G2E-T2M: GLOBAL-TO-EVENT HIERARCHICAL CONDITIONING FOR COMPOSITIONAL TEXT-TO-MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-motion (T2M) models have achieved strong caption-level alignment, yet they often fail to realize individual actions in complex multi-event descriptions. Through extractive event-level analysis of standard T2M benchmarks, we observe that descriptions containing more constituent events are generally associated with longer motions and degraded generation performance. Motivated by these observations, we propose **G2E-T2M**, a Global-to-Event hierarchical conditioning framework for compositional T2M generation. G2E-T2M first conditions motion features on the full-caption tokens through global cross-attention and then performs independent token-level cross-attention with each constituent event. The resulting event-wise feature updates are adaptively aggregated at each motion position according to motion–event relevance and a weak order prior, separating fine-grained interaction within each event from competition across events. In addition to the standard generation objective, we introduce boundary-free event-wise semantic supervision that aligns each constituent event with semantically compatible local motion regions through multi-scale temporal search without requiring event-level temporal annotations. For event-level evaluation, we introduce Event MM-Dist with average- and worst-event aggregation to measure constituent-event alignment, derive event-order consistency from localized event–motion matches, and employ a VLM-based score to assess event realization. Experiments across HumanML3D, KIT-ML, and SnapMoGen demonstrate consistent improvements in both caption- and event-level semantic alignment while maintaining competitive motion quality across different levels of event complexity.

1 INTRODUCTION

Text-to-motion (T2M) generation seeks to synthesize plausible motion from natural-language descriptions. Recent advances in autoregressive modeling Lu et al. (2025), diffusion models Ho et al. (2020); Karunratanakul et al. (2023), and flow-matching approaches Lipman et al. (2023); Huang & Celiktutan (2026) have substantially improved generation quality on standard benchmarks. However, strong caption-level performance does not necessarily imply complete realization of the described semantics. As illustrated in Fig. 1(a), a generated motion may capture the overall description while omitting or incorrectly realizing individual actions in a complex multi-event sequence.

To investigate this issue at a finer semantic granularity, we adopt the event formulation of Event-T2M Hong et al. (2026), where an event represents a semantically self-contained action or state change that is temporally distinguishable within a motion sequence. We perform an LLM-based extractive decomposition that introduces no textual content beyond the original caption, and analyze HumanML3D Guo et al. (2022), KIT-ML Plappert et al. (2016), and SnapMoGen Guo et al. (2025) from an event-level perspective. As shown in Fig. 1(b–c), descriptions containing more constituent events are generally associated with longer motion sequences and lower R@1 for both ground-truth and generated motions, with a more pronounced degradation for representative T2M methods. These trends identify higher-event descriptions as a challenging generation regime and motivate a closer examination of constituent-event modeling in compositional T2M generation.

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.



Figure 1: Event-level analysis of compositional T2M generation. (a) Qualitative illustration of CMDM Yu et al. (2026) generation for single- and multi-event descriptions, highlighting an omitted action in the multi-event case. (b) Motion sequences become longer as the number of constituent events increases in HumanML3D. (c) R@1 decreases with increasing event count for both ground-truth and generated motions, with a more pronounced drop for representative T2M methods.

Existing efforts toward complex T2M generation include enriching textual conditions with finer-grained descriptions Zhang et al. (2023b); Fan et al. (2025); Wen et al. (2025), composing long motions from sub-motions or motion primitives Barquero et al. (2024); Zhao et al. (2025), and explicitly decomposing complex descriptions into semantic units Qing et al. (2023); Hong et al. (2026); Sun et al. (2026). In particular, Event-T2M Hong et al. (2026) combines global text information with event-level conditioning, demonstrating the value of explicitly modeling constituent events. However, each event is represented by a single event-level embedding, compressing its internal textual structure before motion–event interaction. Beyond making events explicit, two questions remain. First, how can conditioning preserve the internal token-level semantics of each event while separating fine-grained interaction within an event from competition across events? Second, without event-level temporal annotations, how can individual events be semantically associated with the local motion regions in which they are realized? We address these two questions jointly through hierarchical event conditioning and boundary-free local semantic supervision.

To this end, we propose **G2E-T2M**, a **Global-to-Event hierarchical conditioning framework** for compositional T2M generation. G2E-T2M first conditions motion features on the full-caption tokens through cross-attention and then performs independent token-level cross-attention with each constituent event. The resulting event-wise feature updates are adaptively aggregated at each motion position according to motion–event relevance and a weak order prior. This hierarchical design separates fine-grained interaction within each event from competition across events while allowing their contributions to vary across motion positions. The standard generation objective does not explicitly associate individual events with the local motion regions in which they are realized. We therefore introduce **boundary-free event-wise semantic supervision**. As the temporal location and duration of an event are not uniquely determined, we perform multi-scale temporal search over the generated motion and match each event to its most semantically compatible local region using a frozen motion retrieval model Petrovich et al. (2023); Li et al. (2025a); Han et al. (2026a). This provides event-wise local semantic supervision without requiring event-level temporal annotations.

We further evaluate constituent-event realization beyond conventional caption-level metrics. Our event-level evaluation measures semantic alignment at both the average- and worst-event levels, examines the relative temporal order of localized event matches, and assesses event realization from rendered motions using a VLM Cao et al. (2026b). Experiments on HumanML3D, KIT-ML, and SnapMoGen show consistent improvements in text–motion alignment, with event-level evaluations further demonstrating improved constituent-event realization across different levels of event com-

plexity. Our contributions are threefold: (1) We propose **G2E-T2M**, a Global-to-Event hierarchical conditioning framework that preserves token-level semantics within individual events and explicitly separates within-event interaction from cross-event aggregation. (2) We introduce **boundary-free event-wise semantic supervision**, which associates constituent events with semantically compatible local motion regions through multi-scale temporal search without requiring event-level temporal annotations. (3) We conduct extensive experiments and event-level analyses across multiple benchmarks and event-complexity regimes, complemented by semantic matching, temporal-order, and VLM-based diagnostics of constituent-event realization.

2 Related Work

2.1 Text-to-Motion Generation

T2M generation has progressed substantially with increasingly expressive generative models. Early approaches mainly relied on continuous latent-variable representations Petrovich et al. (2022); Guo et al. (2022), followed by diffusion-based generation in continuous or latent spaces Zhang et al. (2024); Tevet et al. (2023); Chen et al. (2023) and discrete motion representations with autoregressive sequence modeling Zhang et al. (2023a); Jiang et al. (2023). More recent studies explore masked motion modeling Guo et al. (2024); Pinyoanuntapong et al. (2024); Guo et al. (2025), structured spatiotemporal modeling Hong et al. (2025); Cao et al. (2026a), and causal generation Yu et al. (2026), substantially improving overall generation quality and text–motion alignment. Complementary to advances in generative architectures, our work focuses on how constituent-event conditions are organized within T2M generation and how event-wise local semantic supervision is incorporated into the training objective.

2.2 Motion and Semantic Decomposition in Complex Motion Generation

Complex motion generation has been explored through both motion-sequence and semantic decomposition. Motion-sequence decomposition simplifies long-sequence synthesis by organizing motion into actions, sub-motions, or motion primitives that are progressively composed over time Zhu et al. (2024); Barquero et al. (2024); Zhao et al. (2025); Zhang et al. (2026). Semantic decomposition, in contrast, structures complex descriptions into finer-grained semantic units. Story-to-Motion Qing et al. (2023) and CoMA Sun et al. (2026) organize long descriptions into structured sub-tasks or plans, while FineMoGen Zhang et al. (2023b) introduces fine-grained spatiotemporal textual conditions. FrankenMotion Li et al. (2026) further models motion semantics at multiple granularities. More closely related to our setting, Event-T2M Hong et al. (2026) decomposes complex captions into semantic events and incorporates both global text information and event-level representations into motion generation. Event-T2M represents each constituent event with a single event-level embedding, whereas G2E-T2M retains its token-level textual representation during motion–event interaction. Our framework further separates independent token-level interaction within each event from adaptive aggregation across events and introduces event-wise local semantic supervision without requiring event-level temporal annotations.

2.3 Motion–Language Alignment and Semantic Supervision

Motion–language alignment aims to learn semantically comparable representations between textual descriptions and motions. TMR Petrovich et al. (2023) establishes a shared text–motion representation space, while CAR Fujiwara et al. (2025) and FineMoLA Wang et al. (2026) further investigate temporal and fine-grained motion–language correspondence. Recent studies also incorporate semantic alignment signals into motion generation. AToM Han et al. (2025) introduces event-level visual rewards, ReAlign Weng et al. (2026) guides diffusion sampling with step-aware rewards, and ATM Han et al. (2026b) imposes global and local semantic alignment objectives during training. These studies demonstrate the effectiveness of semantic alignment signals at different granularities. Our event-wise supervision operates between individual extracted events and candidate local motion regions whose temporal locations and durations are unknown. We address this ambiguity through multi-scale temporal search and compute the semantic objective using a frozen text–motion representation model, without requiring event-level temporal annotations.

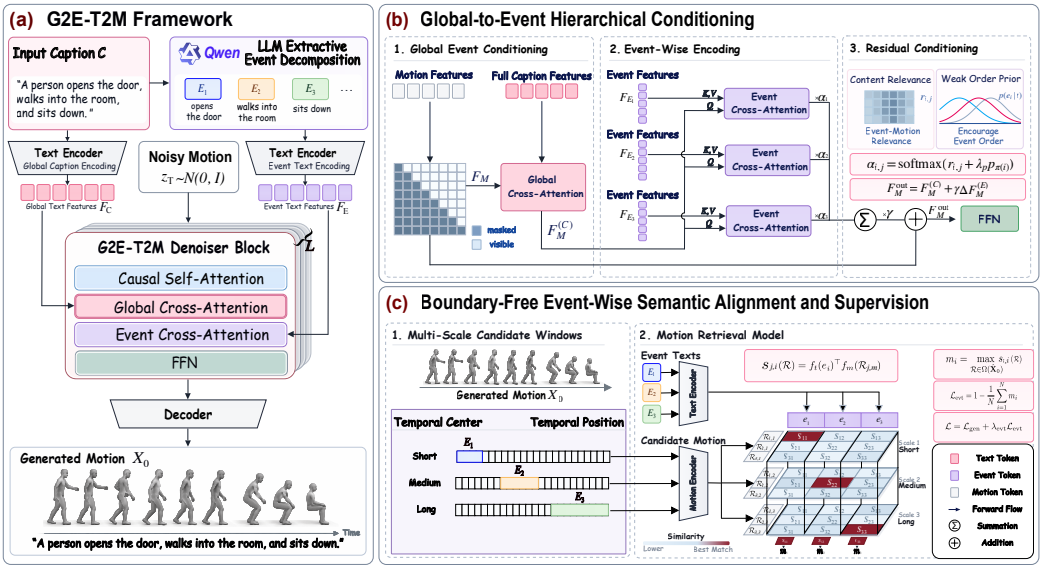

Figure 2: Overview of the proposed G2E-T2M framework. G2E-T2M combines Global-to-Event hierarchical conditioning with boundary-free event-wise semantic supervision to strengthen constituent-event realization without requiring event-level temporal annotations.

3 METHOD

We propose **G2E-T2M**, a Global-to-Event hierarchical conditioning framework for compositional T2M generation. We first formulate the multi-event generation problem and provide an overview of our G2E-T2M in Sec. 3.1. Sec. 3.2 then introduces Global-to-Event hierarchical conditioning, which progressively incorporates full-caption context and constituent-event semantics through a dual residual cross-attention scheme with adaptive event aggregation. Sec. 3.3 further presents boundary-free event-wise semantic supervision based on multi-scale local motion matching, providing local semantic supervision for constituent events without requiring temporal boundary annotations.

3.1 PROBLEM FORMULATION AND OVERVIEW

Problem Definition. Given a natural language description C, we obtain its constituent events $[e_1, \ldots, e_N]$ through extractive decomposition, preserving their order in the original caption. In this work, compositional T2M refers to realizing multiple constituent events described within a single caption, rather than explicitly evaluating systematic generalization to unseen event combinations. Our goal is to generate a motion sequence $X = [x_1, \ldots, x_T] \in \mathbb{R}^{T \times D}$ that realizes the constituent events in their intended temporal order. Unlike multi-prompt motion generation, which assumes predefined text-to-segment assignments, standard T2M benchmarks provide only caption–motion pairs without event-level temporal annotations. We therefore formulate compositional generation as $p_\theta(X \mid C, e_1, \ldots, e_N)$, where C provides the global context and $\{e_i\}_{i=1}^N$ provides fine-grained event semantics. The challenge is to generate a coherent motion sequence that preserves the semantics and ordering of the constituent events.

