MoRe-DVC: Motion Retrieval-Augmented Generation for Detailed Video Captioning

Vipul Baghel **Ravi Hegde**
Department of Electrical Engineering
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India
{baghelvipul,hegder}@iitgn.com

Abstract

Despite significant advances in large vision–language models (Video-LLMs), accurately narrating fine-grained human motion remains difficult because fluent text can still contain incorrect body parts, directions, or temporal orderings. We introduce Motion Retrieval-Augmented Generation for Detailed Video Captioning (MoRe-DVC), a pose-conditioned framework for event-grounded motion narration. Its primary contribution is an event-boundary-aware, graph-conditioned retrieval mechanism that uses refined temporal structure and kinematic similarity to select relationally compatible evidence. We further instantiate a deterministic, domain-specific physical-faithfulness reward that checks generated text against closed-form symbolic kinematics derived from the same pose sequence, thereby constraining decoder-side motion-language inconsistencies. MoRe-DVC combines Pose–Orientation Encoding (POEM), event-boundary refinement and graph construction, Graph-Conditioned Hybrid Retrieval (GCHR), a structured motion memory bank, and retrieval-conditioned generation. Across BoFiT, HumanML3D, and FineMotion, the method consistently improves motion-captioning quality and physical faithfulness. Controlled same-backbone experiments isolate the architectural gains, while additional reliability, open-stack, graph-isolation, reward-isolation, proposal-robustness, and efficiency analyses are provided in the Appendix.

1 Introduction

Video understanding has evolved from recognizing isolated actions toward generating language that explains what happens, when it happens, and how it unfolds over time. Within this broader agenda, *fine-grained* and *motion-centric* video captioning focuses on videos whose dominant semantic content is carried by articulated human motion rather than static appearance alone. In such settings, useful narration must capture body-part coordination, pose transitions, motion phases, and temporally localized events, making the task central to dense video captioning, sports narration, instructional understanding, and long untrimmed video summarization [Zhou et al., 2024, Kim et al., 2024, Islam et al., 2024, Wu et al., 2025c, Zhao et al., 2025, Davids et al., 2025]. Its importance is therefore both scientific and practical: in egocentric and procedural videos, narration must verbalize ordered sub-actions and local motion changes [Xu et al., 2024, Jung et al., 2023, Islam et al., 2024]; in sports, it must describe event semantics together with motion pattern and context [Xi et al., 2025c, 2026, 2025a,b, Davids et al., 2025]; and in fine-grained motion understanding, it must resolve subtle differences in trajectory, symmetry, and rotational dynamics that generic video-language models often fail to capture [Zhao et al., 2025, Wu et al., 2025b].

Despite recent progress, existing approaches still suffer from three broad limitations. First, many methods remain fundamentally appearance-centric or text-centric: they exploit RGB features, language priors, or retrieved captions effectively, yet they do not explicitly represent articulated motion

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

as a structured signal with joint trajectories, higher-order dynamics, rotational changes, and body-part interactions [Zhou et al., 2024, Kim et al., 2024, Islam et al., 2024, Wu et al., 2025c]. Second, they often lack temporal precision: fine-grained narration requires not only recognizing an event but grounding it to the correct temporal span, whereas dense-captioning pipelines frequently inherit noisy proposals or coarse partitions, causing caption-boundary mismatch and temporal drift [Zhou et al., 2024, Wu et al., 2025c, Islam et al., 2024, Kim et al., 2024]. Third, retrieval-augmented captioning commonly retrieves captions, video–text pairs, or external videos using broad semantic similarity, rather than motion-native evidence such as kinematic similarity, motion-phase compatibility, body-part saliency, or graph-structured relations between neighboring events [Zhang et al., 2021, Chen et al., 2023, Xu et al., 2024, Jung et al., 2023, Kim et al., 2024, 2025, Jeon et al., 2025, Ma et al., 2026]. Consequently, a caption can be linguistically plausible while remaining biomechanically wrong, e.g., by reversing left/right motion, naming an incorrect joint, or permuting event order.

We address **pose-conditioned, event-grounded dense motion narration**, rather than fully end-to-end RGB dense video captioning. MoRe-DVC operates downstream of frozen upstream estimators: raw video is used only to obtain the 3D pose sequence when poses are not already provided, while coarse event proposals are produced by a pose-based temporal segmenter. No raw RGB frames are consumed by MoRe-DVC downstream of pose extraction. This separation lets us study the conditioning-to-generation gap directly: a decoder may still override correct pose evidence with language priors, producing physically incorrect text even when the input motion is valid.

