EVENT-AWARE PROTOTYPE MATCHING FOR 3D HUMAN MOTION RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-motion retrieval is a challenging task that aims to search relevant motion sequences based on natural language descriptions. The typical solution is to directly align the global sequence-level and sentence-level features for text-motion matching during learning. However, motion sequences often involve multiple action events, which are difficult to capture effectively with just a single embedding. In addition, different motion sequences may share similar actions, which leads to high semantic overlap between unpaired samples and thereby introduces semantic ambiguity during alignment. To this end, we propose an Event-Aware Prototype Matching (EAPM) approach, which models text-motion correspondence as an action events matching procedure. Specifically, to decompose events from motion sequence and text, we design an event-aware prototype aggregation method that adaptively aggregates local features into a set of action event prototypes. We further design a variance loss to encourage these prototypes to focus on distinct contents, thereby effectively capturing diverse action semantics. Finally, we present a prototype-based event soft-alignment approach to modulate confidence estimation for events via intra-modal self-matching probability, which enables more fine-grained text-motion alignment at the event level. Extensive experiments show that our method significantly outperforms existing methods in text-to-motion retrieval and other challenging tasks, such as human interaction recognition and motion temporal localization.

1 INTRODUCTION

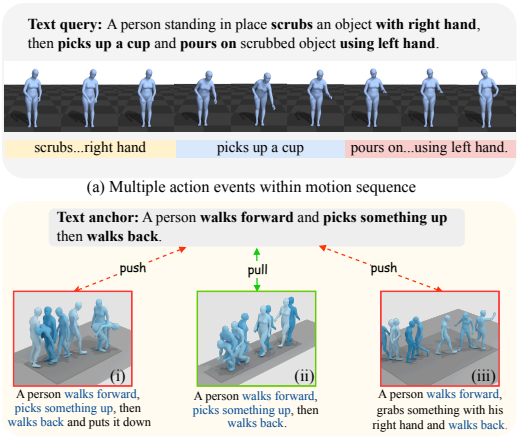
With the rapid development of motion generation tools and methods (Zhou et al., 2025; Karunratanakul et al., 2023; Tevet et al., 2023; Kapon et al., 2024), millions of 3D motion data advent to the world. As a result, efficiently retrieving specific motion sequences from this vast repository has become an increasingly critical need. Text-motion retrieval (TMR) (Petrovich et al., 2023; Yang et al., 2024; Yan et al., 2023; Petrovich et al., 2022; Guo et al., 2022) is a typical multimodal retrieval task that aims to retrieve semantically similar motion sequences based on a given natural language description. It has wide-ranging applications in fields such as animation production, human-computer interaction, and virtual reality, where precise motion retrieval is crucial for enhancing user experience and system performance.

Following the paradigms of conventional multimodal retrieval, *e.g.*, text-image (Ma et al., 2022) and text-video retrieval (Arandjelović et al., 2016), several studies have introduced text-motion retrieval (TMR) (Petrovich et al., 2023; Messina et al., 2023; Yu et al., 2024; Fujiwara et al., 2024; Yang et al., 2024). TMR (Petrovich et al., 2023) introduces the task of text-motion retrieval, establishes a series of evaluation benchmarks with varying difficulty levels, and introduces a joint synthesis and retrieval framework. MoPa (Yu et al., 2024) proposes to represent 3D human motion data as "motion patches" and extend Vision Transformers (ViT) (Kong et al., 2022) to establish a cross-modal latent space between motion and language. CAR (Fujiwara et al., 2024) introduces the concept of action events by decomposing textual descriptions into sequential action events. They propose a simple but effective solution to improve the temporal alignment between language and motion, using shuffled event descriptions as negative text samples to refine the motion-language model.

Although existing works (Yu et al., 2024; Petrovich et al., 2023; Yang et al., 2024; Yan et al., 2023; Yin et al., 2024; Petrovich et al., 2022; Guo et al., 2022) in text-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to-motion retrieval have achieved promising retrieval performance, they typically represent both query text and motion sequences as a single embedding for cross-modal matching. However, in TMR scenes, motion sequences often comprise multiple action events (Fujiwara et al., 2024) with unique action semantics, making it challenging for a single embedding to fully capture the complex semantics of these action events. As illustrated in Fig. 1a, the text-motion pair involves three successive action events, *i.e.*, "scrub...with right hand", "pick up", and "pours on...using left hand". These action events occur sequentially, with only subtle variations in hand movements. Existing methods often fail to accurately model the intricate semantics of these action events, owing to their reliance on simple representation approaches.



(a) Multiple action events within motion sequence

(b) Overlapping actions across different sequences

Figure 1: The illustration of our motivations. In subfigure (a), the sequence comprises multiple events with intricate semantics, exhibiting subtle variations. In subfigure (b), all three motion sequences exhibit significant semantic overlap, as highlighted by color. However, existing methods completely push away the text and the motion sequence (i)/(iii).

In addition, overlapping actions often occur across different motion sequences, leading to significant semantic correlations between unpaired samples. As evidenced in Fig. 1b, given a text anchor, the motion sequence (ii) is the only paired sample in the batch. However, the motion sequences (i) and (iii) exhibit a high overlap in action with sequence (ii), as all contain the actions "walk forward" and "walk back". Existing methods ignore this and rely on simple global contrastive learning, which completely pushes away the text with motion sequences (i) and (iii). This may introduce semantic ambiguity during the alignment process, ultimately impairing the model's representational capacity.

Based on our observations, we find that more than 92% of the motion sequences in the HumanML3D benchmark dataset (Guo et al., 2022) contain multiple action events, as also introduced in CAR (Fujiwara et al., 2024). Based on this, we argue that modeling motion-text correspondence as an action event (decoupled sequential actions) matching procedure is essential. In this paper, we propose an Event-Aware Prototype Matching (EAPM) framework for 3D human motion retrieval. Firstly, to decompose events from both motion and text modalities, we introduce an event-aware prototype aggregation method that adaptively aggregates word and pose features into a set of event prototypes. We further introduce a variance loss to encourage these prototypes to focus on distinct actions, thereby capturing rich action semantics. After that, we present a prototype-based event soft-alignment method that modulates confidence estimation for events via intra-modal self-matching probability during the alignment process and achieves more fine-grained text-motion alignment at the event level. As a result, this method effectively alleviates semantic ambiguity and enhances the robustness of the model. Additionally, a global alignment is also introduced to enable more comprehensive semantic matching, which also provides additional supervision for the learning of event prototypes.

Our contributions are summarized as follows:

- We propose a novel Event-Aware Prototype Matching (EAPM) method for 3D human motion retrieval, which models text-motion correspondence as an action event matching procedure.
- We introduce an event-aware prototype aggregation method that adaptively aggregates word and motion features into multiple event prototypes. A variance loss is incorporated to encourage these prototypes to focus on distinct content, capturing diverse action semantics.
- We propose a prototype-based event soft-alignment to modulate confidence estimation for events via intra-modal self-matching probability, which enables more fine-grained alignment at the event level and alleviates semantic ambiguity during learning.

