MSCode: Advancing Human Motion-Language Understanding via Modality-Shared Codebook

Haoyu Shi
Inner Mongolia University
Hohhot, China
shihaoyu@mail.imu.edu.cn

Huaiwen Zhang*
Inner Mongolia University
Hohhot, China
huaiwen.zhang@imu.edu.cn

Abstract

Recently, human motion understanding has been a prominent area of research due to its critical importance in many fields. The key to advancing this understanding lies in the precise alignment between motion and linguistic modalities. Existing methods mainly follow two paradigms: global contrastive alignment and vocabulary space-based alignment. However, motion sequences exhibit sequential spatiotemporal dynamics while text conveys abstract semantics, leading to a fundamental mismatch in semantic levels and granularities. This undermines cross-modal alignment and results in suboptimal downstream performance. To alleviate this, we introduce a modality-shared codebook that enables unified representation learning and precise alignment of motion and linguistic modalities. Each codeword in the codebook is regularized to encode cross-modality shared semantics, and we leverage sparse activation and distribution consistency loss to enforce matched motion and text are represented by the same set of codewords. Additionally, we introduce a locality-aware Gaussian encoder to refine pose features and design a hard-negative guided loss to strengthen alignment discriminability. Extensive experiments across various language-motion evaluation, including text-motion retrieval, text-motion grounding, and motion caption, demonstrate that our model significantly surpasses current state-of-the-art methods.

1 Introduction

Human motion understanding is an expanding frontier, holding substantial implications for applications ranging from film production and gaming to virtual reality and robotics. As a long-standing research hotspot, human motion understanding has led to the development of various tasks, including human motion-text retrieval (Petrovich et al., 2023; Yu et al., 2024; Messina et al., 2023; Fujiwara et al., 2024; Lyu et al., 2025; Shi and Zhang, 2024, 2025; Yang et al., 2024b,a; Zhang et al.), motion captioning (Jiang et al., 2023; Guo et al., 2022b), and motion grounding (Yan et al., 2023), all of which have made significant progress in recent years.

The core of human motion understanding lies in precise alignment between motion and language representations. Existing methods (Petrovich et al., 2023; Messina et al., 2023; Yu et al., 2024) typically encode global motion features and sentence embeddings into a latent space, then adopt contrastive learning to enforce the alignment of representations between matched motion-text pairs (as shown in Fig. 1(a)). However, motion sequences inherently contain intricate semantic details (e.g., subtle body movements), which pose a considerable challenge to achieving precise motion-text alignment when relying solely on global embeddings. To address this, the ToHL method (Lyu et al., 2025) attempts to align motion and text modalities in a shared lexical vocabulary space for unified motion-language understanding, as illustrated in Fig. 1(b). However, this vocabulary space exhibits an extremely high feature dimension (i.e., 30,522, the vocabulary size of BERT (Devlin et al., 2019a)), which impairs the efficiency of motion-text matching. Moreover, such static lexical representations struggle to effectively capture the dynamic nature of motion sequences.

In this paper, we argue that above models overlook a critical issue: the semantic information conveyed by motions and text inherently differs in granularity. For instance, a motion sequence of "walk forward quickly" portrays various attributes, such as step length, body posture, motion dynamics, and walk speed. In contrast, the corresponding textual description (e.g., "a person is quickly walking forward") is generally more abstract and compact. Existing methods fail to explicitly align the semantics of motion and text at the same granu-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 38116–38132

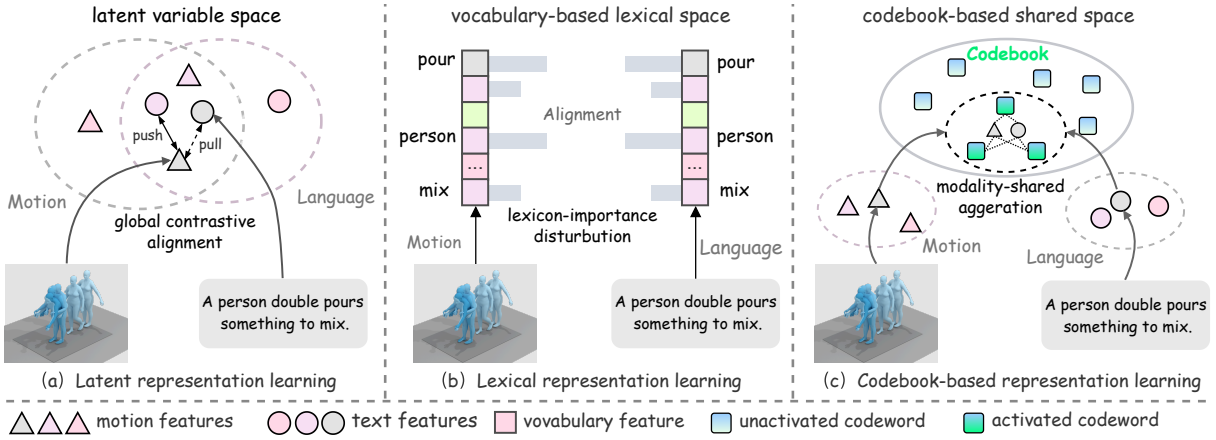

Figure 1: Comparison of different feature representation learning methods. (a) Latent representation learning directly aligns motion features and text features in a latent variable space via global contrastive alignment. (b) Lexical representation learning achieves cross-modal connection by aligning the lexicon-importance distributions of motion and text representations in a vocabulary-based lexical space. (c) Our proposed method introduces a modality-shared codebook for representation learning and alignment. Both motion and language features are activated by the same set of codewords, enabling unified semantic granularities and distributions of the two modalities.

larity level, hindering effective multimodal representation learning and potentially leading to performance degradation.

To address the above challenge, we introduce a modality-shared codebook that enables unified representation learning and precise alignment of motion and linguistic modalities. The codebook consists of several learnable codewords that encode cross-modal shared semantic features. Both motion and text are represented as the combination of shared codewords in a unified semantic space so that the semantic granularities are unified (see Figure 1(c)). To avoid noisy activation and strengthen consistent codeword activation across motion and text, we further propose sparse activation and a distribution consistency loss, enforcing matched motion-text pairs to share the same small set of codewords. Additionally, we observe that vanilla self-attention transformer encoder tends to drive pose embeddings toward excessive similarity, which impairs the model's ability to capture fine-grained motion semantics. To mitigate this issue, we propose a locality-aware gaussian encoder that constrains each pose to focus on its adjacent poses instead of the entire sequence, thereby ensuring high-quality pose features for codeword aggregation. Finally, a hard-negative guided contrastive loss is reformulated to mine more discriminative representations to build a better-aligned representation space. Our contributions can be summarized as follows:

- We introduce a modality-shared codebook for representation learning and alignment between motion and language, which unifies the granularity and distribution of two modalities.
- We propose a Locality-aware Gaussian Encoder to obtain high-quality pose features for codeword aggregation, and design a hard-negative guided contrastive loss to enable a well-aligned representation space.
- Comprehensive evaluations across several language-motion tasks demonstrate that our model achieves state-of-the-art performance, affirming the efficacy of the proposed method.

2 Related work

2.1 Human Motion Understanding

The rapid proliferation of 3D human motion data has made human motion understanding an increasingly critical area in computer vision research. In the motion-language domain, this understanding encompasses key tasks such as text-motion retrieval text-to-motion grounding, and motion captioning. Text-motion retrieval involves identifying the most semantically relevant 3D human motion sequences from extensive databases based on natural language queries. Existing works, including TMR (Petrovich et al., 2023), TEMOS (Petrovich et al., 2022), MotionCLIP (Tevet et al., 2022), DTL (Yan et al., 2023), MoPa (Yu et al., 2024) and CAR (Fujiwara et al., 2024), tackle this task by constructing a cross-modal embedding space via CLIP-style

contrastive learning. For a more fine-grained understanding, text-to-motion grounding focuses on localizing semantically relevant motion segments within untrimmed sequences. Motion captioning aims to generate descriptive captions for human motions.TM2T (Guo et al., 2022b) encodes motion sequences and uses a translation network to align motion and text tokens. MotionGPT (Jiang et al., 2023) models human motions as a foreign language, enabling descriptions via an expanded vocabulary.