To address this gap, we propose **MoRe-DVC** (**Mo**tion-**Re**trieval **D**etailed **V**ideo **C**aptioning), a hierarchical retrieval-augmented framework. Its central contribution is event-boundary-aware **graph-conditioned retrieval**: refined motion events are linked into a relational graph and used to retrieve evidence that is compatible not only semantically and kinematically, but also chronologically. We additionally instantiate a **domain-specific structured faithfulness reward** $\mathcal{L}_{\text{faith}}$ that compares decoder outputs with deterministic symbolic kinematics from POEM. The underlying policy-gradient mechanism is established in prior structured-reward work; our contribution lies in the closed-form kinematic instantiation and its coupling with graph-conditioned retrieval. Additional controlled experiments in the Appendix isolate both effects.

Contributions.

- We introduce an **event-boundary-aware graph-conditioned retrieval** mechanism that retrieves motion evidence using semantic similarity, kinematic compatibility, and chronological/hierarchical event-graph relations rather than isolated frame or text similarity.
- We instantiate a deterministic **kinematic faithfulness reward** $\mathcal{L}_{\text{faith}}$ that checks generated language against auditable symbolic pose descriptors, and show in controlled analyses that it complements rather than substitutes for graph-conditioned retrieval.
- We evaluate the framework across BoFiT, HumanML3D, and FineMotion using benchmark-native protocols and same-backbone controls. Additional multi-seed, significance, open-weight, efficiency, memory-size, proposal-source, expression-diversity, and causal ablation analyses are reported in the Appendix.

2 Related Work

2.1 Dense Video Captioning and Motion LLMs

Recent advancements in video captioning tackle long-form and fine-grained complexities via memory, hierarchical, and equalized architectures, alongside context-aware sports summarization [Zhou et al., 2024, Kim et al., 2024, Islam et al., 2024, Wu et al., 2025c, Ohkawa et al., 2025, Zhao et al., 2025, Xi et al., 2025c,a,b, 2026, Davids et al., 2025]. Concurrently, motion-language modeling has evolved from reciprocal translation [Guo et al., 2022b, Radouane et al., 2024b,a] to unified foundation frameworks treating human motion as a distinct LLM modality for multi-granularity alignment [Jiang et al., 2023, Zhang et al., 2024, Wu et al., 2024, 2025b, Li et al., 2025, Wu et al., 2025d, Zhu et al., 2026]. These methods provide strong motion-language baselines, but most do not retrieve evidence conditioned on an explicit temporal event graph.

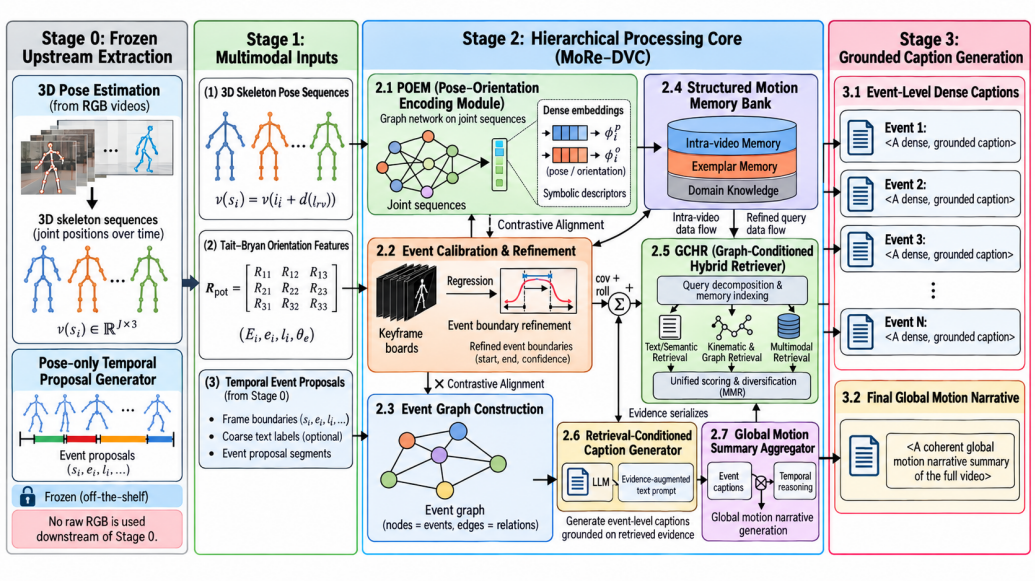

Figure 1: Overview of MoRe-DVC. Stage 0 consists of frozen upstream extraction: a 3D pose estimator (only when dataset poses are unavailable) and a pose-based temporal proposal generator. MoRe-DVC itself consumes pose/orientation features and coarse proposals; no raw RGB frames are used downstream.

