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

Supplementary Material

1. Additional Results

1.1. More Qualitative Results

Figure 1 presents additional qualitative comparisons between AToM and the baseline models, highlighting AToM's superior performance.

1.2. Number of Iterations for Fine-tuning

We increased the number of iterations for fine-tuning, with the results presented in Figure 2. In the general task, which includes data from three sub-tasks, increasing fine-tuning iterations has a mixed impact on the evaluated metrics. During the early stages, performance improves across most metrics, including reductions in MM Dist and FID, along with higher Top-1/Top-2 accuracy and Diversity, reflecting enhanced sample quality, diversity, and text-motion alignment. However, exceeding 30 iterations tends to lead to overfitting, resulting in degraded generalization, as seen in increased MM Dist and fluctuations in Top-1/Top-2 accuracy. This suggests that prolonged fine-tuning reduces the model's ability to generalize across distributions and may lead to mode collapse, thereby negatively impacting diversity and FID. Optimal performance is observed between 20-30 iterations, where the balance between quality and generalization is most effectively maintained.

1.3. Hyper-parameter β of IPO

The effect of the IPO hyper-parameter β on alignment and quality metrics is illustrated in Figure 3. The optimal performance across most metrics is achieved at $\beta = 0.10$, where MM Dist and FID are minimized, and Top-1/Top-2 accuracy and Diversity reach their maximum values, indicating enhanced alignment, sample quality, and diversity. However, increasing β beyond 0.10 leads to increased MM Dist and FID, likely caused by an overemphasis on alignment objectives. In contrast, smaller β values (e.g., $\beta = 0.05$) fail to adequately align the model, resulting in suboptimal performance. These findings highlight the inherent trade-offs between alignment, quality, and diversity, underscoring the importance of setting $\beta = 0.10$ to effectively balance these competing objectives.

1.4. Preference Accuracy Comparison: GPT-4V vs. Contrastive Encoders on Human Preference Datasets

In the introduction, we noted that prior works, such as Mao *et al*. [2], have utilized contrastive pre-trained text and motion encoders from Guo *et al*. [1] to construct reward models. Following this approach, we conducted an evaluation to compare the alignment accuracy of the contrastive encoders and our proposed method against human preferences.

To assess the ability of GPT-4V and the contrastive encoders to align with human preferences, we evaluated their alignment accuracy using the human preference dataset provided by InstructMotion [3]. For each pair in the dataset (excluding those marked as "skipped"), GPT-4V was prompted to evaluate the motions and determine which one performed better, providing its preference directly. For the contrastive encoders, we encoded both the motion and text features separately and calculated the Euclidean distance between the two. The motion in the pair with the smaller distance to the text feature was considered the preferred sample. This setup allowed us to systematically compare the two approaches' alignment performance with human annotations.

As shown in Table 1, among the total of 2216 pairs, GPT-4V achieved an alignment accuracy of 69.77% , with 1546 aligned pairs. In contrast, the contrastive encoders exhibited a lower alignment accuracy of 66.11%, with 1465 aligned pairs out of the same total. These results highlight the superior capability of GPT-4V in capturing human preferences compared to the contrastive encoders, underscoring its potential for more effective human-centric applications.

Model	Aligned Pairs	Total Pairs	Accuracy (%)
GPT-4V	1546	2216	69.77
CE-based	1465	2216	66.11

Table 1. Alignment quality comparison between GPT-4V and contrastive encoder-based methods (denoted as CE-based)

1.5. GPT-4V Finetuned v.s. Constrastive Encoders Finetuned

To better demonstrate that our method outperforms approaches leveraging contrastive pre-trained encoders, we utilized these encoders to label motions generated in our temporal sub-task. The labeled preference data was then used to fine-tune MotionGPT, and a comparative analysis was conducted against AToM. As presented in Table 2, AToM consistently outperforms the approach based on contrastive encoders across key metrics, including MM Dist, R-precision, FID, and MultiModality, demonstrating its superior capability in alignment quality and generation variety. Although AToM exhibits slightly lower performance in the Diversity metric, its overall advantage across other critical metrics underscores its effectiveness and robustness compared to the contrastive encoder-based approach.

