FineMoLA: Towards Fine-Grained Motion-Language Alignment from Clip-Level Supervision

Tongyan Wang[† 1] Zhenguyan Li[† 1] Muhan Lin[1] Shengyang Luo [1] Yifan Shen [2] Aniket Bera[1]
Baijian Yang[1] Yingjie Victor Chen[1]

[1]Purdue University [2]University of Illinois Urbana-Champaign

† Equal Contribution

Abstract

Text-conditioned human motion generation has made rapid progress with the emergence of large-scale motion–language datasets. However, even datasets with rich long-form descriptions typically provide supervision only at the clip level, without explicit temporal correspondence between motion frames and language. This limits fine-grained motion–text grounding and temporally precise generation. We propose FineMoLA, a weakly supervised framework that learns fine-grained frame–phrase correspondence directly from clip-level annotations. Our method first segments long-form descriptions into action-bearing phrases, and then formulates motion–language alignment as an optimal transport problem, which naturally models many-to-many relations between motion frames and text under global constraints. With entropic regularization and Sinkhorn iterations, FineMoLA efficiently infers pseudo frame-level alignments without human labeling. Experiments on SnapMoGen demonstrate that the learned alignments outperform baselines in motion–text grounding. Website: finemola.github.io.

1. Introduction

Recent years have witnessed rapid progress in text-conditioned 3D human motion generation, driven by advances in diffusion [32] and transformer-based modeling [8, 9]. These models can synthesize diverse and realistic motion sequences from natural language prompts, enabling applications in character animation, AR/VR, embodied agents, and human–computer interaction.

A key catalyst behind recent progress is the emergence of large-scale motion–language datasets. Early benchmarks such as KIT-ML [23] and HumanML3D [8] established the standard text-to-motion setting with paired clip-level captions and standardized evaluation. However, these captions are typically short and describe an entire motion clip coarsely, limiting fine-grained controllability and semantic grounding. More recent datasets [11, 15, 24, 33, 35] broaden the supervision spectrum with richer long-form descriptions, hierarchical text annotations, frame-level action labels, and temporally localized captions. Together, these developments highlight a growing demand for *fine-grained temporal correspondence* between motion and language, especially for long-form narratives and multi-action descriptions.

Among existing datasets, SnapMoGen [10] stands out for its high-quality motion–language supervision. Its motions are collected using professional mocap suits, and its annotations are primarily written by human annotators. In contrast to earlier datasets that rely on short action labels or simple captions, SnapMoGen provides long-form descriptions that often cover multiple actions and transitions within a single motion clip. Such expressive supervision offers a strong foundation for studying fine-grained motion–text correspondence. However, the annotations remain at the clip level, without explicit frame-to-token alignment between motion and language. Learning such fine-grained correspondence could benefit motion understanding tasks such as temporal localization and dense motion captioning, and provide stronger supervision for fine-grained text-to-motion generation with temporal structure [34, 35, 39]. At the same time, obtaining dense open-vocabulary alignment at scale remains challenging due to the inherently many-to-many relationships between words and motion over time.

As illustrated in Fig. 1, we aim to recover fine-grained frame–phrase correspondence from long-form clip-level annotations without requiring manual temporal labels. To bridge this gap, we propose a weakly supervised framework that formulates frame–token correspondence as an optimal transport (OT) problem [5, 21]. Optimal transport naturally models many-to-many soft alignments under global marginal constraints, making it well-suited for motion-text grounding. To enable efficient and differentiable learn-

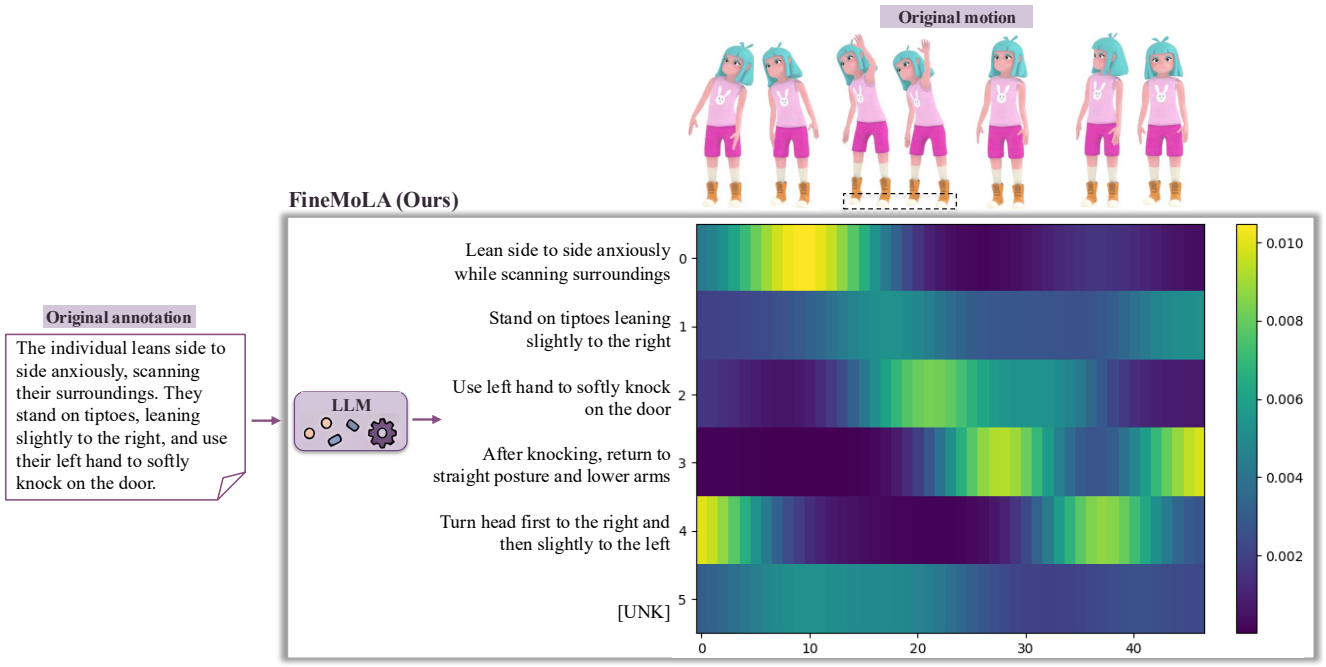

Figure 1. While the SnapMoGen dataset provides human motions with long-form text annotations, FineMoLA finds the fine-grained alignment matrix between motion frames and LLM-detected action-bearing phrases in the text without manual labeling. Our method includes an [UNK] token to absorb frames that are not explicitly described by any specific phrase. The dashed black box highlights that the character does stand on tiptoes in those frames.

ing, we adopt entropic regularization and solve the resulting transport plan using Sinkhorn iterations [5]. This formulation aligns with a broader trend in multimodal learning where fine-grained token-level alignment objectives improve cross-modal consistency and robustness [17, 36, 40]. Leveraging abundant clip-level supervision in SnapMoGen, our method produces pseudo frame-level alignments that support temporally grounded motion–language modeling.