2.2 Retrieval-Augmented Generation and Faithfulness

Retrieval enhances video captioning by providing external priors, evolving from sentence-level augmentation [Zhang et al., 2021, Chen et al., 2023] to cross-modal memory and hierarchical retrieval [Kim et al., 2024, 2025]. In the motion domain, ReMoGPT [Yu et al., 2025] uses part-level retrieval, while MG-MotionLLM [Wu et al., 2025b] models multiple motion granularities and FineMotion [Wu et al., 2025a] supplies dense spatial-temporal body-part annotations. Our retrieval contribution differs by conditioning evidence selection on refined event-graph structure as well as semantic and kinematic similarity.

Structured rewards for hallucination reduction are not new. MOCHa applies multi-objective reinforcement learning to caption hallucination [Ben-Kish et al., 2024]; VideoRFT introduces a learned semantic-consistency reward for video reasoning [Wang et al., 2025]; and CANOE uses rule-based reinforcement learning to improve contextual faithfulness [Si et al., 2025]. We therefore do not claim novelty for policy-gradient reward optimization itself. Instead, $\mathcal{L}_{\text{faith}}$ is a domain-instantiated structured reward whose target is a closed-form kinematic vector computed from pose, making the constraint explicit and auditable. The Appendix isolates the reward while holding graph-conditioned retrieval fixed and quantifies its effect on both physical errors and linguistic diversity.

3 Methodology

3.1 Problem Formulation

Let $\mathbf{P} = \{P_t\}_{t=1}^{T_v}$ be a pose sequence, $P_t \in \mathbb{R}^{J \times d_p}$ with $d_p \in \{2, 3\}$, and let $\mathbf{B} = \{B_t\}_{t=1}^{T_v}$ denote orientation features derived from the same motion. A frozen upstream pose-based temporal segmenter provides coarse event proposals $\mathcal{E} = \{p_i\}_{i=1}^{N}$, where $p_i = (s_i^p, e_i^p, r_i)$ contains a start frame, end frame, and coarse event label. The default proposal source is kinematic change-point detection over smoothed angular-velocity/motion-energy profiles followed by a lightweight pose-based action classifier; alternative proposal sources are evaluated in Appendix C.7. Raw RGB enters only the initial pose estimator when required and is outside the claimed scope of MoRe-DVC.

Formally, the task is

$$\mathcal{F}_\theta : (\mathbf{P}, \mathbf{B}, \mathcal{E}) \mapsto \left(\{(\hat{s}_i, \hat{e}_i, \hat{\mathbf{y}}_i)\}_{i=1}^{N},\ \hat{y}^{\text{global}} \right), \tag{1}$$