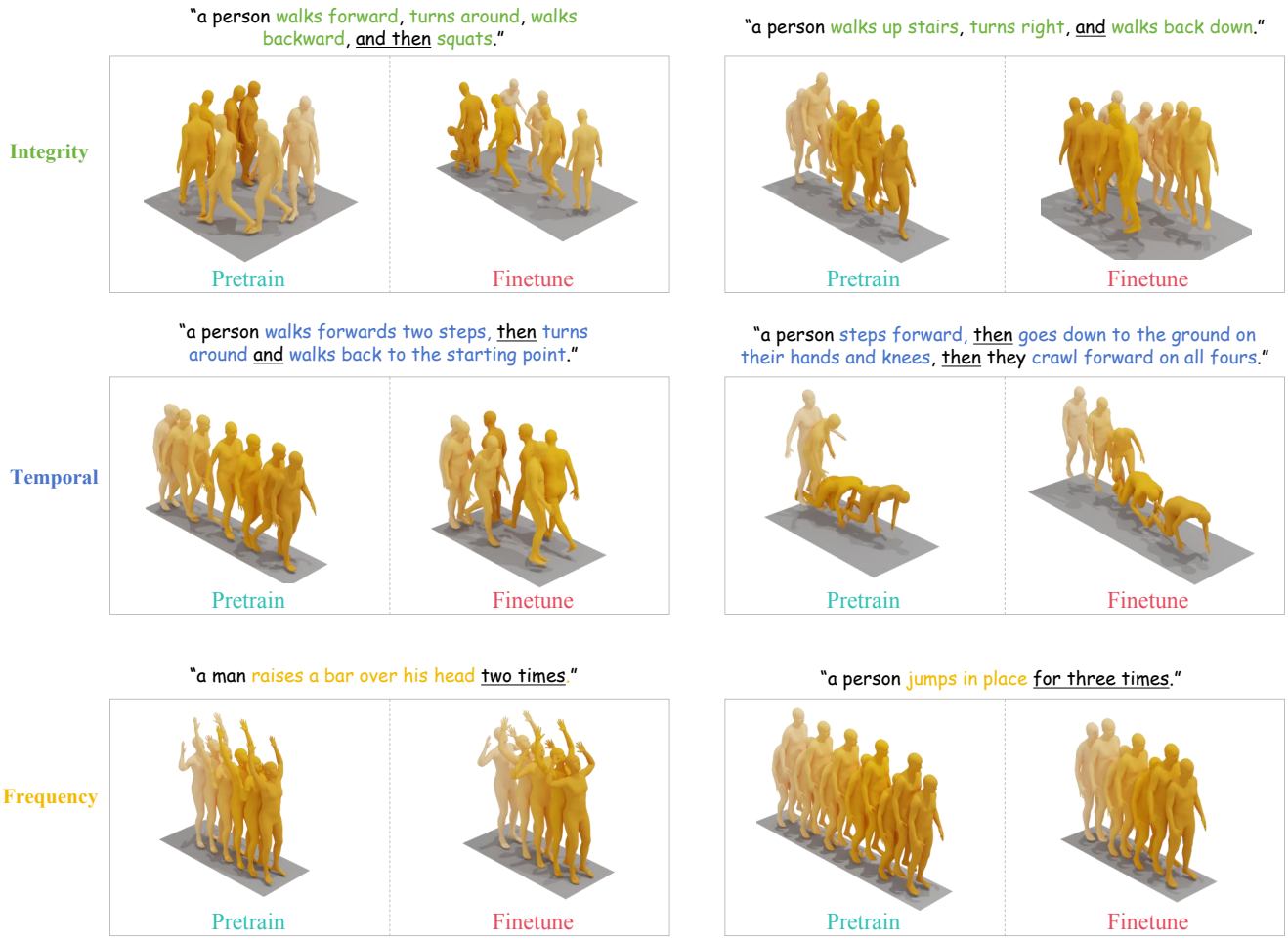


Figure 1. Generated qualitative samples comparison of pretrained model MotionGPT and finetuned model AToM.

Table 2. Comparison of methods AToM (ours) and contrastive encoder-based method.