- Our method significantly improves text-to-motion retrieval performance and demonstrates considerable potential for other challenging applications, such as human interaction recognition and motion temporal localization.

2 RELATED WORK

3D Human Motion Retrieval. In recent years, 3D human motion retrieval (Petrovich et al., 2023; Yu et al., 2024; Yin et al., 2024; Yang et al., 2024; Li et al., 2025) has become increasingly popular, with the aim of retrieving semantically relevant motion sequences based on a given natural language. In general, existing TMR methods (Petrovich et al., 2023; 2022; Yu et al., 2024; Yin et al., 2024; Liu et al., 2020) follow text-to-image retrieval, aggregating query and motion sequences to learn global embeddings in common space. Yu *et al.* (Yu et al., 2024) extend Vision Transformer (ViT) (Dosovitskiy et al., 2021) to build motion-language models by representing 3D human motion data as "motion patches." Fujiwara *et al.* (Fujiwara et al., 2024) propose a framework to achieve a better temporal correspondence between language and motion. However, existing methods primarily focus on aligning the global embeddings, which is inadequate to achieve precise semantic matching between intricate motion sequences and their corresponding text.

Cross-Modal Semantic Alignment. Cross-modal alignment (Edstedt et al., 2024; Li et al., 2022a; Wang et al., 2021; 2023; Chun et al., 2021; Zhang et al., 2023b;a; Pham et al., 2024; Lee et al., 2018) is a significant task in multi-modality scenarios, aiming to bring instances from different modalities with similar semantics closer in a common space for precise cross-modal matching. Vision-language alignment, in particular, has proven highly effective in aligning instances across multiple modalities. Some common approaches Fu et al. (2021); Li et al. (2022b); Ma et al. (2022) are to embed cross-modal data into a shared semantic space, using cosine similarity to facilitate cross-modal interactions for global alignment. Other approaches (Huang et al., 2024; Faghri et al., 2018) utilize uni-modal pre-trained models to generate soft-label supervision signals, enhancing cross-modal robustness alignment. Additionally, other methods (Li et al., 2024; 2023; 2022c) incorporate external knowledge from pre-trained models, bringing supplementary insights that strengthen cross-modal alignment. However, the semantic ambiguity of the overlapping actions across motion sequences poses challenges for directly applying these methods to 3D motion scenarios.

3 METHOD

In this section, we present Event-Aware Prototype Matching (EAPM), a motion-language model for 3D human motion retrieval, which enables more comprehensive text-motion matching from both sequence and event perspectives. We first introduce the Multi-modal Feature Encoder in Sec. 3.1. Then, we introduce the Event-Aware Prototype Aggregation in Sec. 3.2 and the Prototype-based Event Soft-Alignment in Sec. 3.3. Lastly, we describe the training and inference procedures in Sec. 3.4. The details are shown in Fig. 2.

3.1 MULTI-MODAL FEATURE ENCODER

The motion sequence comprises a series of 3D human poses $\boldsymbol{m} = (p_1, p_2, \ldots, p_N), p \in \mathbb{R}^{d_p}$, where N represents the length of a motion sequence and d_p is the feature dimension. The text description in our model is a sequence of words $\boldsymbol{t} = (t_1, t_2, \ldots, t_L)$, where L is the number of words.

Spatial-Temporal Motion Encoder. Firstly, each pose p is transformed to $j = (j_1, j_2, \ldots j_J)$, J is the joint number. The human body can be denoted as a graph structure, with $G = \{\text{Nodes}, \text{Edges}\}$. The joint $\{j_i\}_{i=1}^{J}$ represent nodes and their correlations are defined by an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{J \times J}$, with elements set to 1 if joints are connected, and 0 otherwise. We employ a multi-layer GCN (Kipf & Welling, 2017) to effectively encode the spatial relationships among body joints. The graph convolutional layer l takes the matrix $\boldsymbol{H}^{(l)} \in \mathbb{R}^{J \times |H|}$ as input, with $|H|$ being feature dimension. The output of l-th layer is computed by :

$$\boldsymbol{H}^{(l+1)} = \mathrm{BN}\left(\tanh\left(\boldsymbol{A}^{(l)}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right)\right) \tag{1}$$

where BN denotes batch normalization (Ioffe & Szegedy, 2015). Subsequently, we employ mean pooling to aggregate spatial-aware motion feature, $\boldsymbol{m}^i = \mathrm{MeanPool}(\boldsymbol{H}_1^{(L)}, \ldots, \boldsymbol{H}_N^{(L)})$, L is the lth

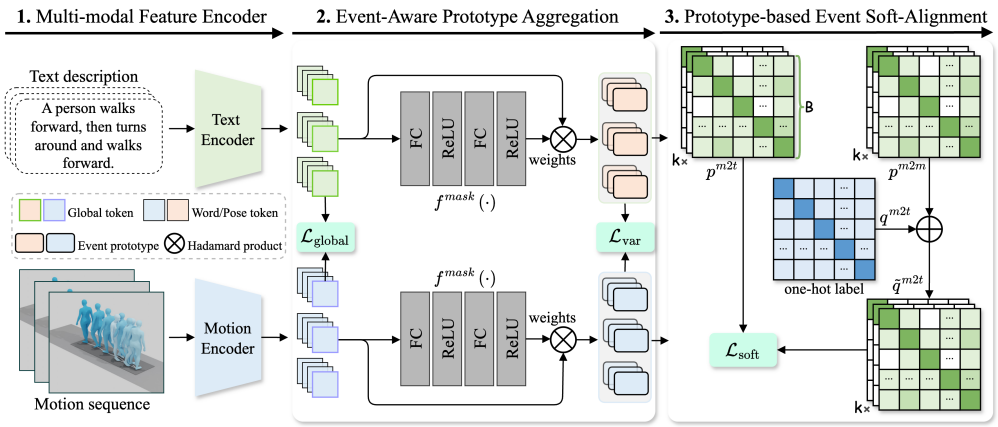


Figure 2: Overall architecture of EAPM. Specifically, we obtain text and motion embeddings from their respective encoders and perform global alignment on the global embeddings. Then, we design an event-aware prototype aggregation method that adaptively aggregates local features into a set of event prototypes. Finally, we introduce a prototype-based event soft alignment to modulate confidence estimation for events via intra-modal self-similarity, which alleviates semantic ambiguity and achieves fine-grained alignment at the event level.