To summarize, our contributions are as follows:

- We introduce FineMoLA, a weakly supervised framework that learns fine-grained motion–language correspondence from clip-level annotations. By formulating frame-to-token alignments as an optimal transport problem, our method learns fine-grained correspondences for long-form textual descriptions, enabling dense motion–text grounding without requiring manual frame-level labels.
- Experiments on the SnapMoGen dataset demonstrate that the learned alignments provide stronger motion–text grounding than baselines.

2. Related Work

2.1. Human Motion Datasets with Text Annotation

The rapid progress of text-driven human motion modeling has been closely tied to the emergence of motion datasets with language annotation [3, 10, 33, 34]. Early benchmarks such as KIT-ML [23] and HumanML3D [8] established the standard text-to-motion setting by providing paired motion clips and clip-level natural language descriptions, enabling shared embedding learning, retrieval-based training, and standardized evaluation [20].

Subsequent datasets broadened the supervision spectrum beyond coarse clip-level captions. BABEL [24] augments mocap sequences with sequence- and frame-level action labels, supporting temporally localized understanding. Motion-X [16] provides expressive whole-body motion together with sequence-level semantics and video-derived frame-level pose descriptions. More recently, SnapMoGen [10] introduces long-form narrative descriptions with improved temporal continuity, while MotionLib [33] scales motion–text supervision to millions of samples with hierarchical annotations. Dense Motion Captioning further introduces CompMo, a large-scale dataset with temporally bounded multi-action sequences and grounded captions [35]. In parallel, FrankenMotion provides temporally aware part-level prompts for compositional motion modeling [15].

Despite these advances, most motion–language datasets still provide supervision either at the clip level or through limited predefined labels, leaving many-to-many frame-level (or token-to-frame) correspondence under-explored.

This gap motivates learning objectives that explicitly model fine-grained temporal correspondence between motion and language [14, 35], complementary to efforts that focus on stronger generators or clip-level supervision [9, 18, 32, 37].

2.2. Human Motion Understanding

Human motion understanding has long been studied in computer vision through video-based tasks such as action recognition and temporal localization, where models align motion with action labels or textual descriptions [1, 4, 7, 19]. In contrast, our work focuses on structured human motion data (e.g., mocap sequences), where the temporal evolution of poses is directly available and can be paired with language supervision.

In the motion domain, early understanding models largely relied on shared embedding learning and retrieval-style objectives. MotionCLIP [31] aligns motion latents with the CLIP [25] embedding space to exploit strong language priors, while TMR [20] combines contrastive learning with a motion synthesis objective for text–motion retrieval. Cross-dataset studies [3] further show that such retrieval performance is sensitive to shifts in motion distributions and annotation styles, motivating robustness and transfer-aware evaluation.

More recent works move beyond global alignment toward fine-grained motion–language understanding. Dense Motion Captioning [35] requires temporally localizing and describing multiple actions within long motion sequences, while unified motion–language models such as MG-MotionLLM [34] and UniMotion [14] further connect motion understanding, localization, and generation. At the same time, advances in generation and control, including ReMoDiffuse [38], FineMoGen [39], CoMo [12], and recent text-conditioned generators [2, 6, 9, 13, 18, 22, 30, 37, 41], highlight the growing demand for fine-grained temporal correspondence between motion and language. Our work addresses this need by learning dense motion–text alignment from clip-level supervision.

2.3. Semi-supervised Learning Using Optimal Transport

Optimal transport (OT) provides a principled tool to establish soft correspondences between two sets of elements under marginal constraints, and has been widely used for matching and alignment problems [5, 21, 29]. Recent work has incorporated OT into representation learning and robust alignment. For instance, OT-based matching has been used to mitigate data poisoning in CLIP-style [25] pre-training by aligning samples through transport plans [40]. OT has also been applied to understand and generalize CLIP by modeling relationships between sets of features via transport [27, 36]. In this work, we focus on *fine-grained correspondences* in the motion-language domain.

3. Preliminary

3.1. Optimal Transport

Optimal Transport (OT) [29] provides a principled framework for measuring the discrepancy between two distributions while explicitly accounting for the underlying geometry. Assume two uniform discrete distributions $\mathbf{a}$ and $\mathbf{b}$, supported on T_m and T_t elements, respectively. We denote the set of admissible transport plans by

$$U(\mathbf{a},\mathbf{b}) = \left\{ \mathbf{P} \in \mathbb{R}_+^{T_m \times T_t} \;\middle|\; \mathbf{P}\mathbf{1}_{T_t} = \mathbf{a},\ \mathbf{P}^\top \mathbf{1}_{T_m} = \mathbf{b} \right\}, \tag{1}$$

where

$$\mathbf{a} = \frac{1}{T_m}\mathbf{1}_{T_m}, \qquad \mathbf{b} = \frac{1}{T_t}\mathbf{1}_{T_t}. \tag{2}$$

The marginal constraints ensure that $\mathbf{P}$ redistributes the mass of $\mathbf{a}$ onto $\mathbf{b}$. Given a cost matrix $\mathbf{C} \in \mathbb{R}^{T_m \times T_t}$ measuring the cost of transporting unit mass between elements, the Kantorovich formulation of OT is defined as

$$\min_{\mathbf{P} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle \tag{3}$$

where $\langle \mathbf{P}, \mathbf{C} \rangle = \sum_{i,j} \mathbf{P}_{ij}\mathbf{C}_{ij}$.

In contrast to pointwise matching, OT naturally supports many-to-many correspondence and enforces global consistency via marginal constraints. These properties make OT particularly suitable for modeling fine-grained alignment under weak supervision, such as motion–text correspondence.

3.2. Entropic-Regularized Optimal Transport and Sinkhorn Algorithm

Directly solving Eq. (3) requires linear programming, which is computationally expensive and unsuitable for end-to-end learning. To address this issue, entropic regularization augments the OT objective with an entropy term:

$$\mathbf{P}^* = \arg\min_{\mathbf{P} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle \ + \ \varepsilon \sum_{i,j} P_{ij}\big(\log P_{ij} - 1\big) \tag{4}$$

where $\varepsilon > 0$ controls the strength of the regularization. The entropic regularization renders the problem strictly convex and enables efficient optimization via the Sinkhorn algorithm.