2.2 Vector-Quantization and Codebook.

The codebook is a key component in vector quantization (van den Oord et al., 2018), widely used in both understanding (Bao et al., 2022) and generation (Kalakonda et al., 2022; Tevet et al., 2022) tasks. During quantization, encoder features are replaced by their nearest-neighbor codewords from the codebook before being reconstructed by the decoder. Inspired by this, current most text-to-motion generation methods (Guo et al., 2022b; Jiang et al., 2023; Deichler et al., 2025; Jin et al., 2026) universally adopt this paradigm, employing a codebook to transform continuous motion representations into a sequence of codebook indices. For instance, MotionGPT (Deichler et al., 2025) employs separate text and motion codebooks to achieve multimodal modeling for text-to-motion generation. In contrast to existing methods, we employ a unified codebook to embed both textual and motion features, thereby explicitly constructing a shared multimodal latent space that facilitates unified cross-modal alignment.Furthermore, we reformulate the computational mechanism of the codebook to support the exploration of fine-grained cross-modal semantic correspondences.

3 Methodology

3.1 Model Overview

The overview of our proposed method is shown in Fig. 2. We begin by introducing the feature encoding of text queries and motion sequences. We then introduce our codebook-based representation learning to represent cross-modal shared semantic features, thereby unifying the semantic granularity of these two modalities. Next, we design a locality-aware Gaussian encoder to capture high-quality pose features crucial for effective codeword aggregation. Finally, we refine the learning process with a hard-negative guided contrastive loss, which mines more discriminative representations, ultimately building a better-aligned latent space.

3.1.1 Text query encoding.

Given a sentence containing N words, we utilize a pre-trained DistilBERT (Devlin et al., 2019b) model with a projection head to extract initial word embeddings, which are denoted as:

$$\mathcal{T} = \{w_1, w_2, \ldots, w_N\} \in \mathbb{R}^{N \times D} \quad (1)$$

where N is the number of words in the sentence and D denotes the feature dimension.

3.1.2 Motion sequence encoding.

3D human motion is defined as a sequence of 3D human poses, $\mathcal{P} = (p_1, \ldots, p_M) \in \mathbb{R}^{M \times d}$, where M is the pose number, d is the pose dimension. Each p in the sequence is a detailed representation of the articulated human body, including join positions, rotations, foot contact, etc. We use our proposed Gaussian encoder to encode the poses into a sequence of embeddings:

$$\mathcal{M} = \{m_1, m_2, \cdots, m_M\} \in \mathbb{R}^{M \times D} \quad (2)$$

where M is the number of poses, and D is the feature dimension.

3.2 Codebook-based Representation Learning

To unify semantic granularity and establish a common multi-modal space, we introduce a modality-shared codebook for representation learning. This codebook serves as common bases for both motion and text representations. As a result, the granularities of cross-modal semantics are explicitly unified. In addition, it encodes the shared semantic knowledge inherent in both modalities, serving as an informative prior to guide motion and text encoders in learning discriminative cross-modal embeddings.

3.2.1 Modality-shared Feature Aggregation.

Assume that we have a batch of motion-text pairs $\{(m_i, t_i)\}_{i=1}^{B}$, where m_i, t_i represent the i-th motion sequence and its caption respectively, and B is the batch size. This aggregator uses a set of N shared codewords to represent motion feature $\mathcal{M}^C$ and text feature $\mathcal{T}^C$. Specifically, it can be represented as:

$$\mathcal{M}_c = \sum_{k=1}^{N} w_{i,k}^{(m)} z_k, \mathcal{T}_c = \sum_{k=1}^{N} w_{i,k}^{(t)} z_k \quad (3)$$

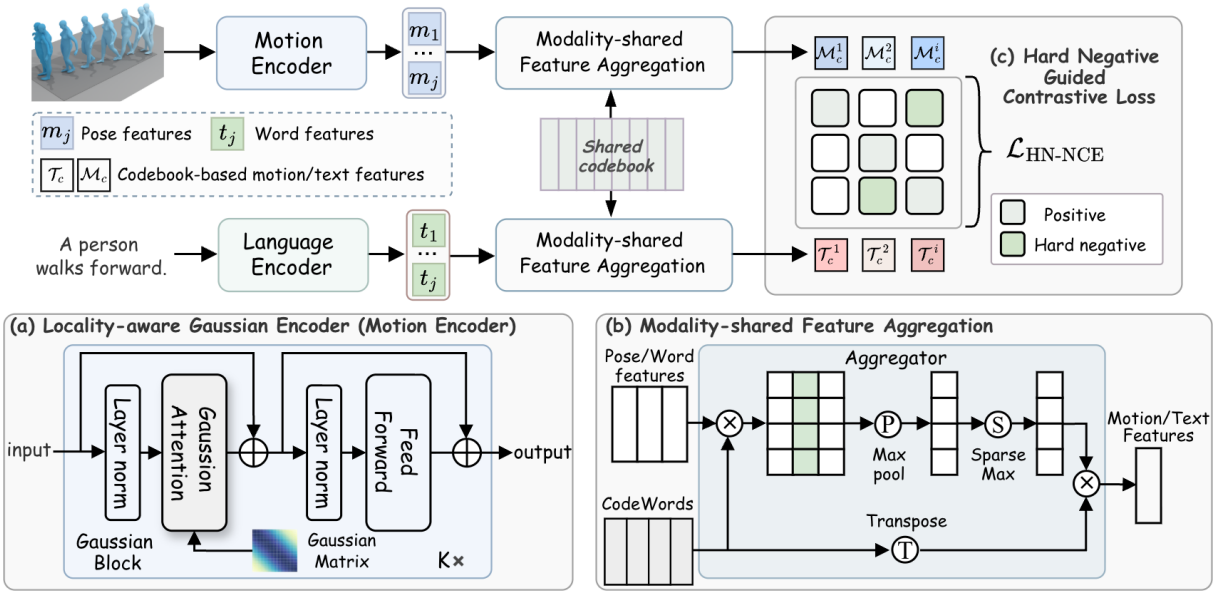

Figure 2: The overall architecture of MSCode. Our method consists of three key components: the Locality-aware Gaussian Encoder, the Modality-shared Codebook, and the Hard Negative Guided Contrastive Loss.

where $\{z_k \mid z_k \in \mathbb{R}^D, k = 1, 2, \cdots N\}$ represents the set of learnable cross-modal shared codewords, and $w_{i,k}^{(m)}$, $w_{i,k}^{(t)}$ denote the aggregation weights corresponding to the motion sequence m_i and the text query t_i, respectively. To capture rich semantics within the feature aggregation, we design the pipeline to calculate the aggregation weights $w_{i,k}^{(m)}$ and $w_{i,k}^{(t)}$. Specifically, during the motion feature aggregation, we define the relevance score $s_{i,k}^{(m)}$ between the motion sequence and each codeword z_k as:

$$s_{i,k}^{(m)} = \max_j < m_{i,j}, z_k > /\eta \tag{4}$$

where $m_j \in \mathbb{R}^D$ is the j-th pose embedding in motion sequence, η is a scaling factor, and $< \cdot, \cdot >$ is the inner product function.

Sparse Activation. The relevance scores $s_{i,:}^{(m)}$ are next normalized by the Sparsemax function (Martins and Astudillo, 2016), which works similarly to Softmax but encourages most of the elements in the probability distribution to be precisely zero:

$$w_{i,:}^{(m)} = \text{Sparsemax}(s_{i,:}^{(m)}) \tag{5}$$

where $w_i^{(m)}$ is the weight of the i-th token with respect to the motion sequence. By the sparse constraints, the motion feature $\tilde{m}_i$ can be represented by only a few relevant codewords, reducing noisy activations and improving the interpretability of the model. Similarly, the text feature $\mathcal{T}_c$ is aggregated in the same way.

3.2.2 Activation Distribution Consistency.

To refine the semantic correspondence established through the codebook aggregation, we propose an auxiliary loss that enforces consistency at the level of semantic codebook selection. This constraint explicitly minimizes the distributional difference between the codebook activation vectors of matched motion $w_{i,:}^{(m)}$ and text $w_{i,:}^{(t)}$, promoting semantic co-selection. We use Symmetric KL-Divergence (Kullback and Leibler, 1951) to measure the distributional disparity between the matched pairs, which is effective for sparse probability vectors:

$$\mathcal{L}_{\text{ADC}} = \sum_{i=1}^{B} \left(\text{KL}\left(w_{i,:}^{(m)} \| w_{i,:}^{(t)}\right) + \text{KL}\left(w_{i,:}^{(t)} \| w_{i,:}^{(m)}\right) \right) \tag{6}$$

This auxiliary loss compels matched motion-text samples to exhibit highly consistent activation over the semantic codebook, further strengthening the fine-grained cross-modal semantic alignment.

3.3 Locality-aware Gaussian Encoder

Motion sequences often consist of multiple actions with only subtle differences yet convey completely distinct semantics. Traditional transformer architectures, which rely on vanilla self-attention mechanisms, typically facilitate global interactions across the entire sequence, leading to excessively similar features. To address this, we propose a Locality-aware Gaussian Encoder (illustrated in Fig. 2(a)), which utilizes parallel multi-scale Gaussian blocks to capture discriminative pose features for code-

words aggregation.