GCN layer. These pooled features are then fed into a Transformer encoder (Vaswani et al., 2023) to capture temporal dependencies.

$$\boldsymbol{m}^s = \text{TransEncoder}(\text{Norm}(\boldsymbol{m}^i) + \textbf{pos}) \tag{2}$$

Here, the motion embeddings $\boldsymbol{m}^s \in \mathbb{R}^{F \times d_r}$, Norm is the normalized operator, and **pos** is the position embeddings.

Linguistic Encoder. For a text query $\boldsymbol{t}$, we adopt a pre-trained frozen DistillBERT (Devlin et al., 2019) to extract the text features $\boldsymbol{t}^i$. We also employ a temporal encoder to better capture linguistic temporal relationships, obtaining the final text embeddings $\boldsymbol{t}^s \in \mathbb{R}^{L \times \mathrm{d}_r}$, where L represents the length of the text embeddings. This linguistic encoder has the same structure as the motion temporal encoder.

Inspired by the leverage of class tokens in BERT (Devlin et al., 2019) and ViT (Kong et al., 2022), the [CLS] token is prepended to both the word tokens and the pose tokens to obtain global representations $\boldsymbol{t}^g$ and $\boldsymbol{m}^g$.

3.2 Event-Aware Prototype Aggregation

After extracting the motion features and text features, most existing methods only calculate the similarity between the global representation $\boldsymbol{m}^g$ and $\boldsymbol{t}^g$. However, such similarity is unreliable for text-motion matching due to the multiple action events present in the motion sequence, which are difficult to capture by a single embedding. Relying solely on global matching may result in poor retrieval performance. To address this, we propose an event-aware prototype aggregation method, which adaptively captures action events from both motion sequences and text, enabling more fine-grained text-motion matching at the action event level.

Event Prototype Aggregator. For text and motion sequence, we need to obtain K event prototype pairs for fine-grained event matching. There are several approaches for generating multiple visual prototypes, such as transformer decoders, graph neural networks, LSTMs, *etc*. However, unlike images or videos, 3D motion data is focused on human-centric themes and lacks the irrelevant background or object elements that may complicate prototype generation. We find that a mask-based approach for generating event prototypes is both effective and computationally efficient. Specifically, for prototype aggregation at motion level, the k-th event prototype is computed as:

$$\boldsymbol{m}_k^o = \sum_{j=1}^{N} \boldsymbol{m}_j^s \cdot f^{mask}\left(\boldsymbol{m}_j^s\right)[k] \tag{3}$$

where $\boldsymbol{m}_j^s$ represents the j-th token, and $f^{\text{mask}}(\cdot)$ is a mask-generating function, which includes two Fully Connected (FC) layers and ReLU functions used to predict sparse mask weights. $f^{mask}(\boldsymbol{m}_j^s)[k]$ represents the contribution of the j-th token feature to the k-th event prototype. Consequently, each event prototype $\boldsymbol{m}_k^o$ is computed as a weighted sum of all token features, with the weights determined by the mask-generating function. Then, a projection linear layer followed by normalization is applied to obtain the final event prototype pairs.

$$\boldsymbol{m}^p = \text{Norm}(\text{LN}(\boldsymbol{m}^o)), \boldsymbol{t}^p = \text{Norm}(\text{LN}(\boldsymbol{t}^o)) \tag{4}$$

Variance Loss. Define mask $\boldsymbol{w}_j^k = f^{mask}(\boldsymbol{m}_j^s)[k]$, which represents the mask weight on j-th token for k-th event prototype. Considering that different prototypes may capture the redundant and overlapping semantic information to each other. We'd like all event prototypes to pay attention to different token positions. Inspired by (Bardes et al., 2022), we introduce a variance loss on mask values to ensure that event prototypes are diverse and focus on distinct action events.

$$\mathcal{L}_{var} = \frac{1}{N}\sum_{j=1}^{N} \max(0, \gamma - D(\boldsymbol{w}_j, \epsilon)) \tag{5}$$

where γ is a constant target value for standard deviation; $D(\cdot)$ is the regularized standard deviation, defined as:

$$D(x, \epsilon) = \sqrt{\text{Var}(x) + \epsilon} \tag{6}$$

where ϵ is a small scalar to prevent numerical instabilities. The variance loss forces the variance of all prototype mask values for each token to be large. This ensures that each event prototype captures unique, non-overlapping information, thereby improving the event representation of both motion sequence and text.

3.3 Prototype-based Event Soft-Alignment

In TMR scenes, different motion-text pairs contain similar actions, resulting in event prototypes aggregated by the EAP module with high semantic overlap. This will unavoidably introduce semantic ambiguity during alignment, remarkably degrading the performance. For the i-th prototype pair, its calculation is:

$$\tilde{q}_i = (1-\alpha)q_i + \frac{\alpha}{B-1}(\mathbb{1} - q_i) \tag{7}$$

where $\mathbb{1}$ is the all-ones vector, q_i is the real matching probability, and α controls smoothing intensity. Despite the potential benefits of label smoothing in providing confidence to events with similarities, it is limited in its ability to distinguish similar action events only with weak and fixed confidence. To address this limitation and further enhance the alignment process, we propose leveraging the intra-modal self-similarity to modulate the confidence estimation and refine the confidence assigned to individual event prototypes. Specifically, we assign higher certainty to events that align well within the modality, while reducing the confidence for events that deviate from this intramodality consistency. This approach recognizes that the event features within the same modality should demonstrate a certain degree of similarity.

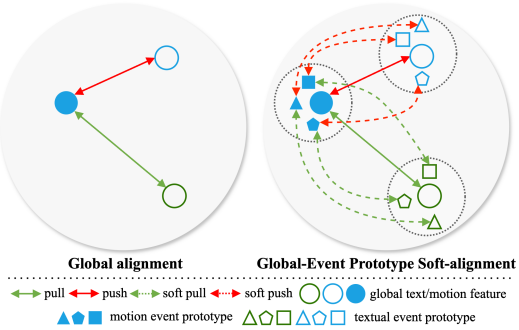


Figure 3: The illustration of our alignment approach. The motion sequence consists of multiple action events, and relying solely on global alignment is insufficient. Our approach performs both global and event alignment.