Overview. As illustrated in Fig. 2, G2E-T2M decomposes the input caption C into ordered constituent events $[e_1, \ldots, e_N]$ and independently encodes the full caption and each event. Global-to-Event hierarchical conditioning incorporates global context and constituent-event semantics through dual residual cross-attention with adaptive event aggregation. To provide explicit local semantic supervision for constituent events, we further introduce boundary-free event-wise semantic supervision, which associates events with semantically compatible local motion regions through multi-scale temporal search.

3.2 GLOBAL-TO-EVENT HIERARCHICAL CONDITIONING

Given the full caption C and its ordered constituent events $[e_1, e_2, \ldots, e_N]$, we independently encode them using a pretrained text encoder $\mathcal{E}$ Sanh et al. (2019); Radford et al. (2021); Ni et al. (2022), yielding token-level representations $\mathbf{F}_C \in \mathbb{R}^{T_C \times d}$ and $\mathbf{F}_{E_i} \in \mathbb{R}^{T_i \times d}$, where T_C and T_i denote the token lengths of the full caption and event e_i, respectively. A straightforward alternative is to concatenate $[\mathbf{F}_C; \mathbf{F}_{E_1}; \ldots; \mathbf{F}_{E_N}]$ and condition motion features through a single cross-attention operation. However, such flat conditioning requires the same attention space to simultaneously model fine-grained semantics within individual events and resolve competition across events (as detailed in Appendix A.2). We instead adopt a Global-to-Event hierarchy that explicitly separates these two levels of interaction. Specifically, let $\mathbf{F}_M \in \mathbb{R}^{L \times d}$ denote the motion features. Given a textual feature $\mathbf{F}_T$, we incorporate textual information through standard cross-attention Vaswani et al. (2017):

$$\mathrm{CA}(\mathbf{F}_M, \mathbf{F}_T) = \mathrm{softmax}\left(\frac{(\mathbf{F}_M\mathbf{W}_q)(\mathbf{F}_T\mathbf{W}_k)^\top}{\sqrt{d_a}}\right)(\mathbf{F}_T\mathbf{W}_v)\mathbf{W}_o, \tag{1}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$, and $\mathbf{W}_o$ are learnable projection matrices. The motion features serve as queries, while the textual features provide keys and values. We implement the Global-to-Event hierarchy through dual residual cross-attention conditioning:

$$\mathbf{F}_M^{(C)} = \mathbf{F}_M + \mathrm{CA}_C(\mathbf{F}_M, \mathbf{F}_C), \quad \Delta\mathbf{F}_{M,i}^{(E)} = \mathrm{CA}_E(\mathbf{F}_M^{(C)}, \mathbf{F}_{E_i}). \tag{2}$$

Full-caption cross-attention first conditions the motion features on the complete description, yielding $\mathbf{F}_M^{(C)}$ as the contextualized motion representation. Each event then independently interacts with $\mathbf{F}_M^{(C)}$ through token-level cross-attention, producing an event-wise feature update $\Delta\mathbf{F}_{M,i}^{(E)}$. Importantly, event tokens remain unpooled when computing the event-wise feature updates; event-level summaries are introduced for estimating the cross-event aggregation weights described below.

To further capture the relative progression of constituent events, we model their position-dependent contributions along the motion sequence. Let $\mathbf{F}_{M,l}^{(C)}$ denote the feature at the l-th temporal position of $\mathbf{F}_M^{(C)}$. We compute a content-dependent relevance score $r_{l,i} = \mathbf{q}_l^\top \bar{\mathbf{k}}_i / \sqrt{d_a}$ to characterize its semantic compatibility with event e_i, where $\mathbf{q}_l = \mathbf{F}_{M,l}^{(C)}\mathbf{W}_q^r$ and $\bar{\mathbf{k}}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{F}_{E_i,t}\mathbf{W}_k^r$ denotes the mean key representation of the valid tokens in event e_i. We further introduce a weak Gaussian order prior $\pi_{l,i} = \exp\!\left(-\frac{(p_l-\mu_i)^2}{2\sigma_p^2}\right)$, where $p_l = (l-\frac{1}{2})/L$ and $\mu_i = (i-\frac{1}{2})/N$ denote the normalized temporal position of the l-th motion feature and the relative position of the i-th event, respectively, while σ_p controls the prior bandwidth. The event-level responses are then adaptively aggregated as

$$\alpha_{l,i} = \mathrm{softmax}_i\left(r_{l,i} + \lambda_p \pi_{l,i}\right), \qquad \Delta\mathbf{F}_{M,l}^{(E)} = \sum_{i=1}^{N} \alpha_{l,i}\Delta\mathbf{F}_{M,i,l}^{(E)}. \tag{3}$$

Here, λ_p controls the strength of the order prior. The aggregated event response is subsequently fused with the globally conditioned motion representation through a residual connection, $\mathbf{F}_M^{\mathrm{out}} = \mathbf{F}_M^{(C)} + \gamma\Delta\mathbf{F}_M^{(E)}$, where γ controls the contribution of event conditioning. The position-wise weights $\alpha_{l,i}$ adapt event contributions to the evolving motion context, while the Gaussian-shaped bias provides an order-aware bias toward the intended event progression. Together, this formulation supports flexible temporal allocation of constituent events without requiring explicit event boundaries or predefined motion intervals.

3.3 BOUNDARY-FREE EVENT-WISE SEMANTIC ALIGNMENT AND SUPERVISION

The standard generation objective supervises the generated motion at the sequence level but does not explicitly associate individual constituent events with the local motion regions in which they are realized. We therefore introduce an event-wise semantic objective that encourages each event to have a semantically compatible local realization.

Direct event-level supervision, however, would require knowing where each event occurs in the motion sequence. Such temporal annotations are unavailable in standard T2M datasets, which provide

only caption–motion pairs. Moreover, event realization is temporally non-unique: the same event may occur at different locations and over different durations while preserving its semantics. We therefore formulate event supervision as a boundary-free semantic matching problem over multiple temporal locations and scales.

We first estimate the clean motion latent as $\hat{\mathbf{z}}_0 = \mathbf{z}_t + (1-t)\mathbf{v}_\theta$, and decode it into motion space as $\hat{\mathbf{X}}_0 = \mathcal{D}(\hat{\mathbf{z}}_0)$. To account for the unknown temporal location and duration of each event, we construct a multi-scale set of local motion regions. Specifically, we uniformly sample temporal locations along the motion sequence and extract candidate windows at multiple temporal scales. We denote the resulting candidate set as

$$\Omega(\hat{\mathbf{X}}_0) = \{\mathcal{R}_j\}_{j=1}^{J}, \tag{4}$$

where each $\mathcal{R}_j$ corresponds to a candidate region at a particular temporal location and scale. Regions extending beyond the motion sequence are clipped to the valid range. The resulting candidate set then covers different temporal locations and durations without prescribing event-level boundaries.

To establish event–motion alignment, we employ a frozen text–motion representation model Petrovich et al. (2023); Li et al. (2025a); Han et al. (2026a) with text encoder f_t and motion encoder f_m. Their outputs are ℓ_2-normalized such that the inner product measures cosine similarity. For each event e_i, we compute its semantic similarities with all candidate regions $\mathcal{R}_j \in \Omega(\hat{\mathbf{X}}_0)$:

$$s_{:,i}(\mathcal{R}) = f_t(e_i)^\top f_m(\mathcal{R}), \qquad m_i = \max_{\mathcal{R}\in\Omega(\hat{\mathbf{X}}_0)} s_{:,i}(\mathcal{R}), \qquad \mathcal{L}_{\text{evt}} = 1 - \frac{1}{N}\sum_{i=1}^{N} m_i. \tag{5}$$

This best-region matching provides a direct semantic learning signal for each event without requiring an explicit event-to-frame assignment. We combine it with the original generation objective, weighted by λ_{evt}:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{evt}}\mathcal{L}_{\text{evt}}. \tag{6}$$

4 Experiments

4.1 Experiment Settings

Standard Benchmarks and Multi-event Subsets. We conduct experiments on the official splits of HumanML3D Guo et al. (2022), KIT-ML Plappert et al. (2016), and SnapMoGen Guo et al. (2025). Following standard practice, motions are represented in a root-relative coordinate system using root velocities and local joint features. Following Event-T2M Hong et al. (2026), we apply LLM-based extractive event decomposition to the test captions. For HumanML3D and KIT-ML, we construct increasingly compositional subsets according to the number of constituent events, including $E \geq 2$, $E \geq 3$, and $E \geq 4$. We additionally evaluate on the full SnapMoGen test set to assess performance on longer and more compositionally complex motions. Detailed dataset statistics and subset construction procedures are provided in Appendix A.1.

Evaluation Metrics. Following standard T2M evaluation protocols Guo et al. (2022; 2025), we report **R-Precision** (R@1, R@2, and R@3) and **MM-Dist** for text–motion alignment, **FID** for distributional quality, and **Diversity** for generation variability. To assess constituent-event realization beyond caption-level alignment, we further introduce event-level metrics based on multi-scale temporal matching. Using the standard benchmark text–motion evaluator Guo et al. (2022; 2025), independently of the retrieval model used for training supervision, **Event MM-Dist** measures each event by its minimum text–motion distance over candidate local motion regions. We report its average (**Avg-E**) and worst-event (**Worst-E**) variants, and derive **Order** from the locations of the best-matched regions as a coarse measure of relative event ordering. Details are provided in Appendix A.3. We additionally report **VLM-E** using a frozen vision-language model to assess event occurrence directly from rendered motions, with details provided in Appendix A.4. Lower Avg-E and Worst-E, and higher Order and VLM-E, indicate better event-level performance.

Implementation Details. We adopt CMDM Yu et al. (2026) as the base generator due to its extensible Causal-DiT architecture for long-sequence motion modeling. Its backbone consists of

Table 1: Quantitative comparison on the HumanML3D and KIT-ML multi-event test subsets ($E \geq 2$). Results are reported as mean performance with 95% confidence intervals ($\pm$) over 20 runs. For Diversity ($\rightarrow$), values closer to GT are better. The best result is highlighted in **bold**, and the second-best is underlined.