Gaussian Attention. Initially, we project the input $\mathcal{P} \in \mathbb{R}^{M\times d}$ into query, key, and value matrices via learnable parameters W_q, W_k, and W_v. We perform scaled dot-product attention on the query and key matrix to obtain an attention score matrix. Then we design a Gaussian matrix $W^g \in \mathbb{R}^{M\times M}$ to perform element-wise product over the attention score matrix. After putting this result through a softmax function to determine attentional distributions over the value matrix, we get the output of the Gaussian attention module:

$$\begin{aligned} \mathcal{M}^{attn} &= \text{Softmax}\left(W^g \odot \frac{\mathcal{P}W_q\left(\mathcal{P}W_k\right)^\top}{\sqrt{d_k}}\right)\mathcal{P}W_v, \\ W^g(i,j) &= \frac{1}{2\pi}\exp\left(-\frac{(j-i)^2}{\sigma^2}\right) \end{aligned} \tag{7}$$

where d_k is the dimension of queries and keys, σ^2 is the variance of the Gaussian distribution and $\odot$ indicates the element-wise product function.

Multi-Scale Aggregation. Following the Gaussian attention module, $\mathcal{M}^{attn}$ is fed into a Feed-Forward Network (FFN) to obtain Gaussian block output $\mathcal{M}_i^{out}$. Similar to the vanilla Transformer block, residual connections and Layer Normalization are employed in both the Gaussian attention and the FFN. Furthermore, considering that actions typically exhibit variable temporal durations given, we deploy K Gaussian blocks in parallel, each with a different variance σ_k^2, which effectively constructs multi-scale temporal receptive fields. The final motion embedding $\mathcal{M}$ is then obtained by averaging the outputs of these parallel outputs:

$$\mathcal{M} = \frac{1}{K}\sum_{k=1}^{K}\text{GB}(\mathcal{M}^{out}, \sigma_k^2) \tag{8}$$

where $\text{GB}(\cdot, \sigma_k^2)$ denotes a Gaussian block with variance σ_k^2. In practice, we set $K = 4$ and adopt Gaussian blocks with small, medium, large, and infinite variances respectively.

3.4 Hard Negative Guided Contrastive Loss

Contrastive learning significantly benefits from in-batch hard negative samples. For vision-language tasks, several contrastive-based works (Robinson et al., 2021; Gutmann and Hyvärinen, 2012) treat negative samples equally or rely on resampling or manually crafting hard negative instances to improve alignment, which generally involves more training costs. In this work, we devise a simple and effective re-weighting approach to force the model to pay more attention to hard negative samples during training. It is defined as follows,

$$\begin{aligned} \mathcal{L}_{\text{HN\_CLAP}} = &-\sum_{i=1}^{B}\log\frac{e^{\langle m_i,t_i\rangle/\tau}}{e^{\langle m_i,t_i\rangle/\tau}+\sum_{j,j\neq i}\alpha_{i,j}e^{\langle m_i,t_j\rangle/\tau}} \\ &-\sum_{i=1}^{B}\log\frac{e^{\langle t_i,m_i\rangle/\tau}}{e^{\langle t_i,m_i\rangle/\tau}+\sum_{j,j\neq i}\beta_{i,j}e^{\langle t_i,m_j\rangle/\tau}} \end{aligned} \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denote cosine similarity, and $\alpha_{i,j}$, $\beta_{i,j}$ is the difficulty scores for unpaired samples, they are designed so that difficult negative pairs (with higher similarity) are emphasized, and easier pairs are ignored. Thus, the model will be forced to learn a more discriminative feature space to distinguish confusable pairs for fine-grained alignment. It is defined as,

$$\alpha_{i,j} = \frac{(B-1)e^{\gamma<m_i,t_j>/\tau}}{\sum_k e^{\gamma<m_i,q_k>/\tau}}, \beta_{i,j} = \frac{(B-1)e^{\gamma<t_i,p_j>/\tau}}{\sum_k e^{\gamma<t_i,p_k>/\tau}} \tag{10}$$

where γ is a scaling ratio, the larger it is, the more importance we attach to the hard negative samples as the distribution of $\alpha_{i,:}$, $\beta_{i,:}$ can be sharper.

4 Experiments

4.1 Experimental Setup

Datasets. Our model's performance is evaluated on two commonly used public datasets: HumanML3D (Guo et al., 2022a) and KIT-ML (Plappert et al., 2016). HumanML3D (Guo et al., 2022a) is a large-scale dataset having 14,616 motion sequences paired with 44,970 detailed textual descriptions. Following prior research (Guo et al., 2022a), we use a standard split of 23,384 samples for training, 1,460 for validation, and 4,383 for testing. KIT-ML (Plappert et al., 2016) provides 3,911 motion sequences along with 6,278 corresponding text inputs. For this dataset, we utilize 4,888 samples for training, 300 for validation, and 830 for testing.

Evaluation Metrics. For text-to-motion retrieval, we use standard metrics, including Recall at various ranks (R@1, R@5, etc.), and the median rank (MedR) for both text-to-motion and motion-to-text retrieval tasks; for text-to-motion grounding, following prior work (Yan et al., 2024), we adopt two protocols (Normal, Assigned) and use IoU/mIoU as metrics. The Assigned protocol accounts for repeated or similar actions within a single motion for evaluation. For motion captioning, we employ the linguistic metrics as previous works (Jiang et al., 2023), including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al.,

Methods	Text-to-motion retrieval				Motion-to-text retrieval			
	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
T2M	1.80	7.12	12.47	81.00	2.92	8.36	12.95	81.50
TEMOS	2.12	8.26	13.52	173.00	3.86	9.38	9.38	183.25
MotionCLIP	2.33	12.77	18.14	103.00	5.12	12.46	19.02	91.42
DTL	2.69	11.36	17.71	73.00	2.33	10.31	17.48	76.00
TMR	5.68	20.34	30.94	28.00	9.95	23.56	32.69	28.50
MoPa	6.25	20.96	31.29	28.00	10.26	25.31	35.98	23.50
CAR	6.55	22.99	34.60	22.00	11.18	25.52	36.38	21.50
ToHL	6.78	23.10	33.89	22.00	10.76	26.20	36.95	21.50
MSCode	**7.69**	**24.20**	**35.86**	**21.50**	**12.38**	**26.98**	**38.45**	**21.25**

Table 1: Results of text-to-motion retrieval and motion-to-text retrieval on HumanML3D (Guo et al., 2022a).

Methods	Text-to-motion retrieval				Motion-to-text retrieval			
	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
T2M	3.37	16.87	27.71	28.00	4.94	16.14	25.30	28.50
TEMOS	7.11	24.10	35.66	24.00	11.69	26.63	36.39	26.50
MotionCLIP	4.87	20.09	31.57	26.00	6.55	25.48	34.97	23.00
DTL	6.77	23.18	37.24	18.00	9.11	25.26	38.02	17.00
TMR	7.23	28.31	40.12	17.00	11.20	28.07	38.55	18.00
MoPa	8.05	30.11	42.69	16.00	12.65	29.64	40.33	16.00
CAR	8.59	30.88	43.59	16.00	12.61	30.55	40.27	16.00
ToHL	8.96	31.25	44.43	15.50	12.89	30.87	44.96	15.50
MSCode	**10.61**	**33.25**	**45.21**	**14.50**	**13.80**	**31.59**	**42.88**	**14.50**

Table 2: Results of text-to-motion retrieval and motion-to-text retrieval on KIT-ML (Plappert et al., 2016).

2014), and BERTScore (Zhang* et al., 2020). The best evaluation results are highlighted in "**bold**".

4.2 Experimental results

To validate our model's capabilities, we conduct experiments across the following tasks. Importantly, for the motion caption task, we equip our model with a language decoder for autoregressive caption generation and adopt a loss function that maximizes the log-likelihood of the predicted token distribution during training.

Results of Text-to-Motion Retrieval. To evaluate the alignment between text and motion features, we conduct text-motion retrieval benchmarks on the HumanML3D and KIT-ML datasets. As summarized in Tab. 1 and 2, our method consistently outperforms all previous approaches across all evaluation metrics on both datasets. This indicates that our model effectively captures action semantics through our codebook-based representation learning, achieving more precise cross-modal semantic alignment for the language-motion model.