Event Soft-Alignment. In this subsection, we introduce our event soft-alignment, as illustrated in Fig. 3. Specifically, for the i-th motion–text pair (M_i, T_i), we can obtain K event prototype pairs $\{(\boldsymbol{m}_{i,k}^p, \boldsymbol{t}_{i,k}^p)\}_{k=1}^K$. Then intra-modal event self-matching probability vectors of the k-th event pair, denoted as $p_{i,k}^{m2m} = \{p_{i,j,k}^{m2m}\}_{j=1}^B$ and $p_{i,k}^{t2t} = \{p_{i,j,k}^{t2t}\}_{j=1}^B$, can be calculated as:

$$p_{i,j,k}^{m2m} = \frac{\exp\left(\tilde{\text{s}}(\boldsymbol{m}_{i,k}^p, \boldsymbol{m}_{j,k}^p)/\tau\right)}{\sum_{j=1}^B \exp\left(\tilde{\text{s}}(\boldsymbol{m}_{l,k}^p, \boldsymbol{m}_{j,k}^p)/\tau\right)}, \quad p_{i,j,k}^{t2t} = \frac{\exp\left(\tilde{\text{s}}(\boldsymbol{t}_{i,k}^p, \boldsymbol{t}_{j,k}^p)/\tau\right)}{\sum_{j=1}^B \exp\left(\tilde{\text{s}}(\boldsymbol{t}_{l,k}^p, \boldsymbol{t}_{j,k}^p)/\tau\right)} \tag{8}$$

Methods	Text-to-Motion						Motion-to-Text						Rsum ↑
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	
TEMOS (Petrovich et al., 2022)	2.12	4.09	5.87	8.26	13.52	173.00	3.86	4.54	6.94	9.38	14.00	183.25	72.58
T2M (Guo et al., 2022)	1.80	3.42	4.79	7.12	12.47	81.00	2.92	3.74	6.00	8.36	12.95	81.50	63.57
DTL (Yan et al., 2023)	2.30	4.63	6.66	10.06	16.40	76.00	2.64	4.91	7.34	10.95	17.21	76.00	83.10
HSA (Yang et al., 2024)	7.30	12.66	16.51	22.88	33.80	23.00	10.13	13.46	19.32	25.46	35.45	22.50	196.97
TMR (Petrovich et al., 2023)	8.92	12.04	16.33	22.06	33.37	25.00	9.44	11.84	16.90	22.92	32.21	26.00	229.97
CAR (Fujiwara et al., 2024)	10.10	13.50	18.96	26.42	37.28	19.50	11.18	13.85	20.19	26.65	36.37	20.50	214.55
MoPa (Yu et al., 2024)	10.80	14.98	20.00	26.72	38.02	19.00	11.25	13.86	19.98	26.86	37.40	20.50	219.81
LaMP* (*Liet al.*, 2025)	3.65	7.56	10.90	16.56	26.26	30.00	4.51	8.71	11.76	17.34	27.84	30.00	135.09
EAPM	**11.98**	**16.96**	**23.25**	**30.69**	**42.70**	**15.00**	**13.01**	**16.00**	**21.86**	**30.21**	**40.51**	**17.00**	**247.17**
EAPM‡	**12.41**	**17.05**	**23.37**	**31.20**	**42.99**	**14.00**	**13.35**	**16.82**	**22.91**	**31.13**	**42.04**	**16.00**	**253.26**

Table 1: Performance comparison with the state-of-the-art methods on HumanML3D (Guo et al., 2022).* denotes reproduced using official checkpoints. ‡ denotes using inverted dual softmax (Cheng et al., 2021) or QB-Norm (Bogolin et al., 2022) for post-processing.

Methods	Text-to-Motion						Motion-to-Text						Rsum ↑
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	
TEMOS (Petrovich et al., 2022)	7.11	13.25	17.59	24.10	35.66	24.00	11.69	15.30	20.12	26.63	36.39	26.50	207.84
T2M (Guo et al., 2022)	3.37	6.99	10.84	16.87	27.71	28.00	4.94	6.51	10.72	16.14	25.30	28.50	129.39
DTL (Yan et al., 2023)	6.77	13.28	16.67	23.18	37.24	18.00	9.11	14.32	20.31	25.26	38.02	18.00	204.16
HSA (Yang et al., 2024)	9.54	15.78	20.23	27.99	41.48	17.00	11.70	14.63	22.14	29.01	41.09	16.50	233.59
TMR (Petrovich et al., 2023)	10.05	13.87	20.74	30.03	44.66	14.00	11.83	13.74	22.14	29.39	38.55	16.00	221.80
CAR (Fujiwara et al., 2024)	13.87	19.63	25.78	37.22	53.01	11.00	13.01	14.69	24.93	32.53	46.74	14.00	281.41
MoPa (Yu et al., 2024)	14.02	21.08	28.91	34.10	50.00	10.50	13.61	17.26	27.54	33.33	44.77	13.00	287.20
LaMP* (*Liet al.*, 2025)	6.25	12.50	17.76	24.86	42.47	13.00	6.39	14.49	18.89	27.56	44.03	13.00	215.20
EAPM	**16.14**	**23.01**	**31.45**	**42.53**	**54.70**	**8.00**	**15.06**	**18.19**	**28.55**	**35.30**	**48.80**	**11.00**	**313.73**
EAPM‡	**16.27**	**23.25**	**32.53**	**42.29**	**56.87**	**7.50**	**15.15**	**17.95**	**28.74**	**36.39**	**49.40**	**11.00**	**318.60**

Table 2: Performance comparison with the state-of-the-art methods on KIT-ML (Plappert et al., 2016).* denotes reproduced using official checkpoints. ‡ denotes using inverted dual softmax (Cheng et al., 2021) or QB-Norm (Bogolin et al., 2022) for post-processing.

where $\tilde{\mathrm{s}}(\cdot,\cdot)$ indicates that if the cosine similarity is greater than β, it is preserved; otherwise, it is set to 0. We then leverage event self-matching probability vectors to modulate the confidence estimation assigned to each event prototype, thereby supervising motion-to-text and text-to-motion correspondences. The formulations are as follows:

$$\tilde{q}_{i,k}^{m2t} = (1-\alpha)q_i^{m2t} + \frac{\alpha}{B-1}p_{i,k}^{m2m}, \quad \tilde{q}_{i,k}^{t2m} = (1-\alpha)q_i^{t2m} + \frac{\alpha}{B-1}p_{i,k}^{t2t} \tag{9}$$

where $q_i^{m2t} = \{q_{i,j}^{m2t}\}_{j=1}^{B}$ and $q_i^{t2m} = \{q_{i,j}^{t2m}\}_{j=1}^{B}$ are the probability of real event matching for the motion-text pair i. Next, for the k-th motion–text event pair, the event soft-alignment loss is calculated as:

$$\mathcal{L}_{\text{soft}}^{k} = \frac{1}{B}\sum_{i=1}^{B}\left(H\left(\tilde{q}_{i,k}^{m2t}, p_{i,k}^{m2t}\right) + H\left(\tilde{q}_{i,k}^{t2m}, p_{i,k}^{t2m}\right)\right) \tag{10}$$

where $H(\cdot,\cdot)$ denotes the cross-entropy function. We can obtain the final soft alignment loss of the cross-modal event in pairs of event prototypes K by: $\mathcal{L}_{\text{soft}} = \sum_{k=1}^{K}\mathcal{L}_{\text{soft}}^{k}$.