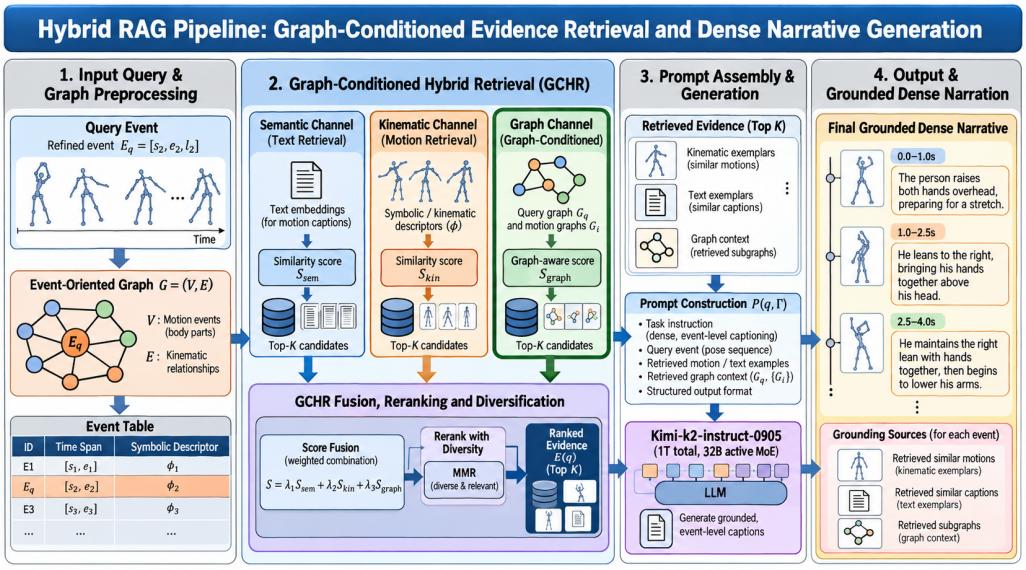

Figure 2: Graph-conditioned evidence retrieval and dense narrative generation. GCHR combines semantic, kinematic, and event-graph compatibility before diversity-aware selection and prompt serialization.

where $[\hat{s}_i, \hat{e}_i]$ is a refined temporal event span, $\hat{\mathbf{y}}_i$ is its dense caption, and $\hat{y}^{\text{global}}$ is a sequence-level narrative. This definition separates our task from end-to-end RGB captioning and makes explicit that event grounding and motion-to-text generation are optimized downstream of frozen preprocessing. A compact notation table is provided in Appendix A.1.

3.2 Architectural Overview

MoRe-DVC is organized around two contribution-bearing mechanisms: POEM-based symbolic motion representation and Graph-Conditioned Hybrid Retrieval (GCHR), coupled with the faithfulness reward. POEM maps pose/orientation trajectories to dense representations and deterministic symbolic descriptors. A boundary refiner adjusts coarse proposal endpoints and connects refined events into an event graph. A structured motion memory bank stores past intra-video context, training-set exemplars, and domain rules. GCHR retrieves a diverse evidence set using semantic, kinematic, graph, and phase compatibility. A retrieval-conditioned language generator then produces event captions, and a summary aggregator composes them into the global narrative.

Pose–Orientation Encoding (POEM). For a refined event $[\hat{s}_i, \hat{e}_i]$, POEM computes

$$\phi_i = \Phi(P_{\hat{s}_i:\hat{e}_i}, B_{\hat{s}_i:\hat{e}_i}) = [v_i, a_i, \omega_i, \text{sal}(i), \text{sym}(i), \text{phase}(i), \text{contact}(i)]. \tag{2}$$

Raw poses are root-centered, scale-normalized, and confidence-weighted for missing joints. Spatial trajectories are smoothed with a Savitzky–Golay filter before differentiation [Savitzky and Golay, 1964]; 3D orientation follows standard Tait–Bryan conventions [Diebel, 2006]. Saliency is the temporal variance of joint angular velocity, while symmetry uses mirrored left/right trajectory distance. These features are intentionally deterministic so that the generated text can be checked against auditable motion quantities. Exact preprocessing constants and feature definitions are given in Appendix B.4.

Boundary refinement and event graph. For each coarse proposal, local pose features around (s_i^p, e_i^p) predict temporal offsets $(\Delta s_i, \Delta e_i)$, giving $\hat{s}_i = s_i^p + \Delta s_i$ and $\hat{e}_i = e_i^p + \Delta e_i$. Refined events form a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_G)$ whose edges encode chronological adjacency, overlap, and kinematic compatibility. The graph need not be a chain: concurrent actions can create parallel relations. Proposal-source dependence is evaluated explicitly in Appendix C.7.

Motion memory and Graph-Conditioned Hybrid Retrieval (GCHR). The memory is $\mathcal{M} = \mathcal{M}^{\text{intra}} \cup \mathcal{M}^{\text{ex}} \cup \mathcal{M}^{\text{dom}}$, containing causal intra-video history, training-only motion exemplars, and

domain knowledge. To avoid future leakage, intra-video memory at event time t contains only earlier entries. For query q_i, a memory item d_j is scored only over modalities available to both:

$$S(q_i, d_j) = \frac{\sum_{m \in \mathcal{A}(q_i, d_j)} \alpha_m \, \mathrm{Sim}_m(f_m(q_i), f_m(d_j))}{\sum_{m \in \mathcal{A}(q_i, d_j)} \alpha_m}. \tag{3}$$