Results of Text-to-Motion Grounding. We perform zero-shot evaluation on the text-to-motion grounding task to validate the model's fine-grained understanding capability. This evaluation is implemented on the HumanML3D (restore) dataset, introduced in Sec. A.1 of the appendix. Notably, during the entire training process, the start/end labels of temporal segments are completely not seen by the model. As shown in Tab. 3, our method outperforms the existing method. This further demonstrates the effectiveness of our model and its ability to fine-grained cross-modal alignment.

Settings	Methods	HumanML3D(restore) Dataset				
		IoU@0.3↑	IoU@0.5↑	IoU@0.7↑	IoU@0.9↑	mIoU↑
Normal	TMR	67.15	44.01	29.03	5.29	40.40
	DTL	64.26	42.28	28.53	4.60	39.87
	HSA	70.15	46.98	33.26	7.98	42.45
	ToHL	71.69	45.56	33.02	10.55	44.80
	CAR	73.12	46.07	35.84	11.49	44.94
	MSCode	**75.63**	**48.10**	**37.89**	**13.18**	**46.85**
Assigned	TMR	84.26	62.25	41.85	7.17	53.60
	DTL	80.69	58.76	38.81	5.63	52.16
	HSA	87.63	72.62	55.14	13.36	61.25
	ToHL	89.23	73.02	57.56	16.98	63.14
	CAR	89.87	74.61	59.74	18.36	64.25
	MSCode	**91.65**	**77.11**	**61.59**	**20.31**	**67.09**

Table 3: Results of text-to-motion grounding. We employ the temporal pyramid method (Gao et al., 2017), using a sliding window ranging from 20 to 200 frames with a stride of 5 frames to evaluate IoU.

HumanML3D					
Methods	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	Bert Score↑
TM2T	48.9	7.00	38.1	16.8	32.2
MotionGPT	48.2	12.47	37.4	29.2	32.4
ToHL	49.7	13.62	39.2	53.1	33.1
MSCode	**51.2**	**15.29**	**41.6**	**59.5**	**34.2**
KIT-ML					
Methods	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	Bert Score↑
TM2T	35.1	6.2	28.7	28.9	30.4
MotionGPT	34.5	7.3	27.4	36.8	30.5
ToHL	43.4	8.9	35.2	65.3	31.2
MSCode	**44.8**	**11.2**	**37.9**	**70.7**	**32.3**

Table 4: Results of motion captioning on HumanML3D (Guo et al., 2022a) and KIT-ML (Plappert et al., 2016).

Results of Motion Captioning. To evaluate the semantic capturing ability of our model, we conduct motion captioning benchmarks on the HumanML3D and KIT-ML datasets. We compare our approach with recent methods, including TM2T (Guo et al., 2022b), MotionGPT (Jiang et al., 2023) and ToHL (Lyu et al., 2025). As illustrated in Table 4, our method outperforms these method in generating text descriptions for specified motions. This further demonstrates our model's superior capacity for human motion understanding and semantic capture.

4.3 Ablation Studies and Analyses

Module Gain. To better verify the effectiveness of our component, we provide a comprehensive ablation study in Tab. 5. The incorporation of the shared codebook substantially boosts the model's performance, underscoring the effectiveness of our core design. Moreover, integrating a locally-aware gaussian encoder yields substantial performance improvements across all datasets. Ultimately, by

MC	LG	HN	Text-to-motion retrieval			Motion-to-text retrieval		
			R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
✗	✗	✗	5.82	20.18	31.55	9.87	24.03	32.55
✓	✗	✗	6.88	23.24	35.09	11.76	26.03	37.62
✗	✓	✗	6.38	21.76	33.71	11.09	25.17	35.34
✓	✓	✗	7.32	23.87	35.65	12.12	26.43	38.01
✓	✓	✓	**7.69**	**24.20**	**35.86**	**12.38**	**26.98**	**38.45**

Table 5: The effectiveness of each component: Modality-shared codebook(**MC**), Locality-aware Gaussian encoder (**LG**), and Hard Negative Guided Contrastive Loss(**HN**).

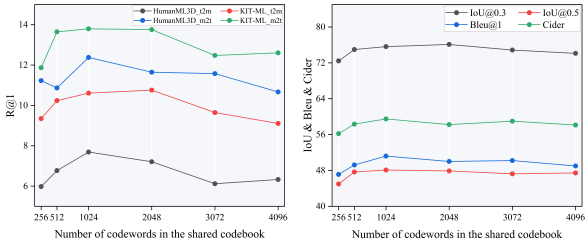

Figure 3: Model performance on text-motion retrieval (left) and grounding and captioning (right) tasks with different sizes of codebook on HumanML3D dataset.

incorporating a hard negative loss to enhance contrastive learning, the method achieves state-of-the-art performance on both datasets.

Size of Codebook. We conduct ablation studies on the size of our modality-shared codebook. As shown in Fig. 3(Left), the retrieval performance drops significantly when the codebook size increases from 1024 to 4096. We argue that a larger codebook introduces noisy activations and irrelevant information during aggregation, hindering accurate retrieval. However, fewer codewords, while effective for retrieval tasks, exhibit limitations in the fine-grained grounding task (Fig. 3(Right)). This may be because each codeword must convey multiple semantics within a smaller codebook, thereby disturbing the pose-word interaction while seeking fine-grained alignment. Finally, motion captioning achieves optimal performance with a codebook size of 1024.

Gaussian Block. We select four types of Gaussian blocks characterized by low, medium, high, and infinite variance to perceive action moments of varying durations. In this subsection, we investigate the impact of these Gaussian blocks. We construct four variants (i.e., w/o low, w/o medium, w/o high, and w/o infinite). Tab. 6 reports the performance of these variants across different groups. Notably, all variants exhibit inferior performance compared to the complete configuration, indicating that all four types of Gaussian blocks contribute to the overall performance.

Methods	Text-to-motion retrieval			Motion-to-text retrieval		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
complete	7.69	24.20	35.86	12.38	26.98	38.45
w/o low	6.83	21.57	32.19	10.92	24.35	35.62
w/o medium	7.01	22.14	33.05	11.28	24.97	36.18
w/o high	7.15	22.68	33.74	11.54	25.42	36.79
w/o infinite	7.38	23.36	34.82	11.96	26.13	37.53

Table 6: The ablation studies of the Gaussian block with different types.

Methods	Text-to-motion retrieval			Motion-to-text retrieval		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Effect of Sparse Activation						
Softmax	5.97	21.46	30.02	10.09	22.39	33.09
Sparsemax	7.69	24.20	35.86	12.38	26.98	38.45
Effect of Activation Distribution Consistency (ADC)						
w/o ADC	7.48	23.82	35.01	12.14	26.43	37.68
w/ ADC	7.69	24.20	35.86	12.38	26.98	38.45

Table 7: The ablation studies of the modality-shared feature aggregation.

Sparse Constraints. We further explore the design of activation functions for weight normalization during feature aggregation. As reported in Tab. 7, adopting the Softmax activation leads to an obvious performance drop on all evaluation tasks. We argue that this degradation stems from the dense weight allocation of Softmax, which tends to activate irrelevant codewords and introduces unwanted noise into the learned feature representations. In comparison, Sparsemax with inherent sparsity constraints consistently yields better results across all tasks. Such improvement arises from its ability to unify semantic granularity by representing cross-modal semantics using a shared set of representative codewords.

Activation Distribution Consistency. We further analyze the effectiveness of our proposed activation distribution consistency (ADC). As shown in Tab. 7, we observe a slight degradation in model performance when ADC is removed. This decline may be attributed to the inherent discrepancies between two modalities, which lead to inconsistent codeword activations and subsequently undermine the cross-modal alignment.

4.4 Visualization Results

Visualization of the learned codewords. To explicitly show the cross-modal correspondence learned by our modality-shared codebook, we visualize the similarity of codewords to textual words and pose features (Fig. 4). In the first example, the codeword #684 exhibits high similarity to "run"

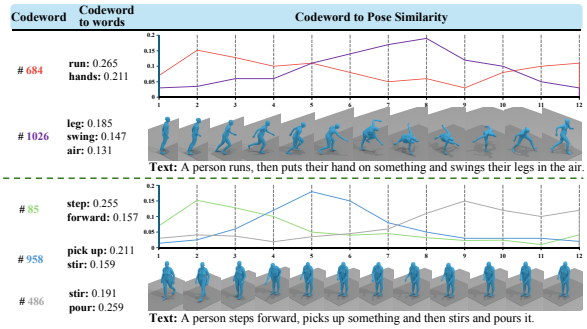

Figure 4: Visualization of codewords' role in connecting text modality and motion modality.