3.4 Training and inference

The global cross-modal correspondence is more comprehensive and complementary to the action event perspective. Therefore, the commonly used Info-NCE loss (Radford et al., 2021) is utilized to calculate the global alignment loss $\mathcal{L}_{\text{global}}$. During the training phase, the model is trained in an end-to-end manner, and the overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{soft}} + \beta\mathcal{L}_{\text{var}} \tag{11}$$

where β is the weight assigned to the variance loss. In the inference phase, the similarity between query text T and motion sequence M is computed by:

$$\mathrm{S}(T,M) = \cos(\boldsymbol{m}^g, \boldsymbol{t}^g) + \frac{1}{K}\sum_{k=1}^{K}\cos(\boldsymbol{m}_k^p, \boldsymbol{t}_k^p) \tag{12}$$

which considers both global and event-level correspondence for a more comprehensive text-motion matching.

Methods	K	Variance Loss	Text-to-motion				Time(s) ↓
			R@1 ↑	R@5 ↑	R@10 ↑	Rsum ↑	
Baseline	0	✗	9.04	22.16	35.85	67.05	34.81
w/ Part	3	✗	9.12	22.98	35.70	67.80	36.12
w/ Cluster	3	✗	10.32	27.74	39.12	77.18	51.38
w/ Decoder	3	✗	11.24	28.17	40.87	80.28	47.35
EAPM (ours)	3	✗	11.31	29.52	40.94	81.77	37.92
	3	✓	**11.98**	**30.69**	**42.70**	**85.37**	38.21
	1	✓	10.64	28.36	39.35	78.35	**37.15**
	2	✓	11.75	29.94	42.18	83.87	38.01
	4	✓	10.25	25.62	38.15	74.02	38.93
	10	✓	9.11	23.54	37.19	69.84	41.36
	#tokens	✓	8.44	21.02	34.29	63.75	45.54

Table 3: In-depth analysis of event prototype generation. We report the result at the text-to-motion retrieval on HumanML3D.

Methods	Text-to-motion			Motion-to-text		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
Baseline	9.04	22.16	35.85	10.02	23.04	33.14
w/ EAP	10.45	25.17	37.15	11.32	26.90	37.07
w/ PES	10.96	26.65	27.91	12.54	27.01	38.23
w/ EAP, PES	**11.98**	**30.69**	**42.70**	**13.01**	**30.21**	**40.51**

Table 4: The effectiveness of each component. We quantify the contribution of key components: Event-Aware Prototype Aggregation (EAP) and Prototype-based Event Soft-Alignment (PES).

4 EXPRIMENTS

4.1 EXPERIMENT SETTINGS

Datasets. We evaluate on two text-annotated 3D human motion datasets: HumanML3D (Guo et al., 2022) and KIT-ML (Plappert et al., 2016). HumanML3D draws motion sequences from AMASS (Mahmood et al., 2019) and HumanAct12 (Guo et al., 2020); following TMR (Petrovich et al., 2023), we use 23,384/1,460/4,380 training/validation/test motion samples, with approximately three descriptions per sequence. KIT-ML contains 3,911 motion recordings and 6,278 descriptions; its corresponding splits contain 4,888/300/830 motion samples.

Evaluation Metric. The evaluation metrics are the recall R@K (the percentage of ground truth in the retrieved top-K lists, K=1,2,3,5,10) and Rsum (sum of all R@K). Additionally, we compute the median rank (MedR) of our results, where lower MedR values indicate more accurate retrievals. To enable a fair comparison with the baselines, we adopt the experimental settings proposed in previous work (Yu et al., 2024). Specifically, a text sample is randomly chosen as the matching text during training, and the first text in the test set is used to report evaluation performance.

Implementation Details. We train our model for 100 epochs with a batch size of 256 and a learning rate of 5e-5, using the Adam (Kingma & Ba, 2017) optimizer. For both datasets, we set the maximum motion sequence length to 200. The representation dimension d_r of the common space is set to 256. We set the temperature parameter τ to 0.07, following CLIP (Radford et al., 2021). The hyperparameters α and β are set to 0.1 and 0.8. The variance loss parameter γ, ϵ are set as 0.75, 1e-4. The variance loss weight β is set as 0.5.

4.2 COMPARISON WITH THE STATE-OF-THE-ART

In this subsection, we compare our proposed method with recent state-of-the-art methods (Petrovich et al., 2022; Fujiwara et al., 2024; Petrovich et al., 2023; Yu et al., 2024; Yang et al., 2024; Yan et al., 2023; Guo et al., 2022) on the HumanML3D (Guo et al., 2022) and KIT-ML (Plappert et al., 2016) datasets. In Tab. 1 and Tab. 2, we show the results of our model against state-of-the-art methods on the HumanML3D and KIT-ML datasets. Our model achieves the best performance across all evaluation metrics at both text-to-motion (t2m) and motion-to-text (m2t) retrieval. Specifically, EAPM outperforms the previous best method, MoPa (Yu et al., 2024), by R@1 1.18% at t2m and R@1 1.76% at m2t retrieval on HumanML3D. Additionally, for both datasets, EAPM obtains higher Rsum performance than the previous best method (219.81 *vs.* 247.17 on HumanML3D and 287.20 *vs.* 313.73 on KIT-ML). Our results again surpass methods with post-processing operations like dual softmax (Cheng et al., 2021) and QB-Norm (Bogolin et al., 2022).

4.3 ABLATION STUDIES

Effectiveness of each component. We conduct experiments to verify the effectiveness of the proposed components in our methods. **w/ EAP** refers to training our model using only the Event-Aware Prototype Aggregation with the traditional InfoNCE loss for model optimization. **w/ PES** refers to training our model using only soft alignment on global features, which does not consider fine-grained action events. In Tab. 4, it can be observed that these two components have a positive impact on retrieval performance. The integration of **EAP** and **PES** results in a further boost to retrieval per-

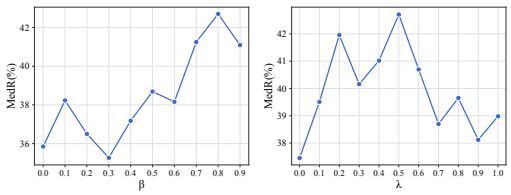


Figure 4: The ablation studies of the parameter β and the loss balance factor λ. We report the MedR results at text-to-motion retrieval on the HumanML3D dataset.

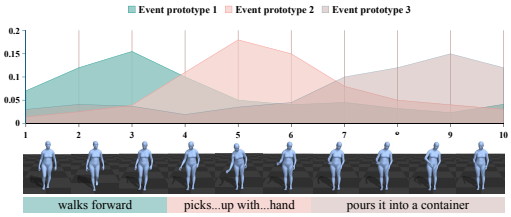


Figure 5: Visualization of event prototypes. We sample 10 motion frames from the motion sequence and show the mask weights to event prototypes in a line graph. The weight values are normalized to the range [0, 0.2] for clearer representation.