4. Method

Suppose we are given a paired motion–text dataset $\mathcal{D} = \{(m^{(j)}, t^{(j)})\}_{j=1}^{N}$. For simplicity, we omit the sample index (j) when there is no ambiguity. Each sample consists of a motion sequence $m \in \mathbb{R}^{T'_m \times D_m}$ and a text sequence

$t \in \mathbb{R}^{T_t' \times D_t}$. Our goal is to learn fine-grained motion–text alignment. We model the alignment between a motion sequence and a text sequence as an optimal transport plan $\mathbf{P}$, where each entry P_{uv} measures the semantic correspondence between the u-th motion frame and the v-th text token. We define the transport plan in Sec. 4.1 and then provide an intuitive analysis of this formulation in Sec. 4.2. Finally, we describe the feature extraction process in Sec. 4.3 and the CLIP-style contrastive training objective in Sec. 4.4.

4.1. Optimal Transport Cost

Our model consists of a motion feature extractor E_m and a text feature extractor E_t. Given a motion sequence m and a text sequence t, E_m produces a sequence of fine-grained motion features

$$Z^m = \{z_1^m, \dots, z_{T_m}^m\}, \tag{5}$$

and E_t produces a sequence of text features

$$Z^t = \{z_1^t, \dots, z_{T_t}^t\}. \tag{6}$$

Note that $T_m \neq T_m'$ and $T_t \neq T_t'$ due to down-sampling in the encoders. Let $s(\cdot, \cdot)$ denote the cosine similarity between ℓ_2-normalized features.

We define the transport cost matrix $\mathbf{C} \in \mathbb{R}^{T_m \times T_t}$ by

$$\mathbf{C}_{uv} = 1 - s(z_u^m, z_v^t). \tag{7}$$

Intuitively, a smaller cost indicates a stronger semantic correspondence between the u-th motion frame and the v-th text token.

Based on the admissible transport set $U(\mathbf{a}, \mathbf{b})$ defined in the previous subsection, we define the OT cost for a motion–text pair (m, t) as

$$\mathcal{J}_{\mathrm{OT}}(m, t) = \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle. \tag{8}$$

This OT cost measures the minimum transport cost required to align the motion and text features under the marginal constraints. A lower value of $\mathcal{J}_{\mathrm{OT}}(m, t)$ indicates better fine-grained motion–text alignment. As in Sec. 3.2, we follow Eq. (4) to obtain approximated optimal transport plan $\mathbf{P}'$. The estimated optimal transport cost is

$$\mathcal{J}_{\mathrm{EOT}} = \langle \mathbf{P}', \mathbf{C} \rangle \tag{9}$$

Although $\mathbf{P}'$ is a function of $\mathbf{C}$, we treat it as a constant during backpropagation for simplicity. Under this approximation, $\mathcal{J}_{\mathrm{EOT}}$ remains differentiable with respect to $\mathbf{C}$.

4.2. Analysis: a Case Study

To better understand how the proposed objective encourages fine-grained alignment, we consider an idealized case in which each motion frame is described by exactly one text token.

Assumption 1 (Perfect Frame–Token Correspondence)**.** *Assume that the motion and text sequences have equal lengths, i.e., $T_m = T_t$. Further assume that there exist encoder parameters θ_m and θ_t such that, for each paired sample (m, t), there exists a bijection $\pi : \{1, \dots, T_m\} \to \{1, \dots, T_t\}$ satisfying*

$$s\left(z_u^m, z_{\pi(u)}^t\right) = 1, \quad \forall u \in \{1, \dots, T_m\}, \tag{10}$$

where z_u^m and z_v^t denote the u-th and v-th output features of $E_m(m; \theta_m)$ and $E_t(t; \theta_t)$, respectively.

Theorem 1 (Zero OT Cost under Perfect Alignment)**.** *Under Assumption 1, the OT objective $\mathcal{J}_{OT}(m, t)$ admits a feasible transport plan with zero cost. Specifically, the transport plan defined by*

$$P_{uv} = \frac{1}{T_m} \mathbb{1}\{v = \pi(u)\} \tag{11}$$

is feasible and achieves

$$\mathcal{J}_{OT}(m, t) = 0. \tag{12}$$

Proof. See supplementary. □

In practice, motion-text alignment is generally many-to-many rather than bijective. Nevertheless, Theorem 2 shows that the proposed objective is consistent with perfect fine-grained alignment in an idealized setting, which helps explain why minimizing $\mathcal{J}_{\mathrm{OT}}$ can encourage meaningful motion–text correspondence in practice.

4.3. Feature Extraction

The weakly supervised learning framework is illustrated in Fig. 2. Our architecture adopts a dual-stream design to learn representations from textual descriptions and motion sequences. The original expressive text annotation is first segmented into multiple **action-bearing phrases** using an LLM (details are in the supplementary material). These phrases are then concatenated into a sequence and fed into the text feature extractor. The text feature extractor E_t leverages a frozen T5-base [26] backbone to extract latent semantic features, which are subsequently projected into a latent space through an additional MLP. To facilitate temporal alignment, we introduce an **action unit aggregator** that groups tokens belonging to the same phrase and pools their embeddings into compact **action unit features**. Some motion segments may not be explicitly described by any action-bearing phrase in the text. To avoid imposing spurious correspondences, we do not force every motion frame to align with an action-unit representation alone and allow a motion frame to correspond to an `[UNK]` token. We introduce a **global feature** mechanism that aggregates sequence-level information and provides holistic semantic context for

