UniMo: Unified Motion Generation and Understanding with Chain of Thought

Guocun Wang[1*], Kenkun Liu[2*], Jing Lin[3], Guorui Song[1], Jian Li[1†], Xiaoguang Han[2,4,5†]
[1]SIGS, Tsinghua University
[2]SSE, The Chinese University of Hong Kong (Shenzhen)
[3]MMLab, Nanyang Technological University
[4]FNii-Shenzhen
[5]Guangdong Provincial Key Laboratory of Future Networks of Intelligence
{wang-gc25, sgr24, l-j21}@mails.tsinghua.edu.cn, kenkunliu@link.cuhk.edu.cn, jing026@e.ntu.edu.sg, hanxiaoguang@cuhk.edu.cn

Abstract

Existing 3D human motion generation and understanding methods often exhibit limited interpretability, restricting effective mutual enhancement between these inherently related tasks. While current unified frameworks based on large language models (LLMs) leverage linguistic priors, they frequently encounter challenges in semantic alignment and task coherence. Moreover, the next-token prediction paradigm in LLMs is ill-suited for motion sequences, causing cumulative prediction errors. To address these limitations, we propose UniMo, a novel framework that integrates motion-language information and interpretable chain of thought (CoT) reasoning into the LLM via supervised fine-tuning (SFT). We further introduce reinforcement learning with Group Relative Policy Optimization (GRPO) as a post-training strategy that optimizes over groups of tokens to enforce structural correctness and semantic alignment, mitigating cumulative errors in motion token prediction. Extensive experiments demonstrate that UniMo significantly outperforms existing unified and task-specific models, achieving state-of-the-art performance in both motion generation and understanding.

Code — https://github.com/GuocunWang/UniMo

1 Introduction

Human motion generation and understanding is a pair of inverse problems. The formal aims to synthesize human motion from given input like textual instructions, and is a popular research field due to its various applications in animation, controllable video generation, gaming, virtual reality and etc (Hu 2024; Zhu et al. 2024; Qiu et al. 2025; Wang et al. 2024a; Song et al. 2025). By contrast, human motion understanding is to interpret the actions and intentions in the form of text descriptions from the given human motion. Previous methods (Chen et al. 2023; Zhang et al. 2024a; Ouyang et al. 2025; Guo et al. 2024) usually treat them as two tasks and address them separately. Despite their state-of-the-art performance in each single task, the mutually-inverse property of the two tasks is overlooked and these methods merely translate one modality to the other. Some efforts (Wu et al. 2024a; Jiang et al. 2023; Wang et al. 2024c) attempt to unify the two tasks into a single LLM-based model, but their performance in each single task is still unsatisfactory.

There are several reasons accounting for the situation. The first is that the given textual instruction is usually quite simple and coarse, so there might be many human motions could follow the instruction. In other words, the correspondence between the text and motion is too weak. Therefore, existing methods can hardly learn to generate fine-grained human actions. Second, unlike natural language that has strong sequential causality, human motion is essentially temporal sequence with weaker causality. Thus, the next-token-prediction paradigm is not the optimal training manner.

Therefore, to address the above mentioned issues, we propose the UniMo, the first unified model for motion generation and understanding with chain of thought (CoT). The model is built upon a pretrained open-sourced LLM, i.e. Qwen2.5-3B-Instruct (Qwen et al. 2025), to exploit their compressed knowledge. The key ingredients of of our model includes motion-consistent CoT data curation, CoT-guided motion generation and understanding, and RL-based joint training with task-specific rewards.

In the aspect of CoT data, MotionR1 (Ouyang et al. 2025), as a concurrent work, exploits the existing LLM and merely extend the original texts from the motion-language dataset like HumanML3D (Guo et al. 2022a) into step-by-step CoT data. Although simple and format-aligned, they cannot ensure the constructed CoT data is consistent with the corresponding motions. Using this data may lead to severe hallucination for trained model. By contrast, we render the motions via 3D Engine like Blender into videos, and then exploit the existing VLM model like Qwen2.5-VL (Bai et al. 2025) to describe the human motion in the rendered videos. The output descriptions are used as the CoT data, which are consistent with the original human motions.

With motion-consistent CoT data, we exploit them for CoT-guided unified modeling. For the task of text-to-motion (T2M) generation, instead of directly predicting motion tokens with input text, our model requires to first plan the motion generation process in the format of step-by-step CoT,

*These authors contributed equally.
†Corresponding author.