The available channels include semantic, kinematic, graph, and phase compatibility. GCHR then applies maximum marginal relevance to select a diverse top-K evidence set:

$$d_k^\star = \arg \max_{d_j \in \mathcal{M} \setminus \mathcal{R}_{k-1}} \left[\lambda S(q_i, d_j) - (1 - \lambda) \max_{d_r \in \mathcal{R}_{k-1}} \mathrm{Sim}_{\mathrm{red}}(d_j, d_r) \right]. \tag{4}$$

The univariate graph-conditioning toggle requested by the reviewers is reported in Appendix C.6.

Retrieval-conditioned generation and summary aggregation. The retrieved evidence is serialized as $\Gamma_i = \mathrm{Serialize}(\boldsymbol{\phi}_i, \mathcal{R}_K^\star, \mathcal{N}(i))$ and conditions event-caption generation, $\pi_\theta(\hat{\mathbf{y}}_i \mid p_i, \Gamma_i) = \prod_{t=1}^{L_i} \pi_\theta(\hat{y}_{i,t} \mid \hat{\mathbf{y}}_{i,<t}, p_i, \Gamma_i)$. A softmax-weighted aggregation of event features and generated captions then conditions the final global narrative.

Kinematic Faithfulness Reward ($\mathcal{L}_{\mathrm{faith}}$). A frozen rule-based parser $E_\psi(\hat{\mathbf{y}}_i)$ maps generated text into the same canonical kinematic slots as $\boldsymbol{\phi}_i$. The discrepancy

$$c_i^{\mathrm{faith}} = \| E_\psi(\hat{\mathbf{y}}_i) - \boldsymbol{\phi}_i \|_2^2 \tag{5}$$

penalizes decoder-side disagreements in body part, direction, temporal order, and coarse magnitude. The target is deterministic rather than an independent sensor measurement: its purpose is to prevent the autoregressive decoder from overriding pose-conditioned evidence with language priors. With $r_i^{\mathrm{faith}} = -c_i^{\mathrm{faith}}$, we use the REINFORCE surrogate

$$\mathcal{L}_{\mathrm{faith}}^{\mathrm{PG}} = -\frac{1}{N} \sum_{i=1}^{N} (r_i^{\mathrm{faith}} - b_i) \sum_{t=1}^{L_i} \log \pi_\theta(\hat{y}_{i,t} \mid \hat{\mathbf{y}}_{i,<t}, x_i, p_i, \mathcal{M}). \tag{6}$$

This is a domain-specific instantiation of structured-reward optimization rather than a new policy-gradient mechanism. Because rule parsing can penalize unusual but correct wording, we quantify lexical diversity and manual over-constraint rates in Appendix C.9.

3.3 Learning Objectives and Training Pipeline

The multi-task objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cap}} + \alpha \mathcal{L}_{\mathrm{ret}} + \beta \mathcal{L}_{\mathrm{bnd}} + \gamma \mathcal{L}_{\mathrm{faith}}^{\mathrm{PG}} + \zeta \mathcal{L}_{\mathrm{ord}}. \tag{7}$$

The captioning loss is standard token-level negative log-likelihood,

$$\mathcal{L}_{\mathrm{cap}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{L_i} \log \pi_\theta(\hat{y}_{i,t} \mid \hat{\mathbf{y}}_{i,<t}, \mathcal{C}), \tag{8}$$

while boundary regression uses

$$\mathcal{L}_{\mathrm{bnd}} = \frac{1}{N} \sum_{i=1}^{N} \left[\rho(\hat{s}_i - s_i^*) + \rho(\hat{e}_i - e_i^*) \right]. \tag{9}$$

With midpoint $\hat{t}_i = (\hat{s}_i + \hat{e}_i)/2$, temporal ordering is regularized by

$$\mathcal{L}_{\mathrm{ord}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \max\{0, \Delta - (\hat{t}_{i+1} - \hat{t}_i)\}. \tag{10}$$

The retrieval alignment objective, optimization schedule, and all constants are given in Appendix A.3 and Appendix B.4.