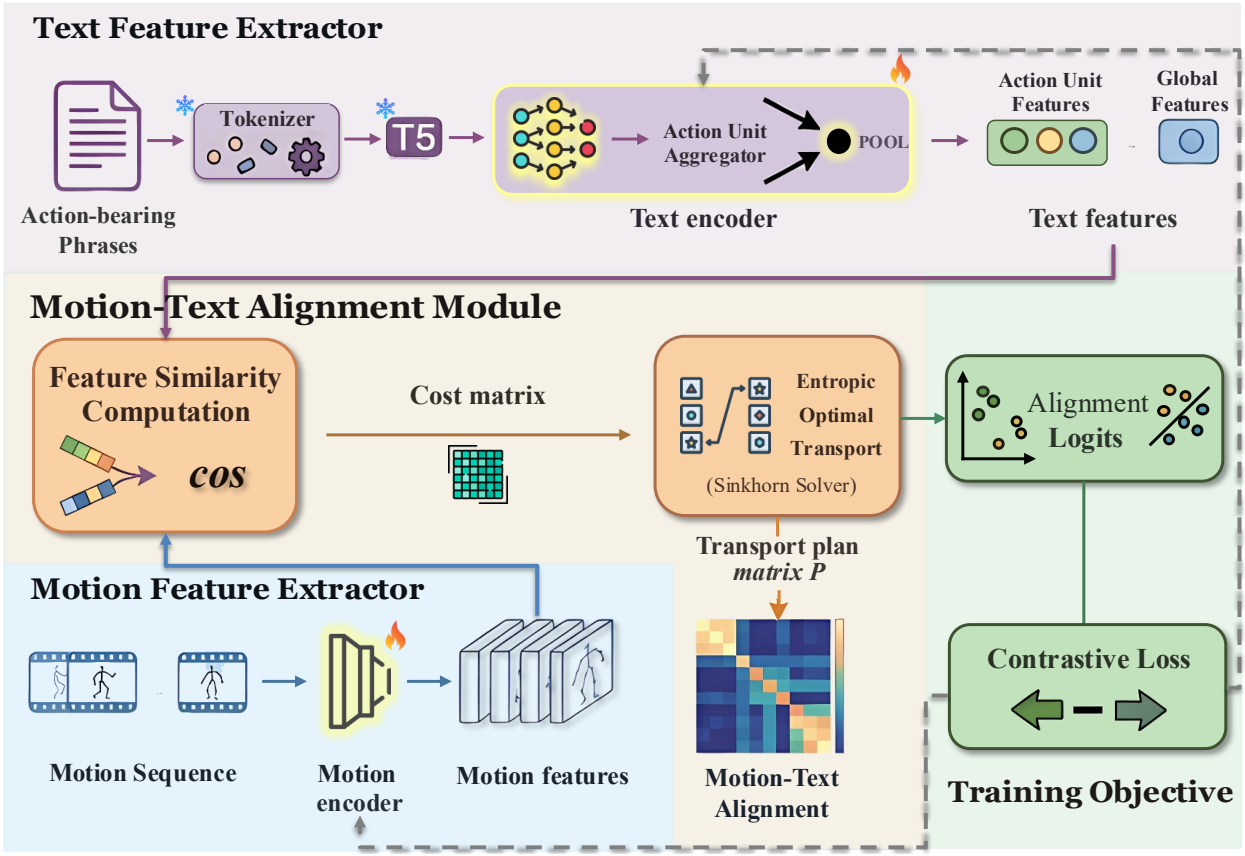

Figure 2. Overview of FineMoLA. The text stream first encodes the action-bearing phrases with a frozen T5 backbone. Then the action unit aggregator aggregates the features into action unit features and the global feature, which form a text feature set. In parallel, the motion stream extracts temporal motion features using a temporal-convolution-based encoder. Given the motion and text features, we compute a feature-similarity-based cost matrix and solve an entropic optimal transport problem with the Sinkhorn algorithm to obtain pairwise alignment logits. These logits are then used in a CLIP-style symmetric contrastive objective for training.

the contrastive objective. This global feature is appended to the action unit features to form the **text features** Z^t in Algorithm 1. On the motion stream E_m, we finetune the pre-trained VQ-VAE encoder [10], composed of temporal convolutional layers, to extract **motion features** Z^m from motion sequences.

4.4. Training Objective

As illustrated in Algorithm 1, the training process is conceptually similar to CLIP [25], in that both methods learn cross-modal representations through batch-wise contrastive supervision over paired motion–text samples. We highlight in green the steps that differ from standard CLIP training.

Given a mini-batch of paired samples $\mathcal{D}$, CLIP first computes a global representation for each sample in the two modalities and then forms a batch-wise similarity matrix, where the diagonal entries correspond to matched pairs and the off-diagonal entries correspond to mismatched pairs. A symmetric cross-entropy objective is subsequently applied to encourage high similarity for ground-truth pairs and low similarity for mismatched pairs.

Our training pipeline follows the same high-level paradigm, but differs in how the pairwise alignment score is computed. Instead of directly measuring similarity between two global embeddings, we first compute a fine-grained motion–text alignment cost using the entropy-regularized OT objective in Eq. (4). Specifically, for each pair $(m^{(i)}, t^{(j)})$ in the batch, we compute

$$L_{ij} = -\mathcal{J}_{\mathrm{EOT}}(m^{(i)}, t^{(j)}), \tag{13}$$

where a larger L_{ij} indicates stronger cross-modal alignment.

Using these pairwise alignment logits, we define the motion-to-text matching probability as

$$p(t^{(j)} \mid m^{(i)}) = \frac{\exp(L_{ij}/\tau)}{\sum_{k=1}^{B} \exp(L_{ik}/\tau)}, \tag{14}$$

and analogously define the text-to-motion matching proba-

Algorithm 1 CLIP-style Motion–Text Training with Sinkhorn-based Alignment

Require: Paired dataset $\mathcal{D} = \{(m^{(j)}, t^{(j)})\}_{j=1}^{N}$; motion feature extractor $E_m(\cdot;\theta_m)$; text feature extractor $E_t(\cdot;\theta_t)$
Ensure: Trained parameters θ_m, θ_t

for each training epoch **do**
  for each mini-batch $\{(m^{(i)}, t^{(i)})\}_{i=1}^{B}$ **do**
    for $i = 1$ to B **do**
      $Z^{m,(i)} \leftarrow E_m(m^{(i)};\theta_m)$
      $Z^{t,(i)} \leftarrow E_t(t^{(i)};\theta_t)$
      ℓ_2-normalize all feature vectors in $Z^{m,(i)}$ and $Z^{t,(i)}$
    end for
    for $i = 1$ to B **do**
      for $j = 1$ to B **do**
        Construct the cost matrix $\mathbf{C}^{(i,j)}$ from $Z^{m,(i)}$ and $Z^{t,(j)}$ according to Eq. (7)
        Compute the Sinkhorn transport plan $P^{(i,j)} \in U(\mathbf{a}, \mathbf{b})$ according to Eq. (4).
        Compute the pairwise alignment logit $L_{ij} = -\mathcal{J}_{\mathrm{EOT}}(m^{(i)}, t^{(j)})$ using Eq. (8)
      end for
    end for
    Compute motion-to-text probabilities $p(t^{(j)} \mid m^{(i)})$ by row-wise softmax over L
    Compute text-to-motion probabilities $p(m^{(i)} \mid t^{(j)})$ by column-wise softmax over L
    Compute the symmetric contrastive loss $\mathcal{L} = \frac{1}{2}(\mathcal{L}_{m2t} + \mathcal{L}_{t2m})$
    Update θ_m, θ_t by minimizing $\mathcal{L}$
  end for
end for

bility as

$$p(m^{(i)} \mid t^{(j)}) = \frac{\exp(L_{ij}/\tau)}{\sum_{k=1}^{B} \exp(L_{kj}/\tau)}, \tag{15}$$

where τ is a temperature parameter and B is the batch size.

We then optimize the model using a symmetric contrastive loss:

$$\mathcal{L}_{m2t} = -\frac{1}{B}\sum_{i=1}^{B} \log p(t^{(i)} \mid m^{(i)}), \tag{16}$$

$$\mathcal{L}_{t2m} = -\frac{1}{B}\sum_{i=1}^{B} \log p(m^{(i)} \mid t^{(i)}), \tag{17}$$

and the final training objective is

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{m2t} + \mathcal{L}_{t2m}\right). \tag{18}$$

Compared with CLIP, the key difference is that our pairwise logits are not produced by a single global embedding similarity. Instead, they are derived from a fine-grained Sinkhorn-based transport cost between motion frames and text units, allowing the model to capture local temporal correspondence while still benefiting from batch-wise contrastive supervision.

5. Experiments

5.1. Setting and Implementation Details

Dataset. We conduct our experiments on the SnapMoGen dataset [10], a comprehensive collection of high-fidelity human motion capture data paired with rich textual descriptions. The dataset comprises 34,750 training sequences and 2,012 validation sequences. All motions are originally recorded at 30 FPS. Following SnapMoGen [10], we downsample the sequences by a factor of 4, resulting in an effective frame rate of 7.5 FPS for both the motion embeddings and the alignment computation. We choose a set of 30 motion-text pairs from the test set and manually label the ground truth alignment matrix $\mathbf{P}^{gt} \in \mathbb{R}^{T_m \times T_t}$, where T_m and T_t denote the number of downsampled motion frames and atomic action segments, respectively. These manual annotations are used solely for quantitative evaluation and are never used during training. The details of the labeling process are deferred to the supplementary.

Evaluation Metrics. We evaluate the alignment accuracy by measuring the discrepancy between the transport matrix $\mathbf{P}$ and the ground-truth annotations $\mathbf{P}^{gt}$. Specifically, we compute the normalized ℓ_1 distance between the two matrices:

$$d(\mathbf{P}, \mathbf{P}^{gt}) = \frac{1}{T_m \times T_t}\sum_{i=1}^{T_m}\sum_{j=1}^{T_t} \left|\mathbf{P}_{ij} - \mathbf{P}^{gt}_{ij}\right| \tag{19}$$

where both matrices are first normalized via the Sinkhorn-Knopp algorithm to enforce doubly stochastic constraints. This ensures they satisfy the predefined row and column marginals (i.e., $\sum_j \mathbf{P}_{ij} = 1/T_m$ and $\sum_i \mathbf{P}_{ij} = 1/T_t$).

Implementation details. Our framework is implemented in PyTorch and trained on a single NVIDIA L40 GPU. Training for 500 epochs takes approximately 12 hours. We use AdamW as the optimizer with an initial learning rate of 2×10^{-4} and a batch size of 128. For text encoding, we adopt a frozen T5-base backbone followed by five trainable residual MLP blocks, which project the text features into a 512-dimensional embedding space. For Sinkhorn-based alignment, we set the entropy regularization parameter to

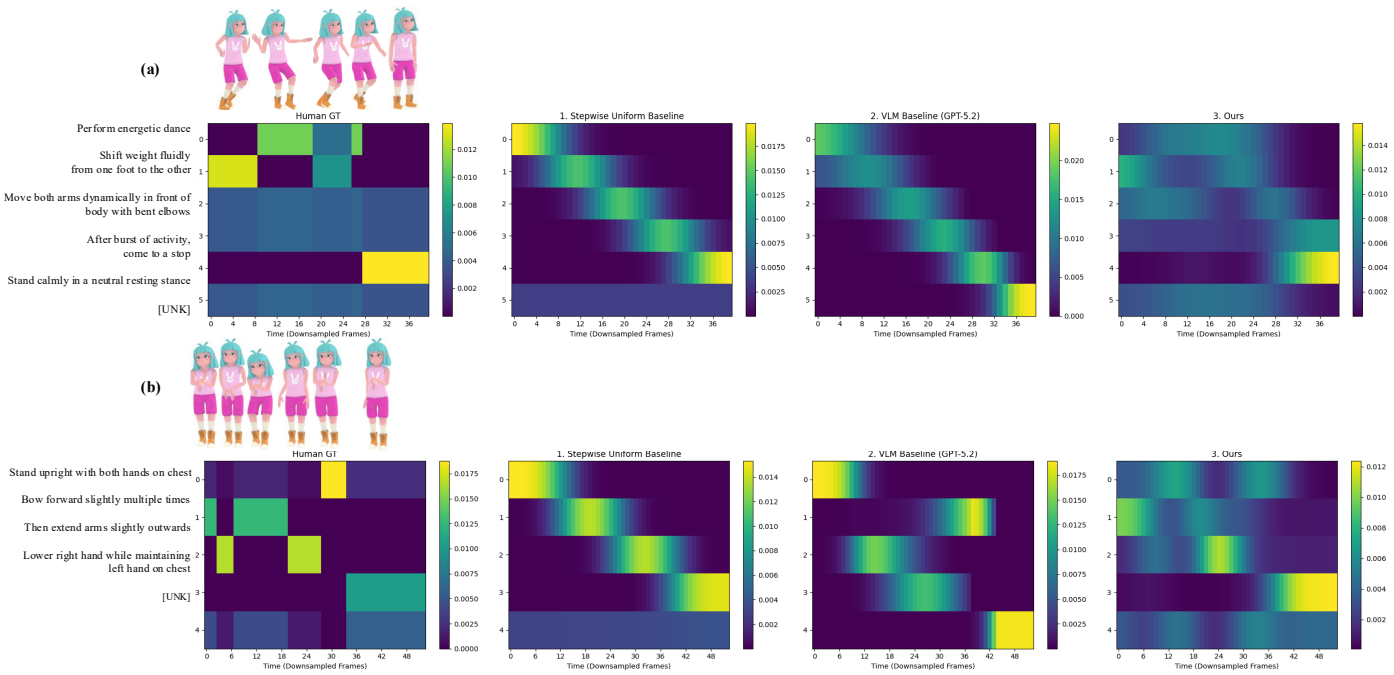

Figure 3. Qualitative comparison with baselines. We plot the ground truth transport plan and those produced by different methods. Our method produces transport plans that are structurally similar to the ground truth.

$\varepsilon = 0.1$ and use 50 forward iterations, which we find sufficient for stable convergence in practice. The contrastive temperature is set to $\tau = 0.07$. In practice, T_m is typically around 50, while T_t is typically around 5.

5.2. Comparison with Baselines

Stepwise Uniform Baseline. To demonstrate the necessity of dynamic alignment, we construct a stepwise uniform baseline, which inherently assumes a naive, sequential progression between textual actions and motion segments. Specifically, given a motion sequence of T_m frames and N textual actions, this baseline constructs an initial binary matrix by uniformly allocating exactly $\lfloor T_m/N \rfloor$ consecutive frames to each action index in chronological order. Note that the resulting matrix is not strictly block-diagonal; applying the Sinkhorn-Knopp algorithm enforces marginal distribution constraints, which naturally softens the hard boundaries and diffuses the alignment probabilities toward adjacent frames.

VLM Baseline. To establish a competitive baseline against state-of-the-art vision-language reasoning, we leverage GPT-5.2 [28] as a zero-shot temporal segmenter. We provide rendered frames sampled at 2.0 FPS and segmented action-bearing phrases to the VLM and ask it to predict the alignment. Details are provided in the supplementary material.

Results. Table 1 presents the quantitative evaluation of our proposed model. The results demonstrate that FineMoLA achieves a significant relative error reduction over the VLM and stepwise uniform baselines. These results demonstrate the necessity of our dedicated data-driven alignment framework.

5.3. Ablation

Impact of the Global Feature Mechanism. A key component of our alignment module is the global feature, which acts as an `[UNK]` token to absorb transitional or semantically ambiguous frames. In practice, frames near action boundaries often do not correspond strongly to any specific text token. Without such a token, optimal transport tends to force these frames onto neighboring action tokens, leading to blurred boundaries. During evaluation, we reassign the accumulated probability of the `[UNK]` token strictly to the preceding valid action token. By introducing the global feature, our model can naturally capture these uncertain frames and preserve cleaner temporal segmentation. As shown in Table 1, this design consistently improves alignment performance over the variant without a global token.

MLP vs. Attention-based Text Encoders. As shown in Tab. 1, replacing the lightweight MLP with a self-attention-based encoder increases the error. We conjecture that self-attention mixes excessive global context into each action token, blurring semantic boundaries and weakening temporal correspondence. In contrast, the MLP better preserves local

Table 1. Quantitative results. Values are reported in normalized ℓ_1 distance ($\times 10^{-2}$), where lower is better. GF indicates whether to use the global feature. MLP Enc. determines whether to use an MLP-based text encoder or a self-attention-based text encoder.

Method	GF	MLP Enc.	temperature	Normalized $\ell_1 \downarrow$
Stepwise Uniform Baseline				0.347
VLM Baseline (GPT-5.2)				0.296
Ours (GF + MLP encoder)	✓	✓	0.07	**0.225**
Ours (GF + MLP encoder)	✓	✓	0.01	0.230
Ours (GF + MLP encoder)	✓	✓	0.03	0.237
Ours (GF + attention encoder)	✓	×	0.07	0.275
Ours (MLP encoder)	×	✓	0.07	0.239

phrase-level semantics for precise alignment.

Sensitivity Analysis of Entropic Regularization To evaluate the robustness of OT alignment, we conduct a sensitivity analysis on the entropic regularization parameter ε. In the entropic OT framework, ε controls the trade-off between transport cost minimization and the sparsity of the alignment matrix P. As shown in Figure 4, we monitor two primary convergence metrics: (1) Marginal Error,

$$E_{marg} = \sum_i \left| \sum_j P_{ij} - a_i \right| + \sum_j \left| \sum_i P_{ij} - b_j \right| \quad (20)$$

which quantifies the satisfaction of the double-stochasticity constraints in Eq. (1), and (2) Delta P,

$$\Delta P^{(k)} = ||P^{(k)} - P^{(k-1)}||_1 = \sum_i \sum_j \left| P_{ij}^{(k)} - P_{ij}^{(k-1)} \right| \quad (21)$$

which measures the stability of the transport plan between successive iterations. Our results indicate that setting $\varepsilon = 0.01$ significantly increases the number of iterations required to reach numerical stability. Conversely, $\varepsilon = 0.1$ achieves a balance between alignment precision and training efficiency. We therefore use 50 iterations in all experiments.

5.4. Qualitative Results

Qualitative results of the frame-level text-motion alignment are illustrated in Figure 3. Compared to the VLM and stepwise uniform baselines, the transport matrix $\mathbf{P}$ generated by our model shows superior structural correspondence with the ground truth $\mathbf{P}^{gt}$. Notably, for captions containing action-bearing phrases that encompass extended temporal spans, our model effectively captures these long-range dependencies. For instance, in sample (a), the first three action phrases in $\mathbf{P}$ demonstrate a relatively uniform probability distribution across the corresponding frames. This highlights the effectiveness of our framework in modeling complex many-to-many alignments between motion and text, a challenging scenario where the VLM baseline falls short.

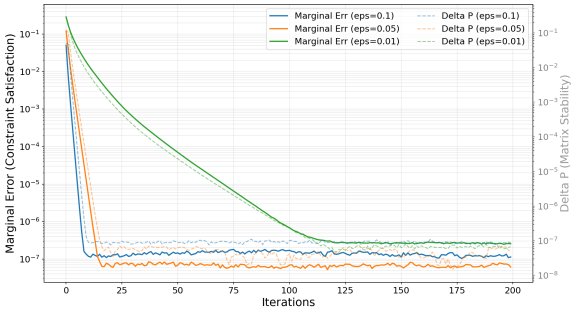

Figure 4. Sinkhorn Convergence Analysis. The plot illustrates the impact of different entropic regularization parameters (ε) on the convergence rates of Marginal Error (representing constraint satisfaction) and Delta P (representing pointwise matrix stability). The results are obtained using a randomly initialized cost matrix C.

6. Conclusion

This paper introduced FineMoLA, a weakly supervised approach for recovering fine-grained frame–phrase correspondence from long-form clip-level motion annotations. Our key idea is to cast motion–language alignment as an entropic optimal transport problem, which provides a principled way to model many-to-many correspondence between motion frames and textual units under weak supervision. Combined with LLM-based phrase decomposition and CLIP-style contrastive training, the proposed framework learns temporally grounded motion–text alignment without manual frame-level labels. Experiments on SnapMoGen show that FineMoLA yields more accurate motion–text grounding than existing baselines, and benefits from the proposed global feature mechanism. We hope this work encourages future research on fine-grained motion–language supervision.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[2] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[3] Léore Bensabath, Mathis Petrovich, and Gul Varol. A cross-dataset study for text-based 3d human motion retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1932–1940, 2024. 2, 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, page 4, 2021. 3

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 1, 2, 3

[6] Bowen Dang, Lin Wu, Xiaohang Yang, Zheng Yuan, and Zhixiang Chen. Segmo: Segment-aligned text to 3d human motion generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6946–6955, 2026. 3

[7] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 3

[8] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 1, 2

[9] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1, 3

[10] Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts. *arXiv preprint arXiv:2507.09122*, 2025. 1, 2, 5, 6

[11] Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. Amd: Autoregressive motion diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2022–2030, 2024. 1

[12] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*, pages 180–196. Springer, 2024. 3

[13] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 3

[14] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. In *2025 International Conference on 3D Vision (3DV)*, pages 240–249. IEEE, 2025. 3

[15] Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. Frankenmotion: Part-level human motion generation and composition. *arXiv preprint arXiv:2601.10909*, 2026. 1, 2

[16] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 2

[17] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. *arXiv preprint arXiv:2401.16702*, 2024. 2

[18] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27859–27871, 2025. 3

[19] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *European Conference on Computer Vision*, pages 645–662. Springer, 2022. 3

[20] 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*, pages 9488–9497, 2023. 2, 3

[21] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019. 1, 3

[22] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 3

[23] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 1, 2

[24] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 722–731, 2021. 1, 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 5

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and

Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 4

[27] Liangliang Shi, Jack Fan, and Junchi Yan. Ot-clip: Understanding and generalizing clip via optimal transport. In *Forty-first International Conference on Machine Learning*, 2024. 3

[28] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 7

[29] Justin Solomon. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*, 3, 2018. 3

[30] Shanlin Sun, Gabriel De Araujo, Jiaqi Xu, Shenghan Zhou, Hanwen Zhang, Ziheng Huang, Chenyu You, and Xiaohui Xie. Coma: Compositional human motion generation with multi-modal agents. *arXiv preprint arXiv:2412.07320*, 2024. 3

[31] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 3

[32] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 3

[33] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu. Scaling large motion models with million-level human motions. *arXiv preprint arXiv:2410.03311*, 2024. 1, 2

[34] Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27849–27858, 2025. 1, 2, 3

[35] Shiyao Xu, Benedetta Liberatori, Gül Varol, and Paolo Rota. Dense motion captioning. In *Thirteenth International Conference on 3D Vision*, 2025. 1, 2, 3

[36] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2, 3

[37] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 3

[38] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 3

[39] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36:13981–13992, 2023. 1, 3

[40] Tong Zhang, Kuofeng Gao, Jiawang Bai, Leo Yu Zhang, Xin Yin, Zonghui Wang, Shouling Ji, and Wenzhi Chen. Pre-training clip against data poisoning with optimal transport-based matching and alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9836–9849, 2025. 2, 3

[41] Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 3

FineMoLA: Towards Fine-Grained Motion-Language Alignment from Clip-Level Supervision

Supplementary Material

7. Proof of Theorem 1

Theorem 2 (Zero OT Cost under Perfect Alignment)**.** *Under Assumption 1, the OT objective $\mathcal{J}_{OT}(m,t)$ admits a feasible transport plan with zero cost. Specifically, the transport plan defined by*

$$P_{uv} = \frac{1}{T_m}\mathbb{1}\{v = \pi(u)\} \tag{22}$$

is feasible and achieves

$$\mathcal{J}_{OT}(m,t) = 0. \tag{23}$$

Proof. Under Assumption 1, for each motion frame u, there exists a unique text token $\pi(u)$ such that $s(z_u^m, z_{\pi(u)}^t) = 1$. By Eq. 7, this implies that $\mathbf{C}_{u,\pi(u)} = 0$. Now consider the transport plan

$$P_{uv} = \frac{1}{T_m}\mathbb{1}\{v = \pi(u)\}. \tag{24}$$

Since π is a bijection and $T_m = T_t$, each row and each column of P contains exactly one nonzero entry equal to $1/T_m$. Therefore, P satisfies the marginal constraints and is feasible. The total transport cost is

$$\sum_{u=1}^{T_m}\sum_{v=1}^{T_t} P_{uv}\mathbf{C}_{uv} = 0, \tag{25}$$

which is the minimum possible value. Hence, $\mathcal{J}_{OT}(m,t) = 0$. □

8. Caption Segmentation

To bridge the gap between long-form narrative descriptions and local temporal motion segments, we first decompose each caption into multiple **action-bearing phrases** using an LLM-based caption segmentation pipeline, as illustrated in Fig. 5. Specifically, GPT-4o-mini rewrites each caption into a sequence of semicolon-separated phrases, where each phrase describes a temporally coherent action while preserving the original temporal order. The prompt is illustrated in Fig. 6.

The segmented phrases are concatenated and processed by a tokenizer and a frozen T5 encoder to obtain token-level embeddings. We then apply the **action unit aggregator**, which identifies token spans corresponding to each phrase using separator tokens ("`;`") and the end-of-sequence token. For each span, token embeddings are pooled along the sequence dimension to produce compact **action unit features**.

In addition, we compute a **global features** by pooling over the entire token sequence. This global feature acts as an `[UNK]` token that absorbs motion frames not explicitly described by any phrase. The final text feature set, therefore, consists of the action unit features together with the global feature, which are used for the Sinkhorn-based motion–text alignment described in the main paper.

9. Ground Truth Annotation

To quantitatively evaluate the temporal alignment performance, we manually label the ground truth alignment matrix $\mathbf{P}^{gt} \in \mathbb{R}^{T_m \times T_t}$. The annotation process involves manual segment-level boundary identification: for each action segment j, human annotators specify the temporal interval $[t_{start}, t_{end}]$ in seconds. These intervals are subsequently mapped to discrete frame indices based on the video rendering frame rate. We construct $\mathbf{P}^{gt}$ initially as a binary indicator matrix, where $p_{i,j} = 1.0$ if frame i falls within the temporal scope of action j. To map it to $U(\mathbf{a}, \mathbf{b})$, we apply the Sinkhorn-Knopp algorithm to iteratively normalize the binary matrix. This yields a soft, doubly stochastic probability distribution in $U(\mathbf{a}, \mathbf{b})$.

10. VLM Baseline

For the VLM baseline, motion videos are first downsampled to a fixed temporal resolution of 2.0 FPS. For motion sequences spanning 30 to 40 seconds, this sampling strategy yields approximately 60 to 80 frames, which provides sufficient visual density to capture macroscopic action transitions. We utilize a structured system prompt that configures the LLM as a "specialized human motion analyst." The model receives the sequence of visual frames alongside the full unsegmented contextual caption and the explicitly parsed action sequence. To mitigate temporal hallucinations, which are prevalent in LLM-based video reasoning, we overlay the original frame indices onto the top-left corner of each sampled image. These indices serve as absolute temporal anchors, enabling the model to ground its predictions in visible spatial markers. Finally, the model outputs the localized start and end frames in a structured JSON format to ensure reliable temporal grounding.

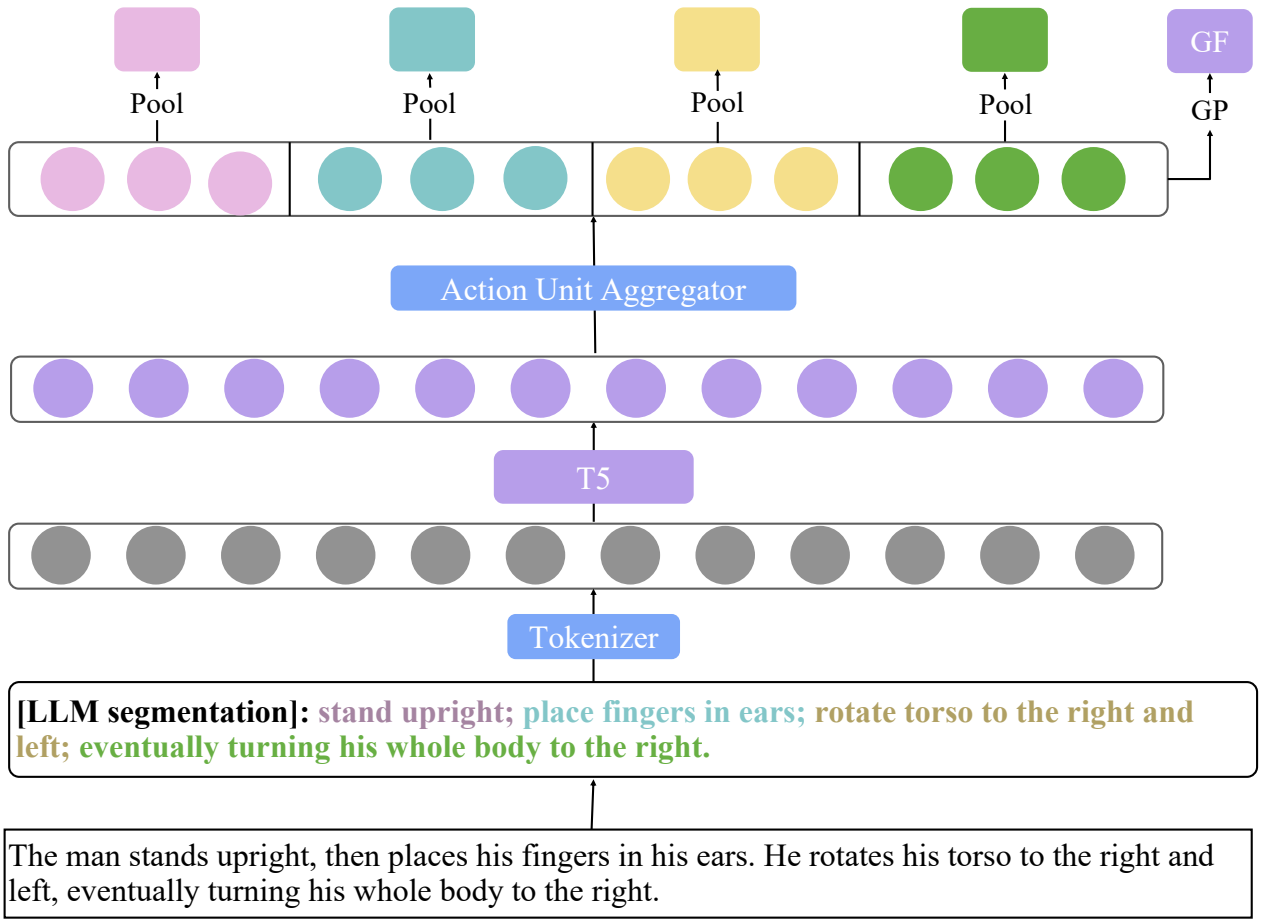

Figure 5. Overview of the **action unit aggregator**. A long-form motion caption is first segmented into multiple **action-bearing phrases** using an LLM. The concatenated phrase sequence is tokenized and encoded by a frozen T5 backbone to obtain token-level embeddings. The action unit aggregator then groups tokens belonging to the same phrase and pools them into compact **action unit features**. An additional **global feature** (GF) is obtained by global pooling (GP) over the entire sequence, which acts as an `[UNK]` token to absorb motion frames that are not explicitly described by any phrase.

System Prompt You are an expert in Motion Capture (MoCap) and computer animation. Your task is to segment complex motion captions into "Atomic Action Units" to facilitate temporal alignment for motion generation models (e.g., SnapMoGen).

Segmentation Principles:

1. **Atomization with Temporal Context:** Break down complex sentences into basic, physically continuous atomic actions.
2. **Preserve Sequence Markers:** Keep important temporal markers like "then", "soon after", "finally", or "subsequently" at the beginning of descriptions to provide directionality for alignment.
3. **Handle Concurrency:** If actions happen simultaneously (e.g., "while", "during"), include concurrency keywords in the description (e.g., "wave right hand while walking") to signify temporal overlap.
4. **Body-part Decoupling:** Separate concurrent movements of different body parts if they are distinct actions.
5. **Standardization:** Output must be a structured JSON array of objects.

Output Format Requirement: Always return a pure JSON object containing a key `"results"` which is a list of objects, each with `"caption_index"` and `"atomic_actions"` (list of objects with `"id"` and `"description"`).

Few-Shot Examples

Example Input Captions to Split:

- "The person walks forward a few steps, pauses, and then steps to her left with her left foot, standing still. She raises her arms slightly in front of her at stomach height before lowering them. She repeats the side step with her left foot and turns her head to the left."
- "The person stands with hands raised while slowly nodding their head, then starts to move his hands down while tilting his body forward."
- "The person advances quickly, bending at the waist during the walk to pick up an object, then finally stands upright while gazing at the object."

Example Segmentation Results:

```
{
  "results": [
    {
      "caption_index": 1,
      "atomic_actions": [
        {"id": 1, "description": "walk forward a few steps"},
        {"id": 2, "description": "pause and stand still"},
        {"id": 3, "description": "then step to the left with left foot"},
        {"id": 4, "description": "standing still again"},
        {"id": 5, "description": "raise arms slightly to stomach height"},
        {"id": 6, "description": "soon after, lower arms"},
        {"id": 7, "description": "repeat side step with left foot"},
        {"id": 8, "description": "simultaneously turn head to the left"}
      ]
    },
    {
      "caption_index": 2,
      "atomic_actions": [
        {"id": 1, "description": "stand with hands raised"},
        {"id": 2, "description": "slowly nod head while standing"},
        {"id": 3, "description": "then start to move hands down"},
        {"id": 4, "description": "tilt body forward while moving hands"}
      ]
    },
    {
      "caption_index": 3,
      "atomic_actions": [
        {"id": 1, "description": "advance quickly"},
        {"id": 2, "description": "bend at the waist while walking to pick up object"},
        {"id": 3, "description": "then finally stand upright"},
        {"id": 4, "description": "gaze at the object while standing"}
      ]
    }
  ]
}
```

Figure 6. The complete system prompt and contextual few-shot examples provided to the LLM for atomic action units segmentation.