Method	MM Dist↓	Top-1↑	Top-2↑	Top-3↑	FID↓	Diversity↑	MultiModality↑
AToM (ours)	$5.576^{\pm.027}$	$0.232^{\pm.004}$	$0.354^{\pm.006}$	$0.432^{\pm.005}$	$0.539^{\pm.031}$	$8.793^{\pm.074}$	$3.736^{\pm.147}$
CE-based	$5.664^{\pm.028}$	$0.188^{\pm.005}$	$0.307^{\pm.006}$	$0.392^{\pm.005}$	$0.556^{\pm.039}$	$8.843^{\pm.096}$	$3.724^{\pm.131}$

1.6. Influence of Preference Dataset Volume on Model Performance

The impact of preference pair quantity on alignment and quality metrics is depicted in Figure 4. At lower volumes (e.g., 2000 pairs), metrics such as MM Dist and FID are minimized, indicating better alignment and motion quality, while Diversity and MM Modality are relatively high, suggesting balanced performance. However, as the volume increases, MM Dist and FID worsen, and Top-1/Top-2 accuracy decreases significantly, likely due to over-fitting to a larger but potentially noisy set of preferences, which degrades generalization. The Diversity and MM Modality metrics exhibit fluctuations, with notable drops at intermediate volumes (e.g., 10,000 pairs) and partial recovery at higher volumes (14,000 pairs). These observations highlight the trade-off between data volume and model performance, where excessively large preference datasets may introduce noise, reducing alignment and diversity, and emphasizing the need for careful curation and optimal dataset sizing.

2. Details of `MotionPrefer` Construction

2.1. GPT-4 Instruction for Prompt Construction

We instruct GPT-4 to generate motion-event-based prompt. The designed instruction for three tasks are as follows:

The distribution of prompts with varying numbers of motion events is shown in the Table 4.

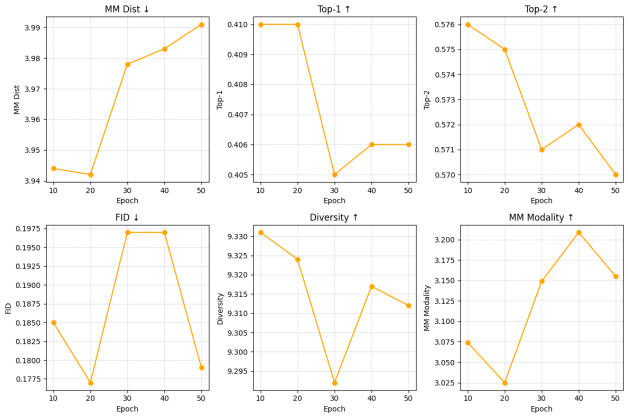


Figure 2. Impact of different epoch numbers on alignment and quality metrics.

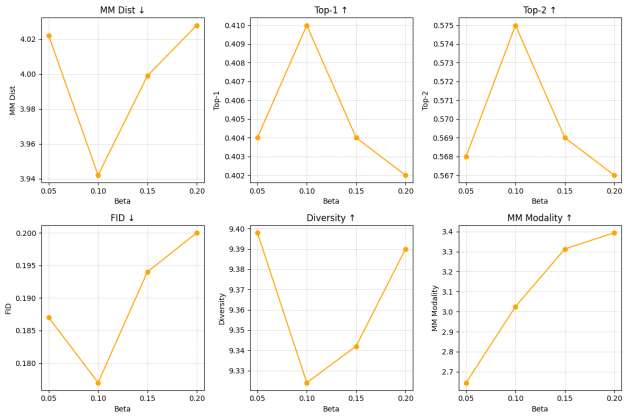


Figure 3. Impact of β on alignment and quality metrics.

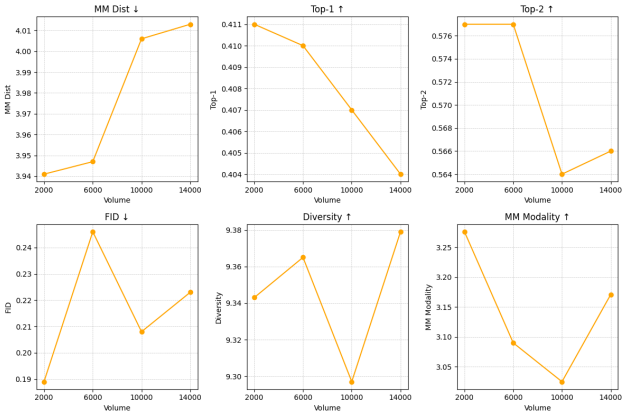


Figure 4. Impact of preference pair quantity on alignment and quality metrics.