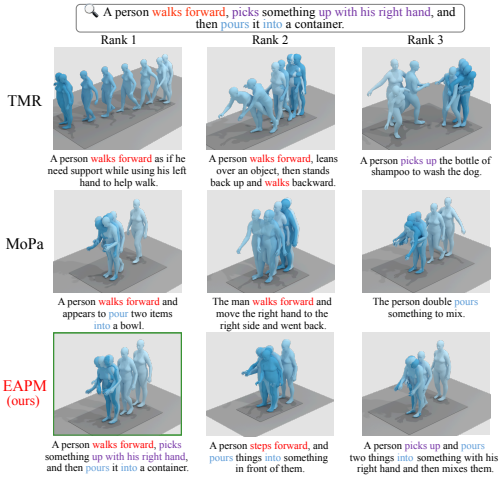


Figure 6: Visualization of retrieval results. We present the top-3 motion sequences retrieved by our method EAPM, as well as by the state-of-the-art MoPa (Yu et al., 2024) and TMR (Petrovich et al., 2023). The successful retrieval result is highlighted by the green border. We also highlight the semantically similar verbs to the query with the same color.

formance and achieves state-of-the-art performance, with a 2.94% and 2.99% improvement in R@1 at both text-to-motion retrieval and motion-to-text retrieval.

In-depth analysis of event prototype generation. We conduct several experiments to provide an in-depth analysis of our event prototype generation approach. Firstly, we design three variants to study whether our approach is sufficiently effective. (1) **Part:** We divide the token features into several segments, with segment centers as event prototypes. (2) **Cluster:** Following prior work in the video field (Arandjelović et al., 2016), the shared cluster centers are used to obtain event prototypes. (3) **Decoder:** We initialize a set of queries and input them with local features into a two-layer decoder to generate event prototypes. The "#tokens" denotes each token as a prototype. As shown in Tab. 3, our method outperforms other generation strategies by a clear margin, with lower test time consumption. We observe that with K=3 and variance loss, our method outperforms the baseline by 18.32 on Rsum. When K is greater than 3, performance degrades as K increases. An appropriate number of prototypes enables the model to learn representative event features. When too many prototypes are introduced, the learned event features become noise rather than representative.

Hyperparametric sensitivity study. In Fig. 4, we conduct experiments to analyze the effects of the parameter β and the loss balance factor λ. First, for the parameter β, the best performance is achieved when β is 0.8. The retrieval performance is poor when β is small. There may be some noise during the initial training phase of the model. Second, as the loss balance factor λ increases, MedR initially increases and then decreases. The proposed method with $\lambda = 0.5$ performs the best.

4.4 Qualitative Results

Visualization of event prototypes. In Fig. 5, we present the mask weights for 10 motion frames to event prototypes according to our mask-generating function. These weights are displayed in a line graph. The text description corresponding to the motion sequence is "A person walks forward, picks something up with their right hand, and then pours it into a container". We can find that event prototype 1 exhibits larger weights in the first three frames, indicating that it likely corresponds to the semantics described, *i.e*, "walks forward". Similarly, event prototypes 2 and 3 correspond to other action events. The weights of different event prototypes across the frames are quite different,

demonstrating that the model is capable of learning temporal relationships and showing that the event prototypes can capture diverse action semantics.

Visualization of retrieval results. In Fig. 6, we present the retrieval results for text-to-motion retrieval of the proposed method EAPM and the state-of-the-art MoPa (Yu et al., 2024) and TMR (Petrovich et al., 2023). The given text query contains multiple action events. For each method, we showcase the top-3 retrieved motions along with their corresponding ground-truth text labels. For the MoPa and TMR results, we observe that all three retrieval methods 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. 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.

5 APPLICATIONS

5.1 HUMAN INTERACTION RECOGNITION

To further validate the model's representational capacity, we conduct experiments on the multi-person interaction recognition task using the InterHuman Dataset (Liang et al., 2024). This dataset includes a diverse set of 3D motion sequences that involve interactions between two individuals. To obtain the features of the interactions, we apply a shared motion encoder to each person and concatenate the motion features before the projection head. As shown in Tab. 5, our proposed method outperforms the baselines, which rely solely on global features for human interaction recognition. By leveraging event prototypes, our method is able to recognize such complex, multi-action interactions more precisely.

Methods	Text-to-motion			Motion-to-text		
	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
EAPM	**11.23**	**24.17**	**35.87**	**10.87**	**24.58**	**34.91**

Table 5: Results of human interaction recognition.

5.2 MOTION TEMPORAL LOCALIZATION

Furthermore, we demonstrate the effectiveness of our proposed method in the motion temporal localization task using the BABEL (Punnakkal et al., 2021) dataset. Notably, this is implemented in a zero-shot manner (*i.e.*, the model is not trained for this task). 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. The IoU is calculated by selecting the window size and position that exhibits the maximum similarity to the text query. As shown in Tab. 6, our method outperforms the other retrieval baselines (Petrovich et al., 2023; Fujiwara et al., 2024) by a significant margin. This further validates the representation learning capability of our model. Our model leverages event-level alignment during training, incorporating a comprehensive understanding of local features to enhance semantic learning and improve the perception of action events.

Methods	BABEL Dataset				
	IoU@0.3 ↑	IoU@0.5 ↑	IoU@0.7 ↑	IoU@0.9 ↑	mIoU ↑
TMR	25.65	14.49	5.75	1.10	17.19
CAR	27.09	15.97	7.55	1.92	17.70
EAPM	**32.24**	**20.03**	**10.54**	**3.49**	**23.21**

Table 6: Results of motion temporal localization.

6 CONCLUSION

In this paper, we propose an Event-Aware Prototype Matching (EAPM) framework for 3D human motion retrieval. We design an event-prototype aggregation method that adaptively aggregates pose and word features into a set of event prototypes from both the motion sequence and the text. We introduce a variance loss function to ensure that different prototypes focus on distinct, non-overlapping action events. We propose a prototype-based event soft-alignment approach that modulates confidence estimation for samples via intra-modal self-similarity during the alignment process and achieves fine-grained text-motion alignment at the event level. Extensive experiments on two widely used datasets demonstrate the efficacy of our method at text-to-motion retrieval and other challenging motion tasks.

REFERENCES

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2016.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation, 2022.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss, 2021.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

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. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19790–19800, 2024.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2018. URL https://arxiv.org/abs/1707.05612.

Zheren Fu, Yan Li, Zhendong Mao, Quan Wang, and Yongdong Zhang. Deep metric learning with self-supervised ranking. In *AAAI Conference on Artificial Intelligence*, 2021. URL https://api.semanticscholar.org/CorpusID:235306316.

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models, 2024. URL https://arxiv.org/abs/2407.15408.