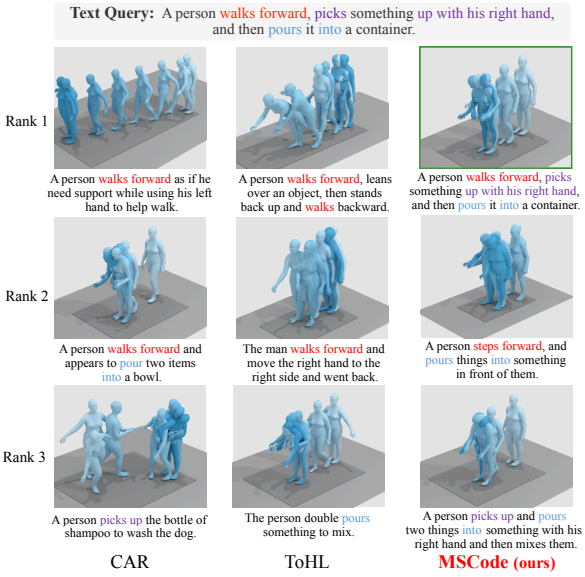

Figure 5: Visualization of the retrieval results. The text below each motion sequence is the ground truth.

(0.265) and maintains consistent pose similarity throughout the sequence, suggesting that it encodes holistic semantics. In contrast, codeword #1026 prioritizes the tokens "swing" (0.147) and "leg" (0.185); here, pose similarity is concentrated in the latter half of the sequence. In the second example, where the motion sequence comprises multiple subtle sub-actions, each codeword has high relevance values to the sub-action, further validating its comprehensive understanding ability.

Visualization of the retrieval results. In Fig. 5, we present the retrieval results for text-to-motion retrieval of the proposed method MSCode and the CAR (Fujiwara et al., 2024) and ToHL (Lyu et al., 2025). The given text query contains multiple actions. For each method, we showcase the top-3 retrieved motions along with their corresponding ground-truth text labels. For the CAR and ToHL results, we observe that all three retrieval methods

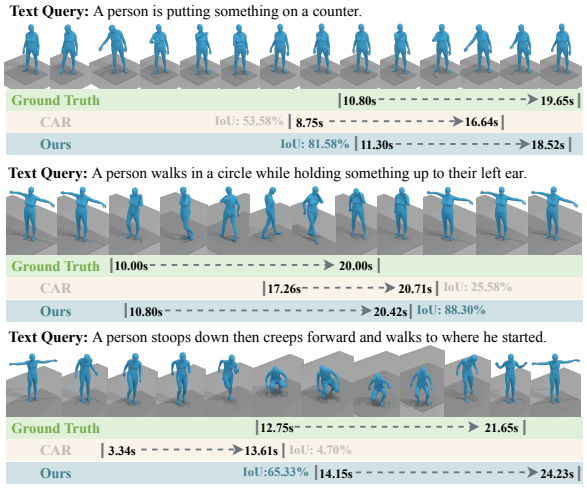

Figure 6: Visualization of the ground-truth moment and predictions by CAR and our proposed method on the HumanML3D(Restore) dataset.

capture only part of the semantics that align with the query text. While they successfully retrieve the action "walks forward", they fail to capture the more intricate actions. This may be due to the use of global features, which capture only the average semantics of the sequence, limiting their ability to capture more intricate action. Additionally, vocabulary-based representations struggle to capture motion dynamics. In contrast, our proposed method successfully retrieves the correct motion sequence, with the rank two and rank three retrieval results also capturing most of the actions.

Visualization of the motion grounding results. In Fig. 6, we present three examples of text-to-motion grounding. For a fair comparison, we provide the visualization results of our method and CAR (Fujiwara et al., 2024), both of which are evaluated under the zero-shot setting. From the first example in the figure, we can observe that when the text query contains fewer actions, both methods perform well with relatively high IoU values. However, when the query includes multiple actions, CAR exhibits poor performance; in contrast, our method can maintain the high IoU. This demonstrates our model's fine-grained understanding capability to perceive detailed action semantics.

5 Conclusion

In this paper, we propose MSCode, a framework based on modality-shared codebook, to address the semantic granularity mismatch between motion and text in human motion-language understanding. By integrating a modality-shared codebook, locality-

aware gaussian encoder, and hard-negative guided contrastive loss, MSCode unifies cross-modal semantic representation, refines pose features, and enhances alignment discriminability. Experiments on multiple datasets and tasks (text-motion retrieval, grounding, captioning) show that MSCode outperforms state-of-the-art methods, validating the effectiveness of its core components.

Limitations

While MSCode effectively mitigates the semantic granularity mismatch between motion and language and achieves state-of-the-art performance across multiple tasks, it still has several limitations for future improvements.

The first limitation lies in the fixed-size design of the modality-shared codebook, which introduces an inherent trade-off across diverse task requirements. As demonstrated in our ablation studies, a moderate codebook size (e.g., 1024) optimizes retrieval performance by balancing semantic coverage and noise resistance, whereas fine-grained motion grounding necessitates a larger codebook to capture subtle spatiotemporal dynamics. A static codebook fails to dynamically adapt to the varying semantic granularity demands of different tasks, resulting in suboptimal performance for scenario-specific applications.

Another practical challenge concerns the computational efficiency of the codebook aggregation process. While MSCode outperforms vocabulary-based methods (e.g., ToHL) in efficiency, the sparse activation mechanism and distribution consistency loss still introduce non-negligible computational overhead during training, particularly as the codebook size scales up.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62576179, 62576246, 62506178, 62532004, and 62276257, in part by the National Natural Science Foundation of Inner Mongolia under Grant 2025JQ012.

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. Beit: Bert pre-training of image transformers. *Preprint*, arXiv:2106.08254.

Anna Deichler, Jim O'Regan, Teo Guichoux, David Johansson, and Jonas Beskow. 2025. Grounded gesture generation: Language, motion, and space. *Preprint*, arXiv:2507.04522.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. 2024. Chronologically accurate retrieval for temporal grounding of motion-language models. In *Proc. of the European Conf. on Computer Vision (ECCV)*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. *Preprint*, arXiv:1705.02101.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022a. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5142–5151. IEEE.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022b. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. *Preprint*, arXiv:2207.01696.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2021–2029. ACM.

Michael U. Gutmann and tevet2022motionclipAapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. 13.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *Preprint*, arXiv:2306.14795.

Chuhao Jin, Rui Zhang, Qingzhe Gao, Haoyu Shi, Dayu Wu, Yichen Jiang, Yihan Wu, and Ruihua Song. 2026. Sentiavatar: Towards expressive and interactive digital humans. *arXiv preprint arXiv:2604.02908*.

Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. 2022. Action-gpt: Leveraging large-scale language models for improved and generalized action generation. *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 31–36.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Solomon Kullback and Richard A Leibler. 1951. On information and regression. *The Annals of Mathematical Statistics*, 22(1):79–86.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pages 1–21.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*.

Guangtao Lyu, Chenghao Xu, Jiexi Yan, Muli Yang, and Cheng Deng. 2025. Towards unified human motion-language understanding via sparse interpretable characterization. In *The Thirteenth International Conference on Learning Representations*.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: archive of motion capture as surface shapes. In *ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5441–5450. IEEE.

André F. T. Martins and Ramón Fernandez Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. *Preprint*, arXiv:1602.02068.

Nicola Messina, Jan Sedmidubský, Fabrizio Falchi, and Tomás Rebok. 2023. Text-to-motion retrieval: Towards joint understanding of human motion data and natural language. In *ACM SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2420–2425. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Mathis Petrovich, Michael J. Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*.

Mathis Petrovich, Michael J. Black, and Gül Varol. 2023. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*.

Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT motion-language dataset. *Big Data*, 4(4):236–252.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. 2021. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. *Preprint*, arXiv:2010.04592.

Haoyu Shi and Huaiwen Zhang. 2024. Modal-enhanced semantic modeling for fine-grained 3d human motion retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10114–10123.

Haoyu Shi and Huaiwen Zhang. 2025. Sequence-event semantic consistent learning for text-to-motion retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8740–8749.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. *Preprint*, arXiv:2203.08063.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. Neural discrete representation learning. *Preprint*, arXiv:1711.00937.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Sheng Yan, Mengyuan Liu, Yong Wang, Yang Liu, Chen Chen, and Hong Liu. 2024. Mlp: Motion label prior for temporal sentence localization in untrimmed 3d human motions. *Preprint*, arXiv:2404.13657.

Sheng Yan, Yang Liu, Haoqiang Wang, Xin Du, Mengyuan Liu, and Hong Liu. 2023. Cross-modal retrieval for motion and text via droptriple loss. In *ACM Multimedia Asia 2023*, MMAsia '23. ACM.