2.2. GPT Instruction for Scoring

In this section, we outline the GPT-based instructions for scoring, providing a comprehensive framework for evaluating alignment between motion and description effectively.

Given a label group, $\mathrm{X_{task}}$, the 2-5 labels in it will be described into a motion event group: $\{\texttt{Event}_1, ..., \texttt{Event}_5\}$.
Then, please join the motion event group with conjunction to form a motion prompt, defined as "$\texttt{Event}_1, \texttt{Conjunction}_1, \texttt{Event}_2, \texttt{Conjunction}_2, ..., \texttt{Conjunction}_4, \texttt{Event}_5$".

Conjunction list: $\mathrm{X_{conj}}$

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 3. GPT instruction for prompt construction in integrity task.

Motion events	Prompt
2	250
3	1500
4	500
5	250

Table 4. Motion event number distribution.

Given a label group, $\mathrm{X_{task}}$, the 3 labels in it will be described into a motion event group: $\{\texttt{Event}_1, \texttt{Event}_2, \texttt{Event}_3\}$.
Then, please join the motion event group with conjunction to form a motion prompt, defined as "$\texttt{Event}_1, \texttt{Conjunction}_1, \texttt{Event}_2, \texttt{Conjunction}_2, \texttt{Event}_3$".

Conjunction list: $\mathrm{X_{conj}}$

Please ensure ...

Table 5. GPT instruction for prompt construction in temporal task.

Given a label group, $\mathrm{X_{task}}$, the single label in it will be described into a motion event: $\{\texttt{Event}_1\}$.
Then, please join the motion event with frequency to form a motion prompt, defined as "$\texttt{Event}_1, \texttt{Frequency}_1$".

Frequency list: $\mathrm{X_{freq}}$

Please ensure ...

Table 6. GPT instruction for prompt construction in frequency task.

Please evaluate the alignment between a given generative motion clip and the corresponding text description ("Input"). The description T, consists of 2-5 motions, and the motion clip is represented as a sequence of frames M.

For example:
T= "A man walks forward, walks backward, and squats" describes three motions."

Scoring Criteria:
5: All described motions appear in the frames.
0: Some motions are missing or incomplete in the frames.

Output Format:
Rating: [0 or 5]
Rationale: [A brief explanation for the rating, no more than 20 words]

Table 7. GPT annotation instruction for integrity task.

Please evaluate the alignment between a given generative motion clip and the corresponding text description ("Input"). The description T, describes 3 motions in temporal order, and the motion clip is represented as a sequence of frames M.

For example:
T= "A person walks forward , then is pushed to their right and then returns to walking in the line."

Scoring Criteria:
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.

Output Format:
Rating: [1 to 5]
Rationale: [A brief explanation for the rating, no more than 20 words]

Table 8. GPT annotation instruction for temporal task.

2.3. Motion Injection Forms in Questioning

There are demonstrations of three different forms of motion injection in questioning, as illustrated in Figure 5 to 7.

3. Human Evaluation

We present an example of the user study for the frequency task in Figure 8.

Please evaluate the alignment between a given generative motion clip and the corresponding text description ("Input"). The description T, describes a repeated motion for several times, and the motion clip is represented as a sequence of frames M.

For example:
T= "a person jumps forward three times."

Scoring Criteria:
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.

Output Format:
Rating: [1 to 3]
Rationale: [A brief explanation for the rating, no more than 20 words]

Table 9. GPT annotation instruction for frequency task.