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

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

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5142–5151. IEEE, 2022.

Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18298–18306, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i16.29789. URL http://dx.doi.org/10.1609/aaai.v38i16.29789.

Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pp. 448–456. JMLR.org, 2015.

Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H. Bermano. Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1965–1974, June 2024.

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*, pp. 2151–2162, 2023.

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

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Mengshu Sun, Wei Niu, Xuan Shen, Geng Yuan, Bin Ren, Minghai Qin, Hao Tang, and Yanzhi Wang. Spvit: Enabling faster vision transformers via soft token pruning, 2022. URL https://arxiv.org/abs/2112.13890.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Computer Vision – ECCV 2018*. Springer International Publishing, 2018.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *ArXiv*, abs/2208.02515, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022b.

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence Yang. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. In *International Conference on Learning Representations*, volume 2025, pp. 84238–84250, 2025.

Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, Ying Jin, and Yufeng Zhang. Image-text retrieval with binary and continuous label supervision, 2022c. URL https://arxiv.org/abs/2210.11319.

Zheng Li, Caili Guo, Xin Wang, Zerun Feng, and Yanjun Wang. Integrating listwise ranking into pairwise-based image-text retrieval, 2023. URL https://arxiv.org/abs/2305.16566.

Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Zhongtian Du. Integrating language guidance into image-text matching for correcting false negatives. *IEEE Transactions on Multimedia*, 26:103–116, 2024. doi: 10.1109/TMM.2023.3261443.

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

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition, 2020.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *ACM Multimedia*, 2022.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *ICCV 2019*. IEEE, 2019.

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

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*, volume 13682 of *Lecture Notes in Computer Science*, pp. 480–497. Springer, 2022.

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

Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14354–14363, 2024.

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

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

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. Learning transferable visual models from natural language supervision, 2021.

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. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Xiaohan Wang, Linchao Zhu, and Yi Yang. T2VLAD: global-local sequence alignment for text-video retrieval. *CoRR*, abs/2104.10054, 2021.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2804–2815, 2023.

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

Yang Yang, Haoyu Shi, and Huaiwen Zhang. 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 '24, pp. 1083–1092, New York, NY, USA, 2024. Association for Computing Machinery.

Kangning Yin, Shihao Zou, Yuxuan Ge, and Zheng Tian. Tri-modal motion retrieval by learning a joint embedding space. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1596–1605, 2024.

Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. Exploring vision transformers for 3d human motion-language models with motion patches. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 937–946, 2024.

Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Robust video-text retrieval via noisy pair calibration. *IEEE Transactions on Multimedia*, 25:8632–8645, 2023a.

Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Debiased video-text retrieval via soft positive sample calibration. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5257–5270, 2023b.

Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pp. 18–38. Springer, 2025.

A APPENDIX

This appendix provides additional method and implementation details (Sec. A.1), additional experimental results (Sec. A.2), further analysis of event prototypes (Sec. A.3), and limitations (Sec. A.4). Meanwhile, we address the following questions and potential concerns:

- **RQ1:** Does EAPM introduce substantial additional computational overhead or resource consumption?
- **RQ2:** Do the learned event prototypes capture distinct and semantically aligned action content across the text and motion modalities?
- **RQ3:** Can a fixed number of event prototypes effectively accommodate motion sequences with varying numbers of action events?

A.1 ADDITIONAL METHOD AND IMPLEMENTATION DETAILS

Methodological distinction from existing methods. Unlike prototype-based methods that use prototypes as shared cluster centers for semantically similar instances, our event prototypes serve as explicit event descriptors. They are adaptively aggregated from word and pose token features to capture action events in the text and motion modalities. Specifically, we introduce a token-adaptive aggregation method together with a variance loss to generate multiple event prototypes that focus on distinct action semantics. We further propose an event soft-alignment method that incorporates intra-modal event self-matching probabilities into the alignment objective. This design does not require additional regularization terms or supervision from external augmented data.

Textual prototype aggregation. We employ a Transformer encoder to obtain the word features $\boldsymbol{t}^s$ and then map them to multiple textual event prototypes $\boldsymbol{t}^o$. Specifically, the k-th textual event prototype is computed as

$$\boldsymbol{t}_k^o = \sum_{j=1}^{L} \boldsymbol{t}_j^s \cdot f^{\text{mask}}\left(\boldsymbol{t}_j^s\right)[k], \tag{13}$$

where L denotes the number of word tokens, $\boldsymbol{t}_j^s$ is the feature of the j-th word token, and $f^{\text{mask}}(\boldsymbol{t}_j^s)[k]$ represents the contribution of the j-th word token to the k-th textual event prototype. The same prototype aggregation process is applied to both the text and motion modalities.

Evaluation protocol. To ensure a fair comparison, we re-implement HSA using the evaluation protocol adopted by MoPa. In particular, a text sample is randomly selected as the matching description during training, while the first description of each motion sequence is used during evaluation. The test data also contain mirroring-augmented samples. Consequently, some queries have two correct matches in the retrieval gallery: the original motion and its mirrored counterpart. The corresponding comparison is reported in Tab. 7.

Post-processing. Following previous retrieval methods, we additionally report results obtained using inverted dual softmax (Cheng et al., 2021) or QB-Norm (Bogolin et al., 2022) as post-processing. Inverted dual softmax reweights the cross-modal similarity matrix through bidirectional normalization, while QB-Norm calibrates the similarity scores using a query bank to reduce the influence of frequently retrieved samples. These operations are applied only to the similarity scores during inference and do not modify the model architecture or training objective.

A.2 ADDITIONAL EXPERIMENTAL RESULTS

A.2.1 COMPARISON WITH HSA AND THE SPATIAL ENCODER

For a fair comparison, we re-implement HSA using the same evaluation protocol as MoPa and EAPM. We also evaluate EAPM-S, a variant of EAPM without the GCN-based spatial encoder, to examine the contribution of spatial modeling. As shown in Tab. 7, EAPM-S outperforms HSA across all reported recall metrics on both datasets. Specifically, it improves R@1 from 10.04 to 11.56 on HumanML3D and from 14.22 to 15.69 on KIT-ML. These results indicate that the advantages of EAPM extend beyond the use of the spatial encoder. Adding the spatial encoder further increases R@1 to 11.98 and 16.14 on HumanML3D and KIT-ML, respectively, with consistent improvements

Method	HumanML3D			KIT-ML		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
HSA	10.04	27.84	38.54	14.22	39.85	51.04
EAPM-S	11.56	29.77	41.98	15.69	42.09	53.47
EAPM	**11.98**	**30.69**	**42.70**	**16.14**	**42.53**	**54.70**

Table 7: The results for HSA and EAPM-S (w/o spatial encoder).