Dataset	Method	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →
		R@1	R@2	R@3			
HumanML3D	GT	$0.523^{\pm.000}$	$0.716^{\pm.000}$	$0.810^{\pm.000}$	$0.000^{\pm.000}$	$2.971^{\pm.000}$	$9.354^{\pm.000}$
	T2M Guo et al. (2022)	$0.440^{\pm.002}$	$0.620^{\pm.004}$	$0.724^{\pm.005}$	$1.179^{\pm.028}$	$3.436^{\pm.015}$	$9.138^{\pm.061}$
	MotionDiffuse Zhang et al. (2024)	$0.493^{\pm.003}$	$0.682^{\pm.005}$	$0.781^{\pm.003}$	$0.892^{\pm.017}$	$3.074^{\pm.012}$	$9.624^{\pm.044}$
	MDM Tevet et al. (2023)	$0.389^{\pm.007}$	$0.577^{\pm.008}$	$0.684^{\pm.008}$	$0.314^{\pm.010}$	$3.722^{\pm.027}$	$9.595^{\pm.081}$
	T2M-GPT Zhang et al. (2023a)	$0.467^{\pm.003}$	$0.649^{\pm.005}$	$0.745^{\pm.007}$	$0.154^{\pm.008}$	$3.292^{\pm.017}$	$9.513^{\pm.058}$
	MoMask Guo et al. (2024)	$0.508^{\pm.004}$	$0.704^{\pm.003}$	$0.793^{\pm.003}$	$\mathbf{0.062}^{\pm.003}$	$3.047^{\pm.007}$	$\underline{9.475}^{\pm.029}$
	SALAD Hong et al. (2025)	$0.572^{\pm.005}$	$0.765^{\pm.005}$	$0.852^{\pm.005}$	$0.100^{\pm.006}$	$2.722^{\pm.013}$	$9.542^{\pm.056}$
	Event-T2M Hong et al. (2026)	$0.556^{\pm.005}$	$0.754^{\pm.004}$	$0.846^{\pm.003}$	$0.112^{\pm.003}$	$2.766^{\pm.006}$	$\mathbf{9.318}^{\pm.022}$
	MoScale Zheng et al. (2026)	$0.525^{\pm.004}$	$0.721^{\pm.004}$	$0.813^{\pm.002}$	$0.068^{\pm.002}$	$2.888^{\pm.008}$	$9.230^{\pm.025}$
	TriC-Motion Cao et al. (2026a)	$0.564^{\pm.002}$	$0.763^{\pm.003}$	$0.850^{\pm.003}$	$0.376^{\pm.007}$	$2.779^{\pm.007}$	$9.584^{\pm.032}$
	CMDM Yu et al. (2026)	$\underline{0.576}^{\pm.007}$	$\underline{0.766}^{\pm.003}$	$\underline{0.853}^{\pm.003}$	$0.098^{\pm.005}$	$\underline{2.706}^{\pm.008}$	$9.572^{\pm.026}$
	G2E-T2M	$\mathbf{0.586}^{\pm.003}$	$\mathbf{0.777}^{\pm.003}$	$\mathbf{0.861}^{\pm.003}$	$\underline{0.065}^{\pm.003}$	$\mathbf{2.653}^{\pm.011}$	$9.485^{\pm.039}$
KIT-ML	GT	$0.299^{\pm.011}$	$0.488^{\pm.013}$	$0.614^{\pm.018}$	$0.000^{\pm.000}$	$3.076^{\pm.000}$	$10.390^{\pm.000}$
	T2M Guo et al. (2022)	$0.245^{\pm.019}$	$0.399^{\pm.018}$	$0.517^{\pm.014}$	$7.531^{\pm.627}$	$4.235^{\pm.132}$	$10.643^{\pm.088}$
	MotionDiffuse Zhang et al. (2024)	$0.316^{\pm.018}$	$0.483^{\pm.022}$	$0.586^{\pm.012}$	$4.653^{\pm.104}$	$3.459^{\pm.029}$	$10.773^{\pm.043}$
	MDM Tevet et al. (2023)	$0.274^{\pm.015}$	$0.447^{\pm.017}$	$0.584^{\pm.021}$	$1.102^{\pm.094}$	$3.523^{\pm.055}$	$\underline{10.526}^{\pm.096}$
	T2M-GPT Zhang et al. (2023a)	$0.293^{\pm.023}$	$0.484^{\pm.016}$	$0.604^{\pm.016}$	$1.149^{\pm.181}$	$3.389^{\pm.090}$	$10.564^{\pm.108}$
	SALAD Hong et al. (2025)	$\underline{0.351}^{\pm.018}$	$0.525^{\pm.016}$	$0.647^{\pm.011}$	$\underline{0.504}^{\pm.021}$	$3.130^{\pm.049}$	$10.555^{\pm.050}$
	Event-T2M Hong et al. (2026)	$0.312^{\pm.018}$	$0.499^{\pm.013}$	$0.614^{\pm.014}$	$0.537^{\pm.028}$	$3.083^{\pm.031}$	$10.593^{\pm.049}$
	MoScale Zheng et al. (2026)	$0.307^{\pm.015}$	$0.504^{\pm.020}$	$0.633^{\pm.020}$	$2.001^{\pm.078}$	$3.288^{\pm.046}$	$10.986^{\pm.060}$
	TriC-Motion Cao et al. (2026a)	$0.348^{\pm.023}$	$\underline{0.528}^{\pm.022}$	$\underline{0.649}^{\pm.029}$	$0.583^{\pm.037}$	$\underline{2.853}^{\pm.052}$	$10.226^{\pm.090}$
	CMDM Yu et al. (2026)	$0.334^{\pm.026}$	$0.515^{\pm.018}$	$0.637^{\pm.011}$	$0.601^{\pm.046}$	$2.940^{\pm.046}$	$\mathbf{10.384}^{\pm.118}$
	G2E-T2M	$\mathbf{0.387}^{\pm.020}$	$\mathbf{0.591}^{\pm.017}$	$\mathbf{0.707}^{\pm.012}$	$\mathbf{0.458}^{\pm.019}$	$\mathbf{2.723}^{\pm.024}$	$10.714^{\pm.066}$

a $4\times$ causal MAC-VAE and an 8-layer Causal-DiT with hidden dimension 512. The full caption and constituent events are independently encoded using a frozen DistilBERT encoder. For boundary-free event-wise supervision, we employ a frozen TMR Petrovich et al. (2023) model and set $\lambda_{\text{evt}} = 0.025$. We optimize the model using AdamW with a peak learning rate of 1.6×10^{-4} and an effective batch size of 2048. During inference, classifier-free guidance is applied with a scale of 3.0, and motion samples are generated using 50-step Euler sampling.

4.2 Experimental Results

Quantitative Comparison. We compare G2E-T2M with representative T2M methods on the multi-event subsets of HumanML3D and KIT-ML, as well as the full SnapMoGen benchmark, with results reported in Tab. 1 and Tab. 2. On HumanML3D, G2E-T2M achieves the best R@1, R@2, R@3, and MM-Dist, while obtaining the second-best FID. On KIT-ML, it achieves the best performance across all R-Precision, FID, and MM-Dist metrics. The advantage further extends to the substantially longer and more compositionally complex motions in SnapMoGen, where G2E-T2M achieves the best R@1, R@2, R@3, and FID, together with the second-best CLIP Score. These results indicate that Global-to-Event conditioning improves semantic alignment across different levels of motion and textual complexity while maintaining competitive generation quality.

Event-level Evaluation. Beyond global text–motion evaluation, we further assess whether individual events are effectively realized in the generated motions using Avg-E, Worst-E, and VLM-E. As shown in Tab. 3, G2E-T2M consistently improves event-level realization across both HumanML3D and KIT-ML, achieving the best Avg-E and VLM-E on both datasets, together with the best or tied-best Worst-E. The consistent improvements under both motion–text matching and VLM-based evaluation indicate that the gains in global semantic alignment are accompanied by stronger realization of individual constituent events. In addition, we further use the temporal locations obtained from best-window matching to coarsely estimate the ordering of constituent events. G2E-T2M achieves the best Order score on both multi-event HumanML3D and KIT-ML, suggesting better preservation of the temporal structure among multiple events. Since the localized windows

Table 2: Quantitative comparison on the SnapMoGen test set.

Method	R@1 ↑	R@2 ↑	R@3 ↑	FID ↓	CLIP Score ↑	Diversity →
GT	$0.943^{\pm.001}$	$0.977^{\pm.001}$	$0.985^{\pm.001}$	$0.001^{\pm.000}$	$0.837^{\pm.000}$	$19.736^{\pm.047}$
MDM Tevet et al. (2023)	$0.503^{\pm.002}$	$0.653^{\pm.002}$	$0.727^{\pm.002}$	$57.783^{\pm.092}$	$0.481^{\pm.001}$	–
T2M-GPT Zhang et al. (2023a)	$0.618^{\pm.002}$	$0.773^{\pm.002}$	$0.812^{\pm.002}$	$32.629^{\pm.087}$	$0.573^{\pm.011}$	–
MoMask Guo et al. (2024)	$0.777^{\pm.002}$	$0.888^{\pm.002}$	$0.927^{\pm.002}$	$17.404^{\pm.051}$	$0.664^{\pm.001}$	–
MoMask++ Guo et al. (2025)	$0.789^{\pm.002}$	$0.894^{\pm.003}$	$0.931^{\pm.002}$	$\underline{14.583}^{\pm.106}$	$\mathbf{0.681}^{\pm.001}$	$19.810^{\pm.021}$
MoScale Zheng et al. (2026)	$0.776^{\pm.003}$	$0.889^{\pm.003}$	$0.927^{\pm.002}$	$16.700^{\pm.072}$	$0.675^{\pm.001}$	$19.948^{\pm.027}$
SALAD Hong et al. (2025)	$0.768^{\pm.005}$	$0.881^{\pm.004}$	$0.921^{\pm.004}$	$25.142^{\pm.186}$	$0.650^{\pm.002}$	$\mathbf{19.735}^{\pm.037}$
Event-T2M Hong et al. (2026)	$0.771^{\pm.005}$	$0.883^{\pm.002}$	$0.922^{\pm.002}$	$19.301^{\pm.046}$	$0.658^{\pm.001}$	$19.877^{\pm.019}$
TriC-Motion Cao et al. (2026a)	$0.743^{\pm.004}$	$0.864^{\pm.005}$	$0.910^{\pm.004}$	$29.134^{\pm.219}$	$0.632^{\pm.001}$	$\underline{19.701}^{\pm.055}$
CMDM Yu et al. (2026)	$\underline{0.824}^{\pm.002}$	$\underline{0.918}^{\pm.004}$	$\underline{0.951}^{\pm.002}$	$15.008^{\pm.074}$	$0.644^{\pm.001}$	$19.895^{\pm.061}$
G2E-T2M	$\mathbf{0.853}^{\pm.003}$	$\mathbf{0.936}^{\pm.002}$	$\mathbf{0.964}^{\pm.002}$	$\mathbf{13.972}^{\pm.046}$	$\underline{0.678}^{\pm.001}$	$19.892^{\pm.019}$

Table 3: Quantitative comparison on the multi-event subsets ($E \geq 2$) of HumanML3D and KIT-ML. Avg-E/Worst-E denote average/worst best-window event MM-Dist, Order denotes temporal-order consistency (%), and VLM-E denotes the VLM-based event observability score. Bold and underline indicate the best and second-best generated results.

Method	HumanML3D ($E \geq 2$)				KIT-ML ($E \geq 2$)			
	Avg-E ↓	Worst-E ↓	VLM-E ↑	Order ↑	Avg-E ↓	Worst-E ↓	VLM-E ↑	Order ↑
GT	$4.306^{\pm.000}$	$5.568^{\pm.000}$	$78.09^{\pm.00}$	$74.02^{\pm.00}$	$4.144^{\pm.000}$	$5.519^{\pm.000}$	$68.95^{\pm.00}$	$68.51^{\pm.00}$
T2M	$4.724^{\pm.014}$	$6.050^{\pm.023}$	$64.85^{\pm.77}$	$65.90^{\pm.40}$	$4.322^{\pm.040}$	$5.779^{\pm.071}$	$60.76^{\pm.63}$	$61.81^{\pm.71}$
MotionDiffuse	$4.568^{\pm.007}$	$5.877^{\pm.010}$	$68.00^{\pm.49}$	$64.26^{\pm.56}$	$4.436^{\pm.012}$	$5.814^{\pm.019}$	$63.42^{\pm.46}$	$66.75^{\pm.27}$
MDM	$4.778^{\pm.016}$	$6.211^{\pm.023}$	$65.04^{\pm.53}$	$62.89^{\pm.26}$	$4.377^{\pm.043}$	$5.672^{\pm.046}$	$63.39^{\pm.00}$	$65.64^{\pm.89}$
T2M-GPT	$4.534^{\pm.011}$	$5.928^{\pm.016}$	$70.22^{\pm.73}$	$64.87^{\pm.36}$	$4.360^{\pm.050}$	$5.783^{\pm.063}$	$62.48^{\pm.65}$	$64.85^{\pm.55}$
MoMask	$4.345^{\pm.007}$	$5.666^{\pm.011}$	$72.53^{\pm.54}$	$66.14^{\pm.31}$	$4.175^{\pm.033}$	$5.561^{\pm.059}$	$67.11^{\pm.40}$	$62.25^{\pm.54}$
SALAD	$\underline{4.275}^{\pm.014}$	$\underline{5.593}^{\pm.011}$	$\underline{75.94}^{\pm.32}$	$\underline{72.04}^{\pm.49}$	$4.194^{\pm.022}$	$5.612^{\pm.032}$	$62.30^{\pm.67}$	$65.17^{\pm.61}$
Event-T2M	$4.358^{\pm.004}$	$5.645^{\pm.006}$	$75.91^{\pm.20}$	$69.34^{\pm.48}$	$4.080^{\pm.016}$	$\mathbf{5.344}^{\pm.038}$	$61.31^{\pm.79}$	$\underline{67.04}^{\pm.70}$
MoScale	$4.365^{\pm.004}$	$5.704^{\pm.006}$	$70.86^{\pm.80}$	$69.26^{\pm.48}$	$4.150^{\pm.024}$	$\underline{5.508}^{\pm.034}$	$66.29^{\pm.32}$	$60.20^{\pm.77}$
TriC-Motion	$4.504^{\pm.005}$	$5.862^{\pm.006}$	$74.50^{\pm.87}$	$71.81^{\pm.53}$	$\underline{4.079}^{\pm.020}$	$5.522^{\pm.039}$	$\underline{67.66}^{\pm.61}$	$66.32^{\pm.71}$
CMDM	$4.345^{\pm.008}$	$5.701^{\pm.012}$	$74.67^{\pm.48}$	$71.13^{\pm.48}$	$4.190^{\pm.017}$	$5.596^{\pm.035}$	$66.59^{\pm.82}$	$66.51^{\pm.57}$
G2E-T2M	$\mathbf{4.270}^{\pm.007}$	$\mathbf{5.564}^{\pm.024}$	$\mathbf{76.03}^{\pm.64}$	$\mathbf{73.12}^{\pm.49}$	$\mathbf{4.057}^{\pm.017}$	$\mathbf{5.344}^{\pm.016}$	$\mathbf{68.18}^{\pm.64}$	$\mathbf{67.76}^{\pm.35}$

may overlap or remain ambiguous even for ground-truth motions, we regard Order as a relative temporal diagnostic rather than an absolute ordering accuracy. Notably, Event-T2M also performs favorably on the event-level metrics, providing further evidence that explicitly modeling constituent events is beneficial for compositional T2M generation.