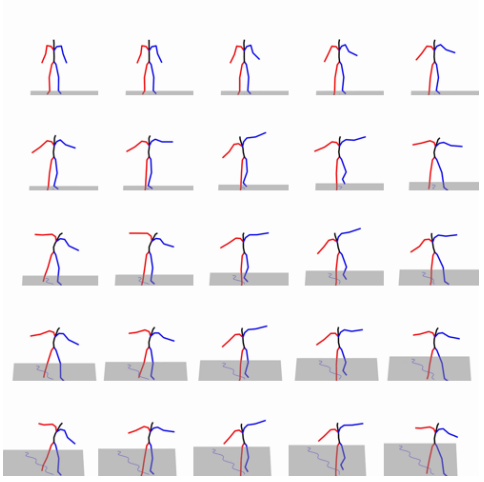


Figure 5. **Full-Image Example**

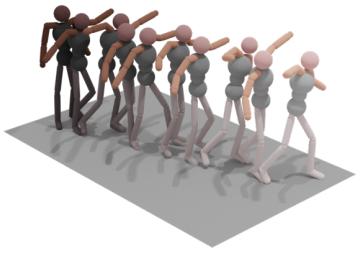


Figure 6. **Trajectory-Image Example**

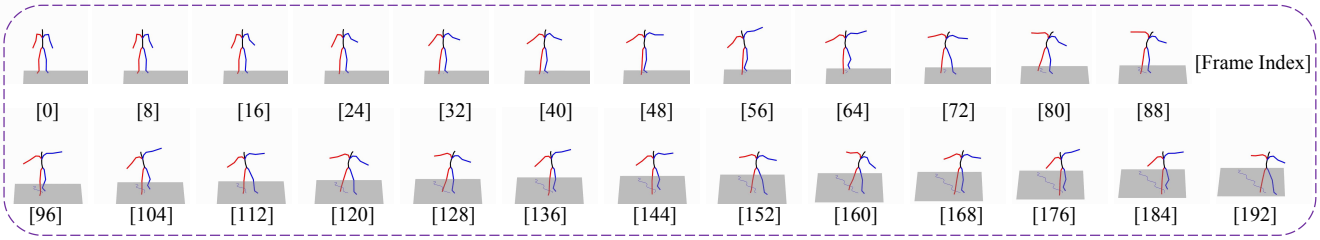


Figure 7. **Frame-by-Frame Example**

Task1: Frenquency

Task 1 Description (**Must Read Before Answering**)

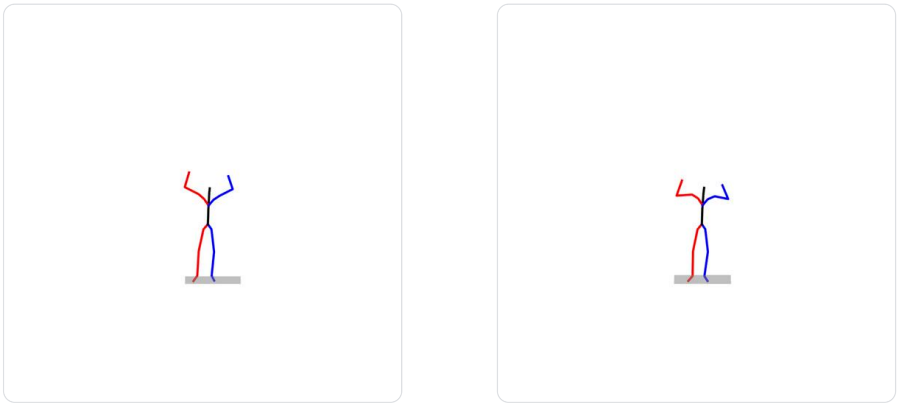
- **Target**: Select the video with higher quality **in terms of action frequency** from each pair of videos.
- **Definition of Action Frequency:**
  - Definition of High Quality: The video should **contain the target action**, and the action **appearance times should match the number described in the PROMPT**.
- Selection Criteria:
  - Select the video that **performs better** in terms of action frequency (vote).
  - If you think the two videos are of comparable quality in terms of action frequency, you can select the "**Equivalent Quality**" option.

1-1

PROMPT:

a person who has his arms raised head high raises his arms above his head and lowers them three times

Video:



questionnaire

Figure 8. **User study example of frequency task**

References

[1] 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. 1

[2] Yunyao Mao et al. Learning generalizable human motion generator with reinforcement learning. *arXiv preprint arXiv:2405.15541*, 2024. 1

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