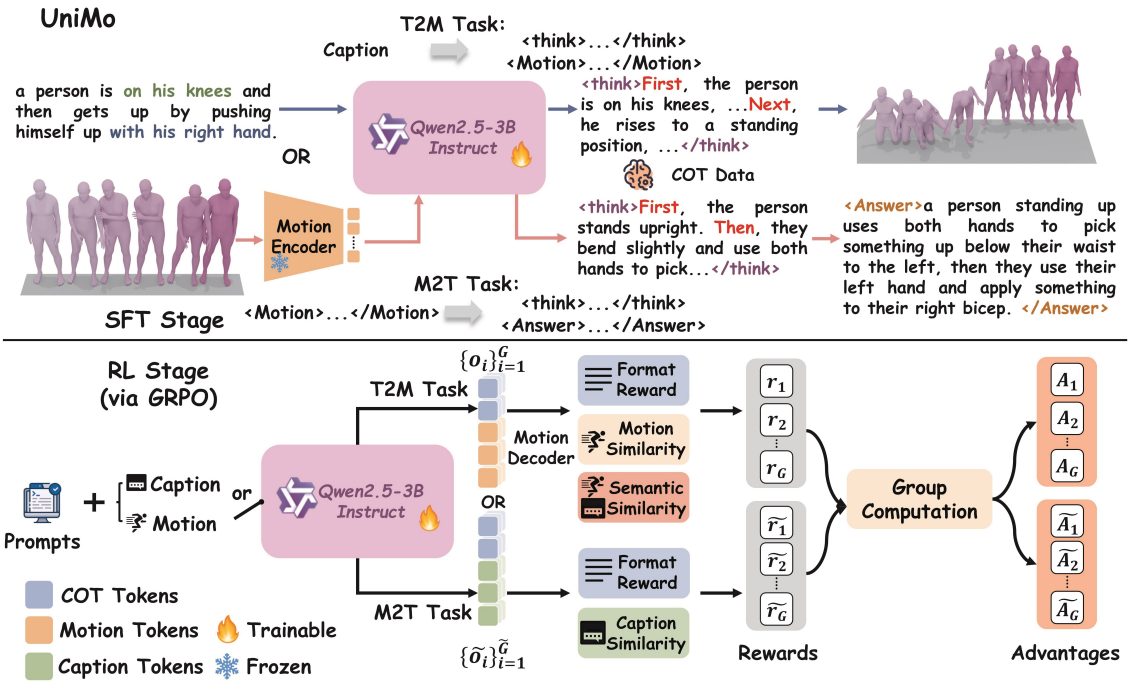

Figure 1: The UniMo is trained in two stages: the SFT stage and the reinforcement learning stage with GRPO. In the SFT stage, the model is teached to perform both T2M and M2T tasks with structured reasoning, i.e. CoT. In the RL stage, the model is further optimized with task-specific rewards, enabling unified and interpretable motion generation and understanding.

then start to generate motion tokens conditioned on the input text and the CoT. Similarly, for the task of the motion-to-text (M2T), the model also need to output the CoT before generate the final text description. By introducing the CoT as a bridge into our unified model, it provides more precise control signal for T2M generation and decompose the complex problem of M2T into easier subproblems.

As mentioned before, next-token-prediction paradigm is not optimal for the motion token prediction, so we resort to RL for post-training. By employing the Group Relative Policy Optimization (GRPO), we can force the model to follow the desired format and design rewards to increase the performance of the two tasks, which optimize multiple tokens simultaneously instead of a single token in next token prediction. Moreover, we jointly train the two tasks together and finally find that they can benefit each other.

With these designs, UniMo finally achieves state-of-the-art performance on both tasks, demonstrating the synergic effect of unifying generation and understanding with CoT for the first time. In all, our contributions are as follows:

- We propose UniMo, an LLM-based framework that unifies both motion generation and understanding with chain of thought reasoning.
- We design task-specific rewards for GRPO to supervise the RL-based joint training of both T2M and M2T tasks, which optimizes over groups of tokens and mitigates cumulative errors from next-token prediction.
- We curate a CoT dataset based on HumanML3D, which bridges the gap between motion and language.
- Our unified model outperforms existing state-of-the-art single-task methods in both tasks, proving that unified modeling can simultaneously improve motion generation and understanding through bidirectional synergy.

2 Related Works

2.1 Human Motion Generation

Earlier works (Ahuja and Morency 2019; Ghosh et al. 2021; Huang et al. 2020; Plappert, Mandery, and Asfour 2018) adopt a deterministic paradigm and usually obtain unsatisfactory performance. Then, the stochastic methods (Guo et al. 2022a,b; Chen et al. 2023; Zhang et al. 2024a; Tevet et al. 2022; Guo et al. 2024) gradually dominate the field and the text input also becomes a main kind of condition due to its flexibility. Representative GAN-based methods (Cai et al. 2018; Wang et al. 2020) are action-conditioned. With increasing popularity of diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Song, Meng, and Ermon 2020), some pioneer works like MDM (Tevet et al. 2022), MotionDiffuse (Zhang et al. 2024a), MLD (Chen et al. 2023) successfully apply diffusion models to the text-to-motion generation task and achieve competitive performance. Besides them, MoMask (Guo et al. 2024) adopts mask modeling for motion tokens and achieves the state-of-the-art performance. There are also some attempts (Zhang et al. 2023; Jiang et al. 2023; Wang et al. 2024c; Jiang et al. 2024; Ouyang et al. 2025) adopting the auto-regressive model. However, these AR-based models are still inferior to other methods in performance.

2.2 Unified Modeling of Motion Generation and Understanding

There are already some multi-modal LLMs (Xie et al. 2024; Chen et al. 2025b; Wang et al. 2024b; Tong et al. 2024; Pan et al. 2025; Deng et al. 2025; Chen et al. 2025a) that attempt to unify the visual understanding and generation in a single model, like GPT-4o, EMU-3 (Wang et al. 2024b) and Bagel (Deng et al. 2025). They excel in diffusion-based models (Esser et al. 2024; Lipman et al. 2022) in the aspects of complex instruction following and multiturn editing. Some previous works (Wang et al. 2024c; Wu et al. 2024a; Jiang et al. 2023; Zhang et al. 2024b) also explore the unified modeling of human motion generation and understanding. However, their performance is inferior to that of single-task models, due to the straggle between the two tasks.

2.3 Chain of Thought for Generative Models

Chain of thought (CoT) has been proven to be the key to unlocking the reasoning ability of LLMs (Wei et al. 2022; Guo et al. 2025a) as CoT can decompose a complex problem into multiple subproblems that are easier to solve. Drawing on the similar idea, some works propose to synthesize images with CoT (Deng et al. 2025; Guo et al. 2025b), i.e., generate detailed generation steps first, and then generate the image according to these detailed descriptions. With CoT, these methods could generate images from indirect prompts, requiring further thinking. MotionR1 (Ouyang et al. 2025), as a concurrent work, is the first to adopt CoT to enhance human motion generation. However, their CoT data is curated via a rewrite manner, which cannot ensure the consistency with original human motion.

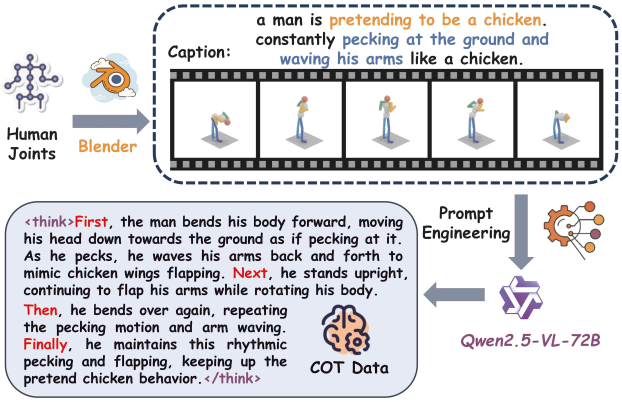

Figure 2: Illustration of the CoT annotation process. Human joint sequences are rendered in the Blender and paired with captions, which are further processed by the Qwen2.5-VL-72B to generate reasoning traces.

3 Method

3.1 Unified Modeling

As illustrated in Figure 1, UniMo incorporates structured, interpretable reasoning through CoT and achieves tight coupling between motion generation and understanding by joint training across tasks. The model is trained in two stages: the SFT stage with CoT annotations as a cold start, and a reinforcement learning stage using GRPO (Shao et al. 2024) with task-specific rewards.

3.2 Motion Representation

To integrate human motion into the LLM framework, we adopt the VQ-VAE as the motion tokenizer, following prior work (Zhang et al. 2023; Jiang et al. 2023; Wu et al. 2024a). This module discretizes continuous motion sequences into token indices, so they can be processed in the same manner as the LLM's tokens. Formally, given a human motion sequence expressed as a time series $\mathbf{m}_{1:T} \in \mathbb{R}^{T \times D}$, where T is the number of frames and D is the dimension of each frame. The encoder E maps this input to a latent feature sequence $\mathbf{z}_{1:T'} \in R^{T' \times d}$, $T' = T/l$, where d is the latent dimension and l is the downsampling factor. Each latent vector $\mathbf{z}_t$ is then quantized with a learnable codebook $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K}$, where K is the number of discrete motion codes and each $\mathbf{c}_k \in \mathbb{R}^d$. The quantized representation can be represented as:

$$\hat{\mathbf{z}}_t = \arg\min_{\mathbf{c}_k \in \mathcal{C}} \|\mathbf{z}_t - \mathbf{c}_k\|_2. \tag{1}$$

To reconstruct the original motion sequence, a decoder D maps the quantized tokens back into the continuous space: $\hat{\mathbf{m}}_{1:T} = D(\hat{\mathbf{z}}_{1:T'})$. The motion tokenizer is trained using a composite loss that includes the reconstruction loss $\mathcal{L}_{\text{recon}}$, the embedding loss $\mathcal{L}_{\text{embed}}$ and the commitment loss $\mathcal{L}_{\text{commit}}$. Following the prior work (Zhang et al. 2023), the objective function is:

$$\begin{aligned} \mathcal{L}_{\text{VQ}} &= \mathcal{L}_{\text{recon}} + \underbrace{\|\mathbf{z}_{1:T'} - \text{sg}[\hat{\mathbf{z}}_{1:T'}]\|_2}_{\mathcal{L}_{\text{embed}}} \\ &\quad + \underbrace{\|\text{sg}[\mathbf{z}_{1:T'}] - \hat{\mathbf{z}}_{1:T'}\|_2}_{\mathcal{L}_{\text{commit}}}, \end{aligned} \tag{2}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator. We employ exponential moving average updates for the codebook, along with a codebook reset strategy. The VQ-VAE is frozen during subsequent training stages.

3.3 Supervised Fine-tuning with CoT

In the SFT stage, UniMo leverages high-quality CoT annotations to bridge the gap between natural language and human motion. As shown in the Figure 2, we first render the motion sequences in the HumanML3D dataset into video clips using Blender, and use the Qwen2.5-VL-72B (Bai et al. 2025) to generate CoT annotations. For each motion-caption pair, the Qwen2.5-VL-72B is prompted with both the rendered video and its original caption as input, and then instructed to produce structured step-by-step reasoning under the format:`<think>{CoT}</think>`. Unlike Motion-R1 (Ouyang et al. 2025), which generates CoT solely from the captions using DeepSeek-R1 (Guo et al. 2025a), this multimodal annotation strategy ensures that the generated CoT is not only aligned with the caption semantics but also grounded in the physical realization of the motion, as observed from the video.

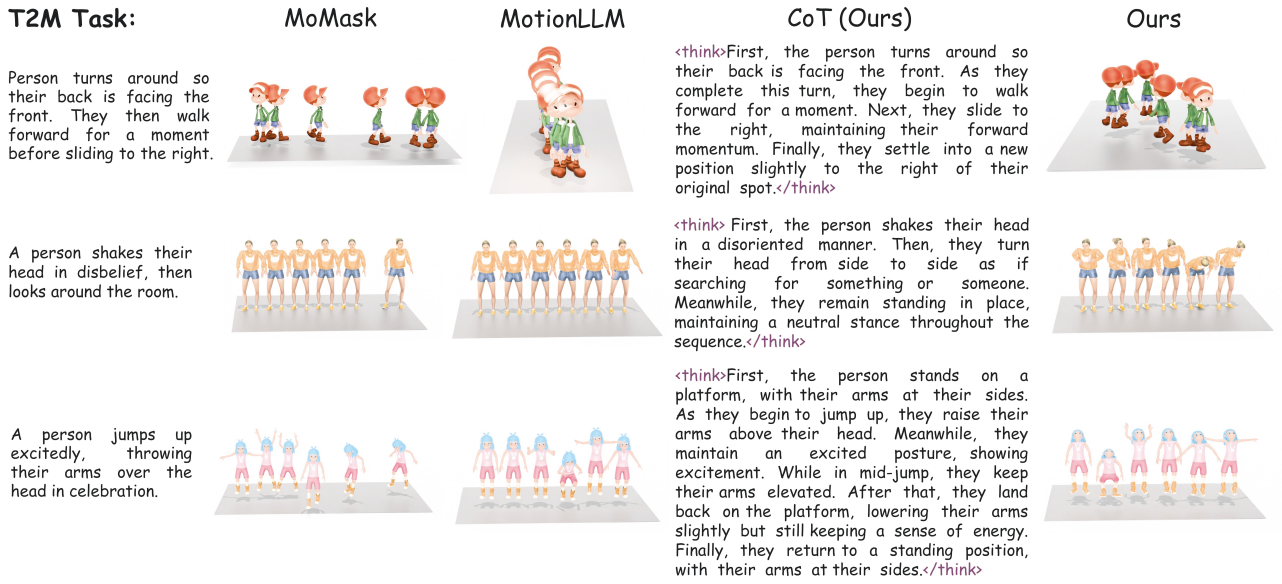

Figure 3: Qualitative comparison of our method with other open-source SOTAs such as MoMask (Guo et al. 2024) and MotionLLM (Wu et al. 2024a) for the text-to-motion task. Our method presents stronger instruction-following capability and can generate sequential actions. We highly recommend the readers to watch the video comparisons in supplementary materials.

The introduction of curated CoT in the SFT stage offers two advantages over previous methods (Ouyang et al. 2025; Wu et al. 2024a; Jiang et al. 2023). First, training on paired motion, caption and CoT data encourages the model to generate reasoning aligned with real motion dynamics, thus enabling it to acquire robust semantic grounding and explicit interpretability within a unified framework. Second, this stage serves as a principled initialization, equipping the model with strong bidirectional mapping and structured reasoning abilities before reinforcement learning.

3.4 Reinforcement Learning with GRPO

Group Relative Policy Optimization. To further align the model outputs with structural and semantic expectations under the paradigm of reinforcement learning, we adopt GRPO (Shao et al. 2024), a variant of PPO (Schulman et al. 2017) tailored for generation tasks. GRPO evaluates multiple sampled completions per input prompt, estimating the advantage function via within-group comparisons, which reduces variance in policy updates and stabilizes the learning process without requiring an explicit critic network. It integrates a clipping range with KL regularization, ensuring policy updates remain stable and consistent with the SFT-trained reference policy. The training objective is formally defined as:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) \quad = \quad & E_c \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)} \hat{A}_i, \right. \right. \\ & \left. \left. \text{clip} \left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon \right) \hat{A}_i \right) \right] \\ & -\beta \cdot D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}), \end{aligned} \tag{3}$$

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, ..., r_G\})}{\text{std}(\{r_1, ..., r_G\})}, \tag{4}$$

where q denotes the input prompt, and o_i represents the i-th sampled output among a group of G completions. The normalized advantage $\hat{A}_i$ is computed from the corresponding reward signals. The ratio $\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}$ measures the relative probability assigned to the same output o_i under the new policy π_θ versus the old policy π_{old}. The clipping parameter ε prevents drastic changes in the policy by restricting this ratio within a conservative range. Additionally, the KL divergence is used to regularize deviations from the reference policy π_{ref}, weighted by the coefficient β.

Format Reward. To ensure format compliance, we introduce the format reward r_{format} that encourages the model to think before answering. For the T2M task, the required format is `<think>{CoT}</think><Motion>{Motion Tokens}</Motion>`. For the M2T task, it is `<think>{CoT}</think><Answer>{Caption}</Answer>`. Answers that strictly follow these formats get the reward of 1, otherwise they get 0.

Motion Similarity Reward. The motion similarity reward r_{motion} is introduced to encourage realistic and coherent motion generation in the T2M task. To encourage the model to produce motions that closely resemble the true temporal dynamics and spatial patterns of the targets, it leverages the pretrained motion encoder f_{motion} from (Guo et al. 2022a) to compute the cosine similarity between the generated motion $\hat{\mathbf{m}}$ and the ground-truth motion $\mathbf{m}$:

$$r_{\text{motion}} = \frac{f_{\text{motion}}(\hat{\mathbf{m}}) \cdot f_{\text{motion}}(\mathbf{m})}{\|f_{\text{motion}}(\hat{\mathbf{m}})\|_2 \cdot \|f_{\text{motion}}(\mathbf{m})\|_2}. \tag{5}$$

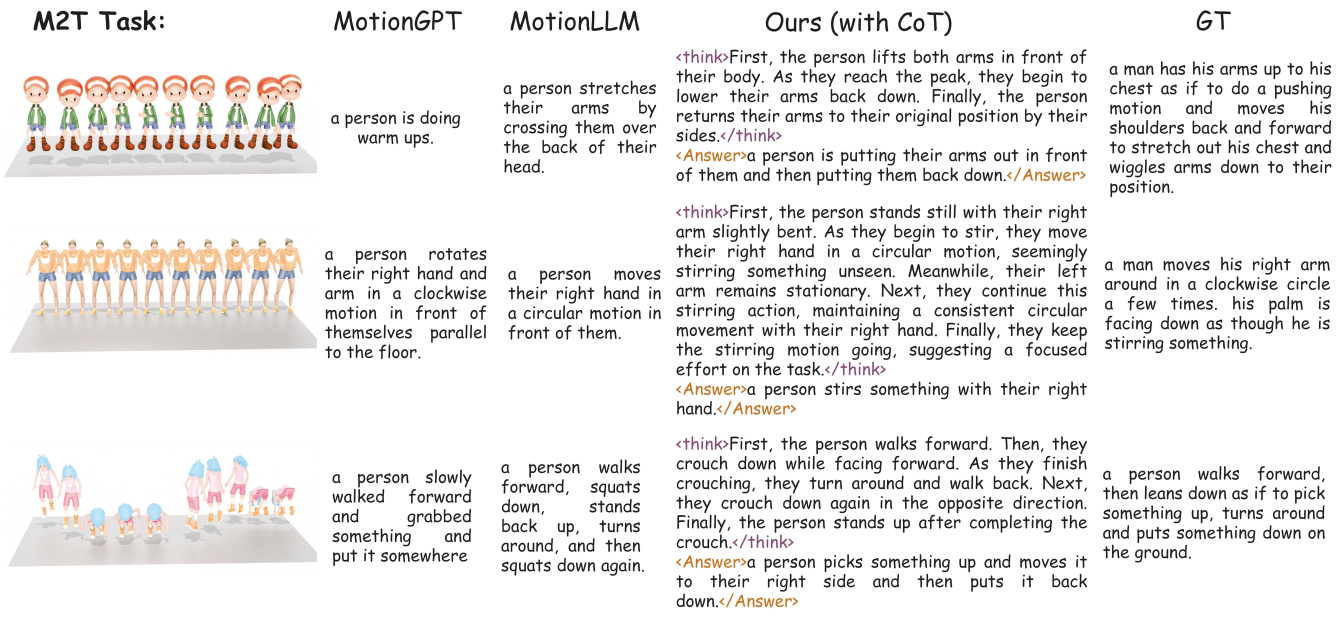

Figure 4: Qualitative comparison of our method with other open-source SOTAs such as MotionGPT (Jiang et al. 2023) and MotionLLM (Wu et al. 2024a) for the motion-to-text task. Our method can more precisely describe complex motions.

Semantic Similarity Reward. To maintain semantic alignment between the input caption T and the generated motion $\hat{\mathbf{m}}$ in the T2M task, the semantic similarity reward $r_{semantic}$ uses pretrained encoders f_{motion} and f_{text} to compute cosine similarity in a shared embedding space, which are from the prior work (Guo et al. 2022a). This ensures generated motions faithfully reflect the input textual semantics. The formula for the semantic similarity reward is as follows:

$$r_{\text{semantic}} = \frac{f_{\text{motion}}(\hat{\mathbf{m}}) \cdot f_{\text{text}}(T)}{\|f_{\text{motion}}(\hat{\mathbf{m}})\|_2 \cdot \|f_{\text{text}}(T)\|_2}. \tag{6}$$

Caption Similarity Reward. In the M2T task, the caption similarity reward $r_{caption}$ is obtained with the CLIP (Radford et al. 2021) text encoder $f^{\text{text}}_{\text{CLIP}}$ as the cosine similarity between the generated caption $\hat{T}$ and the reference caption T. This reward is then doubled to match the combined scale of the r_{motion} and $r_{semantic}$ employed in the T2M task. This process is demonstrated in the following formula:

$$r_{\text{caption}} = 2 \cdot \frac{f^{\text{text}}_{\text{CLIP}}(\hat{T}) \cdot f^{\text{text}}_{\text{CLIP}}(T)}{\|f^{\text{text}}_{\text{CLIP}}(\hat{T})\|_2 \cdot \|f^{\text{text}}_{\text{CLIP}}(T)\|_2}. \tag{7}$$

4 Experiments

4.1 HumanML3D with Curated CoT Annotations

The HumanML3D dataset (Guo et al. 2022a) is a large-scale benchmark designed for evaluating human motion generation and understanding from natural language. It contains over 14,616 human motion clips sourced from AMASS (Mahmood et al. 2019) and HumanAct12 (Guo et al. 2020), paired with nearly 44,970 textual descriptions. Each motion sequence is represented as a series of 3D joint positions over time, providing rich temporal and spatial information for modeling complex human behaviors. In this work, we further augment HumanML3D with curated CoT annotations, providing step-by-step reasoning traces that explicitly link language and motion.

4.2 Evaluation Metrics

For the T2M task, evaluation metrics from the (Guo et al. 2022a) include R-Precision at Top-1, Top-2, and Top-3, FID, MM-Dist, Diversity and MModality, which together reflect the realism and diversity of the generated motions. For the motion-to-text task, we report BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and BertScore (Zhang et al. 2019), providing a comprehensive evaluation of the linguistic and semantic quality of the generated captions.

4.3 Implementation Details

UniMo is trained in two stages on $8\times$A100 GPUs with the global batch size of 8, with Qwen2.5-3B-Instruct (Qwen et al. 2025) serving as the foundational model. During the SFT stage, the model is first adapted to the motion modality by extending the vocabulary with motion tokens and pretrained on the T2M task for 10 epochs. This is followed by joint fine-tuning on both T2M and M2T tasks for another 10 epochs. The Adam optimizer is employed with the learning rate of 1×10^{-4}, following the cosine decay.

In the RL stage, we employ the commonly-used GRPO (Shao et al. 2024), and the training process is conducted for 14,000 steps, with the learning rate of 5×10^{-5}. Each prompt is completed with $G = 8$ samples to compute the rewards, and gradient clipping with a maximum norm of 0.1 is ap-

Methods	R-Precision ↑			FID ↓	MM-Dist↓	Diversity↑	MModality↑
	Top1	Top2	Top3				
MDM (Tevet et al. 2022)	$0.320^{\pm0.005}$	$0.498^{\pm0.004}$	$0.611^{\pm0.007}$	$0.544^{\pm0.044}$	$5.566^{\pm0.027}$	$9.559^{\pm0.086}$	$\mathbf{2.799}^{\pm0.074}$
MLD (Chen et al. 2023)	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	$0.473^{\pm0.013}$	$3.196^{\pm0.010}$	$9.724^{\pm0.082}$	$2.413^{\pm0.072}$
MotionDiffuse (Zhang et al. 2024a)	$0.491^{\pm0.001}$	$0.681^{\pm0.001}$	$0.782^{\pm0.001}$	$0.630^{\pm0.001}$	$3.113^{\pm0.001}$	$9.410^{\pm0.049}$	$1.553^{\pm0.064}$
T2M (Guo et al. 2022a)	$0.457^{\pm0.002}$	$0.559^{\pm0.007}$	$0.740^{\pm0.003}$	$1.067^{\pm0.002}$	$3.340^{\pm0.008}$	$9.188^{\pm0.002}$	$2.090^{\pm0.088}$
TM2T (Guo et al. 2022b)	$0.424^{\pm0.003}$	$0.618^{\pm0.003}$	$0.729^{\pm0.002}$	$1.501^{\pm0.017}$	$3.467^{\pm0.011}$	$8.589^{\pm0.076}$	$\underline{2.424}^{\pm0.079}$
T2M-GPT (Zhang et al. 2023)	$0.491^{\pm0.003}$	$0.680^{\pm0.003}$	$0.775^{\pm0.002}$	$\underline{0.116}^{\pm0.004}$	$3.118^{\pm0.011}$	$9.761^{\pm0.081}$	$1.856^{\pm0.111}$
MotionGPT (Jiang et al. 2023)	$0.492^{\pm0.003}$	$0.681^{\pm0.003}$	$0.778^{\pm0.002}$	$0.232^{\pm0.008}$	$3.096^{\pm0.008}$	$9.528^{\pm0.071}$	$2.008^{\pm0.083}$
MoMask (Guo et al. 2024)	$\underline{0.521}^{\pm0.002}$	$0.713^{\pm0.002}$	$0.807^{\pm0.002}$	$\mathbf{0.045}^{\pm0.002}$	$2.958^{\pm0.008}$	$9.620^{\pm0.064}$	$1.241^{\pm0.064}$
MotionChain (Jiang et al. 2024)	$0.504^{\pm0.003}$	$0.617^{\pm0.002}$	$0.790^{\pm0.003}$	$0.248^{\pm0.009}$	$3.033^{\pm0.010}$	$9.470^{\pm0.075}$	$1.727^{\pm0.014}$
MotionLLM (Wu et al. 2024a)	$0.515^{\pm0.004}$	$0.691^{\pm0.003}$	$0.801^{\pm0.004}$	$0.230^{\pm0.009}$	$2.967^{\pm0.020}$	$9.908^{\pm0.102}$	$2.142^{\pm0.014}$
MotionGPT-2 (Wang et al. 2024c)	$0.496^{\pm0.002}$	$0.691^{\pm0.003}$	$0.782^{\pm0.004}$	$0.191^{\pm0.004}$	$3.080^{\pm0.013}$	$9.860^{\pm0.026}$	$2.137^{\pm0.022}$
Motion-R1 (Ouyang et al. 2025)	$0.515^{\pm0.003}$	$\underline{0.719}^{\pm0.002}$	$\underline{0.818}^{\pm0.002}$	$0.201^{\pm0.004}$	$\underline{2.854}^{\pm0.010}$	$\underline{10.026}^{\pm0.075}$	$2.317^{\pm0.105}$
UniMo (Ours)	$\mathbf{0.539}^{\pm0.003}$	$\mathbf{0.738}^{\pm0.002}$	$\mathbf{0.831}^{\pm0.002}$	$0.177^{\pm0.004}$	$\mathbf{2.768}^{\pm0.010}$	$\mathbf{10.042}^{\pm0.076}$	$1.924^{\pm0.080}$

Table 1: Quantitative results of the T2M task on the HumanML3D dataset. Each evaluation is repeated 20 times with average metrics and 95% confidence intervals. The best scores are highlighted in bold, and the second-best scores are underlined.

Captioning	BLEU@1↑	BLEU@4↑	ROUGE-L↑	CIDEr↑	BertScore↑
TM2T (Guo et al. 2022b)	48.90	8.27	38.1	15.80	32.2
LaMPM2T (Li et al. 2024)	47.8	13.04	37.1	28.9	32.7
MoTe (Wu et al. 2024b)	46.7	11.15	37.4	31.5	30.3
MotionGPT (Jiang et al. 2023)	48.20	12.47	37.4	29.20	32.4
MotionGPT-2 (Wang et al. 2024c)	48.7	13.8	37.6	29.8	32.6
MotionChain (Jiang et al. 2024)	48.10	12.56	33.9	33.70	36.9
MotionLLM (Wu et al. 2024a)	<u>54.53</u>	<u>17.65</u>	<u>48.7</u>	<u>33.74</u>	<u>42.63</u>
UniMo (Ours)	**63.10**	**19.74**	**48.8**	**46.69**	**54.26**

Table 2: Quantitative results of the M2T task on the HumanML3D dataset. The best scores are highlighted in bold, and the second-best scores are underlined.

CoT	r_{motion}	r_{semantic}	r_{caption}	FID↓	Diversity↑	R-Precision↑			B@1↑	B@4↑	R-L↑	CIDEr↑	Bert↑
						Top1	Top2	Top3					
				<u>0.178</u>	9.873	0.384	0.559	0.659	56.64	16.34	44.2	36.56	49.01
	✓		✓	0.218	9.694	0.509	0.706	0.807	61.91	19.15	47.5	45.26	52.94
		✓	✓	0.255	9.937	0.534	**0.738**	**0.834**	60.99	18.86	47.3	44.52	52.68
	✓	✓	✓	0.180	9.771	0.533	0.735	<u>0.832</u>	58.43	16.63	45.7	41.08	51.58
✓				0.292	9.726	0.460	0.642	0.735	55.34	15.31	43.2	32.08	48.16
✓	✓		✓	0.179	<u>9.994</u>	0.518	0.712	0.808	62.75	<u>19.95</u>	<u>48.4</u>	44.45	53.98
✓		✓	✓	0.240	9.822	<u>0.537</u>	<u>0.737</u>	0.829	**63.40**	**20.41**	**48.8**	<u>45.94</u>	**54.42**
✓	✓	✓	✓	**0.177**	**10.042**	**0.539**	**0.738**	0.831	<u>63.10</u>	19.74	**48.8**	**46.69**	<u>54.26</u>

Table 3: Ablation study on the effectiveness of CoT and different GRPO reward components on the HumanML3D dataset.

plied. For stable policy optimization, GRPO utilizes the clipping range of $\varepsilon = 0.2$, and the KL coefficient of $\beta = 0.001$.

4.4 Comparisons with the State of the Arts

Text-to-Motion Results. For the T2M task, we conduct comparisons with a comprehensive set of strong baselines, including diffusion-based methods, transformer-based methods, as well as recent LLM-based approaches. All evaluations follow the same experimental settings as previous studies to ensure fair and consistent comparison, with each method evaluated 20 independent runs and results reported as the average within a 95% confidence interval.

The quantitative results for the motion generation task are presented in Table 1. UniMo outperforms state-of-the-art methods in most metrics with an obvious margin. Although MoMask (Guo et al. 2024) reports a lower FID, it sacrifices the MModality score at 1.241, whereas UniMo achieves a higher MModality score of 1.924. Overall, UniMo demonstrates a robust balance between motion realism and semantic expressiveness. Qualitative comparisons are shown in

Train Task	Stage	R-Precision↑			FID↓	Diversity↑	MM-Dist↓
		Top1	Top2	Top3			
T2M	SFT	0.438	0.613	0.702	0.201	9.748	3.588
T2M	SFT+RL	0.529	**0.739**	**0.832**	0.203	9.780	**2.743**
T2M+M2T	SFT	0.460	0.642	0.735	0.292	9.726	3.360
T2M+M2T	SFT+RL	**0.539**	0.738	0.831	**0.177**	**10.042**	2.768

Table 4: Ablation study of the synergy effect of unified modeling for the T2M task.

Train Task	Stage	BLEU@1↑	BLEU@4↑	ROUGE-L↑	CIDEr↑	BertScore↑
M2T	SFT	54.62	14.74	43.3	31.05	47.88
M2T	SFT+RL	61.91	18.89	47.5	42.13	52.88
T2M+M2T	SFT	55.34	15.31	43.2	32.08	48.16
T2M+M2T	SFT+RL	**63.10**	**19.74**	**48.8**	**46.69**	**54.26**

Table 5: Ablation study of the synergy effect of unified modeling for the M2T task.

Figure 3, our method exhibits stronger instruction-following ability and generates more coherent sequential actions than the open-source methods such as MoMask (Guo et al. 2024) and MotionLLM (Wu et al. 2024a).

Motion-to-Text Results. Table 2 shows the results for the motion captioning task. UniMo achieves the new state-of-the-art results across all evaluation metrics. Compared to the strongest previous baseline MotionLLM (Wu et al. 2024a), UniMo shows the substantial improvement in all metrics. The results demonstrate the effectiveness of UniMo in generating motion captions that are more precise, relevant, and consistent with the input motion sequences. Figure 4 demonstrates the superior capability of our method in generating detailed and semantically aligned motion captions compared to other SOTA methods, i.e. MotionGPT (Jiang et al. 2023) and MotionLLM (Wu et al. 2024a).

4.5 Ablation Studies

The effectiveness of CoT. Experimental results in Table 3 demonstrate that introducing CoT leads to a significant improvement across multiple metrics. In the T2M task, CoT improves semantic alignment: Top-1 R-Precision increases from 0.384 to 0.460. Although FID slightly increases when CoT is used alone, combining CoT with the rewards achieves the lowest FID of 0.177 and the best balance across all metrics, confirming the overall benefit from the introduction of CoT. In the T2M task, BLEU@1 rises from 58.43 without CoT to 63.10 when CoT is applied, and CIDEr improves from 41.08 to 46.69. Similar gains are observed across BLEU@4, ROUGE-L and BertScore, demonstrating that CoT helps the model generate more semantically grounded and accurate motion descriptions.

The effectiveness of RL with task-specific rewards We further study the impact of reinforcement learning with task-specific rewards. As shown in Table 3, adding the motion similarity reward r_{motion} notably reduces the FID from 0.240 to 0.177 when combined with CoT and other rewards. The semantic similarity reward r_{semantic} has a marked impact on R-Precision, particularly boosting Top-1 to 0.539 when used together with CoT, r_{motion} and r_{caption}. The caption similarity reward r_{caption} significantly enhances the quality of generated descriptions in the M2T task, as shown by improvements in all textual metrics.

The synergy effect of unified modeling To investigate the effectiveness of unified modeling, we compare the performance of models trained solely on single tasks (T2M or M2T) with those trained jointly on both tasks. Experimental results are shown in Table 4 and 5. For the task of T2M, without RL, the unified modeling performs better than training T2M solely in all metrics. With RL, the unified modeling is also generally better than the single task of T2M. For the task of M2T, unified modeling is signficantly better than training the single task of M2T in all settings and all metrics. These results demonstrate the benefits of integrating generation and understanding tasks within a unified framework.

5 Conclusion

In this paper, we present UniMo, a unified framework for 3D human motion generation and understanding based on large language models. Our approach integrates CoT reasoning with GRPO to achieve joint modeling of T2M and M2T tasks in a structured and interpretable manner. Building upon the HumanML3D dataset, we construct a new motion-language dataset with motion-consistent CoT annotations. By designing task-specific rewards, UniMo achieves stronger semantic alignment between motion and language, and supports interpretable reasoning in both tasks. Importantly, we adopt RL with task-specific rewards, which enables the model to optimize over groups of tokens and mitigates the cumulative errors in the next-token prediction paradigm of LLMs for the motion token prediction. Extensive experiments demonstrate that UniMo achieves state-of-the-art results on both motion generation and understanding.

Acknowledgments

This work was supported by the Shenzhen Science and Technology Project under Grant KJZD20240903103210014, in part by NSFC with Grant No. 62293482, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Guangdong Provincial Outstanding Youth Fund(No. 2023B1515020055), by Shenzhen Science and Technology Program No. JCYJ20220530143604010, NSFC No.62172348.

References

Ahuja, C.; and Morency, L.-P. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International conference on 3D vision (3DV)*, 719–728. IEEE.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cai, H.; Bai, C.; Tai, Y.-W.; and Tang, C.-K. 2018. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European conference on computer vision (ECCV)*, 366–382.

Chen, J.; Xu, Z.; Pan, X.; Hu, Y.; Qin, C.; Goldstein, T.; Huang, L.; Zhou, T.; Xie, S.; Savarese, S.; et al. 2025a. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18000–18010.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; Zhong, S.; Yu, W.; Nie, X.; Song, Z.; et al. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1396–1406.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1910.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5152–5161.

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

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM international conference on multimedia*, 2021–2029.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, Z.; Zhang, R.; Tong, C.; Zhao, Z.; Gao, P.; Li, H.; and Heng, P.-A. 2025b. Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step. *arXiv preprint arXiv:2501.13926*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Huang, R.; Hu, H.; Wu, W.; Sawada, K.; Zhang, M.; and Jiang, D. 2020. Dance revolution: Long-term dance generation with music via curriculum learning. In *International conference on learning representations*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36: 20067–20079.

Jiang, B.; Chen, X.; Zhang, C.; Yin, F.; Li, Z.; Yu, G.; and Fan, J. 2024. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, 54–74. Springer.

Li, Z.; Yuan, W.; He, Y.; Qiu, L.; Zhu, S.; Gu, X.; Shen, W.; Dong, Y.; Dong, Z.; and Yang, L. T. 2024. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. *arXiv preprint arXiv:2410.07093*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Ouyang, R.; Li, H.; Zhang, Z.; Wang, X.; Zhu, Z.; Huang, G.; and Wang, X. 2025. Motion-R1: Chain-of-Thought Reasoning and Reinforcement Learning for Human Motion Generation. *arXiv preprint arXiv:2506.10353*.

Pan, X.; Shukla, S. N.; Singh, A.; Zhao, Z.; Mishra, S. K.; Wang, J.; Xu, Z.; Chen, J.; Li, K.; Juefei-Xu, F.; et al. 2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Plappert, M.; Mandery, C.; and Asfour, T. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109: 13–26.

Qiu, L.; Gu, X.; Li, P.; Zuo, Q.; Shen, W.; Zhang, J.; Qiu, K.; Yuan, W.; Chen, G.; Dong, Z.; et al. 2025. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

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

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

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Song, G.; Wang, G.; Huang, Z.; Lin, J.; Zhe, X.; Li, J.; and Wang, H. 2025. Towards Fine-Grained Human Motion Video Captioning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 846–855.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

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

Tong, S.; Fan, D.; Zhu, J.; Xiong, Y.; Chen, X.; Sinha, K.; Rabbat, M.; LeCun, Y.; Xie, S.; and Liu, Z. 2024. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*.

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

Wang, T.; Li, L.; Lin, K.; Zhai, Y.; Lin, C.-C.; Yang, Z.; Zhang, H.; Liu, Z.; and Wang, L. 2024a. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9326–9336.

Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Wang, Y.; Huang, D.; Zhang, Y.; Ouyang, W.; Jiao, J.; Feng, X.; Zhou, Y.; Wan, P.; Tang, S.; and Xu, D. 2024c. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*.

Wang, Z.; Yu, P.; Zhao, Y.; Zhang, R.; Zhou, Y.; Yuan, J.; and Chen, C. 2020. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12281–12288.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Q.; Zhao, Y.; Wang, Y.; Liu, X.; Tai, Y.-W.; and Tang, C.-K. 2024a. Motion-agent: A conversational framework for human motion generation with llms. *arXiv preprint arXiv:2405.17013*.

Wu, Y.; Ji, W.; Zheng, K.; Wang, Z.; and Xu, D. 2024b. MoTe: Learning Motion-Text Diffusion Model for Multiple Generation Tasks. *arXiv preprint arXiv:2411.19786*.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; and Shan, Y. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14730–14740.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024a. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6): 4115–4128.

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

Zhang, Y.; Huang, D.; Liu, B.; Tang, S.; Lu, Y.; Chen, L.; Bai, L.; Chu, Q.; Yu, N.; and Ouyang, W. 2024b. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7368–7376.

Zhu, S.; Chen, J. L.; Dai, Z.; Dong, Z.; Xu, Y.; Cao, X.; Yao, Y.; Zhu, H.; and Zhu, S. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, 145–162. Springer.