Yang Yang, Liyuan Cao, Haoyu Shi, and Huaiwen Zhang. 2024a. Multi-instance multi-label learning for text-motion retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5829–5837.

Yang Yang, Haoyu Shi, and Huaiwen Zhang. 2024b. Hierarchical semantics alignment for 3d human motion retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1083–1092.

Yang Yang, Haoyu Shi, and Huaiwen Zhang. 2024c. Hierarchical semantics alignment for 3d human motion retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1083–1092. ACM.

Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. 2024. Exploring vision transformers for 3d human motion-language models with motion patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiahang Zhang, Lilang Lin, Shuai Yang, and Jiaying Liu. Sgar: Structural generative augmentation for 3d human motion retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Experimental Settings

A.1 Datasets

We conduct extensive experiments on three publicly available and widely used datasets for human motion-language understanding tasks, with detailed configurations as follows:

HumanML3D (Guo et al., 2022a) is the largest existing 3D human motion dataset with paired textual annotations. Its motion sequences are collected from two widely adopted motion-capture databases: AMASS (Mahmood et al., 2019) and HumanAct12 (Guo et al., 2020). Consistent with the standard experimental protocol in (Petrovich et al., 2023), we split the dataset into training, validation, and testing subsets, which include 23,384, 1,460, and 4,380 motion clips respectively. Each motion sequence is annotated with approximately three descriptive sentences of different lengths.

KIT Motion-Language(KIT-ML) (Plappert et al., 2016) contains 3,911 recordings of full-body motion paired with 6,278 text descriptions. Each motion sequence is described in one to four texts, with an average description length of approximately 8 words. Consistent with the benchmark setup, we utilize 4,888, 300, and 800 motion sequences for the training, validation, and test sets, respectively.

BABEL (Punnakkal et al., 2021) offers frame-level natural language annotations for the AMASS motion capture dataset (Mahmood et al., 2019). We employ all available "raw_label" textual descriptions (i.e., text queries) in our experiments, including the supplementary annotation files provided by the official release. Notably, we discard queries labeled as "transition", since such descriptions lack consistent and meaningful semantic cues for motion characterization. The original dataset is officially partitioned into 60% training, 20% validation, and 20% test subsets. However, the official test split remains private for benchmark challenge purposes. To conduct a fair evaluation, we consequently use 20% of the official validation set as our held-out test set for performance assessment.

HumanML3D(Restore) is a custom-adapted dataset derived from the original HumanML3D dataset (Guo et al., 2022a), specifically tailored for the task of text-to-motion grounding. The original HumanML3D (H3D) comprises motion segments extracted from the source motion sequences of the AMASS dataset (Mahmood et al., 2019), each paired with a corresponding natural language description. In our adapted dataset, these extracted segments are treated as target moments for the text-to-motion grounding task, while the original source motions (prior to segment extraction) are regarded as new input samples. This modification endows the target segments with explicit temporal context relative to their original motion sequences. To ensure the validity of the grounding task, we filter out samples where the relative length of the target moment (compared to its corresponding source motion) exceeds 80%. Subsequently, data augmentation is applied via left-right mirroring, which is performed on both the motion sequences and their associated text queries to enhance the generalization ability of the model. After undergoing these preprocessing steps, the HumanML3D (Restore) dataset exhibits a 68% difference from the source motions in the BABEL dataset. During the dataset construction process, we observed that a small subset of the restored HumanML3D samples contains source motions that are present in both the training and testing sets. To eliminate potential data leakage and avoid unfair evaluation (i.e., cheating), we re-partitioned the dataset following a strict splitting strategy to ensure the independence of training and testing subsets.

A.2 Baselines

We compare our method with state-of-the-art (SOTA) baselines across text-motion retrieval, motion grounding, and motion captioning tasks. The baselines are described as follows:

- **MotionCLIP** (Tevet et al., 2022): A transformer-based method that uses motion patches to model local motion dynamics and aligns with text via contrastive learning.
- **T2M** (Guo et al., 2022a): A transformer-based method that uses motion patches to model local motion dynamics and aligns with text via contrastive learning.
- **TEMOS** (Petrovich et al., 2022): A VAE-based generative model that learns a joint latent distribution for both text and motion modalities.
- **DTL** (Yan et al., 2023): Addresses the temporal misalignment between fine-grained textual descriptions and motion sequences.
- **MotionGPT** (Jiang et al., 2023): Treats human motion as a "foreign language" and uses

a GPT-style transformer to generate text descriptions from motion sequences.

- **TM2T** (Guo et al., 2022b): A stochastic tokenized model for reciprocal generation of 3D human motions and texts, using a transformer decoder for caption generation.
- **HSA** (Yang et al., 2024c): Uses motion pattern aggregation to capture high-level motion semantics and aligns with text via multi-scale contrastive learning.
- **TMR** (Petrovich et al., 2023): A representative contrastive learning-based method that constructs a cross-modal embedding space for text-motion retrieval with negative filtering.
- **MoPa** (Yu et al., 2024): Extends Vision Transformer (ViT) (Dosovitskiy et al., 2020) to build motion-language models by representing 3D human motion data as "motion patches."
- **CAR** (Fujiwara et al., 2024): Advances temporal alignment through action event decomposition, where textual descriptions are segmented into sequential sub-actions.
- **ToHL** (Lyu et al., 2025): Attempts to align motion and text modalities in a shared lexical vocabulary space for motion-language understanding.

A.3 Implementation Details

All experiments are conducted on a workstation with NVIDIA A100 GPUs by using the PyTorch-1.10 library. We set the batch size as 128 and employ the Adam optimizer (Kingma and Ba, 2017) to optimize the model. We set the temperature parameter τ to 0.07, following CLIP (Radford et al., 2021). The scaling ratio γ is set to 0.6, and the scaling factor η is set to 0.1. For all datasets, we set the maximum motion sequence length and training epochs to 200 and 300, respectively. The pose dimension d is 768, and the representation dimension D of motion and text feature is set to 256. The overall loss function is defined as the weighted sum of $\mathcal{L}_{\text{HN\_CLAP}}$ and $\mathcal{L}_{\text{ADC}}$, where the weight for $\mathcal{L}_{\text{HN\_CLAP}}$ is set to 0.8 and the weight for $\mathcal{L}_{\text{ADC}}$ is 0.2. Importantly, during training on the text-to-motion caption task, we incorporate our model with an autoregressive language decoder to output captions sequentially, and employ a loss function that maximizes the log-likelihood of the predicted token distribution. To maintain the model's efficiency, we do not utilize this language decoder for other tasks, even though it contributes to improving semantic capture.

Settings	Methods	BABEL Dataset				
		IoU@0.3 ↑	IoU@0.5 ↑	IoU@0.7 ↑	IoU@0.9 ↑	mIoU ↑
Normal	TMR	27.09	15.96	7.55	1.92	17.70
	DTL	27.33	15.03	7.69	1.53	17.01
	HSA	29.66	17.67	9.36	3.55	19.61
	ToHL	30.14	18.05	10.22	4.19	21.25
	CAR	30.95	18.48	11.65	4.68	22.01
	MSCode	**32.24**	**20.03**	**12.54**	**5.49**	**23.21**
Assigned	TMR	45.26	29.03	13.63	3.41	31.49
	DTL	41.59	26.18	10.51	2.07	30.41
	HSA	56.54	31.28	13.54	5.22	40.56
	ToHL	58.66	32.61	15.02	6.33	42.33
	CAR	60.21	33.31	16.55	7.19	44.26
	MSCode	**63.39**	**34.21**	**17.88**	**8.55**	**46.17**

Table 8: Results of text-to-motion grounding. We employ the temporal pyramid method (Gao et al., 2017), using a sliding window ranging from 20 to 200 frames with a stride of 5 frames to evaluate IoU.

B More Experimental Results

B.1 Results of Text-to-Motion Grounding on BABEL

We further validate our model's fine-grained understanding ability on the BABEL (Punnakkal et al., 2021) dataset. Notably, this dataset uses action labels (instead of natural language action descriptions) as annotations that focuses more directly on action semantics, making it a stricter test of the model's ability to capture precise action features. As shown in Tab. 8, our method outperforms all existing baselines across both the "Normal" and "Assigned" settings on the BABEL dataset: MSCode achieves superior results on all IoU metrics (IoU@0.3, IoU@0.5, IoU@0.7, IoU@0.9) and mIoU. For instance, under the "Assigned" setting, MSCode reaches 63.39 (IoU@0.3), 34.21 (IoU@0.5), and 46.17 (mIoU), outperforming the baseline (CAR) by 3.18, 0.90, and 1.91, respectively. This result further demonstrates the effectiveness of our model, particularly its capacity for fine-grained cross-modal alignment between poses and action labels.

B.2 Results of Motion Classification

Furthermore, we demonstrate the effectiveness of the semantic spaces generated by our method via motion recognition task. We follow the BABEL 60-class benchmark and pre-processed motion sequences with the same procedure as HumanML3D

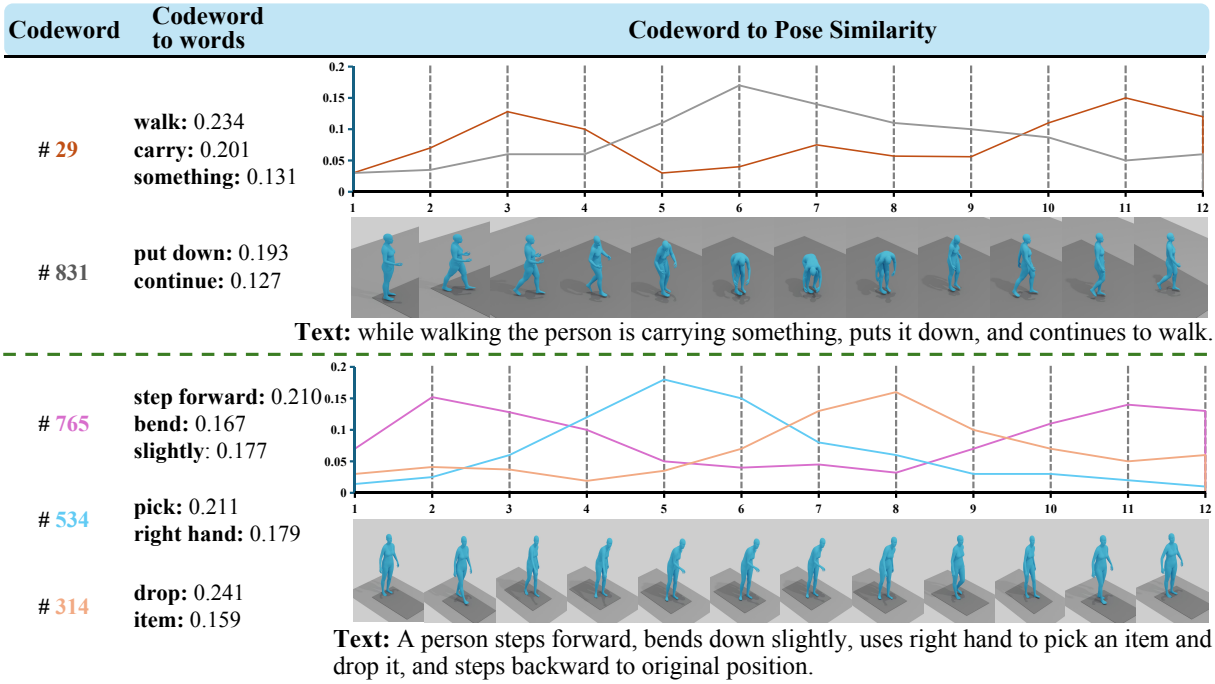

Figure 7: Visualization of codewords' role in connecting text modality and motion modality.

Methods	Training Dataset	Zero-shot	Modality	Top-1 Acc.	Top-5 Acc.
2s-AGCN	BABEL	×	M	41.14	73.18
MotionCLIP	BABEL	×	M+L	40.90	57.71
TMR	HumanML3D	✓	M+L	30.13	41.52
MoPa	HumanML3D	✓	M+L	41.33	68.97
CAR	HumanML3D	✓	M+L	42.83	70.97
MSCode	HumanML3D	✓	M+L	**45.69**	**74.11**

Table 9: Results of zero-shot motion classification. Modality with motion only and motion language are denoted as M and M+L, respectively.

Methods	Params	FLOPs	Inference time	R@1
TMR	49.54M	7.47G	40.858s	5.68
CAR	128.70M	12.64G	57.136s	6.55
ToHL	209.70M	36.62G	75.119s	6.78
MSCode(ours)	95.90M	10.90G	52.986s	7.69

Table 10: **Ablation analysis on model efficiency.** We report the result at text-to-motion retrieval on the HumanML3D datasets. The inference time is for all samples in the test set.

(Guo et al., 2022a). Because this is a zero-shot classification setting, we did not train the model with BABEL and only applied the model trained on HumanML3D to the test data. For the text prompts, the action names in BABEL are formatted as "A person {action}". The cosine distance is calculated between a given motion and all 60 text prompts. As shown in Tab. 9, the results indicate that our framework performs comparably to state-of-the-art supervised methods (Shi et al., 2019; Tevet et al., 2022), thereby demonstrating the well-aligned semantic space obtained by our method.

B.3 Model efficiency

As shown in Tab. 10, we conduct a comprehensive comparison of computational cost and retrieval performance on the HumanML3D dataset. While the baseline method TMR (Petrovich et al., 2023) maintains the lowest computational overhead, our method achieves a substantial leap in retrieval accuracy (improving R@1 from 5.68 to 7.69) with only a marginal increase in complexity. More notably, compared to the state-of-the-art method ToHL, which relies on heavy vocabulary-based alignment, our approach demonstrates superior efficacy in all aspects: it not only outperforms ToHL in retrieval accuracy (7.69 vs. 6.78) but also significantly reduces the computational cost, cutting the parameter count by approximately 54.27% (209.70M → 95.90M) and inference time by 29.46%. This confirms that our method achieves an optimal balance between high performance and model efficiency.

C More Visualization Results

C.1 Visualization of the Learned Codewords

As shown in Fig. 7, we present additional examples of how codewords bridge motion and language, further demonstrating the cross-modal alignment capability of our modality-shared codebook.

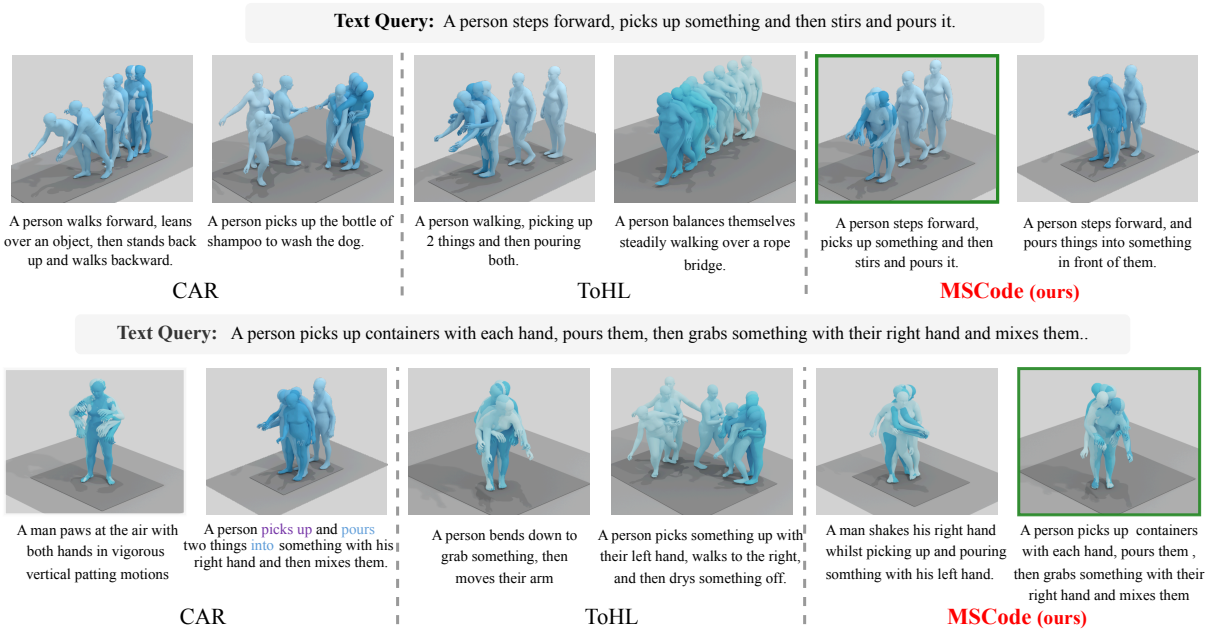

Figure 8: Visualization of the retrieval results. We present the top-2 motion sequences retrieved by our proposed method MESM, as well as by CAR (Fujiwara et al., 2024) and ToHL (Lyu et al., 2025). The text below each motion sequence represents the ground truth. The successful retrieval results are highlighted by the green border.