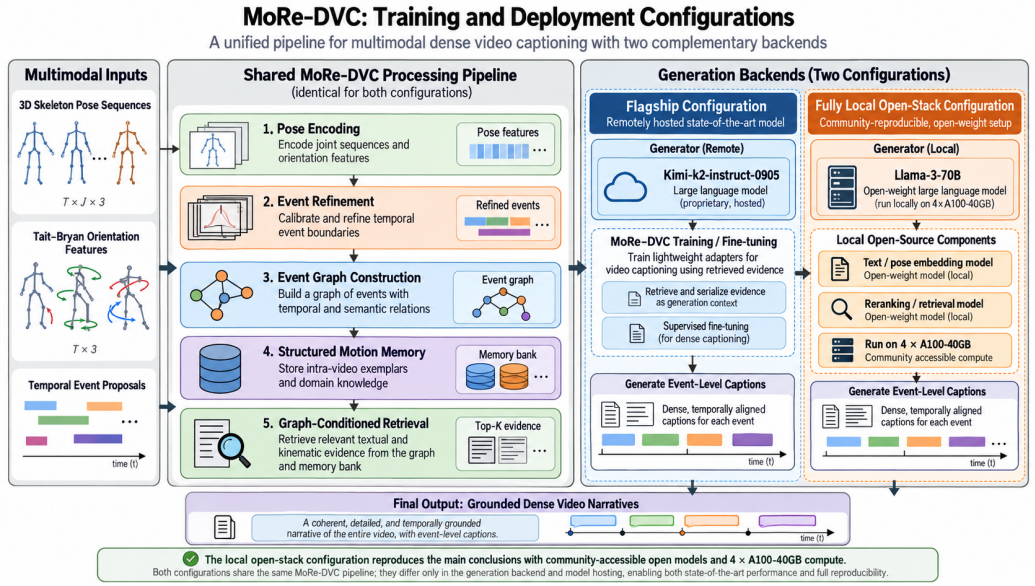

Figure 3: Training and inference organization of MoRe-DVC. Frozen upstream extraction and retrieval services are separated from trainable motion/boundary components and the generator-optimization path. The fully local open-stack configuration is detailed in Appendix C.3.

4 Experiments

4.1 Experimental Setup

Datasets and task alignment. We evaluate on **BoFiT** [Zhao et al., 2025], **HumanML3D** [Guo et al., 2022a], and **FineMotion** [Wu et al., 2025a]. HumanML3D and FineMotion evaluate motion-to-text generation under their native motion-language protocols; BoFiT evaluates fine-grained motion-video narration. General Video-LLM rows in the BoFiT table are retained as benchmark reference points, not as like-for-like architectural controls. A task-alignment summary and baseline reproduction details are provided in Appendix B.2.

Evaluation metrics. We report BLEU, ROUGE, METEOR, CIDEr, and BERTScore for text quality, together with R-Precision and MM-Dist for motion-language alignment and FCE-Motion for fine-grained order-sensitive motion narration [Wu et al., 2025b, Zhao et al., 2025]. Except for MM-Dist and FCE-Motion, metrics are shown on a $[0, 100]$ scale. Definitions and evaluator details appear in Appendix B.3.

Baseline protocol. On HumanML3D and FineMotion, we use official code/checkpoints, standard splits, and the standard motion–text evaluator following MG-MotionLLM. On BoFiT, we follow the official benchmark splits and metric implementation and retain reported scores for shared baselines. The same-backbone controlled ablations use identical Kimi-K2 prompts, pose/proposal inputs, and decoding temperature. Hyperparameters are fixed across datasets. MotionGPT3 and protocol-reproduction checks are reported in Appendix B.2, and multi-run/statistical reliability is reported in Appendix B.5.

Implementation. The flagship configuration uses Kimi-K2-Instruct-0905 as the generator (1T total parameters, 32B active), while the fully local open-weight reproduction uses Llama-3-70B on a single $4\times$A100-40GB node. We distinguish these configurations explicitly because raw cross-scale benchmark rows are not like-for-like comparisons; causal claims rely on the controlled same-backbone experiments below. Full upstream, optimization, and open-stack details are given in Appendix B.4 and Appendix C.3.

Table 1: Motion-to-text performance on HumanML3D. Parameter counts describe the language backbone where reported; NR denotes not reported in the benchmark source. Additional MotionGPT3 results under the identical protocol are provided in Appendix B.2.1.

Methods	Params	R-Precision ↑		MM-Dist ↓	BLEU-4 ↑	BERTScore ