# HMVLM: Human Motion-Vision-Lanuage Model via MoE LoRA

**Lei Hu**[1,2*] **Yongjing Ye**[1*] **Shihong Xia**[1,2†]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{hulei19z, yeyongjing, xsh}@ict.ac.cn

## Abstract

The expansion of instruction-tuning data has enabled foundation language models to exhibit improved instruction adherence and superior performance across diverse downstream tasks. Semantically-rich 3D human motion is being progressively integrated with these foundation models to enhance multimodal understanding and cross-modal generation capabilities. However, the modality gap between human motion and text raises unresolved concerns about catastrophic forgetting during this integration. In addition, developing autoregressive-compatible pose representations that preserve generalizability across heterogeneous downstream tasks remains a critical technical barrier. To address these issues, we propose the Human Motion-Vision-Language Model (HMVLM), a unified framework based on the Mixture of Expert Low-Rank Adaption(MoE LoRA) strategy. The framework leverages the gating network to dynamically allocate LoRA expert weights based on the input prompt, enabling synchronized fine-tuning of multiple tasks. To mitigate catastrophic forgetting during instruction-tuning, we introduce a novel *zero expert* that preserves the pre-trained parameters for general linguistic tasks. For pose representation, we implement body-part-specific tokenization by partitioning the human body into different joint groups, enhancing the spatial resolution of the representation. Experiments show that our method effectively alleviates knowledge forgetting during instruction-tuning and achieves remarkable performance across diverse human motion downstream tasks.

## 1 Introduction

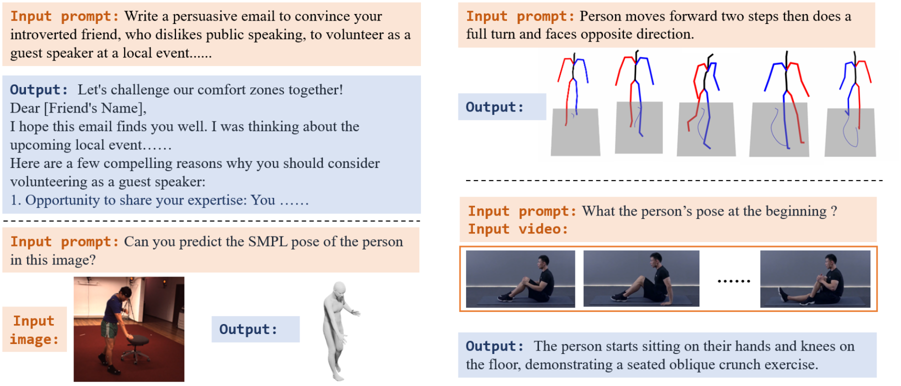
With the capability of encoding semantic information and emotional expression, 3D human motion plays a critical role in virtual reality, embodied intelligence, computer graphics and visions. Recent advances in foundation language models [1, 61, 58, 23] have facilitated multimodal integration. This has stimulated researchers' interest in embedding 3D human motion into these models to address diverse motion-centric tasks, including text-to-motion synthesis[32, 76, 67], motion video understanding[5] and pose estimation [16]. Notably, M3GPT [44] develops a unified vocabulary that integrates text, motion, and music modalities, supporting both text-driven and music-driven motion generation applications.

Although prior work has made progress in motion-centric multimodal modeling, two key issues remain underexplored. First, the effect of incorporating human motion modalities on the foundation

[*] These authors contributed equally to this work.

[†] Shihong Xia is the corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



Input prompt: Write a persuasive email to convince your introverted friend, who dislikes public speaking, to volunteer as a guest speaker at a local event......

Output: Let's challenge our comfort zones together!
Dear [Friend's Name],
I hope this email finds you well. I was thinking about the upcoming local event……
Here are a few compelling reasons why you should consider volunteering as a guest speaker:
1. Opportunity to share your expertise: You ……

Input prompt: Can you predict the SMPL pose of the person in this image?

Input prompt: Person moves forward two steps then does a full turn and faces opposite direction.

Input prompt: What the person's pose at the beginning ?
Input video:

Output: The person starts sitting on their hands and knees on the floor, demonstrating a seated oblique crunch exercise.

Figure 1: HMVLM preserves the original knowledge and dialogue capabilities of the foundation model while supporting a wide range of human-centric downstream tasks.

model's world knowledge is unclear. Dou et al. [12] observed that supervised fine-tuning improves model's instruction-following capabilities with expanding training data, yet simultaneously induces parameter deviation from pre-trained weights, progressively eroding pre-existing knowledge. Although approaches such as temporal continual learning [57] can mitigate catastrophic forgetting in unimodal motion tasks, the substantial modality gap between human motion and text necessitates a deeper examination of their compatibility within the foundation model. Otherwise, catastrophic forgetting may reduce the model to a task-specific generative system with limited dialogue abilities.

Second, how to formulate discrete motion representation compatible with autoregressive architectures in foundation models remains an open research question. Prior methods typically apply temporal convolution to extract motion features along the temporal axis and utilize VQ-VAE architecture to obtain discrete tokens. However, this tokenization paradigm overlooks the spatial information of the pose, limiting the expressiveness of single-frame representations in tasks like pose estimation. Therefore, developing spatially-aware and semantically-grounded tokenization methods for both motion sequences and static poses thus becomes imperative.

To address the first challenge, we observe that supervised instruction-tuning tends to overly focus on new tokens (e.g., motion-related tokens), causing the model to gradually forget its original world knowledge. Therefore, we introduce the Mixture of Expert LoRA framework (MoE LoRA) for multimodal fine-tuning. This framework aims to build a robust Human Motion-Vision-Language Model (HMVLM) for diverse human-centric downstream tasks (as shown in Fig. 1). The gating network dynamically routes task instructions to multiple LoRA expert pairs (LoRA_A/LoRA_B), enabling task-specific adaptation. To avoid knowledge forgetting, we further propose a non-trainable *zero expert* with zero-initialized parameters. We encourage the gating network to select *zero expert* for motion-unrelated tasks, thus preserving the pretrained weights of the foundation model and preventing catastrophic forgetting.

To solve the second issue, we segment the human body into distinct body parts and employ spatial transformers to encode each of them separately. This part-wise encoding, inspired by patch-based tokenization in image processing [11], enhances the resolution of motion or pose tokens while maintaining computational efficiency.

Experimental results show that the proposed HMVLM, built on the MoE LoRA framework, significantly reduces the model's forgetting rate while achieving strong performance in text-to-motion generation, monocular pose estimation, and motion video understanding. The main contributions of this work are:

1. We propose HMVLM, a unified framework that simultaneously supports multiple motion-relevant tasks, including text-to-motion generation, pose estimation, and motion video understanding.

2. We introduce the MoE LoRA architecture for multimodal and multitask fine-tuning of HMVLM, incorporating the novel concept of a *zero expert* to mitigate catastrophic forgetting and preserve foundational knowledge.

3. We design body-part-based tokenizers for pose and motion, improving representation granularity and boosting downstream task performance.

## 2 Related Work

### 2.1 Human Motion Modeling.

Deep learning methods have been extensively employed in conditional human motion generation and modeling tasks. These include deterministic motion prediction from historical states [17, 80, 18], motion completion [25, 54, 48], and motion control [26, 72, 55, 56, 69]. Additionally, deep learning approaches are widely utilized in human pose estimation tasks from RGB images or videos [3, 41, 37, 63, 39, 19, 13, 34]. With growing demands for diverse 3D human motions, probabilistic generation methods have emerged as a prominent research direction[42, 51, 66, 24, 27]. Moreover, leveraging large-scale and uniformly formatted datasets to pre-train a prior model [51, 18] has been shown to be an effective approach for handling multiple motion-related tasks. Among them, probabilistic text-to-motion (T2M), which involves learning cross-modal mappings between textual descriptions and 3D motions, is most related to our work. Early T2M methods focused on aligning modalities by creating a shared representation space for textual and motion features [47, 59]. MotionCLIP [59] embeds motion features into the CLIP latent space using rendered motion images. These methods faced limited motion diversity due to small datasets. Significant advancements followed the release of large-scale datasets such as KIT Motion-Language [49] and HumanML3D [21]. Recent methods including MDM [60], MotionDiffuse [74], ReMoDiffuse [75], and MLD [6] adopted diffusion models for text-guided motion generation, operating in original or compressed motion spaces. Another group of T2M methods employs VQ-VAE to embed motion into discrete latent embeddings, subsequently generating motion sequences autoregressively through Transformer-based architectures. Representative models include TM2T [22], T2M-GPT [73], AttT2M [79], and MoMask [20]. In addition, several works extend T2M approaches with motion editing capabilities, enabling the generation of motions that satisfy both textual descriptions and user-defined geometric constraints [35, 52, 8, 2]. While these methods have achieved impressive results in specific tasks, they primarily focus on human motion modeling or cross-modal learning rather than building robust multimodal frameworks capable of supporting multiple downstream tasks.

### 2.2 Foundation Models and Multi-modal.

Recent advances in foundation language models like ChatGPT [1], BERT [10], Llama [61], Gemma [58], and DeepSeek [23] have shown strong performance in language tasks, with excellent understanding, text generation, and adaptability. These advancements have laid a solid foundation for multimodal research, where instruction tuning has become a central focus, giving rise to frameworks such as vision-language models [43, 64, 46, 40], audio-vision-language models [71, 68], etc.

In the context of human motion, Jiang et al. [32] introduced MotionGPT, which treats 3D human motion as a "foreign language" and constructs a unified vocabulary through motion tokenization to support tasks such as text-to-motion and motion prediction. Zhang et al. [76] adopted Llama-2 as the base model and applied LoRA-based fine-tuning without modifying the word embeddings and prediction head. MotionChain [33] and MotionAgent [67] further enhance motion generation and understanding via multi-round instructions and GPT4-based coordination, respectively. MotionGPT-2 [65] extends MotionGPT by incorporating hand motion to enable whole-body motion generation. Most recently, M3GPT [44] integrates text, motion, and music modalities into a unified framework, supporting diverse cross-modal generation tasks. Although these studies have made progress in integrating human motion into foundational language models, their effects on the models' pre-trained world knowledge remain unexplored.

### 2.3 MoE LoRA Fine-tuning.

Mixture of Experts (MoE) [30] follows a divide-and-conquer strategy by routing inputs to specialized experts, and has been applied across various domains [14, 53, 72, 42, 80]. In foundation models, architectures like Switch Transformers [15] and DeepSeekMoE [9] leverage sparse routing to expand model capacity without increasing inference cost. Recent work [70, 38, 45, 12] integrates MoE with LoRA, showing it can match full fine-tuning performance while enhancing generalization [12].

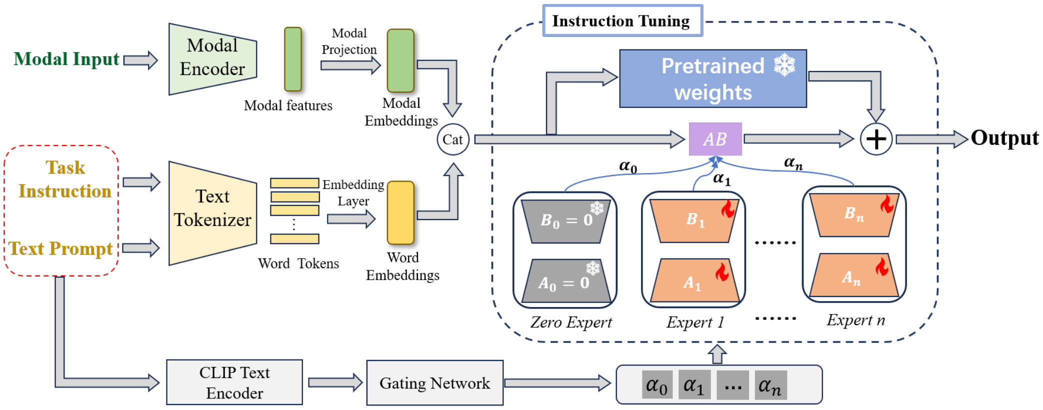


Figure 2: Method overview: task instructions and input prompt are processed by a gating network to produce a mixture weights. Modality-specific inputs are aligned with word embedding via projection layers, and the final outputs are generated through the pre-trained model and the weighted combination of LoRA experts.

Buehler et al. [4] further apply MoE LoRA to bioinformatical tasks such as materials analysis and protein design. Building on these insights, we apply MoE LoRA to HMVLM fine-tuning. Unlike prior approaches that place experts in Transformer feed-forward layers, we introduce multiple LoRA matrix pairs and employ a gating network to route instructions, enabling the model to preserve base model knowledge while adapting to diverse motion-related tasks.

## 3 Method

The overall framework of the proposed Human Motion-Vision-Language Model, based on MoE LoRA, is illustrated in Figure 2. The task instructions and text prompts are encoded using the CLIP text encoder and passed to the gating network $\omega$, which subsequently produces a mixture of expert weights $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_n]$. Simultaneously, modality-specific inputs (e.g., image, video, or motion sequences) are projected into the foundation model's embedding space. The modal embeddings are then combined with the word embeddings and fed into the foundation model. Guided by the semantics of the task instructions and prompts, the LoRA experts are dynamically combined according to the computed weights $\boldsymbol{\alpha}$, enabling task-specific modulation.

### 3.1 LoRA Mixture

We modulate the foundation model's pretrained weights $W$ using multiple LoRA experts:

$$W' = W + \sum_{i=0}^{n} \alpha_i A_i B_i \tag{1}$$

Here, $n$ represents the total number of experts, while $A_i$ and $B_i$ are the corresponding LoRA matrices. We introduce a special *zero expert*, with non-trainable matrices $A_0$ and $B_0$ which are initialized to zero. When the gating network assigns a high weight to $\alpha_0$(i.e., approaching 1), the *zero expert* helps preserve the pre-trained parameters $W$, thereby mitigating catastrophic forgetting. Beyond knowledge preservation, the *zero expert* serves as a shared, general-purpose expert across tasks. Its weight $\alpha_0$ indicates the task's reliance on the foundation model's knowledge, enabling dynamic knowledge fusion and enhancing the synergy between the model and downstream tasks. This design thus provides flexibility and robustness in multimodal, multitask learning.

### 3.2 Multimodal Instruction Format

Given a pre-trained foundation language model $f_\phi(\cdot)$, where $\phi$ denotes the pre-trained parameters, the objective of this work is to construct a HMVLM $f_\psi(\cdot)$ leveraging MoE LoRA and instruction

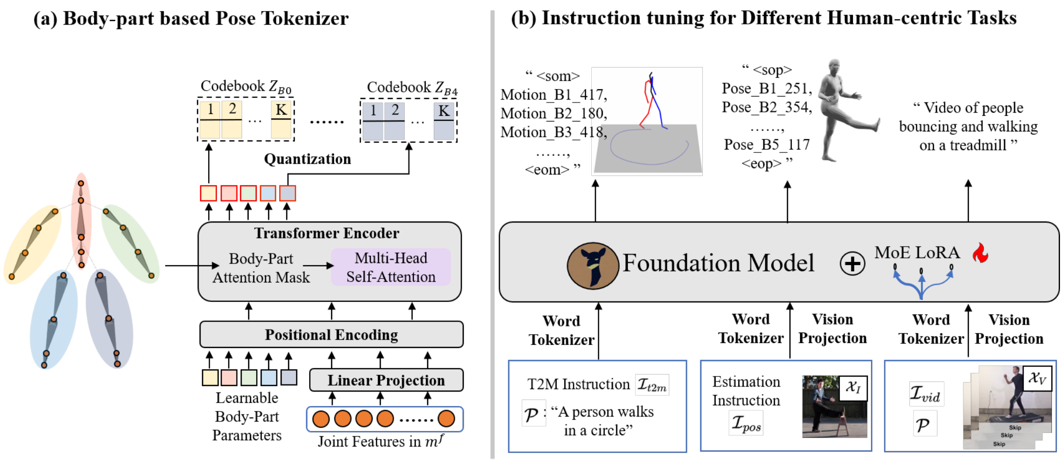

(a) Body-part based Pose Tokenizer
Codebook $Z_{B0}$
1
2
K
Codebook $Z_{B4}$
1
2
K
Quantization
Transformer Encoder
Body-Part Attention Mask
Multi-Head Self-Attention
Positional Encoding
Learnable Body-Part Parameters
Linear Projection
Joint Features in $m^f$
(b) Instruction tuning for Different Human-centric Tasks
" <som> Motion_B1_417, Motion_B2_180, Motion_B3_418, ……, <eom> "
" <sop> Pose_B1_251, Pose_B2_354, ……, Pose_B5_117 <eop> "
" Video of people bouncing and walking on a treadmill "
Foundation Model
MoE LoRA
Word Tokenizer
Word Tokenizer
Vision Projection
Word Tokenizer
Vision Projection
T2M Instruction $\mathcal{I}_{t2m}$
$\mathcal{P}$ : "A person walks in a circle"
Estimation Instruction
$\mathcal{I}_{pos}$
$\mathcal{X}_I$
$\mathcal{I}_{vid}$
$\mathcal{P}$
$\mathcal{X}_V$
Skip
Skip
Skip

Figure 3: (a) Pose/motion tokenization scheme, we introduce learnable body-part parameters into the Transformer to facilitate feature pooling and quantization; (b) instruction tuning for diverse human-centric tasks. The discrete tokens are added to the foundation model's vocabulary, and then instruction tuning guides the model in generating task-related tokens.

tuning. The resulting model $f_\psi$ should not only retain the foundation model's original capabilities and knowledge but also effectively adapt to a diverse set of downstream tasks related to human motion. The general formulation of instruction tuning is as follows:

$$\mathcal{R} = f_\psi(\mathcal{I}, \mathcal{P}, \mathcal{X}) \tag{2}$$

where $\mathcal{I}$ denotes the task instruction, $\mathcal{P}$ represents the input prompt, $\mathcal{X}$ is an optional modality-specific input and $\mathcal{R}$ represents the model responses.

**Text-to-motion generation**: For this task, the model input-output formulation is $\mathcal{R} = f_\psi(\mathcal{I}_{t2m}, \mathcal{P})$, where $\mathcal{I}_{t2m}$ specifies a task-related instruction(e.g., "an AI assistant generates a motion sequence based on user description") and $\mathcal{P}$ contains a concrete user prompt (e.g., "a person walks clockwise in a circle."). The response $\mathcal{R}$ will then be encouraged to contain motion-specific tokens, which can be translated into 3D motion sequences by the motion decoder.

**Pose estimation**: For this task, the formulation is $\mathcal{R} = f_\psi(\mathcal{I}_{pos}, \mathcal{X}_I)$, where $\mathcal{X}_I$ is an image input and $\mathcal{I}_{pos}$ provides the task instructions for pose estimation. The output $\mathcal{R}$ contains pose-relevant tokens, which are then used by the pose decoder to infer the human pose from the input image.

**Motion video understanding**: For this task, the formulation is $\mathcal{R} = f_\psi(\mathcal{I}_{vid}, \mathcal{P}, \mathcal{X}_V)$, where $\mathcal{I}_{vid}$ is the instruction specific to human motion video understanding. $\mathcal{P}$ represents the input prompt and $\mathcal{X}_V$ is the video input.

### 3.3 Multimodal Instruction Tuning

To support pose estimation and T2M tasks, we will pre-train a pose and motion tokenizer (detailed in Sec. 3.4) to discretize the encoding of poses and motions (See Fig. 3 (b)), obtaining the pose vocabulary $V_m$ and the motion vocabulary $V_M$. These vocabularies are merged with the original text vocabulary $V_T$ to form an extended vocabulary $V = [V_T, V_M, V_m]$, while preserving the original text token order.

As illustrated in Fig. 3 (b), MoE LoRA enables joint fine-tuning across multiple human-centric tasks. Given instruction data of the form $(\mathcal{I}, \mathcal{P}, \mathcal{X}, \mathcal{R}_{gt})$, all tasks share the objective of next-token prediction:

$$\mathcal{L}_{fm} = -\mathbb{E}_{R^t_{gt} \in V}[\log p(\mathcal{R}^t_{gt} | \mathcal{I}, \mathcal{P}, \mathcal{X}, \mathcal{R}^{<t}_{gt})] \tag{3}$$

where $\mathcal{R}^t_{gt}$ denotes the ground-truth token at position $t$ in the response sequence and $\mathcal{X}$ is an optional modality-specific input, as described in Sec. 3.2. For image input $\mathcal{X}_I$, we use the pre-trained CLIP ViT-L/14 [50] with the Llava projection layer [43] to align visual features with the model's embedding space. For video input $\mathcal{X}_v$, 8 frames are uniformly sampled and processed similarly.

To preserve the foundation model's world knowledge in motion-unrelated tasks, we supervise the gating network $\omega$ using user prompts from conversation datasets. Specifically, the instruction $\mathcal{I}$ and user prompt $\mathcal{P}$ are first encoded by the CLIP text encoder, and then input into the gating network $\omega$ to obtain the expert weights $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_n]$. To encourage selection of the *zero expert* ($\alpha_0$) in motion-unrelated tasks, we design the loss:

$$\mathcal{L}_{gat} = -\mathbb{E}[\eta * \log p_w(\alpha_0 | \mathcal{I}, \mathcal{P})] \tag{4}$$

Here, $\eta$ is an indicator function such that $\eta = 1$ if the input $(\mathcal{I}, \mathcal{P})$ is unrelated to human motion, and $\eta = 0$ otherwise. This loss helps retain the foundation model's linguistic capabilities and avoid catastrophic forgetting. For human motion-relevant tasks, the gating network dynamically combines experts to enhance the performance of downstream tasks. The final instruction tuning loss is given by $\mathcal{L}_{total} = \mathcal{L}_{fm} + \mathcal{L}_{gat}$.

### 3.4 Pose and Motion Tokenizer

To obtain the vocabularies $V_m$ and $V_M$, and integrate the pose and motion modalities into the foundation language model, prior works [73, 79, 32, 76] typically use VQ-VAE to discretize motion sequences. Specifically, given a motion sequence $M^{1:F} = [m^1, m^2, ..., m^F] \in \mathbb{R}^{F \times D}$, a motion tokenizer $\mathcal{E}$ applies 1D convolutions along the temporal axis to produce latent features $\hat{z}^{1:(F/l)} = \mathcal{E}(M^{1:F})$ ($l$ denotes the temporal compression ratio), followed by quantization:

$$z_i = \mathcal{Q}(\hat{z}_i) := \underset{z_k \in Z}{\arg\min} \|\hat{z}_i - z_k\|^2 \tag{5}$$

Here, $Z = \{z_i\}_{i=1}^{K} \subset \mathbb{R}^S$ is the learned codebook containing $K$ discrete latent vectors, each of dimension $S$. The full tokenization process is thus expressed as $z^{1:(F/l)} = \mathcal{Q}(\mathcal{E}(M^{1:F}))$. However, this classical tokenization focuses solely on temporal encoding by combining discrete codes across time, which limits its capacity to capture spatial granularity. In tasks like pose estimation, where only single-frame input is involved (i.e. $F = 1$ ), the accuracy of discrete encoding relies entirely on the fixed codebook. As a result, its ability to represent pose variations is severely constrained by the codebook size $K$, leading to coarse-grained representations.

To address this limitation, we draw inspiration from patch-based image encoding for spatial modeling [11]. We exploit the body's natural decomposition into limb-based parts. Part-aware modeling has proven effective in motion retargeting [28], motion style transfer [31], and text-to-motion [79], yet most prior work does not compute discrete codes independently for each part.

In this work, we adopt the spatial Transformer architecture from [28] and build body part-based pose and motion tokenizers, with the architecture illustrated in Fig. 3 (a). For a given pose $m^f$, we process each joint feature using a linear projection and go through the positional encoding with the learnable body-part parameters. During the self-attention computation, an attention mask matrix is constructed based on the correspondence between joints and body parts, ensuring that the body part parameters are only associated with joints within that part. Finally, we keep only the outputs corresponding to the body part parameters for pooling the pose embeddings. The spatial modeling process can be formulated as:

$$[\hat{z}^f_{B1}, \hat{z}^f_{B2}, ..., \hat{z}^f_{BN}] = \mathcal{E}_s(m^f) \tag{6}$$

where $N$ denotes the number of body parts. For single-frame pose input, a separate codebook is constructed for each body part, and the embeddings are quantized independently as $z^f_{Bn} = \mathcal{Q}_n(\hat{z}^f_{Bn})$. For motion sequences, the spatial embeddings from $\mathcal{E}_s$ are further compressed along the temporal axis using a temporal convolution module $\mathcal{E}_t$, yielding:

$$(\hat{z}'^1_{Bn}, \hat{z}'^2_{Bn}, ..., \hat{z}'^{F/l}_{Bn}) = \mathcal{E}_t(\hat{z}^1_{Bn}, \hat{z}^2_{Bn}, ..., \hat{z}^F_{Bn}) \tag{7}$$

The tokenization processes for pose and motion are expressed as $\mathcal{Q}_{1:N}(\mathcal{E}_s(m^f))$ and $\mathcal{Q}_{1:N}(\mathcal{E}_t(\mathcal{E}_s(M^{1:F})))$, respectively. Following tokenization, separate decoders $\mathcal{D}_m$ and $\mathcal{D}_M$ are employed to reconstruct the original input as detokenizers. We adopt the training strategy of T2M-GPT [73], with the loss function $\mathcal{L}_M = \mathcal{L}_{rec} + \mathcal{L}_{emb} + \lambda_{com}\mathcal{L}_{com}$. Specifically, $\mathcal{L}_{rec}$ is the reconstruction loss and $\mathcal{L}_{emb}$ is used to update the codebooks, while $\mathcal{L}_{com}$ encourages the body part embeddings to remain close to their assigned codebook vectors.

# 4 Experiments

**Implementation Details.** We use Vicuna-7b-v1.5 [7] as the foundation language model with five LoRA experts (including a *zero expert*), each of rank 8. LoRA adapters are applied to all linear modules, and the gating network is implemented as a two-layer MLP with a hidden dimension of 512. It takes the 512-dimensional text features output by the CLIP model and predicts the weights for five experts. For detailed implementation, please refer to the Appendix.

**Datasets.** We train the gating network with the LMSYS-Chat-1M dataset [77], using 80% of the data for training. For the text-to-motion task, we use HumanML3D [21] and KIT-ML [49] datasets. Notably, the motion tokenizer is trained on the same training splits of HumanML3D and KIT-ML for consistency. For pose estimation and pose tokenizer training, we use the Human3.6M [29] and 3DPW [62] datasets. The MoVid dataset [5] is used for instruction tuning in motion video understanding.

## 4.1 Evaluation on the knowledge preservation

We assess how human motion modalities affect the model's knowledge retention. Specifically, we measure model forgetting by comparing text comprehension performance before and after text-to-motion instruction tuning, using the MT-Bench [78]. MT-Bench evaluates 80 questions across eight topics, each with two-turn dialogues. GPT-4 serves as the judge, scoring responses from 1 (completely incorrect) to 10 (fully correct). We use this to measure performance changes after T2M fine-tuning.

Table 1: Evaluation on dialogue abilities of foundation models before and after text-to-motion tuning

| Methods | FM | Write | Role | Extract | Reason | Math | Code | Stem | Code | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| MotionGPT [65] | Llama2 | 2.70 | 3.63 | 2.25 | 4.00 | 1.50 | 1.15 | 3.50 | 3.10 | 2.73 |
| | Tuned | 1.85 | 2.75 | 1.55 | 2.50 | 1.45 | 1.20 | 3.00 | 2.58 | 2.11 |
| MotionAgent [67] | Gemma2 | 8.83 | 8.65 | 7.90 | 7.25 | 5.80 | 5.30 | 8.93 | 9.70 | 7.79 |
| | Tuned | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Ours | Gemma2 | 8.83 | 8.65 | 7.90 | 7.25 | 5.80 | 5.30 | 8.93 | 9.70 | **7.79** |
| | Tuned | 8.45 | 8.55 | 7.85 | 6.75 | 5.75 | 5.20 | 8.20 | 9.45 | **7.53** |
| Ours | Vicuna | 7.43 | 7.52 | 5.21 | 4.90 | 3.69 | 2.68 | 6.98 | 9.0 | **5.90** |
| | Tuned | 7.75 | 6.20 | 5.80 | 4.50 | 3.00 | 2.45 | 6.45 | 8.15 | **5.54** |
| Ours | Vicuna | 7.43 | 7.52 | 5.21 | 4.90 | 3.69 | 2.68 | 6.98 | 9.0 | 5.90 |
| w/o $\mathcal{L}_{gat}$ | Tuned | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

We conduct a quantitative comparison between our method and two representative baselines, MotionGPT [32] and MotionAgent [67], with results summarized in Table 1. Since these methods are built on different foundation models, absolute MT-Bench scores are not directly comparable. Therefore, we primarily focus on the relative degradation in dialogue performance of each foundation model after instruction-tuning for the text-to-motion task.

As shown, MotionGPT, which applies LoRA solely to the *Query* and *Value* matrices without introducing new motion tokens, preserves part of the original linguistic capability but still exhibits a noticeable 22.71% performance drop (2.73 → 2.11). In contrast, MotionAgent suffers from a drastic 87.16% performance degradation (7.79 → 1.00), with MT-Bench scores across all topics collapsing to 1 (rated as "completely unreasonable" by GPT-4). This severe collapse is primarily attributed to overfitting on the newly introduced motion-specific tokens (e.g., Motion_index), resulting in catastrophic forgetting of linguistic knowledge.

In comparison, our approach achieves effective task decoupling and knowledge preservation through the proposed MoE LoRA framework. Using the same foundation model, Gemma-2-2b-it, as MotionAgent, our method demonstrates only a marginal 3.34% degradation (7.79 → 7.53). This clearly highlights the superior capacity of our framework to retain foundational knowledge while adapting to new modalities and tasks. Moreover, our MoE LoRA framework is model-agnostic and can be seamlessly integrated into the instruction tuning process of various foundation models. When applied to Vicuna-7b-v1.5, the primary foundation model used in our study, fine-tuning with MoE LoRA results in only a 6.10% performance drop (5.90 → 5.54), further demonstrating the broad applicability and robustness of our framework. Additional qualitative results are provided in Appendix A.1.

## 4.2 Evalution on Text-to-Motion Task

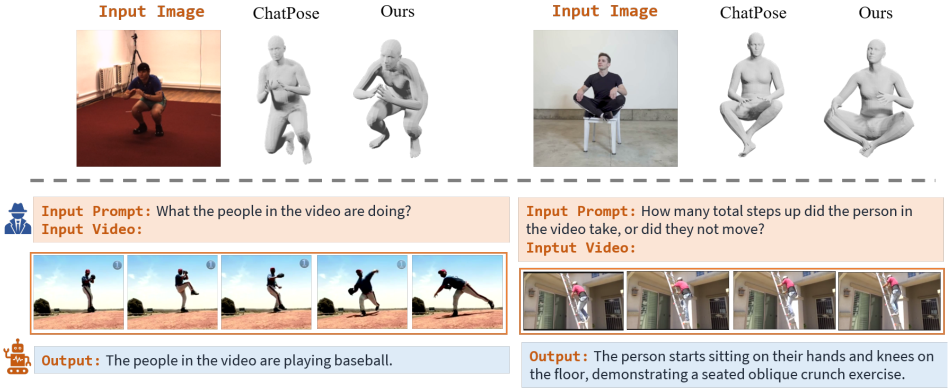
For text-to-motion task, we compare our HMVLM with state-of-the-art methods on HumanML3D dataset. Following prior work [21, 73], we use four evaluation metrics: *R precision*(*Top1-Top3*) and *Multi-modal Distance*(*MM-D*) for text-to-motion retrieval accuracy, *Frechet Inception Distance* (FID) for motion realism, and *Diversity* (Div.) for motion variation.

Table 2: Quantitative results of text-to-motion on the HumanML3D dataset

| Methods | R precision↑ | | | FID.↓ | MM-D.↓ | Div.→ |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| GT | 0.511 ±.003 | 0.703 ±.003 | 0.797 ±.002 | 0.002 ±.000 | 2.974 ±.008 | 9.503 ±.065 |
| TM2T [22] | 0.424 ±.003 | 0.618 ±.003 | 0.729 ±.002 | 1.501 ±.017 | 3.467 ±.011 | 8.589 ±.076 |
| T2M [21] | 0.455 ±.003 | 0.636 ±.003 | 0.736 ±.002 | 1.087 ±.021 | 3.347 ±.008 | 9.175 ±.083 |
| MDM [60] | 0.320 ±.005 | 0.498 ±.004 | 0.611 ±.007 | 0.544 ±.044 | 5.566 ±.027 | 9.559 ±.086 |
| MD [74] | 0.491 ±.001 | 0.681 ±.001 | 0.782 ±.001 | 0.630 ±.001 | 3.113 ±.001 | 9.410 ±.049 |
| MLD [6] | 0.481 ±.003 | 0.673 ±.003 | 0.772 ±.002 | 0.473 ±.013 | 3.196 ±.010 | 9.724 ±.082 |
| T2M-GPT [73] | 0.491 ±.003 | 0.680 ±.003 | 0.775 ±.002 | 0.116 ±.004 | 3.118 ±.011 | **9.761** ±.081 |
| ReMoDiffuse [75] | 0.510 ±.005 | 0.698 ±.006 | 0.795 ±.004 | 0.103 ±.004 | 2.974 ±.016 | 9.018 ±.075 |
| AttT2M [79] | 0.499 ±.003 | 0.690 ±.002 | 0.786 ±.002 | 0.112 ±.006 | 3.038 ±.007 | 9.700 ±.090 |
| MoMask [20] | **0.521** ±.002 | **0.713** ±.002 | **0.807** ±.002 | **0.045** ±.002 | **2.958** ±.008 | 9.620 ±.064 |
| MotionGPT [76] | 0.364 ±.005 | 0.533 ±.003 | 0.629 ±.004 | 0.805 ±.002 | 3.914 ±.013 | 9.972 ±.026 |
| MotionGPT [32] | 0.492 ±.003 | 0.681 ±.003 | 0.733 ±.006 | 0.232 ±.008 | 3.096 ±.008 | 9.528 ±.071 |
| MotionAgent [67] | 0.482 ±.004 | 0.672 ±.003 | 0.770 ±.002 | 0.491 ±.019 | 3.138 ±.010 | 9.838 ±.244 |
| MotionGPT-2 [65] | 0.496 ±.002 | 0.691 ±.003 | 0.782 ±.004 | 0.191 ±.004 | 3.080 ±.013 | **9.860** ±.026 |
| Ours (single task) | **0.502** ± .003 | **0.692** ±.004 | **0.785** ±.002 | **0.123** ±.004 | **3.039** ±.027 | 9.443 ±.132 |
| Ours | 0.463 ± .006 | 0.646 ±.004 | 0.744 ±.001 | 0.156 ±.010 | 3.328 ±.004 | 9.544 ±.161 |

Tab. 2 presents quantitative comparisons. Methods in the upper section are task-specific T2M models trained from scratch, whereas methods in the lower section are multimodal frameworks based on foundation language models, fine-tuned via instruction tuning. Among these, MoMask [20] achieves state-of-the-art results in most metrics due to its cascaded mask Transformer design; however, such specialized approaches often lack scalability and generalizability for broader multimodal tasks.

Among multimodal foundation models, we evaluate our proposed HMVLM in two settings (lower section of Tab. 2). *Ours (single task)* denotes evaluation with only T2M task fine-tuning, using the same MoE LoRA architecture. Our model demonstrates remarkable performance across most metrics, thanks to dynamic expert assignment via MoE and fine-grained body-part tokenization (detailed in Sec. 4.5). Under multi-task fine-tuning (*Ours*), HMVLM remains competitive across all metrics, although its performance decreases compared to the single-task setting. This is because, in a single-task setting, all LoRA experts focus solely on one downstream task within the same parameter budget.



Input Prompt: What the people in the video are doing?
Input Video:

Output: The people in the video are playing baseball.

Input Prompt: How many total steps up did the person in the video take, or did they not move?
Inptut Video:

Output: The person starts sitting on their hands and knees on the floor, demonstrating a seated oblique crunch exercise.

Figure 4: Qualitative results for human pose estimation and human video understanding.

### 4.3 Evaluation on Human Vision Tasks

**Human Pose Estimation.** We follow the evaluation setup of ChatPose [16] using the MPJPE (Mean Per-Joint Position Error) and PA-MPJPE (Procrustes-Aligned MPJPE) as metrics. As shown in Tab. 4, our method surpasses ChatPose—a comparable foundation-model-based approach, in the single-task fine-tuning scenario. This result highlights the advantage of the MoE LoRA architecture in enabling fine-grained expert assignment. Qualitative comparisons with ChatPose, presented in Fig. 4, demonstrate our method's superior accuracy in capturing limb details using examples from the Human3.6M (left) and MoViD (right) datasets, validating the effectiveness of our body-part-based tokenization strategy.

**Motion Video Understanding.** Fig. 4 illustrates an example of our model's performance in human motion video comprehension and reasoning tasks. Our method successfully identifies motion categories (e.g., baseball motion at the top-left) and exhibits spatio-temporal reasoning capabilities. For instance, the bottom-left corner shows the model accurately determining the number of steps climbed, while the right side provides concise descriptions of the character's movements and postures.

### 4.4 Expert Weight Distribution

We analyze the average expert weights from the gating network across different tasks to evaluate the MoE LoRA architecture's task-decoupling capability. Specifically, we extract textual features by inputting each task's instructions and prompts into CLIP's text encoder, then compute the expert weights using the gating network. As shown in Table 3, for the general dialogue task (GD), the loss in Equation 4 encourages the gating network to prioritize the *zero expert*, thereby preserving the pretrained parameters of the foundation model. For other tasks, the gating network dynamically adjusts the expert weight based on the instruction and prompt, reflecting the idea of divide and conquer.

Table 3: Average gating weights across different tasks. GD, T2M, HPE, and HVU denote general dialogue, text-to-motion, human pose estimation, and human video understanding tasks, respectively.

| **Task** | **Zero Expert** | **Expert 1** | **Expert 2** | **Expert 3** | **Expert 4** |
|---|---|---|---|---|---|
| GD | 0.999 | $2.364\times10^{-6}$ | $5.005\times10^{-6}$ | $1.357\times10^{-6}$ | $1.827\times10^{-6}$ |
| T2M | 0.694 | 0.052 | 0.067 | 0.085 | 0.102 |
| HPE | 0.454 | 0.292 | 0.084 | 0.106 | 0.063 |
| HVU | 0.167 | 0.252 | 0.150 | 0.013 | 0.418 |

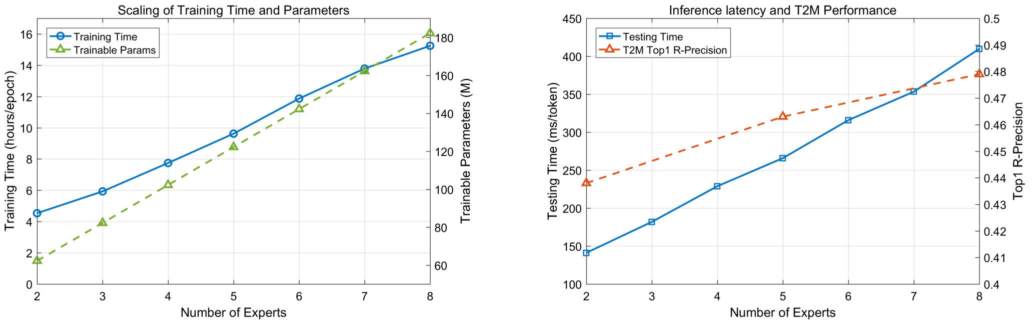
Table 4: Quantitative results of pose estimation on the H3.6M and 3DPW datasets

| Methods | H3.6M | | 3DPW | |
|---|---|---|---|---|
| | MPJPE↓ | PA-MPJPE↓ | MPJPE↓ | PA-MPJPE↓ |
| SPIN [36] | 61.9 | 42.6 | 102.9 | 62.9 |
| HMR 2.0 [19] | **50.0** | **33.6** | **91.0** | **58.4** |
| ChatPose [16] | 126.0 | 82.4 | 163.6 | 81.9 |
| Ours(single task) | **92.8** | **55.3** | **105.3** | **56.24** |
| Ours | 114.7 | 64.8 | 127.7 | 63.3 |

### 4.5 Ablation Study

HMVLM incorporates the MoE LoRA architecture and a body-part-based tokenization strategy. Accordingly, we conduct ablation studies focusing on these two key components and model efficiency.

**Effectiveness of Body Part-based Tokenization.** To evaluate the spatial modeling capability of our proposed body-part-based tokenization, we conduct ablation studies on both motion reconstruction and T2M performance. As shown in Tab. 5, the baseline tokenizer ("W/o BP") corresponds to the standard whole-body motion tokenizer described in Sec. 3.4. Although enlarging the codebook size



(a) (b)

Figure 5: Efficiency analysis of the MoE LoRA model under different numbers of experts. (a) Training time and parameter scaling. (b) Inference latency and T2M performance.

(i.e., $K$ × number of body parts) increases the model capacity, using a single codebook for the entire body fails to enhance spatial expressiveness. In contrast, our body-part-based tokenizer yields clear improvements in R-precision and reconstruction MSE, demonstrating its superior spatial modeling ability. The slightly higher FID may result from combining multiple part-specific codebooks, which increases pose diversity but introduces a minor distribution shift.

Table 5: Abalation study on different tokenizers. $K$ represent the codebook size.

| Methods | R precision(Top-3)↑ | FID.↓ | MM-D.↓ | Div.→ | MSE. ↓ |
|---|---|---|---|---|---|
| Ours (single task) | 0.785 ±.002 | 0.123 ±.004 | 3.039 ±.027 | 9.443 ±.132 | 0.966 |
| W/o BP (K=512) | 0.741 ±.001 | 0.336 ±.012 | 3.291±.003 | 9.000 ±.263 | 1.377 |
| W/o BP (K=512*5) | 0.758 ±.003 | 0.110 ±.008 | 3.232 ±.016 | 9.508 ±.212 | 1.34 |

**Effectiveness of $\mathcal{L}_{gat}$.** We investigate the impact of the $\mathcal{L}_{gat}$ on preserving the foundation model's world knowledge. As shown in Tab. 1, removing the $\mathcal{L}_{gat}$ results in catastrophic forgetting, attributed to modifications in pre-trained parameters and changes in the prediction head that make the model overly focussed on newly introduced motion tokens. These findings support the effectiveness of combining the the MoE LoRA architecture with $\mathcal{L}_{gat}$ for building robust multimodal frameworks while mitigating catastrophic forgetting.

**Efficiency of MoE LoRA.** We evaluate the efficiency of the MoE LoRA architecture under different numbers of experts. As shown in Fig. 5(a), the number of trainable parameters and training time increase nearly linearly with the number of experts, as each LoRA expert shares the same rank. For inference latency, as shown in (b), the LoRA weights must be dynamically combined according to the gating network's output, preventing pre-merging with the pretrained model and causing a moderate rise in latency. We also observe the T2M Top-1 R-precision improves but gradually saturates as experts increase. Considering the trade-off between efficiency and performance, we adopt five experts.

## 5 Discussion

In this paper, we present HMVLM, a MoE LoRA-based multimodal framework designed for a range of human-centric tasks, including motion perception, comprehension, and generation. By leveraging the MoE architecture and the introduction of a *zero expert*, our approach preserves the foundation model's world knowledge and generative capabilities during instruction tuning. Furthermore, we incorporate a spatial Transformer to independently encode body part features into discrete tokens, enabling precise and fine-grained motion and pose representations.

**Limitations and Future Work.** While HMVLM advances the joint modeling of human motion, language and vision, several limitations remain. The modality connections are still learned in a independent pairwise manner, limiting holistic integration. Additionally, domain discrepancies across datasets hinder the model's ability to perform seamless any-to-any generation.

## Acknowledgment

This work was supported by National Key R&D Program of China (NO. 2022YFB3303202) and the Innovation Funding of ICT, CAS under Grant NO. E561010

## 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.

[2] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, October 2016.

[4] Eric L Buehler and Markus J Buehler. X-lora: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2(2), 2024.

[5] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024.

[6] 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.

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[8] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024.

[9] Damai Dai, Chengqi Deng, Chenggang Zhao, Rx Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297, 2024.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[12] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, 2024.

[13] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, March 2024.

[14] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

[15] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[16] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2093–2103, 2024.

[17] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015.

[18] Yang Gao, Po-Chien Luan, and Alexandre Alahi. Multi-transmotion: Pre-trained model for human motion prediction. *arXiv preprint arXiv:2411.02673*, 2024.

[19] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

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

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

[22] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 580–597. Springer, 2022.

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[24] Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. Amd: Autoregressive motion diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2022–2030, 2024.

[25] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[26] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[27] Lei Hu, Zihao Zhang, Yongjing Ye, Yiwen Xu, and Shihong Xia. Diffusion-based human motion style transfer with semantic guidance. In *Computer Graphics Forum*, volume 43, page e15169. Wiley Online Library, 2024.

[28] Lei Hu, Zihao Zhang, Chongyang Zhong, Boyuan Jiang, and Shihong Xia. Pose-aware attention network for flexible motion retargeting by body part. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[29] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[30] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[31] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022.

[32] 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.

[33] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang Yu, and Jiayuan Fan. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, pages 54–74. Springer, 2024.

[34] Boyuan Jiang, Lei Hu, and Shihong Xia. Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14850–14860, 2023.

[35] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023.

[36] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019.

[37] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019.

[38] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *CoRR*, 2024.

[39] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.

[40] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[41] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021.

[42] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020.

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[44] Mingshuang Luo, Ruibing Hou, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. Mgpt: An advanced multimodal, multitask framework for motion comprehension and generation. *CoRR*, 2024.

[45] Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*, 2024.

[46] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

[47] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022.

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

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

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

[51] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021.

[52] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[54] Paul Starke, Sebastian Starke, Taku Komura, and Frank Steinicke. Motion in-betweening with phase manifolds. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–17, 2023.

[55] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019.

[56] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39(4):54–1, 2020.

[57] Jianwei Tang, Jiangxin Sun, Xiaotong Lin, Wei-Shi Zheng, Jian-Fang Hu, et al. Temporal continual learning with prior compensation for human motion prediction. *Advances in Neural Information Processing Systems*, 36:65837–65849, 2023.

[58] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[59] 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.

[60] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[62] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018.

[63] Luyang Wang, Yan Chen, Zhenhua Guo, Keyuan Qian, Mude Lin, Hongsheng Li, and Jimmy S Ren. Generalizing monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[64] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.

[65] 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.

[66] Zhiyong Wang, Jinxiang Chai, and Shihong Xia. Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE transactions on visualization and computer graphics*, 27(1):14–28, 2019.

[67] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. *arXiv preprint arXiv:2405.17013*, 2024.

[68] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[69] Yongjing Ye, Libin Liu, Lei Hu, and Shihong Xia. Neural3points: Learning to generate physically realistic full-body motion for virtual reality users. In *Computer Graphics Forum*, volume 41, pages 183–194. Wiley Online Library, 2022.

[70] Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *CoRR*, 2023.

[71] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

[72] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

[73] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[74] 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.

[75] 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.

[76] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7368–7376, 2024.

[77] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024.

[78] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

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

[80] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6447–6456, 2022.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the proposed framework (HMVLM, MoE LoRA) and experimental scope (MT-benchmark, R-precision, MPJPE), which align with the content presented in the paper (Sections 3, 5, 6, and Supplements).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Section 5 discuss limitations, including challenges in modality connections and any-to-any generation.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please refer to Section 3.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the experimental datasets Section 4, and evaluation metrics in each task. Appendix D provides hyperparameter and data processing details.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code will be released.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental settings are detailed in Section 4 and the Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 1, 2, 3, 4 and appendix tables report results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the hardware used (single A800 80G GPU) in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves algorithmic development and evaluation on standard benchmarks. It does not involve obviously ethically sensitive applications.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on the technical contributions and does not include a specific discussion of broader positive or negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The paper proposes a new human motion-relevant multimodal framwork. Neither the model nor the standard benchmark data used appear to pose a high risk for misuse necessitating specific release safeguards beyond standard open-source practices.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper properly cites the sources for existing assets like baseline methods and data sources.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The primary new asset is the Appendix and supplementary video for the proposed methods.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The research does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The research does not involve human subjects, therefore IRB approval is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# Appendix

This appendix provide qualitative results(Sec.A), additional experiments(Sec.B),visulization(Sec.C) and implementation details(Sec.D).

**Video.** We also provide the supplementary video to showcase our comparisons with SOTA methods and application examples of our approach, including text-to-motion pose estimation, and human motion video understanding.

## A Qualitative Results

### A.1 Qualitative Comparison of Knowledge Retention

Fig. 6 presents a qualitative comparison of the knowledge retention in foundation models across different methods, including MotionAgent [67], MotionGPT [76], and our method. The input prompt is sampled from the MT-Bench writing topic.

The left side of Fig 6 shows the foundation model outputs before instruction tuning on the text-to-motion task. All of the models generate well-structured, relevant emails with reasonable suggestions, fully satisfying the prompt requirements. The right side of the figure shows the results after instruction tuning. MotionAgent nearly completely loses the dialog capability of its foundation model, with outputs overwhelmed by repetitive "<Motion_index>" tokens. This is caused by the introduction of a new motion vocabulary, modifications to the prediction head, and the application of LoRA matrices across all Transformer modules (Query, Key, Value, and Projection), resulting in severe overfitting to the motion-related task and catastrophic forgetting of the original abilities. MotionGPT, which preserves the original prediction head and applies LoRA only to the Query and Value matrices, retains a portion of the foundation model' knowledge. However, its responses are noticeably more terse, suggesting a certain degree of knowledge degradation. Moreover, Tab. 2 in the main text shows that this fine-tuning strategy cannot achieve superior downstream task performance.

In contrast, our method adopts the MoE LoRA architecture, which enables expert routing through a gating network even when LoRA is applied across all linear modules. The *zero expert* mechanism allows the model to fallback to the original parameter path when handling normal dialogue tasks, effectively preserving the foundational knowledge and achieving a good balance between avoiding model forgetting and enhancing downstream task performance.

### A.2 Qualitative Comparison of Text-to-Motion

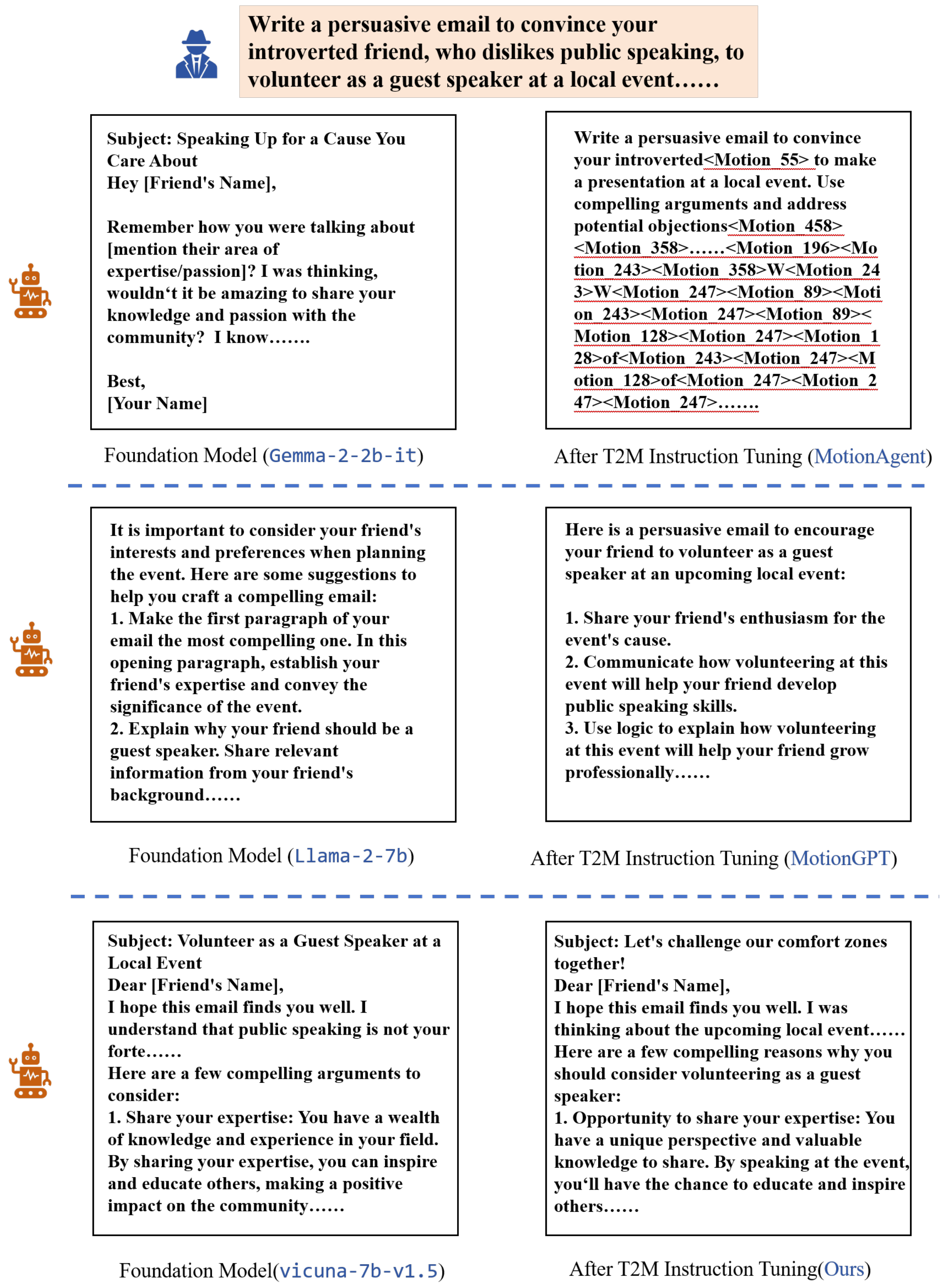
We compare our method with state-of-the-art foundation model-based T2M approaches on the HumanML3D test set. We use their publicly released pretrained models and LoRA parameters. The results are shown in Fig 7, poses or motions with semantic errors are highlighted with red boxes. In the left example, the prompt requires a squat followed by running in place, but neither MotionAgent nor MotionGPT generates the squatting motion. More qualitative results can be found in supplementary video.

### A.3 Qualitative Comparison of Pose Estimation

We conducted qualitative comparisons on the 3DPW dataset, following its official training and test split. we compare our method with ChatPose [16], as shown in Fig. 8. The results demonstrate that our approach outperforms ChatPose in body pose accuracy and semantic consistency.

### A.4 Human Motion Video Understanding

**Human motion RGB video understanding.** We provide some examples of human motion video understanding. The results are shown in Fig. 9, all the video are extracted from MoVid dataset [5]. For more qualitative results, please refer to our supplementary video.



Write a persuasive email to convince your introverted friend, who dislikes public speaking, to volunteer as a guest speaker at a local event……

Subject: Speaking Up for a Cause You Care About
Hey [Friend's Name],

Remember how you were talking about [mention their area of expertise/passion]? I was thinking, wouldn't it be amazing to share your knowledge and passion with the community? I know…….

Best,
[Your Name]

Foundation Model (`Gemma-2-2b-it`)

Write a persuasive email to convince your introverted<Motion_55> to make a presentation at a local event. Use compelling arguments and address potential objections<Motion_458><Motion_358>……<Motion_196><Motion_243><Motion_358>W<Motion_243>W<Motion_247><Motion_89><Motion_243><Motion_247><Motion_89><Motion_128><Motion_247><Motion_128>of<Motion_243><Motion_247><Motion_128>of<Motion_247><Motion_247><Motion_247>…….

After T2M Instruction Tuning (MotionAgent)

It is important to consider your friend's interests and preferences when planning the event. Here are some suggestions to help you craft a compelling email:
1. Make the first paragraph of your email the most compelling one. In this opening paragraph, establish your friend's expertise and convey the significance of the event.
2. Explain why your friend should be a guest speaker. Share relevant information from your friend's background……

Foundation Model (`Llama-2-7b`)

Here is a persuasive email to encourage your friend to volunteer as a guest speaker at an upcoming local event:

1. Share your friend's enthusiasm for the event's cause.
2. Communicate how volunteering at this event will help your friend develop public speaking skills.
3. Use logic to explain how volunteering at this event will help your friend grow professionally……

After T2M Instruction Tuning (MotionGPT)

Subject: Volunteer as a Guest Speaker at a Local Event
Dear [Friend's Name],
I hope this email finds you well. I understand that public speaking is not your forte……
Here are a few compelling arguments to consider:
1. Share your expertise: You have a wealth of knowledge and experience in your field. By sharing your expertise, you can inspire and educate others, making a positive impact on the community……

Foundation Model(`vicuna-7b-v1.5`)

Subject: Let's challenge our comfort zones together!
Dear [Friend's Name],
I hope this email finds you well. I was thinking about the upcoming local event……
Here are a few compelling reasons why you should consider volunteering as a guest speaker:
1. Opportunity to share your expertise: You have a unique perspective and valuable knowledge to share. By speaking at the event, you'll have the chance to educate and inspire others……

After T2M Instruction Tuning(Ours)

Figure 6: Qualitative results of question-answering dialogues from foundation language models with different methods, before and after instruction tuning on text-to-motion tasks.

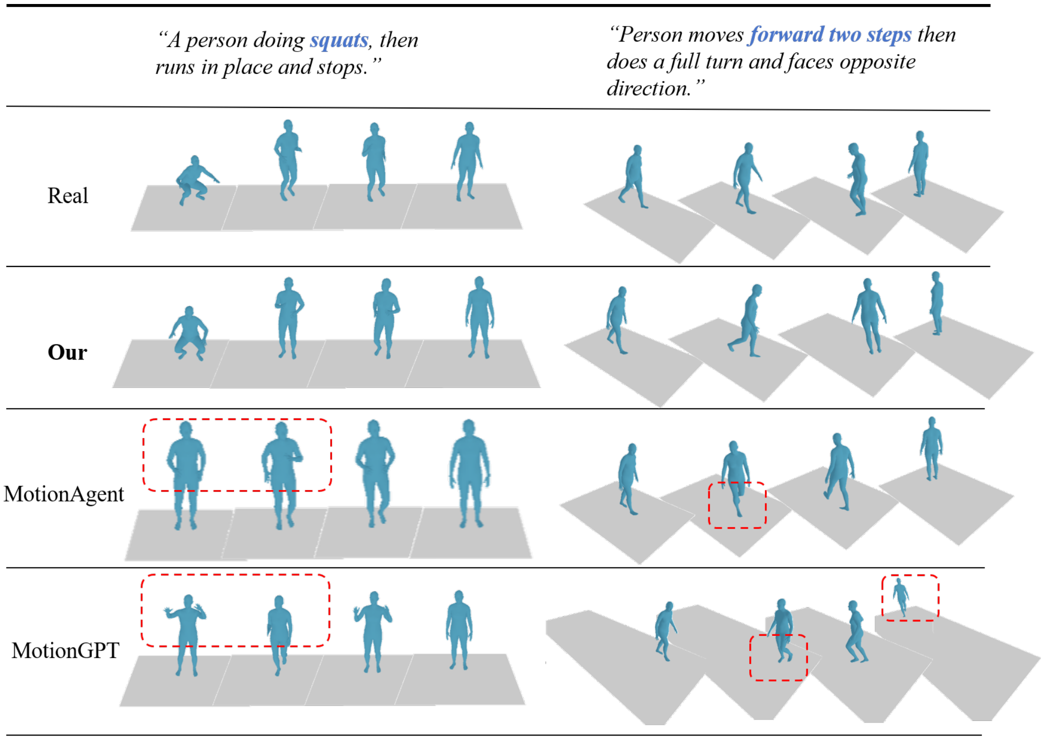

*"A person doing **squats**, then runs in place and stops."*
*"Person moves **forward two steps** then does a full turn and faces opposite direction."*
Real
**Our**
MotionAgent
MotionGPT

Figure 7: Qualitative comparison on text-to-motion task. The provided state-of-the-art methods are under the same training and inference setting on HumanML3D [21]. The red box highlights the poses that do not match the prompt semantics.

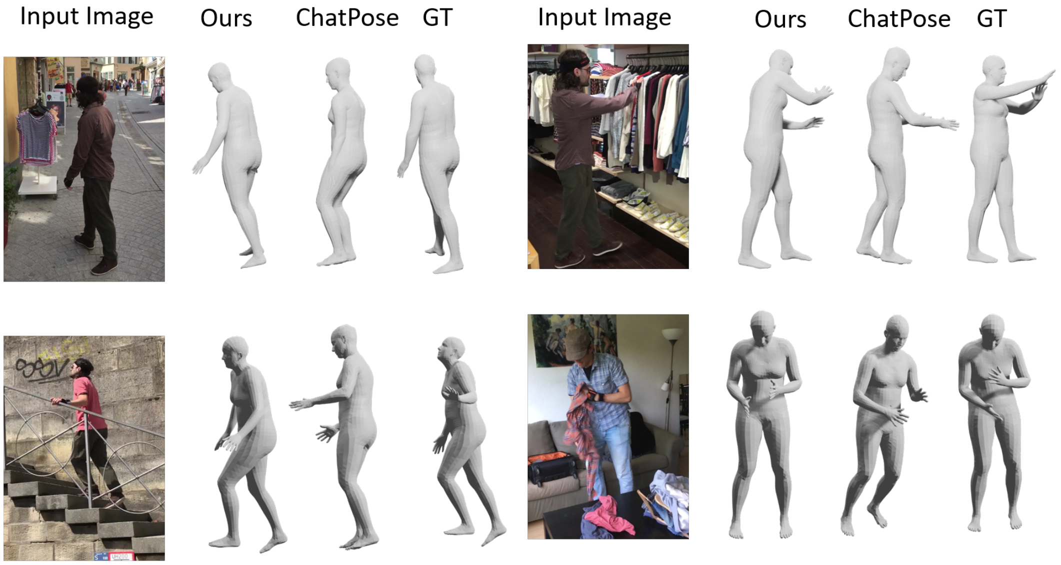
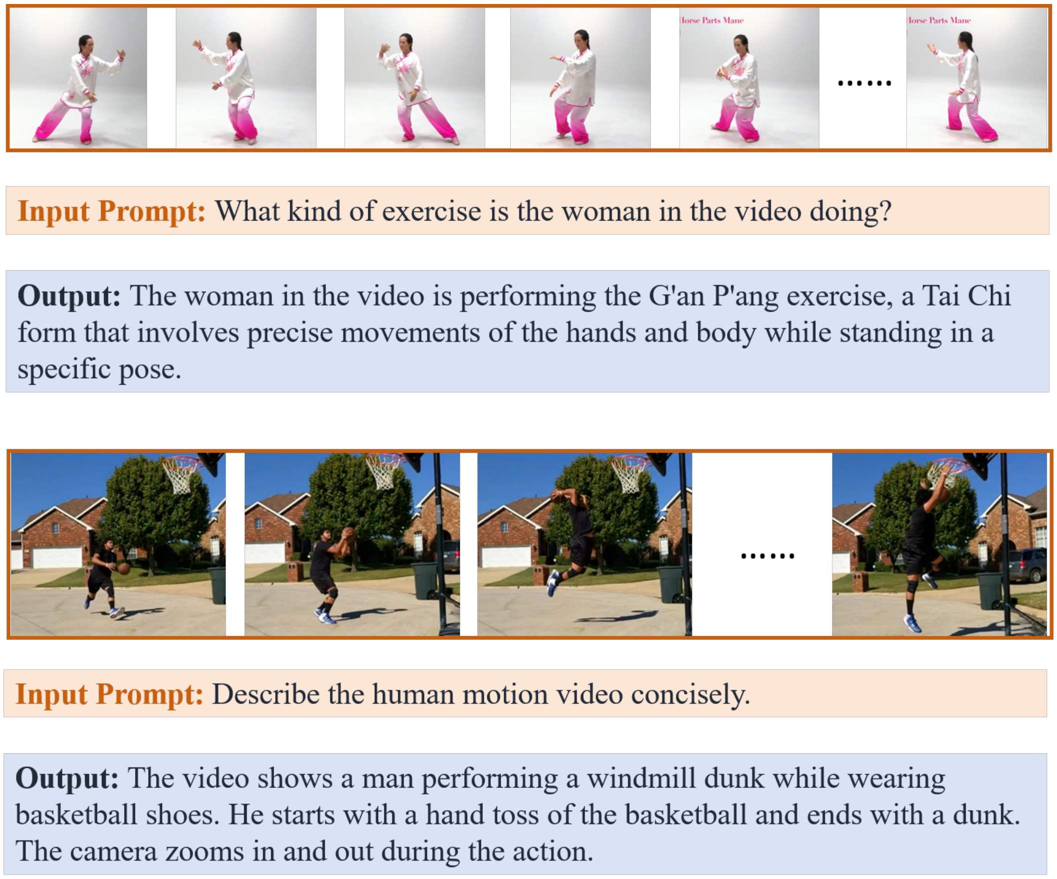


Figure 8: Qualitative comparison with the ChatPose [16] on the 3DPW test set.



**Input Prompt:** What kind of exercise is the woman in the video doing?

**Output:** The woman in the video is performing the G'an P'ang exercise, a Tai Chi form that involves precise movements of the hands and body while standing in a specific pose.

**Input Prompt:** Describe the human motion video concisely.

**Output:** The video shows a man performing a windmill dunk while wearing basketball shoes. He starts with a hand toss of the basketball and ends with a dunk. The camera zooms in and out during the action.

Figure 9: Examples of human motion RGB video understanding.

**Semantic cycle invariance testing.** We combine text-to-motion generation with video understanding tasks to evaluate whether the semantic consistency is preserved after the cycle. As shown in the qualitative results in Fig. 10, we first generate a 3D motion based on the input text prompt, then render the 3D motion sequence into a 2D character video. The model is then tasked with understanding the motion video (using a fixed prompt: "Describe the rendered human motion video concisely"), and finally, we examine the semantic alignment between the input and output texts. In the semantic cycle test, we find that the motion semantics are largely preserved, but errors and hallucinations still occur. Our video understanding strategy involves sampling frames from the video and feeding them into the foundation model. However, if key poses are missed during sampling or represent only a small portion of the entire motion sequence, semantic inaccuracies may arise. For example, in the left part of Figure 10, kneeling is mistakenly identified as squatting.

## B Additional Experiments

The evaluation results on the KIT-ML dataset [49] are shown in Table 6. This section adopts the same evaluation metrics as used for the HumanML3D dataset. Similarly, Our method with single task instruction tuning outperforms other foundation model-based text-to-motion approaches across most metrics.

## C Visualization

As shown in Fig. 11, we visualize the forgetting effects on foundation language models before and after the text-to-motion task for our method (Vicuna-7B-v1.5 based), our method (Gemma-2-2B-it based), MotionGPT, and MotionAgent under the MT_Bench [78] benchmark. This corresponds

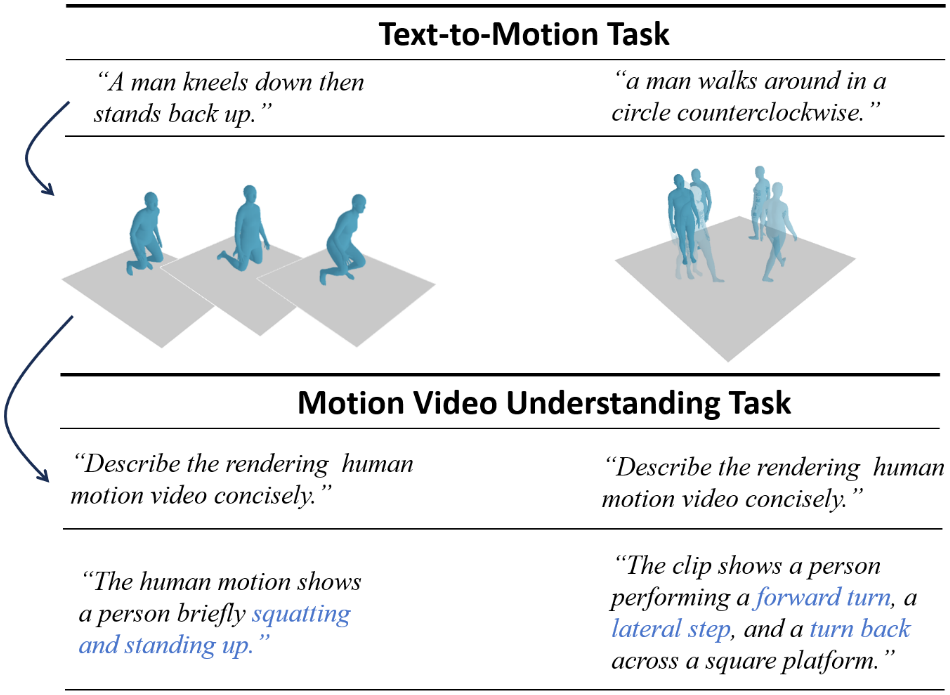

Text-to-Motion Task
"A man kneels down then stands back up."
"a man walks around in a circle counterclockwise."
Motion Video Understanding Task
"Describe the rendering human motion video concisely."
"Describe the rendering human motion video concisely."
"The human motion shows a person briefly squatting and standing up."
"The clip shows a person performing a forward turn, a lateral step, and a turn back across a square platform."

Figure 10: Semantic loop invariance testing.

to Tab. 1 in the main text. MT_Bench covers eight topics, including writing, roleplay, extraction, reasoning, and more. In Fig. 11, greater overlap of the circles indicates less forgetting in the foundation models. It can be observed that our method (based on Gemma-2-2B-it) and MotionAgent exhibit significant differences in their anti-forgetting ability under the same foundation model, which is attributed to our MoE LoRA architecture.

## D Implementation Details

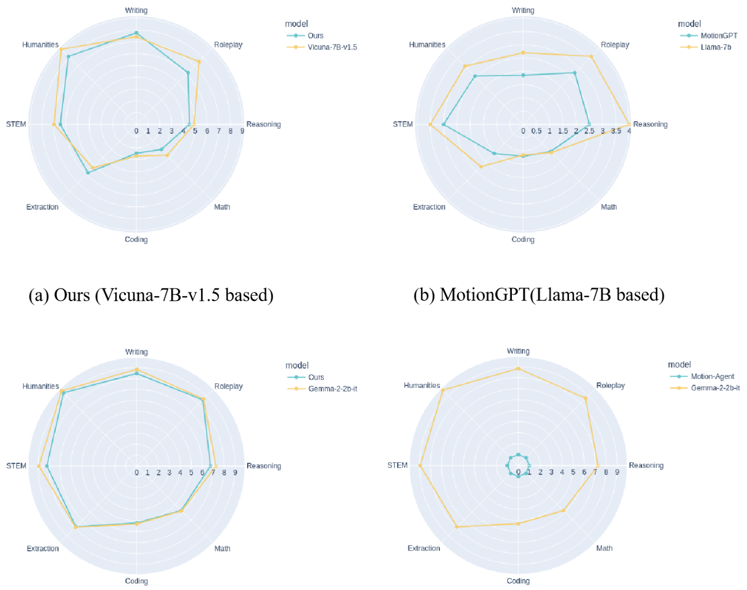
We conduct experiment on a single NVIDIA A800 80G GPU. In the pose and motion tokenizer, the number of body parts is set to 5, corresponding to the torso and four limbs with each part having an embedding dimension of $S = 512$. The codebook size for each body part is fixed at $K = 512$ and temporal compression ratio is set to $l = 4$ in motion tokenization. During training, the commitment loss coefficient $\lambda_{com}$ is set to 0.02. We use the AdamW optimizer with hyperparameters $[\beta_1, \beta_2] = [0.9, 0.99]$ and a learning rate of $2 \times 10^{-4}$.

During the instruction tuning stage, simultaneous fine-tuning on three human motion-related tasks took a total of 120 hours. We used a batch size of 32 and a micro batch size of 2. AdamW is also used for optimization in this stage, with an initial learning rate of $3 \times 10^{-3}$, which is scheduled using Cosine Annealing.

When calculating the MoE LoRA inference latency (as shown in Figure 5 (b)), we use Vicuna-7b-v1.5 as the foundation model. The inference time is the sum of the gating network computation and multimodal inference time. We conduct latency tests with a batch size of 1 and without using caching. Since the model's inference time is related to the number of input tokens, we fix the input token length to 84 when testing latency under different numbers of experts, as this is the average token length in the T2M test set.

Table 6: Quantitative results of text-to-motion on the KIT-ML dataset

| Methods | R precision↑ | | | FID.↓ | MM-D.↓ | Div.→ |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| GT | 0.424 ±.005 | 0.649 ±.006 | 0.779 ±.006 | 0.031 ±.004 | 2.788 ±.012 | 11.080 ±.097 |
| TM2T [22] | 0.280 ±.006 | 0.463 ±.007 | 0.587 ±.005 | 3.599 ±.153 | 4.591 ±.026 | 9.473 ±.117 |
| T2M [21] | 0.361 ±.006 | 0.559 ±.007 | 0.681 ±.007 | 3.022 ±.107 | 3.488 ±.028 | 10.720 ±.145 |
| MDM [60] | 0.164 ±.004 | 0.291 ±.004 | 0.396 ±.004 | 0.497 ±.021 | 9.191 ±.022 | 10.847 ±.109 |
| MD [74] | 0.417 ±.004 | 0.621 ±.004 | 0.739 ±.004 | 1.954 ±.062 | 2.958 ±.005 | **11.100** ±.143 |
| MLD [6] | 0.390 ±.008 | 0.609 ±.008 | 0.734 ±.007 | 0.404 ±.027 | 3.204 ±.027 | 10.800 ±.117 |
| T2M-GPT [73] | 0.416 ±.006 | 0.627 ±.006 | 0.745 ±.006 | 0.514 ±.029 | 3.007 ±.023 | 10.921 ±.108 |
| ReMoDiffuse [75] | 0.427 ±.014 | 0.641 ±.004 | 0.765 ±.055 | **0.155** ±.006 | 2.814 ±.012 | 10.800 ±.105 |
| AttT2M [79] | 0.413 ±.006 | 0.632 ±.006 | 0.751 ±.006 | 0.870 ±.039 | 3.039 ±.021 | 10.960 ±.123 |
| MoMask [20] | **0.433** ±.007 | **0.656** ±.005 | **0.781** ±.005 | 0.204 ±.011 | **2.779** ±.022 | 10.711 ±.087 |
| MotionGPT [76] | 0.340 ±.002 | 0.570 ±.003 | 0.660 ±.004 | 0.868 ±.032 | 3.721 ±.018 | 9.972 ±.026 |
| MotionGPT [32] | 0.366 ±.005 | 0.558 ±.004 | 0.680 ±.005 | 0.510 ±.016 | 3.527 ±.021 | 10.350 ±.084 |
| MotionAgent [67] | 0.409 ±.006 | 0.624 ±.007 | 0.750 ±.005 | 0.781 ±.026 | 2.982 ±.022 | **11.407** ±.103 |
| MotionGPT-2 [65] | **0.427** ±.003 | 0.627 ±.002 | 0.764 ±.003 | 0.614 ±.005 | 3.164 ±.013 | 11.256 ±.026 |
| Ours (single task) | 0.423 ±.004 | **0.643** ±.003 | **0.769** ±.004 | **0.306** ±.014 | **2.848** ±.016 | 11.175 ±.093 |
| Ours | 0.381 ± .005 | 0.585 ±.003 | 0.680 ±.013 | 0.567 ± .028 | 3.404 ±.019 | 10.595 ±.125 |



(a) Ours (Vicuna-7B-v1.5 based)

(b) MotionGPT(Llama-7B based)

(c) Ours (Gemma-2-2B-it based)

(d) Motion-Agent (Gemma-2-2B-it based)

Figure 11: Visualization of the forgetting levels before and after T2M fine-tuning across different methods.

### D.1 Evaluation Metrics

**R-precision:** This metric evaluates the consistency between the generated motion and the textual description. Specifically, a generated motion is paired with its corresponding ground-truth text and 31 randomly selected unrelated text descriptions to form a candidate set. Features of the motion and all text descriptions are extracted using their respective encoders [21], and pairwise feature distances are computed. These distances are then ranked in ascending order. The probabilities of the ground-truth text appearing in the Top-1, Top-2, and Top-3 positions are reported as the evaluation result. Higher values indicate that the generated motion better aligns with the semantic description.

**FID:** This metric measures the distributional difference between generated motions and real motions from the dataset. A lower FID score indicates that the generated motions are more similar to real samples in terms of overall feature distribution.

**MultiModal Distance (MM-D.):** This metric evaluates semantic alignment by computing the Euclidean distance between the text feature and the corresponding generated motion feature. A smaller value indicates better semantic matching.

**Diversity (Div):** This metric assesses the diversity of motions generated by the model. Specifically, two subsets, each containing 300 randomly selected generated motions, are sampled. The average Euclidean feature distance between the two subsets is calculated. A larger value indicates greater diversity in the generated results.

**MPJPE:** Mean Per Joint Position Error measures the average Euclidean distance between the predicted and ground-truth joint positions of a generated 3D human pose. It is computed by calculating the Euclidean distance for each joint in every frame and then averaging over all joints and frames. This metric reflects the spatial accuracy of the generated motion, with lower values indicating closer alignment to the ground-truth poses.

**PA-MPJPE:** Procrustes Aligned MPJPE calculates the MPJPE after applying a rigid alignment (including rotation, scaling, and translation) to the predicted poses to remove global transformation differences. This metric focuses on the structural correctness of the predicted pose regardless of its absolute position and scale. Lower values indicate better structural alignment with the ground truth.

### D.2 Instruction Templates

Tab. 7 presents the templates used for our task instruction tuning, including text-to-motion, pose estimation and video understanding. The <Motion_Placeholder>, <Image_Placeholder>,<Video_Placeholder>, and <Caption_Placeholder> respectively represent the motion sequence, image input, video input and textual description from the trianing datasets.

Table 7: Examples of instruction templates for each task

| Task | Input | Output |
|---|---|---|
| Text-to-Motion | Generate a sequence of motion tokens matching the following human motion description. | <Motion_Placeholder> |
| | Generate a sequence of motion tokens matching the following human motion description given the initial token <Motion_Placeholder>. | <Motion_Placeholder> |
| | Generate a sequence of motion tokens matching the following human motion description given the last token <Motion_Placeholder>. | <Motion_Placeholder> |
| Pose Estimation | Can you predict the SMPL pose of the person in this image <Image_Placeholder>. | <Pose_Placeholder> |
| | There is a person in the middle of the image, please output this person's SMPL pose <Image_Placeholder>. | <Pose_Placeholder> |
| | What is the human pose in this image? Please respond with SMPL pose <Image_Placeholder>. | <Pose_Placeholder> |
| | What is the person doing in this image? Please output SMPL pose <Image_Placeholder>. | <Pose_Placeholder> |
| | There is a person in the middle of the image, use SMPL to describe the pose <Image_Placeholder>. | <Pose_Placeholder> |
| Video Understanding | Write a terse but informative summary of the following human motion video clip <Video_Placeholder>. | <Caption_Placeholder> |
| | Describe the following human motion video concisely <Video_Placeholder>. | <Caption_Placeholder> |
| | Render a clear and concise summary of the human motion video below <Video_Placeholder>. | <Caption_Placeholder> |
| | Share a concise interpretation of the human motion video provided <Video_Placeholder>. | <Caption_Placeholder> |
| | Relay a brief, clear account of the human motion video shown <Video_Placeholder>. | <Caption_Placeholder> |
| | ... | <Caption_Placeholder> |