies after generative model finetuning.

4.3. Ablation study

Motion Injection Forms in Questioning To evaluate the effectiveness of different motion injection strategies in GPT-4V questioning, we experimented with three methods: (1) **Frame-by-Frame**: The generated motion sequences

Table 6. Ablation studies for motion injection methods, score filtering, and LoRA utilization on the test set.

Metric	(a) Motion Injection			(b) Score Filtering		(c) LoRA Utilization	
	F-b-F	F-I	T-I	w/ Filter	w/o Filter	w/ LoRA	w/o LoRA
MM Dist ↓	$\mathbf{5.576}^{\pm.027}$	$5.657^{\pm.048}$	$5.635^{\pm.036}$	$\mathbf{5.576}^{\pm.027}$	$5.640^{\pm.030}$	$\mathbf{5.576}^{\pm.027}$	$6.425^{\pm.066}$
top-1 ↑	$\mathbf{0.199}^{\pm.005}$	$0.197^{\pm.008}$	$0.191^{\pm.008}$	$\mathbf{0.199}^{\pm.005}$	$0.187^{\pm.005}$	$\mathbf{0.199}^{\pm.005}$	$0.128^{\pm.006}$
top-2 ↑	$\mathbf{0.322}^{\pm.006}$	$0.316^{\pm.007}$	$0.304^{\pm.007}$	$\mathbf{0.322}^{\pm.006}$	$0.305^{\pm.006}$	$\mathbf{0.322}^{\pm.006}$	$0.221^{\pm.007}$
top-3 ↑	$\mathbf{0.412}^{\pm.008}$	$0.408^{\pm.005}$	$0.399^{\pm.007}$	$\mathbf{0.412}^{\pm.008}$	$0.394^{\pm.008}$	$\mathbf{0.412}^{\pm.008}$	$0.296^{\pm.009}$
FID ↓	$\mathbf{0.613}^{\pm.044}$	$0.803^{\pm.063}$	$0.652^{\pm.042}$	$\mathbf{0.613}^{\pm.044}$	$0.693^{\pm.059}$	$\mathbf{0.613}^{\pm.044}$	$2.131^{\pm.144}$
Diversity ↑	$\mathbf{8.926}^{\pm.086}$	$8.866^{\pm.103}$	$8.841^{\pm.081}$	$\mathbf{8.926}^{\pm.086}$	$8.882^{\pm.085}$	$\mathbf{8.926}^{\pm.086}$	$8.582^{\pm.105}$
MModality ↑	$3.495^{\pm.147}$	$3.396^{\pm.128}$	$\mathbf{3.675}^{\pm.143}$	$3.495^{\pm.147}$	$\mathbf{3.564}^{\pm.129}$	$3.495^{\pm.147}$	$\mathbf{6.153}^{\pm.182}$

from MotionGPT [16] were rendered into video then sampled at 8-frame intervals, creating a sequence of images. (2) **Full-Image**: The sampled image sequence (at 8-frame intervals) was arranged into a composite image, with 5 frames per row. (3) **Trajectory-Image**: The motion data, sampled at 8-frame intervals from the motion sequence, were rendered by rigged cylinders to a single image. Injection demos are shown in the Appendix. Table 6 (a) shows that employing the Frame-by-Frame image sequence as the motion representation yielded the highest performance, with MM Dist (5.576), top-1 accuracy (0.199), and FID (0.613). This approach enhanced the fine-tuned model, leading to high-quality output and superior motion-text alignment. In contrast, Full-Image and Trajectory-Image yielded slightly lower performance, possibly due to less detailed frame-level information interpretable by GPT-4V.

Score Filtering As shown in Table 6 (b), using score filtering in preference pairs construction, where only samples rated above three are considered positive-leads, results in clear performance improvements. With filtering, the model achieves an improved MM Dist, lowering the score from 5.640 to 5.576. The top-1 accuracy increases by 6.4%, rising from 0.187 to 0.199, and the FID score decreases by 11.5%, dropping from 0.693 to 0.613, all indicating better alignment and realism in generated motions. A larger faithfulness gap between positive and negative samples enhances the fine-tuning signal, enabling the model to better distinguish high-quality motion-text alignments.

Finetune with LoRA In Table 6 (c), incorporating LoRA[15] yields significant improvements compared to the model without it, though with a slight trade-off in multi-modality. Specifically, using LoRA[15] enhances the model's ability to retrieve motion sequences accurately, reduces the MM Dist score from 6.425 to 5.576, and increases top-1 accuracy by 55.5%, from 0.128 to 0.199. The top-2 and top-3 accuracies also improve significantly, with increases of 45.7% and 39.2%, respectively. The FID score also decreases by 71.2%, from 2.131 to 0.613, indicating better alignment with real motion distributions. Diversity also increases, from 8.582 to 8.926, showing greater diversity in generated motions.

RL Strategy Notably, Figure 5 (FID has been negatively treated) highlights that, among the four RL strategies, IPO shows the best overall performance. As the variant of DPO, IPO was specifically designed to mitigate overfitting[6] associated with the Bradley-Terry(BT) model, resulting in better overall retrieval accuracy, alignment precision, and diversity compared to DPO, KTO, and PPO.

Sampling Steps of Motion Sequence Sampling frames at different intervals impacts performance, as illustrated in Figure 6, with shorter intervals (e.g., 4 and 8) generally resulting in better match distances, higher top-1 accuracy, and lower FID, which indicate improved alignment and generation quality. We choose an interval of 8 frames as it provides strong retrieval precision, which is our primary focus.

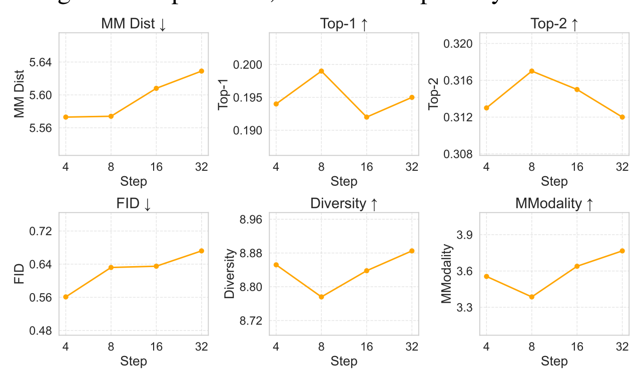

Figure 6. Impact of different frame sampling intervals on alignment and quality metrics.

5. Conclusion

In this paper, we introduce a novel framework, AToM, which utilizes GPT-4Vision feedback to enhance the alignment between text prompts and generated motions in text-to-motion models. AToM comprises three main stages: (1) generating diverse motions from constructed text prompts; (2) evaluating text-motion alignment using GPT-4Vision to construct the high-quality `MotionPrefer` dataset; and (3) fine-tuning the motion generator on `MotionPrefer`. Comprehensive quantitative and qualitative experiments demonstrate that AToM can effectively leverage feedback from Vision-Language Large Models, significantly improving text-motion alignment quality and paving the way for advancements in motion synthesis from textual prompts.

Acknowledgement

This work was partly supported by Shenzhen Key Laboratory of next generation interactive media innovative technology (No: ZDSYS20210623092001004).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE, 2018. 2

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 2

[4] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024. 3

[5] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 3

[6] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024. 2, 5, 8

[7] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 3

[8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. 3

[9] 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. 1, 3

[10] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1396–1406, 2021. 2

[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 (CVPR)*, pages 5152–5161, 2022. 2, 4, 6, 7

[12] Chuan Guo, Xinxin Zuo, Sen Wang, and Li 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. 1

[13] Chuan Guo et al. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910, 2024. 1

[14] Haonan Han, Rui Yang, Huan Liao, Jiankai Xing, Zunnan Xu, Xiaoming Yu, Junwei Zha, Xiu Li, and Wanhua Li. Reparo: Compositional 3d assets generation with differentiable 3d layout alignment, 2024. 2

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 6, 8

[16] 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. 1, 2, 3, 4, 6, 7, 8

[17] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[18] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023. 2

[19] Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1524–1534, 2024. 2

[20] Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. Baton: Aligning text-to-audio model with human preference feedback. In *IJCAI 2024*, 2024. 3

[21] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. *Learning*, 1(2018):1, 2018. 2

[22] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 6715–6724, 2024. 2

[23] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and B Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. *URL: https://github. com/huggingface/peft*, 2022. 6

[24] Yunyao Mao et al. Learning generalizable human motion generator with reinforcement learning. *arXiv preprint arXiv:2405.15541*, 2024. 2, 3

[25] OpenAI. Gpt-4 technical report, 2024. 2

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[27] Massimiliano Pappa et al. Modipo: text-to-motion alignment via ai-feedback-driven direct preference optimization. *arXiv preprint arXiv:2405.03803*, 2024. 2, 3

[28] Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9488–9497, 2023. 2

[29] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[30] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[31] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 3

[32] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 1

[33] Jenny Sheng et al. Exploring text-to-motion generation with human preference. In *CVPR*, pages 1888–1899, 2024. 1, 2, 3, 6

[34] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 3

[35] 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. 6

[36] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[37] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. In *Advances in Neural Information Processing Systems*, pages 20055–20080. Curran Associates, Inc., 2024. 2

[38] 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. 1, 3, 6

[39] 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. 3

[40] 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

[41] Xuying Zhang, Bo-Wen Yin, Yuming Chen, Zheng Lin, Yunheng Li, Qibin Hou, and Ming-Ming Cheng. Temo: Towards text-driven 3d stylization for multi-object meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19531–19540, 2024. 3

[42] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023. 3

[43] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 3