In the first case, codeword #29 aligns with the semantic tokens "walk" (similarity: 0.234) and "carry" (0.201), and its similarity curve peaks in the early motion segments (frames 1–4), exactly where the walking/carrying actions occur in the pose snapshots. Codeword #831, paired with "put down" (0.193) and "continue" (0.127), shows a similarity peak in frames 5–8, corresponding to the put down and continue to walk motion segments. For the lower case, codeword #765 links to "step forward" (0.210) and "bend" (0.167), aligning with the initial stepping/bending frames (2–3); codeword #534 (associated with "pick" (0.211) and "right hand" (0.179)) peaks in frames 4–6 (the picking motion); codeword #314 (paired with "drop" (0.241)) aligns with frames 7–8 (the dropping action). These examples directly verify that each codeword simultaneously maps to text tokens and the corresponding temporal motion segments, demonstrating the codebook's core ability to unify motion and language semantics into a shared space, rather than learning isolated patterns for each modality.

C.2 Visualization of the Retrieval Results

In Fig. 8, we provide additional qualitative comparisons of the text-to-motion retrieval results on the HumanML3D dataset. We compare our proposed MSCode with two baseline approaches: CAR (Fujiwara et al., 2024) and ToHL (Lyu et al., 2025). Correctly retrieved ground-truth sequences are highlighted with a green border. In the first example, the text query specifies: "A person steps forward, picks up something and then stirs and pours it." This description requires capturing sequential, coordinated actions (stepping, picking, stirring, pouring) with fine-grained body motion details. As shown in the visualization, our MSCode successfully retrieves the ground-truth sequence (marked by the green border), the corresponding motion explicitly reflects "stepping forward, picking up an object, stirring, and pouring" in alignment with the text.In contrast, CAR retrieves a sequence of "walking forward, leaning over an object" (missing the core "stir/pour" actions), while ToHL returns a generic "walking, picking up 2 things and pouring" (lacks the specific "stir" step). The second example uses the query: "A person picks up containers with each hand, pours them, then grabs something with their right hand and mixes them." This tests the model's ability to distinguish multi-step, hand-specific actions (picking, pouring, right-hand grabbing/mixing). Here, baseline methods fail to retrieve the ground truth: CAR returns a sequence of "pawing at the air" (semantically irrelevant motion), while ToHL retrieves a sequence of "bending down to grab something" (missing the mixing details). Conversely, our MSCode accurately retrieves the target motion (marked by the green border), where

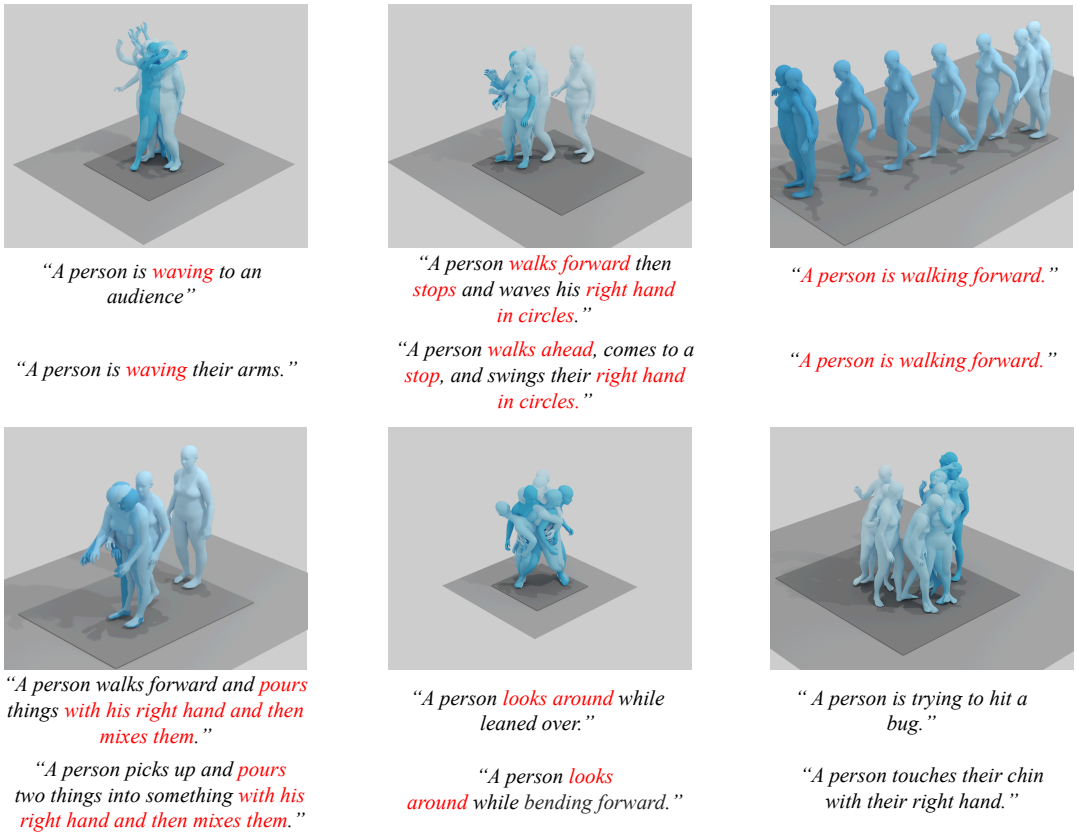

Figure 9: Qualitative results of motion-to-text captioning on HumanML3D test set. The red words highlight the keywords.

the sequence explicitly includes "picking up containers with both hands, pouring, and right-hand mixing." This result highlights the strength of our approach: our alignment strategy enables the model to parse subtle semantic cues (e.g., "each hand," "right hand") and map them to corresponding fine-grained motions capabilities.

C.3 Visualization of Motion Captioning

We present the results of motion captioning in Fig. 9. The captions generated by our method exhibit greater distinction, capturing more detailed and relevant keywords. By comparing the ground-truth (GT) captions with the outputs of our method (denoted as "Ours"), we observe that our generated captions preserve core motion semantics while exhibiting finer granularity: for example, in the case of the "waving to an audience" motion, our model explicitly describes the action as "waving their arms"; for the "walking-forward-and-pouring" sequence, our output expands to "picks up and pours two things into something with his right hand and then mixes them", capturing both the tool (right hand) and the sequential logic (pour then mix). The highlighted keywords (in red) further reflect that our method effectively identifies and retains critical motion details, which are aligned with the key elements of the original motion. These results collectively demonstrate that our approach can generate accurate, detailed, and semantically consistent motion captions, thus verifying the effectiveness of our proposed framework.

D Applications

D.1 Cross-Dataset Motion Retrieval.

In this section, we implement the cross-dataset retrieval, i.e., training the model on HumanML3D first and then testing on Motion-X (Lin et al., 2023) test set, which covers various domains and noise as a more comprehensive benchmark. Specifically, we retrieve the counterparts from a batch following Small protocol (Petrovich et al., 2023). The results are shown in Table 11. As we can see, our method demonstrates a remarkable improvement over previous works, verifying the stronger generalization capacity.

Methods	Text-to-motion retrieval				Motion-to-text retrieval			
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
TMR	20.23	46.18	60.95	8.38	20.21	43.82	59.05	9.24
MoPa	24.00	48.81	64.32	7.50	22.68	46.45	61.74	8.25
MSCode(ours)	**31.05**	**54.77**	**69.22**	**6.81**	**30.13**	**52.92**	**66.77**	**7.46**

Table 11: Cross-dataset motion retrieval results. All models are trained on HumanML3D dataset.

Methods	Text-to-motion retrieval			Motion-to-text retrieval		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
TMR	5.38	15.64	24.40	5.13	15.26	25.65
MoPa	9.51	21.27	32.41	8.26	22.65	32.66
MSCode (ours)	**11.94**	**24.37**	**35.48**	**10.76**	**24.19**	**35.11**

Table 12: Results of human interaction recognition. For all methods, we concatenate the motion features of each person and get the multi-person motion feature through a projection head.

D.2 Human Interaction Recognition

To better verify the representation ability of our model, we carry out experiments on the multi-person interaction recognition task with the Inter-Human Dataset (Liang et al., 2024).This dataset provides a complete set of two-person interaction sequences, including 6,222 training samples and 1,557 test samples. In our feature extraction pipeline, we adopt a shared encoder structure to encode the motion of each person separately, then concatenate the learned features and send them to a projection layer for further processing. As listed in Table 12, our approach achieves better performance than baseline methods that only use global features for human interaction recognition.The performance gain mainly comes from handling complex interactive motions in the dataset, such as the scenario where “two people stand near each other; one lifts arms to hold the other, while the latter pushes back”.Our method can capture these fine-grained, multi-action interactions more accurately and complete the recognition task effectively.