Qualitative Comparison. Figure 3 shows representative multi-event generation results on HumanML3D. Compared with competing methods, G2E-T2M more consistently realizes the full sequence of constituent events, while other methods may omit intermediate or final actions, such as touching the toes, turning clockwise, or returning before squatting. These examples qualitatively demonstrate the advantage of explicitly modeling constituent-event semantics and their temporal progression for complex motion generation.

4.3 Ablation Study

Effect of Global-to-Event Conditioning. Table 4 first examines whether the gains simply arise from additional conditioning capacity. Adding an extra full-caption cross-attention branch already yields moderate improvements over the baseline. Replacing this duplicated global conditioning with the proposed Full→Event hierarchy further improves both caption-level and event-level semantic alignment. In particular, even without auxiliary supervision, Global-to-Event conditioning improves MM-Dist, Avg-E, and Worst-E while achieving the best FID among all variants. These results suggest that the gains are not merely due to additional model capacity, but benefit from explicitly separating global contextual modeling from constituent-event interaction.

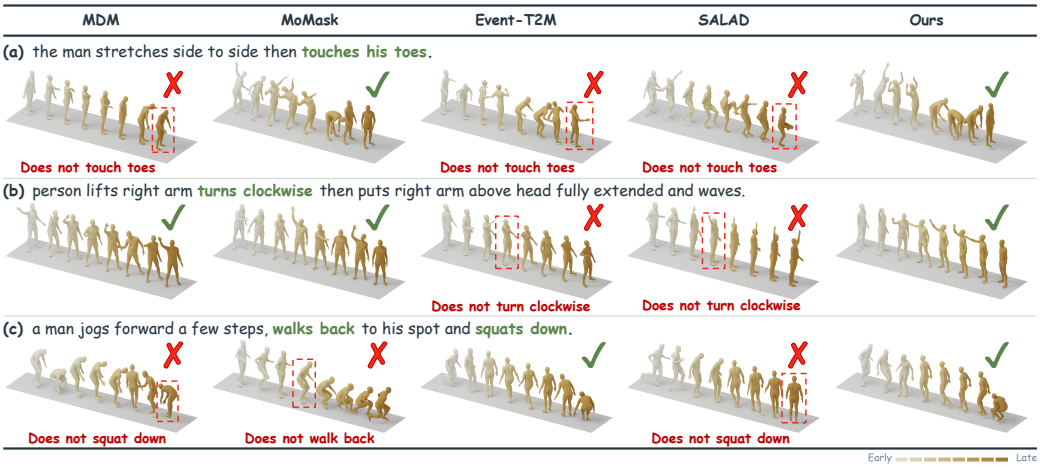

Figure 3: Qualitative comparisons on the HumanML3D dataset.

Table 4: Ablation of Global-to-Event conditioning and auxiliary semantic supervision on HumanML3D ($E \geq 2$). Extra Full CA denotes an additional full-caption cross-attention branch, while Event CA denotes constituent-event cross-attention following full-caption conditioning. Bold and underline indicate the best and second-best results, respectively.

Variant	Extra Full CA	Event CA	CA Order	Aux. Sup.	R@1 ↑	FID ↓	MM-Dist ↓	Avg-E ↓	Worst-E ↓
Baseline			–	None	0.576	0.098	2.706	4.345	5.701
+ Extra Full CA	✓		–	None	0.578	<u>0.062</u>	2.692	4.311	5.597
+ Extra Full CA	✓		–	Event–local	0.579	0.065	2.680	4.283	5.608
+ Event CA		✓	Full→Event	None	0.579	**0.060**	2.665	4.276	5.592
+ Event CA		✓	Full→Event	Caption–full	<u>0.581</u>	0.063	<u>2.657</u>	<u>4.273</u>	<u>5.590</u>
+ Event CA		✓	Full→Event	Event–full	0.580	0.070	2.658	4.270	5.593
G2E-T2M		✓	Full→Event	Event–local	**0.586**	0.065	**2.653**	**4.270**	**5.564**

Effect of Boundary-Free Event-Wise Supervision. We further compare different auxiliary objectives under the same Global-to-Event conditioning architecture. Caption–full supervision provides consistent improvements, while event–full supervision directly aligns each event with the complete motion and yields weaker event-level matching together with a larger FID. In contrast, the proposed event–local objective searches for the best-matched local motion region for each event and achieves the best R@1, MM-Dist, Avg-E, and Worst-E while maintaining competitive FID. This comparison supports the use of boundary-free local matching for constituent events whose temporal locations and durations are not predefined.

5 Conclusion

We introduced **G2E-T2M**, a Global-to-Event hierarchical conditioning framework for compositional T2M generation. G2E-T2M combines dual residual cross-attention with boundary-free event-wise semantic supervision to preserve global compositional context while strengthening constituent-event semantics. Experiments on HumanML3D, KIT-ML, and SnapMoGen demonstrate improved text–motion alignment and competitive motion quality across multi-event and long-form generation settings. Event-level evaluation further shows stronger constituent-event realization and improved temporal organization. Overall, our results highlight the importance of explicitly modeling and evaluating constituent events beyond sentence-level semantics for complex T2M generation.

AI USE DISCLOSURE

Generative AI tools were used to assist with language polishing and literature retrieval and discovery. We also used large language models to generate synthetic annotations through the extractive event decomposition described in this paper. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

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*, pp. 457–469, 2024.

Yiyang Cao, Yunze Deng, Ziyu Lin, Bin Feng, Xinggang Wang, Wenyu Liu, DanDan Zheng, and Jingdong Chen. TriC-Motion: Tri-domain causal modeling grounded text-to-motion generation. In *International Conference on Learning Representations*, pp. 61419–61436, 2026a.

Yong Cao, Chuqiao Li, Xianghui Xie, Gerard Pons-Moll, and Andreas Geiger. NextMotionQA: Benchmarking and judging human motion understanding with vision-language models. *arXiv preprint arXiv:2606.04773*, 2026b. URL https://arxiv.org/abs/2606.04773.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.

Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13336–13348, 2025.

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models. In *Computer Vision – ECCV 2024*, pp. 323–339, 2025. doi: 10.1007/978-3-031-73636-0_19.

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*, pp. 5152–5161, 2022.

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*, pp. 1900–1910, 2024.

Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. SnapMoGen: Human motion generation from expressive texts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99939–99955, 2025. doi: 10.52202/085713-3341.

Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, and Xiu Li. AToM: Aligning text-to-motion model at event-level with GPT-4Vision reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22746–22755, 2025. doi: 10.1109/CVPR52734.2025.02118.

Ke Han, Yueming Lyu, and Nicu Sebe. Motion-aligned word embeddings for text-to-motion generation. In *International Conference on Learning Representations*, pp. 28276–28289, 2026a.

Ke Han, Yueming Lyu, Weichen Yu, and Nicu Sebe. ATM: Enhanced alignment for text-to-motion generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6862–6872, 2026b. doi: 10.1109/WACV61042.2026.00663.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. SALAD: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7158–7168, 2025.

Seong-Eun Hong, Jaeyoung Seon, Juyeong Hwang, JongHwan Shin, and HyeongYeop Kang. Event-T2M: Event-level conditioning for complex text-to-motion synthesis. In *International Conference on Learning Representations*, pp. 37170–37197, 2026.

Guanhe Huang and Oya Celiktutan. Unified number-free text-to-motion generation via flow matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23570–23580, 2026.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20067–20079, 2023. doi: 10.52202/075280-0880.

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.

Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. FrankenMotion: Part-level human motion generation and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16420–16431, 2026.

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*, pp. 84238–84250, 2025a.

Zhengdao Li, Siheng Wang, Zeyu Zhang, and Hao Tang. ReMoMask: Retrieval-augmented masked motion generation. *arXiv preprint arXiv:2508.02605*, 2025b. URL https://arxiv.org/abs/2508.02605.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. ScaMo: Exploring the scaling law in autoregressive motion generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27872–27882, 2025.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, 2022. doi: 10.18653/v1/2022.findings-acl.146.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *Computer Vision – ECCV 2022*, pp. 480–497, 2022. doi: 10.1007/978-3-031-20047-2_28.

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*, pp. 9488–9497, 2023. doi: 10.1109/ICCV51070.2023.00870.

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*, pp. 1546–1555, 2024.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. doi: 10.1089/big.2016.0028.

Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. Story-to-motion: Synthesizing infinite and controllable character animation from long text. In *SIGGRAPH Asia 2023 Technical Communications*, pp. 1–4, 2023. doi: 10.1145/3610543.3626176. Article 28.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

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. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Shanlin Sun, Jiaqi Xu, Gabriel de Araujo, Shenghan Zhou, Hanwen Zhang, Ziheng Huang, Chenyu You, and Xiaohui Xie. CoMA: Compositional human motion generation with multi-modal agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9206–9214, 2026. doi: 10.1609/aaai.v40i11.37878.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tongyan Wang, Zhengyuan Li, Muhan Lin, Shengyang Luo, Yifan Shen, Aniket Bera, Baijian Yang, and Yingjie Victor Chen. FineMoLA: Towards fine-grained motion-language alignment from clip-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5510–5519, 2026.

Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, Jinkun Guo, Dongyue An, Han Liu, Yanyu Tong, Chao Zhang, Qing Guo, Juan Chen, Qiao Zhang, Youyi Zhang, Zihao Yao, Cheng Zhang, Hong Duan, Xiaoping Wu, Qi Chen, Fei Cheng, Liang Dong, Peng He, Hao Zhang, Jiaxin Lin, Chao Zhang, Zhongyi Fan, Yifan Li, Zhichao Hu, Yuhong Liu, Linus, Jie Jiang, Xiaolong Li, and Linchao Bao. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025. URL https://arxiv.org/abs/2512.23464.

Wanjiang Weng, Xiaofeng Tan, Junbo Wang, Guo-Sen Xie, Pan Zhou, and Hongsong Wang. ReAlign: Text-to-motion generation via step-aware reward-guided alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10621–10629, 2026. doi: 10.1609/aaai.v40i13.38035.

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

Qing Yu, Akihisa Watanabe, and Kent Fujiwara. Causal motion diffusion models for autoregressive motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38366–38375, 2026.

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*, pp. 14730–14740, 2023a.

Jianrong Zhang, Hehe Fan, and Yi Yang. Towards decompositional human motion generation with energy-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30650–30660, 2026.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-grained spatio-temporal motion generation and editing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13981–13992, 2023b. doi: 10.52202/075280-0616.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. doi: 10.1109/tpami.2024.3355414.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *International Conference on Learning Representations*, pp. 23569–23592, 2025.

Zhiwei Zheng, Shibo Jin, Lingjie Liu, and Mingmin Zhao. Next-scale autoregressive models for text-to-motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16376–16386, 2026.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2024. doi: 10.1109/TPAMI.2023.3330935.

A APPENDIX

This appendix provides additional experimental and implementation details supporting the main paper. We first describe dataset preparation, extractive event annotation, and the construction of the evaluation cohorts in Sec. A.1. We then present further ablations of the proposed global-to-event conditioning and event-wise supervision in Sec. A.2, followed by details of the multi-scale event-matching protocol in Sec. A.3 and the VLM-based event evaluation in Sec. A.4. We then analyze performance across different levels of event complexity in Sec. A.5. Finally, we provide pseudocode of the complete training procedure in Sec. A.6 and additional qualitative comparisons in Sec. A.7.

A.1 DATA PREPARATION AND EVALUATION DETAILS

Datasets and counting conventions. We conduct experiments on the official splits of HumanML3D Guo et al. (2022), KIT-ML Plappert et al. (2016), and SnapMoGen Guo et al. (2025). For HumanML3D, we retain samples belonging to the official split IDs, while for KIT-ML we use the released HumanML-style preprocessed data. Table 5 summarizes the basic statistics of the three datasets, including their motion representations and the proportion of captions containing multiple events. Caption counts follow the released annotation records and therefore include multiple descriptions, timed sub-captions, and mirrored counterparts. For SnapMoGen, we exclude four empty annotation slots from the 4,004 test records returned by the evaluation loader, resulting in 4,000 valid records used for generation and evaluation. The motion counts in Table 5 refer to the base motion sequences before mirrored augmentation, whereas the motion-length distributions in Fig. 4 are computed over the preprocessed motion records used in our data pipeline and therefore include mirrored counterparts where applicable.

Table 5: Dataset characteristics and corpus-level event statistics.

Dataset	Motions	Frame rate (fps)	Joints	Feature dimension	Multi-event captions (%)
HumanML3D	14,614	20	22	263	46.41
KIT-ML	3,009	12.5	21	251	25.11
SnapMoGen	20,450	30	24	296	99.28

Extractive event annotation. We decompose each nonempty caption into an ordered sequence of nonoverlapping, contiguous text spans, where each span describes an observable action or body-state transition. The decomposition is strictly extractive: every event is copied directly from the original caption, without paraphrasing or introducing inferred content. Sequential predicates are separated when they describe distinct events, whereas concurrent actions and manner expressions are kept together when further splitting would require rewriting the text or leave incomplete fragments. If a caption does not admit a reliable decomposition, we keep the full caption as a single event. We use these decompositions as weak text-level annotations rather than human-verified action labels or temporal boundaries. For all three datasets, event annotations are generated using Qwen3.8-27B. We subsequently verify that all extracted spans occur verbatim in the source caption, do not overlap, and preserve their original order.

Decomposition examples. Table 6 provides representative examples of the extractive event decomposition. Each constituent event is copied directly from a contiguous span of the original caption without paraphrasing or introducing additional semantic content. Sequential actions are separated when they correspond to distinct motion events, while explicitly concurrent actions are retained together when further splitting would break the original semantic structure.

The last example illustrates the treatment of concurrent semantics. Although the description contains multiple motion components, "walks forward while swinging both arms" is retained as a single constituent event because the actions are explicitly concurrent and form one semantically coherent textual span.

Distributional characteristics. Figure 4 shows clear differences among the three datasets. In particular, SnapMoGen contains substantially more multi-event descriptions and generally longer

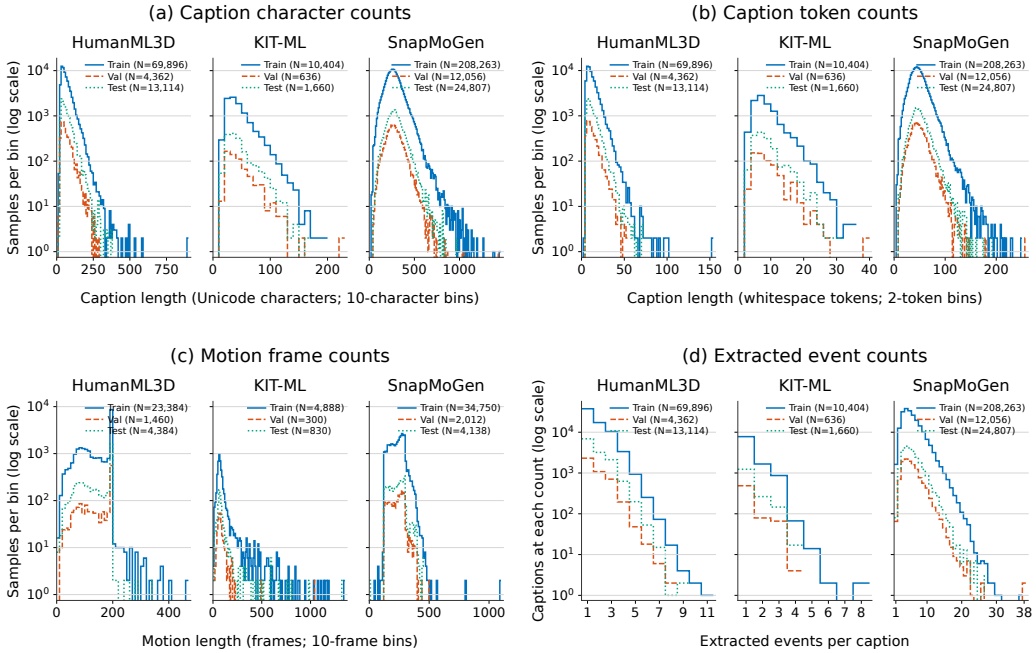

Figure 4: Distributions of (a) caption character counts, (b) caption token counts, (c) motion frame counts, and (d) extracted event counts across HumanML3D, KIT-ML, and SnapMoGen. Caption-based panels count annotation records, whereas the motion-length panel counts preprocessed motion records, including mirrored counterparts where applicable. Frequencies are shown on a logarithmic scale.

Table 6: Examples of extractive event decomposition. All constituent events are contiguous spans copied directly from the original caption.

Original caption	**Extracted constituent events**
A person raises the arms and then goes down on the knees.	[“raises the arms”, “goes down on the knees”]
A person walks forward, turns left, and waves the right hand.	[“walks forward”, “turns left”, “waves the right hand”]
A person bends down to touch the toes and then stretches upward.	[“bends down to touch the toes”, “stretches upward”]
A person walks forward while swinging both arms.	[“walks forward while swinging both arms”]

captions, while HumanML3D and KIT-ML contain a larger proportion of short or single-event descriptions. Event count is not determined by text length alone, since it also depends on the semantic structure of a caption and the granularity of the extractive decomposition. We therefore report caption length and event count separately. Note that the token statistics in Fig. 4(b) are based on whitespace tokenization and should not be interpreted as the subword sequence lengths seen by individual text encoders.

Evaluation cohorts. The corpus-level statistics above should be distinguished from the fixed cohorts used for generation and evaluation. For HumanML3D, our evaluation set contains 4,640 records, with 2,308, 1,098, and 330 captions satisfying $E \geq 2$, $E \geq 3$, and $E \geq 4$, respectively. The KIT-ML multi-event cohort contains 168 records, including 88 two-event, 70 three-event, and 10 four-event captions. For SnapMoGen, we evaluate on 4,000 test records, of which 3,991 contain at least two extracted events. Among them, 3,900 records form complete 100-way retrieval pools used for retrieval-based evaluation.

A.2 Further Ablations

Single vs. hierarchical cross-attention. A straightforward alternative to our hierarchical design is to concatenate the full-caption and constituent-event tokens into a shared memory and inject them through a single cross-attention operation. We therefore compare this Single CA (Token) design with hierarchical Dual CA under the same token-level text representation and parameter budget.

Table 7: Comparison of global–event conditioning structures on HumanML3D ($E \geq 2$). C denotes the full caption and $\{e_1, \ldots, e_K\}$ denotes the constituent events. Results are reported with 95% confidence intervals.

Method	Textual Condition	MM-Dist ↓	R@1 ↑	FID ↓
Baseline	$[C]$	$2.706^{\pm.008}$	$0.576^{\pm.007}$	$0.098^{\pm.005}$
Single CA (Token)	$[C, e_1, \ldots, e_K]$	$2.889^{\pm.011}$	$0.507^{\pm.004}$	$0.092^{\pm.003}$
Ours (Hierarchical Dual CA)	$[C] \rightarrow [e_1, \ldots, e_K]$	$\mathbf{2.653}^{\pm.011}$	$\mathbf{0.586}^{\pm.003}$	$\mathbf{0.065}^{\pm.003}$

As shown in Table 7, simply concatenating constituent-event tokens with the full caption does not improve generation: Single CA (Token) underperforms the caption-only baseline in text–motion alignment despite using additional event-level information. In contrast, hierarchical Dual CA consistently improves over both variants across the reported metrics. Since Single CA and Dual CA use the same token-level textual representation and parameter budget, their difference mainly reflects the conditioning structure. These results support explicitly separating global contextual conditioning from event-specific semantic modeling rather than attending to all textual tokens within a shared memory.

Pooled vs. token-level event representations. We further examine the effect of preserving token-level semantics within each constituent event. Specifically, we compare our token-level event representation with a pooled-event variant that compresses each event into a single vector before event-specific conditioning. Both variants use the same hierarchical conditioning structure and event-local auxiliary supervision, differing only in the representation of individual events.

Table 8: Effect of event representation granularity on HumanML3D ($E \geq 2$). Both variants use the same hierarchical conditioning structure and event-local auxiliary supervision.

Event Representation	R@1 ↑	FID ↓	MM-Dist ↓	Avg-E ↓	Worst-E ↓
Pooled Event	$0.584^{\pm.002}$	$0.070^{\pm.002}$	$2.661^{\pm.003}$	$4.290^{\pm.001}$	$5.624^{\pm.002}$
Token-level Event	$\mathbf{0.586}^{\pm.003}$	$\mathbf{0.065}^{\pm.003}$	$\mathbf{2.653}^{\pm.011}$	$\mathbf{4.270}^{\pm.007}$	$\mathbf{5.564}^{\pm.024}$

As shown in Table 8, retaining token-level event representations yields consistent improvements over pooled-event conditioning. While the difference in R@1 is relatively small, token-level conditioning provides clearer gains in FID and MM-Dist, as well as in both Avg-E and Worst-E. In particular, the improvements in event-level metrics indicate that preserving the internal token-level semantics of each event provides useful information for local event realization that may be lost when the event is compressed into a single representation. These results support our use of token-level event representations for fine-grained event-specific conditioning.

Robustness to supervision models. Our event-wise supervision relies on a frozen motion–text model for local semantic matching. We therefore examine whether its effectiveness depends on the choice of supervision model. Specifically, we replace TMR Petrovich et al. (2023) with LaMP (Li et al., 2025a), ReMoMask (Li et al., 2025b), MotionPatches (Yu et al., 2024), and CAR (Fujiwara et al., 2025), while keeping all other settings unchanged. As shown in Table 9, the five variants yield similar results across both caption- and event-level metrics. These results suggest that the proposed event-wise supervision is not strongly dependent on a particular motion–text representation model.

Table 9: Effect of the frozen motion–text retrieval model used for event-wise supervision. All other settings are kept unchanged.

Supervision model	R@1 ↑	FID ↓	MM-Dist ↓	Avg-E ↓	Worst-E ↓
LaMP	0.584	0.068	2.655	4.273	5.599
ReMoMask	0.584	0.068	2.659	4.267	5.591
MotionPatches	0.585	0.066	2.658	**4.265**	**5.587**
CAR	**0.588**	0.066	**2.648**	4.269	5.592
TMR	0.586	**0.065**	2.653	4.270	5.564

A.3 Multi-Scale Event Matching

Evaluation protocol. For event-level evaluation, we use the frozen HumanML3D text–motion evaluator of Guo et al. (Guo et al., 2022), which is separate from the motion–text representation model used for event-wise supervision during training. Specifically, training uses TMR (Petrovich et al., 2023) to construct the local semantic objective, whereas Event MM-Dist is computed entirely with the benchmark evaluator. This separation avoids evaluating Event MM-Dist in the same representation space used to construct the auxiliary training objective.

Following the multi-scale temporal matching formulation in Sec. 3.3, we uniformly sample temporal centers $\{\tau_j\}_{j=1}^{J}$ along the motion sequence and construct candidate regions at multiple temporal scales $\{\rho_m\}_{m=1}^{M}$. In our implementation, the candidate window lengths are 32, 48, 64, and 80 frames, with adjacent temporal centers separated by eight frames. Regions extending beyond the motion sequence are clipped to the valid range, following Eq. 4. Each constituent event is independently matched to the candidate region with the minimum text–motion embedding distance.

Avg-E is obtained by averaging the minimum distances over all constituent events in a caption, while **Worst-E** takes their maximum and therefore reflects the least well-matched event. We additionally compute **Order** from the temporal centers of the selected regions as the percentage of event pairs whose matched locations follow their textual order. Since events are matched independently without imposing disjoint temporal intervals, their selected regions may overlap, particularly for concurrent or temporally adjacent events.

Matching examples. Figure 5 illustrates representative matching results for sequential and concurrent events. In Fig. 5(a), the selected regions follow the running–backward-step–crouching sequence. In Fig. 5(b), concurrent body-part events are matched to overlapping regions. This behavior is expected because the matching procedure does not enforce mutually exclusive temporal intervals across constituent events.

Training–evaluation matching comparison. The training objective and Event MM-Dist follow the same multi-scale temporal matching formulation but determine local matches in different representation spaces. Figure 6 illustrates this distinction by applying TMR and the benchmark evaluator to the same generated motion using identical temporal centers and window scales. The two evaluators show varying degrees of agreement across events: their selected regions partially overlap for some events while differing substantially for others. This indicates that Event MM-Dist does not simply reproduce the local matching signal used for event-wise supervision. Since no event-level temporal boundaries are available, the selected regions should be interpreted as representation-dependent semantic matches rather than ground-truth event intervals.

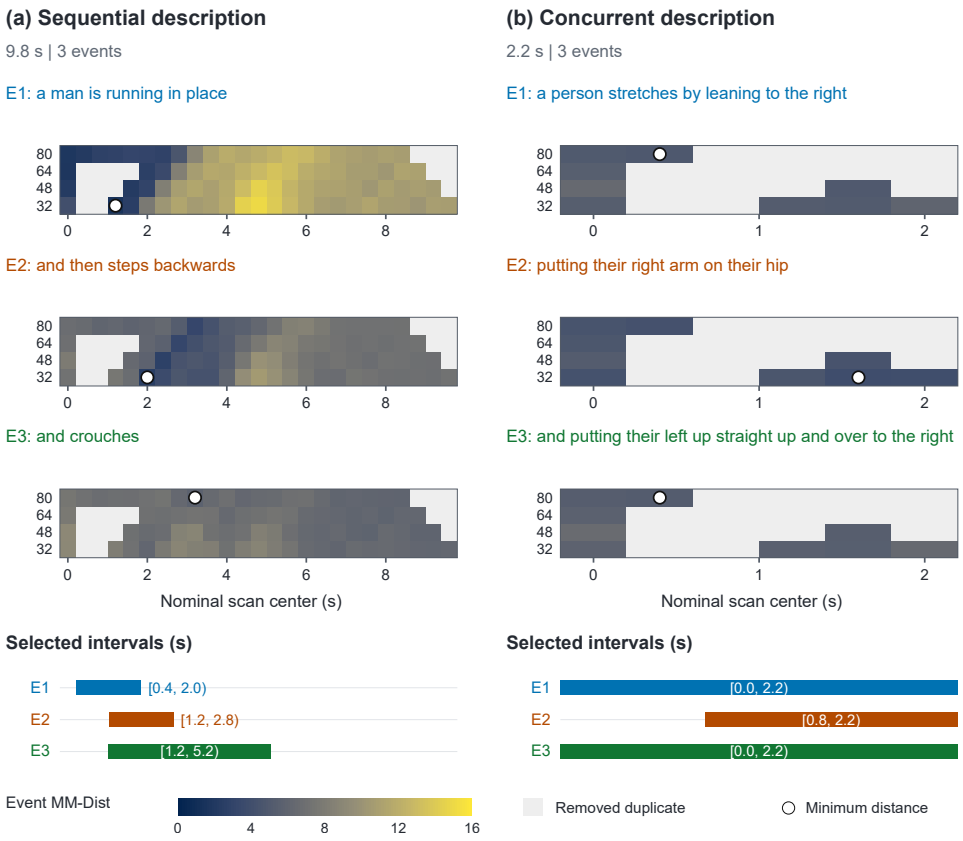

Figure 5: Examples of multi-scale event–motion matching for (a) sequential and (b) concurrent events. Heatmaps show text–motion distances over candidate temporal locations and scales, with circles marking the selected minimum-distance regions. Bars show the corresponding temporal intervals.

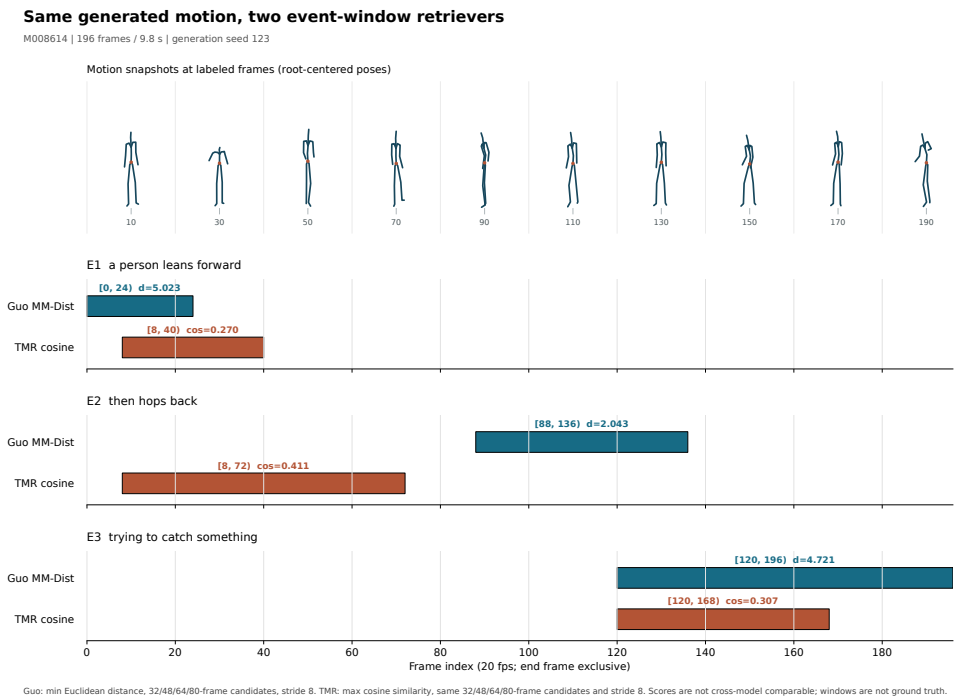

Figure 6: Comparison of event-region matching by TMR and the benchmark evaluator on the same generated motion. Despite identical temporal centers and window scales, the two representation spaces may retrieve different local regions. The selected regions represent semantic matches rather than ground-truth event boundaries.

A.4 VLM-BASED EVENT EVALUATION

A.4.1 EVENT OBSERVABILITY: VLM-E

VLM-E evaluates whether the constituent events in a caption are visually observable in the generated motion. Each query contains an anonymous rendered motion, the full caption, and its extracted events. The VLM judges each event as `observed`, `not_observed`, or `uncertain`.

Evaluation setup. For HumanML3D, we sample 300 captions from the 2,308-caption $E \geq 2$ subset, including 100, 100, 60, and 40 captions with $E = 2$, $E = 3$, $E = 4$, and $E \geq 5$, respectively. Inverse-inclusion weights $w_n = N_h/n_h$ account for the stratified sampling. For KIT-ML, all 168 multi-event captions are evaluated with $w_n = 1$. All methods use the same captions and five random seeds for motion generation, while GT uses one fixed motion per caption. Method identities are hidden during evaluation.

Rendering and VLM judgment. Motions are rendered at their native frame rates using a fixed oblique orthographic view and static per-sequence framing. Left and right limbs are distinguished by color, with a start marker and root trajectory included for spatial reference. We use the frozen Qwen3.8-27B vision-language model Qwen Team (2026) with greedy decoding and video sampling at 4 FPS. An event is marked as `observed` only when supported by visible motion evidence, and `uncertain` is used when the evidence is insufficient. For KIT-ML, the prompt additionally states that concurrent events need not occur sequentially.

Score. Let $y_{m,r,n,i}$ denote the VLM judgment for event i of caption n, generated by method m with seed r. For a caption containing E_n events, we compute

$$\begin{aligned} v_{m,r,n} &= \frac{1}{E_n} \sum_{i=1}^{E_n} \mathbb{I}[y_{m,r,n,i} = \texttt{observed}], \\ \text{VLM-E}^{(m,r)} &= \frac{\sum_n w_n v_{m,r,n}}{\sum_n w_n} \times 100. \end{aligned} \tag{7}$$

Both `not_observed` and `uncertain` are treated as unobserved. We report the mean and sample standard deviation over the 5 generation seeds.

Temporal judgments. VLM-E measures event observability rather than temporal ordering. In our initial evaluation, the VLM's order judgments coincided with successful detection of all events in all 5,577 fully observed cases, making the two signals difficult to separate. We therefore examine temporal judgments separately below, withholding the full caption and expected event order from the VLM.

A.4.2 TEMPORAL ASSESSMENT UNDER QUERY PERMUTATION

Evaluation setup. We select four explicitly sequential two-event descriptions, with two from each dataset, and evaluate GT, CMDM, and G2E-T2M for each description, resulting in 12 motion videos. The VLM receives only the video and two event queries labeled A and B; the full caption, method identity, and expected order are withheld. Each video is queried twice with the A/B assignment reversed while keeping the video unchanged. The responses are then mapped back to the same event identities. This paired evaluation tests whether the predicted temporal relation is stable to the presentation order of the event queries.

Results. As shown in Fig. 7, three of the 12 videos yield stable judgments consistent with the described order, seven provide insufficient evidence under both query permutations, and two change their judgments after the query order is reversed. These results show that VLM-based temporal judgments are affected by both event observability and query presentation. We therefore use VLM-E only for event observability and report temporal Order separately using the matched-window centers described in Appendix A.3.

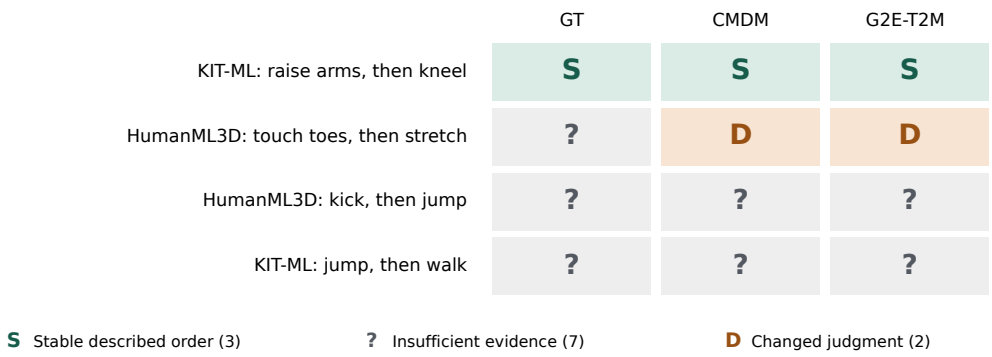

Figure 7: Temporal-order judgments for the 12 evaluated motion videos under paired event queries. S denotes a stable judgment consistent with the described order, ? denotes insufficient evidence under both query permutations, and D denotes a changed judgment after reversing the query assignment.

A.4.3 Prompt Examples

We provide the key instructions used for VLM-E and the temporal diagnostic below. The complete prompts and response schemas are included with the supplementary source.

VLM-E prompt excerpt.

Evaluate only events that are visibly supported by the video. Do not infer an event from the caption alone. Mark an event as `observed` only when visible evidence is present, and use `uncertain` when the visual evidence is insufficient. Concurrent events do not need to occur sequentially.

Example input.
Caption: "A person raises the arms and then goes down on the knees."
Events: ["raises the arms", "goes down on the knees"].

Event status.
`observed | not_observed | uncertain`

Temporal-diagnostic prompt excerpt.

Inspect only the anonymous motion video. The two event queries are presented in arbitrary order; their A/B labels do not indicate temporal order. Determine the relation only from visible motion. If either event cannot be reliably observed, return `insufficient_evidence`. Concurrent or substantially overlapping events should be treated as overlapping rather than forced into a sequential relation.

Paired queries.
Original: A = "raises the arms"; B = "goes down on the knees".
Swapped: A = "goes down on the knees"; B = "raises the arms".

Temporal relation.
`A_before_B | B_before_A | overlap | repeated_or_ambiguous | insufficient_evidence`

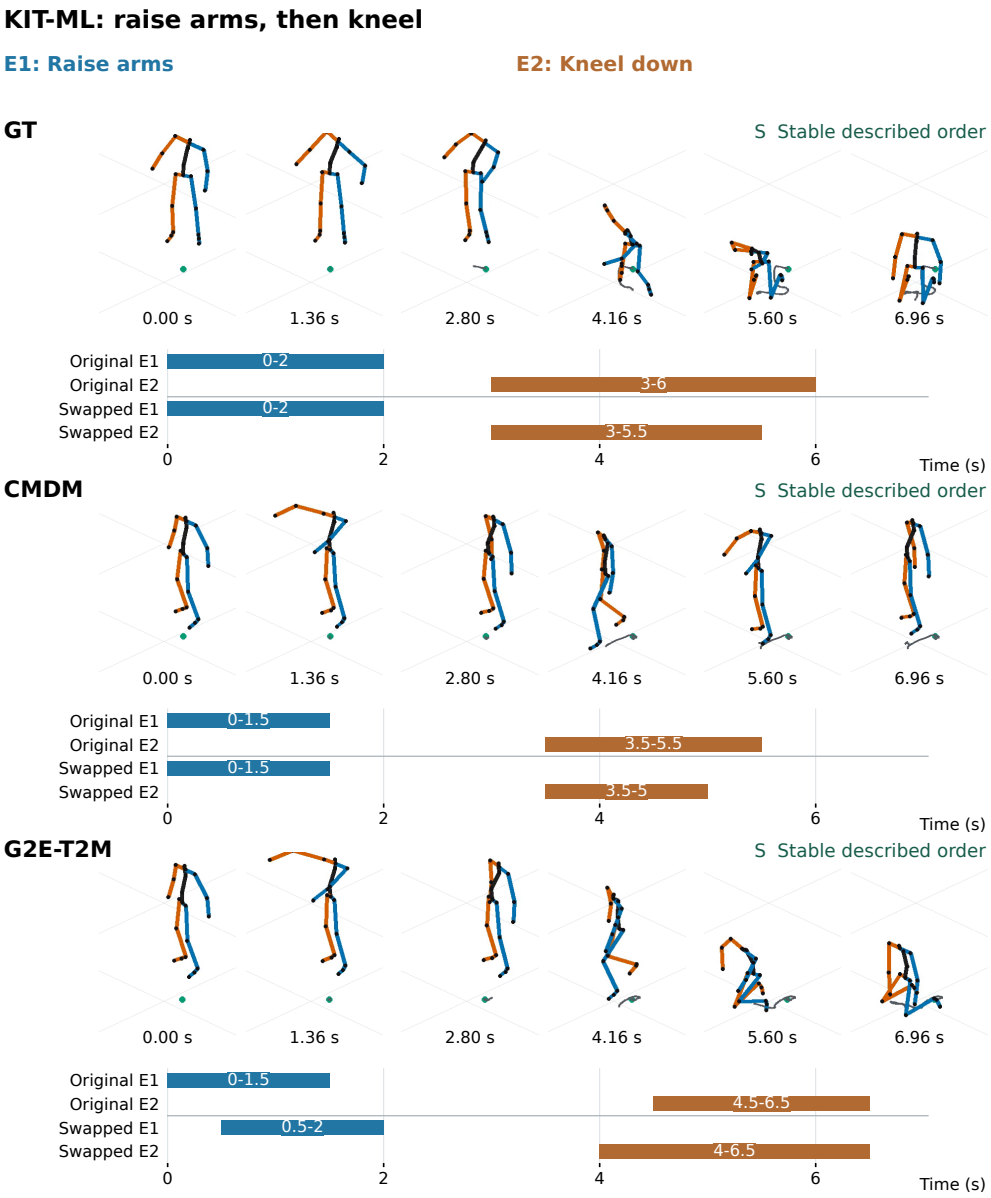

Figure 8: Stable VLM temporal localization under query swapping. The motion depicts arm raising followed by kneeling. Six uniformly sampled frames are shown for each method, together with the temporal intervals predicted for the original and swapped event queries.

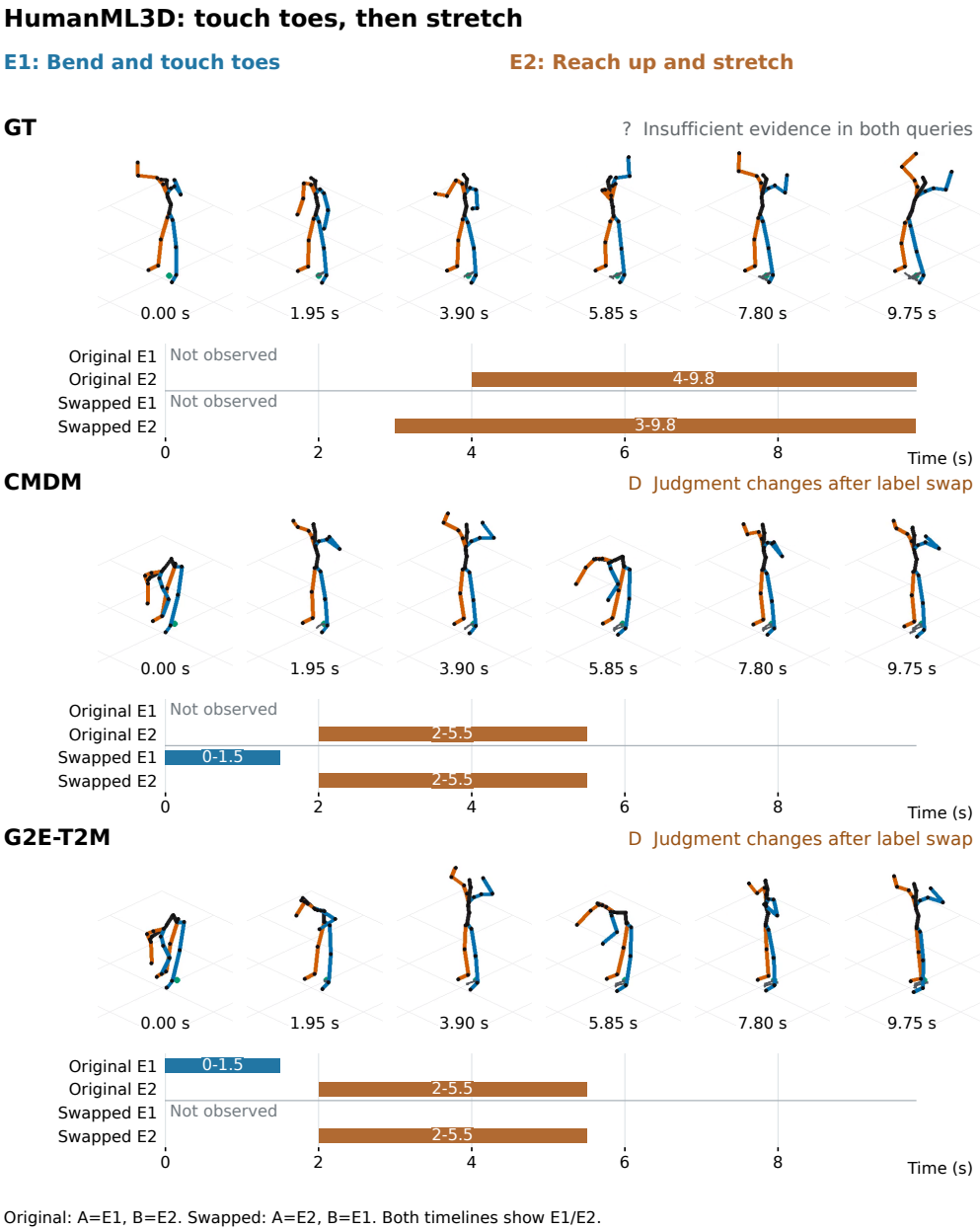

Figure 9: Query-sensitive VLM temporal localization. The motion depicts touching the toes followed by stretching. The same video is evaluated with reversed A/B event assignments, resulting in different temporal predictions. The displayed intervals and visibility decisions are VLM predictions rather than human annotations.

A.5 Performance across Event Complexity

Table 10 reports results on the full HumanML3D test set and the nested $E \geq 2$, $E \geq 3$, and $E \geq 4$ subsets. These subsets contain increasingly larger numbers of constituent events and provide a direct view of model performance on more compositionally complex descriptions.

Performance generally degrades as the event-count threshold increases. Across representative baselines, R@1 decreases while MM-Dist and event-level distances tend to increase from $E \geq 2$ to $E \geq 4$. This trend is observed across models with different generation paradigms, indicating that multi-event descriptions remain challenging for existing T2M methods. Because the subsets are nested and also differ in motion and caption composition, these results characterize performance across event complexity rather than isolating event count as a causal factor.

G2E-T2M shows consistent gains on the multi-event subsets. It achieves the lowest MM-Dist on $E \geq 2$, $E \geq 3$, and $E \geq 4$, while obtaining the best R@1 and event-level distances on $E \geq 2$ and remaining competitive as the event threshold increases. On $E \geq 4$, SALAD performs better on R@1, Avg-E, and Worst-E, whereas G2E-T2M obtains the lowest FID and MM-Dist. Together with the $E \geq 2$ and $E \geq 3$ results, this shows that the proposed event-aware conditioning remains effective as descriptions contain more constituent events. The $E \geq 4$ subset contains only 330 samples, so its distribution-based metrics should be interpreted with the smaller subset size in mind.

Table 10: Evaluation on the HumanML3D full test set and nested subsets defined by event-count thresholds. For the Full set, generated-method results are reported as mean ± sample standard deviation over 20 runs; for the event-complexity subsets, results are reported as mean ± 95% confidence interval. Bold and underline indicate the best and second-best reported means among generated methods within each subset, respectively. Avg-E and Worst-E denote the mean and maximum event MM-Dist after best-window matching.

Subset	Method	R@1 ↑	FID ↓	MM-Dist ↓	Avg-E ↓	Worst-E ↓
Full	GT	$0.515^{\pm.000}$	$0.000^{\pm.000}$	$2.984^{\pm.000}$	$4.060^{\pm.000}$	$4.688^{\pm.000}$
	T2M Guo et al. (2022)	$0.460^{\pm.007}$	$1.015^{\pm.037}$	$3.329^{\pm.022}$	$4.458^{\pm.016}$	$5.117^{\pm.021}$
	MotionDiffuse Zhang et al. (2024)	$0.501^{\pm.006}$	$0.782^{\pm.017}$	$3.004^{\pm.013}$	$4.330^{\pm.006}$	$4.981^{\pm.008}$
	MDM Tevet et al. (2023)	$0.417^{\pm.006}$	$0.473^{\pm.020}$	$3.641^{\pm.017}$	$4.373^{\pm.012}$	$5.086^{\pm.015}$
	T2M-GPT Zhang et al. (2023a)	$0.495^{\pm.004}$	$0.117^{\pm.009}$	$3.091^{\pm.014}$	$4.207^{\pm.011}$	$4.900^{\pm.013}$
	MoMask Guo et al. (2024)	$0.529^{\pm.002}$	$\mathbf{0.043}^{\pm.003}$	$2.939^{\pm.007}$	$4.032^{\pm.008}$	$4.689^{\pm.007}$
	SALAD Hong et al. (2025)	$0.579^{\pm.006}$	$0.067^{\pm.004}$	$2.650^{\pm.012}$	$\underline{3.976}^{\pm.014}$	$\underline{4.617}^{\pm.020}$
	Event-T2M Hong et al. (2026)	$0.571^{\pm.005}$	$0.099^{\pm.002}$	$2.734^{\pm.006}$	$4.067^{\pm.004}$	$4.707^{\pm.005}$
	MoScale Zheng et al. (2026)	$0.540^{\pm.004}$	$\underline{0.046}^{\pm.002}$	$2.826^{\pm.008}$	$4.033^{\pm.009}$	$4.699^{\pm.010}$
	TriC-Motion Cao et al. (2026a)	$0.567^{\pm.005}$	$0.315^{\pm.009}$	$2.708^{\pm.004}$	$4.171^{\pm.010}$	$4.846^{\pm.009}$
	CMDM Yu et al. (2026)	$\underline{0.583}^{\pm.007}$	$0.088^{\pm.004}$	$\underline{2.636}^{\pm.007}$	$4.009^{\pm.011}$	$4.683^{\pm.011}$
	G2E-T2M	$\mathbf{0.594}^{\pm.003}$	$0.052^{\pm.002}$	$\mathbf{2.588}^{\pm.006}$	$\mathbf{3.945}^{\pm.006}$	$\mathbf{4.603}^{\pm.008}$
$E \geq 2$	GT	$0.523^{\pm.000}$	$0.000^{\pm.000}$	$2.971^{\pm.000}$	$4.306^{\pm.000}$	$5.568^{\pm.000}$
	T2M Guo et al. (2022)	$0.440^{\pm.002}$	$1.179^{\pm.028}$	$3.436^{\pm.015}$	$4.724^{\pm.014}$	$6.050^{\pm.023}$
	MotionDiffuse Zhang et al. (2024)	$0.493^{\pm.003}$	$0.892^{\pm.017}$	$3.074^{\pm.012}$	$4.568^{\pm.007}$	$5.877^{\pm.010}$
	MDM Tevet et al. (2023)	$0.389^{\pm.007}$	$0.314^{\pm.010}$	$3.722^{\pm.027}$	$4.778^{\pm.016}$	$6.211^{\pm.023}$
	T2M-GPT Zhang et al. (2023a)	$0.467^{\pm.003}$	$0.154^{\pm.008}$	$3.292^{\pm.017}$	$4.534^{\pm.011}$	$5.928^{\pm.016}$
	MoMask Guo et al. (2024)	$0.508^{\pm.004}$	$\mathbf{0.062}^{\pm.003}$	$3.047^{\pm.007}$	$4.345^{\pm.007}$	$5.666^{\pm.011}$
	SALAD Hong et al. (2025)	$0.572^{\pm.005}$	$0.100^{\pm.006}$	$2.722^{\pm.013}$	$\underline{4.275}^{\pm.014}$	$\underline{5.593}^{\pm.011}$
	Event-T2M Hong et al. (2026)	$0.556^{\pm.005}$	$0.112^{\pm.003}$	$2.766^{\pm.006}$	$4.358^{\pm.004}$	$5.645^{\pm.006}$
	MoScale Zheng et al. (2026)	$0.525^{\pm.004}$	$0.068^{\pm.002}$	$2.888^{\pm.008}$	$4.365^{\pm.004}$	$5.704^{\pm.006}$
	TriC-Motion Cao et al. (2026a)	$0.564^{\pm.002}$	$0.376^{\pm.007}$	$2.779^{\pm.007}$	$4.504^{\pm.005}$	$5.862^{\pm.006}$
	CMDM Yu et al. (2026)	$\underline{0.576}^{\pm.007}$	$0.098^{\pm.005}$	$\underline{2.706}^{\pm.008}$	$4.345^{\pm.008}$	$5.701^{\pm.012}$
	G2E-T2M	$\mathbf{0.586}^{\pm.003}$	$\underline{0.065}^{\pm.003}$	$\mathbf{2.653}^{\pm.011}$	$\mathbf{4.270}^{\pm.007}$	$\mathbf{5.564}^{\pm.024}$
$E \geq 3$	GT	$0.502^{\pm.000}$	$0.000^{\pm.000}$	$2.925^{\pm.000}$	$4.287^{\pm.000}$	$5.812^{\pm.000}$
	T2M Guo et al. (2022)	$0.410^{\pm.006}$	$1.414^{\pm.046}$	$3.466^{\pm.023}$	$4.726^{\pm.014}$	$6.341^{\pm.021}$
	MotionDiffuse Zhang et al. (2024)	$0.448^{\pm.005}$	$1.255^{\pm.026}$	$3.260^{\pm.013}$	$4.591^{\pm.010}$	$6.157^{\pm.019}$
	MDM Tevet et al. (2023)	$0.352^{\pm.006}$	$0.340^{\pm.010}$	$3.824^{\pm.030}$	$4.829^{\pm.017}$	$6.560^{\pm.033}$
	T2M-GPT Zhang et al. (2023a)	$0.420^{\pm.008}$	$0.193^{\pm.019}$	$3.463^{\pm.019}$	$4.564^{\pm.013}$	$6.251^{\pm.016}$
	MoMask Guo et al. (2024)	$0.471^{\pm.008}$	$\mathbf{0.093}^{\pm.006}$	$3.152^{\pm.009}$	$4.385^{\pm.009}$	$5.981^{\pm.018}$
	SALAD Hong et al. (2025)	$\underline{0.546}^{\pm.007}$	$0.144^{\pm.012}$	$2.789^{\pm.011}$	$\underline{4.271}^{\pm.018}$	$\mathbf{5.835}^{\pm.030}$
	Event-T2M Hong et al. (2026)	$0.521^{\pm.005}$	$0.156^{\pm.008}$	$2.844^{\pm.011}$	$4.331^{\pm.005}$	$5.888^{\pm.008}$
	MoScale Zheng et al. (2026)	$0.487^{\pm.005}$	$0.098^{\pm.003}$	$2.981^{\pm.012}$	$4.418^{\pm.006}$	$6.048^{\pm.014}$
	TriC-Motion Cao et al. (2026a)	$0.529^{\pm.006}$	$0.357^{\pm.011}$	$2.846^{\pm.014}$	$4.518^{\pm.008}$	$6.182^{\pm.012}$
	CMDM Yu et al. (2026)	$0.543^{\pm.005}$	$0.116^{\pm.009}$	$\underline{2.746}^{\pm.015}$	$4.338^{\pm.014}$	$5.979^{\pm.017}$
	G2E-T2M	$\mathbf{0.553}^{\pm.006}$	$\underline{0.096}^{\pm.007}$	$\mathbf{2.703}^{\pm.012}$	$\mathbf{4.267}^{\pm.009}$	$\underline{5.883}^{\pm.013}$
$E \geq 4$	GT	$0.484^{\pm.000}$	$0.000^{\pm.000}$	$3.112^{\pm.000}$	$4.282^{\pm.000}$	$5.994^{\pm.000}$
	T2M Guo et al. (2022)	$0.364^{\pm.013}$	$2.212^{\pm.146}$	$3.769^{\pm.038}$	$4.778^{\pm.022}$	$6.607^{\pm.030}$
	MotionDiffuse Zhang et al. (2024)	$0.420^{\pm.009}$	$1.361^{\pm.047}$	$3.420^{\pm.023}$	$4.649^{\pm.020}$	$6.392^{\pm.040}$
	MDM Tevet et al. (2023)	$0.330^{\pm.017}$	$0.955^{\pm.086}$	$4.134^{\pm.058}$	$4.895^{\pm.038}$	$6.864^{\pm.059}$
	T2M-GPT Zhang et al. (2023a)	$0.388^{\pm.010}$	$0.556^{\pm.059}$	$3.613^{\pm.028}$	$4.606^{\pm.024}$	$6.537^{\pm.047}$
	MoMask Guo et al. (2024)	$0.429^{\pm.009}$	$\underline{0.227}^{\pm.008}$	$3.380^{\pm.022}$	$4.419^{\pm.017}$	$6.227^{\pm.034}$
	SALAD Hong et al. (2025)	$\mathbf{0.495}^{\pm.010}$	$0.321^{\pm.017}$	$3.027^{\pm.023}$	$\mathbf{4.266}^{\pm.017}$	$\mathbf{6.039}^{\pm.029}$
	Event-T2M Hong et al. (2026)	$0.455^{\pm.009}$	$0.311^{\pm.012}$	$3.137^{\pm.024}$	$4.412^{\pm.012}$	$6.188^{\pm.020}$
	MoScale Zheng et al. (2026)	$0.430^{\pm.007}$	$0.312^{\pm.010}$	$3.322^{\pm.021}$	$4.494^{\pm.014}$	$6.358^{\pm.029}$
	TriC-Motion Cao et al. (2026a)	$0.468^{\pm.011}$	$0.639^{\pm.031}$	$3.149^{\pm.038}$	$4.565^{\pm.020}$	$6.473^{\pm.035}$
	CMDM Yu et al. (2026)	$0.467^{\pm.014}$	$0.234^{\pm.014}$	$\underline{3.005}^{\pm.027}$	$4.367^{\pm.021}$	$6.246^{\pm.036}$
	G2E-T2M	$\underline{0.470}^{\pm.012}$	$\mathbf{0.202}^{\pm.011}$	$\mathbf{2.978}^{\pm.026}$	$\underline{4.298}^{\pm.015}$	$\underline{6.133}^{\pm.024}$

A.6 PSEUDOCODE OF G2E-T2M

Algorithm 1 summarizes the training procedure of G2E-T2M. Given a full caption and its constituent events, the model first injects global caption semantics and then introduces event-specific conditioning through token-level cross-attention and adaptive relevance estimation. The relevance weights are estimated across constituent events at each motion position and combined with a weak temporal order prior. During training, boundary-free event-wise semantic supervision further associates each event with its most semantically compatible local motion region. This auxiliary branch is used only during training and requires no event-level temporal boundary or duration annotation.

Algorithm 1 Pseudocode of G2E-T2M

Input: Motion–caption pair (x, c), constituent events $\{e_i\}_{i=1}^{N}$, weights $\lambda_{\text{evt}}, \lambda_p$, order-prior bandwidth σ_p, and residual scale γ
Output: Overall training objective $\mathcal{L}$

$F_C \leftarrow \mathcal{E}(c), \quad F_{E_i} \leftarrow \mathcal{E}(e_i),\ i = 1, \ldots, N$ ▷ token-level text encoding
$z \leftarrow \mathcal{M}(x), \quad \epsilon \sim \mathcal{N}(0, I), \quad t \sim \mathcal{U}(0, 1)$
$z_t \leftarrow (1-t)\epsilon + tz$
$X \leftarrow \mathrm{InputProj}(z_t, t)$
for each backbone block b **do**
  $X \leftarrow \mathrm{SelfAttn}_b(X, t)$
  $U \leftarrow X$
  if b is selected for G2E conditioning **then**
    $U \leftarrow X + \mathrm{CA}_b^C(X, F_C)$ ▷ full-caption conditioning
    for $i = 1, \ldots, N$ **do**
      $O_i \leftarrow \mathrm{CA}_b^E(U, F_{E_i})$ ▷ event-specific conditioning
      $\bar{k}_i \leftarrow \frac{1}{T_i} \sum_{t'=1}^{T_i} F_{E_i,t'} W_k^r$
    end for
    for $l = 1, \ldots, L$ **do**
      $q_l \leftarrow U_l W_q^r, \quad p_l \leftarrow (l - \frac{1}{2})/L$
      for $i = 1, \ldots, N$ **do**
        $r_{l,i} \leftarrow q_l^\top \bar{k}_i / \sqrt{d_a}$ ▷ content-aware relevance
        $\mu_i \leftarrow (i - \frac{1}{2})/N$
        $\pi_{l,i} \leftarrow \exp\left(-\frac{(p_l - \mu_i)^2}{2\sigma_p^2}\right)$ ▷ temporal order prior
      end for
      $\alpha_{l,i} \leftarrow \mathrm{Softmax}_i\left(r_{l,i} + \lambda_p \pi_{l,i}\right), \quad i = 1, \ldots, N$ ▷ event weighting
      $\Delta F_{M,l}^{(E)} \leftarrow \sum_{i=1}^{N} \alpha_{l,i} O_{i,l}$
      $U_l \leftarrow U_l + \gamma \Delta F_{M,l}^{(E)}$ ▷ adaptive event aggregation
    end for
  end if
  $X \leftarrow \mathrm{FFN}_b(U, t)$
end for
$v_\theta \leftarrow \mathrm{OutputProj}(X)$
$\widehat{z}_0 \leftarrow z_t + (1-t)v_\theta, \quad \widehat{X}_0 \leftarrow \mathcal{D}(\widehat{z}_0)$
$\mathcal{L}_{\text{gen}} \leftarrow \|v_\theta - (z - \epsilon)\|_2^2$ ▷ generation objective
$\Omega(\widehat{X}_0) \leftarrow \{R_j\}_{j=1}^{J} \leftarrow \mathrm{MultiScaleWindows}(\widehat{X}_0)$ ▷ candidate local regions
for $i = 1, \ldots, N$ **do**
  for $j = 1, \ldots, J$ **do**
    $s_{j,i} \leftarrow f_t(e_i)^\top f_m(R_j)$ ▷ event–region similarity
  end for
  $m_i \leftarrow \max_{j=1,\ldots,J} s_{j,i}$ ▷ best-region matching
end for
$\mathcal{L}_{\text{evt}} \leftarrow 1 - \frac{1}{N} \sum_{i=1}^{N} m_i$ ▷ event-wise semantic supervision
$\mathcal{L} \leftarrow \mathcal{L}_{\text{gen}} + \lambda_{\text{evt}} \mathcal{L}_{\text{evt}}$
return $\mathcal{L}$

A.7 ADDITIONAL QUALITATIVE COMPARISONS ON MULTI-EVENT MOTION GENERATION.



Figure 10: More Qualitative comparisons.