Method	Text-to-motion			Motion-to-text		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
TMR	5.35	17.14	30.21	6.59	14.29	22.91
MoPa	7.35	20.98	34.33	7.90	17.71	25.88
EAPM (ours)	**9.24**	**23.98**	**40.33**	**9.12**	**20.17**	**29.74**

Table 8: Results of zero-shot cross-skeleton recognition.

in R@5 and R@10. This suggests that modeling spatial dependencies among body joints provides complementary information for event-aware text–motion matching.

A.2.2 Zero-Shot Cross-Skeleton Recognition

Beyond the zero-shot motion temporal localization experiments reported in the main paper, we further evaluate EAPM on zero-shot cross-skeleton recognition. This evaluation examines whether the learned text–motion representations transfer across skeleton configurations.

As shown in Tab. 8, EAPM consistently outperforms TMR and MoPa across all reported metrics in both retrieval directions. Compared with MoPa, EAPM improves text-to-motion R@1 from 7.35 to 9.24 and motion-to-text R@1 from 7.90 to 9.12. It also achieves gains of 6.00 and 3.86 percentage points in R@10 for the two directions, respectively. These results suggest that event-level matching supports transferable representations for cross-skeleton text–motion retrieval.

A.2.3 Computational Efficiency **(RQ1)**

We compare the computational cost of EAPM with TMR and MoPa in Tab. 9, considering inference time, memory usage, FLOPs, and parameter count alongside retrieval performance. Compared with MoPa, EAPM reduces inference time from 47.136 s to 44.014 s and memory usage from 8,896 MB to 7,466 MB. It also requires fewer FLOPs (9.21 G versus 12.64 G) and parameters (72.68 M versus 128.90 M), while improving Rsum from 219.81 to 247.17. These results demonstrate that effective event-level matching does not necessarily require a larger or more computationally expensive model.

Compared with TMR, EAPM introduces additional computation, with inference time increasing from 40.858 s to 44.014 s. This approximately 7.7% increase accompanies an absolute Rsum improvement of 61.14 points. Overall, EAPM offers a favorable balance between retrieval performance and computational efficiency.

A.3 Further Analysis

A.3.1 Visualization of Event Prototypes (RQ2)

We visualize the mask weights of textual and motion event prototypes in Fig. 7 to examine how different prototypes aggregate local semantic information. Each weight reflects the contribution of a word or motion token to a particular event prototype.

In the text modality, different prototypes emphasize distinct action words, suggesting that they capture complementary event semantics rather than repeatedly focusing on the same content. In the motion modality, the prototypes assign different weights to motion tokens, allowing each prototype to emphasize the portions of the sequence relevant to its associated action. The correspondence between the emphasized words and motion segments provides qualitative evidence of semantic align-

Method	Time ↓	Memory usage ↓	FLOPs ↓	Params ↓	Rsum ↑
TMR	40.858s	6606MB	7.47G	49.54M	186.03
MoPa	47.136s	8896MB	12.64G	128.90M	219.81
EAPM (ours)	44.014s	7466MB	9.21G	72.68M	**247.17**

Table 9: Comparison of computational cost.

ment at the event level. These visualizations illustrate how prototype aggregation preserves distinct action information within a motion–text pair, supporting fine-grained matching beyond overall semantic similarity.

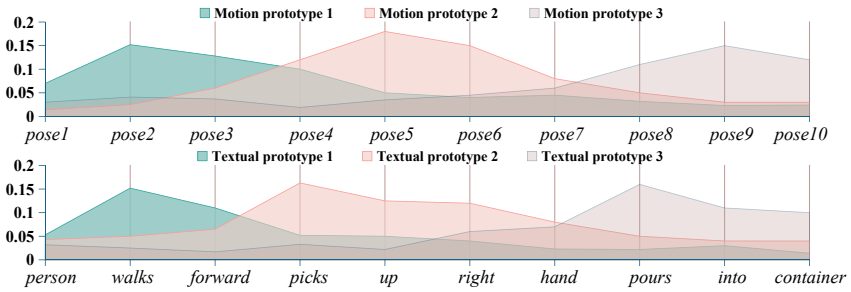


Figure 7: Visualization of textual and motion event prototypes.

A.3.2 Number of Projection Layers

We investigate the effect of the number of projection layers in $f^{\text{mask}}(\cdot)$. As shown in Tab. 10, using two projection layers achieves the best performance on both HumanML3D and KIT-ML. A single projection layer has limited representation capacity, while increasing the number of layers to three does not provide further improvement.

Number	HumanML3D			KIT-ML		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
1	9.25	27.54	38.17	14.24	39.47	50.01
2	**11.98**	**30.69**	**42.70**	**16.14**	**42.53**	**54.70**
3	11.16	29.88	41.53	15.49	41.78	53.11

Table 10: Ablation study on the number of projection layers in $f^{\text{mask}}(\cdot)$.

A.3.3 Number of Event Prototypes (**RQ3**)

The number of event prototypes K is treated as a hyperparameter and determined experimentally. As shown in Tab. 4, using three event prototypes achieves the best performance on the evaluated datasets. This is consistent with the average number of events, which is 3.14 on HumanML3D and 2.88 on KIT-ML. Despite using a fixed number of event prototypes, EAPM remains effective for samples with varying action counts. To further evaluate this, we conduct experiments on two 50-sample HumanML3D subsets: an Origin Set consisting of randomly selected samples and a Diverse Set containing 10 samples for each action count from 1 to 5.

As shown in Tab. 11, EAPM achieves similar performance on the two subsets. This may be attributed to two aspects. First, adaptive prototype aggregation allows event prototypes to capture different action semantics, including closely related sequential actions such as bending down to pick something up. Second, the combination of global and event-level representations supports the modeling of complex motion sequences, with global representations providing complementary semantic information, as evaluated in Tab. 4.

Dataset	Text-to-Motion			Motion-to-Text		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
Origin Set	53.26	84.78	90.13	45.65	82.61	88.04
Diverse Set	52.07	82.14	89.48	44.69	80.59	89.12

Table 11: Results on the Origin Set and Diverse Set of HumanML3D.

A.4 Limitations and Future Work

EAPM achieves effective text–motion retrieval by capturing diverse action semantics through event-level matching, and demonstrates transferability to additional motion understanding tasks. Nevertheless, fine-grained temporal distinctions, such as action duration and execution speed, remain challenging because event prototypes do not explicitly encode these attributes. Future work will incorporate temporal modeling to further improve retrieval for queries involving such details.

In addition, while a fixed number of event prototypes performs well across the evaluated datasets, adapting the prototype number to individual sequences may better accommodate varying event complexity. We will explore adaptive prototype selection to extend the flexibility of the framework.

AI use statement

We used generative AI tools to assist with translation, manuscript drafting, and language editing, and to obtain feedback on the presentation of our methodology. The authors reviewed the AI-assisted content and take full responsibility for the paper's technical claims, results, and final text.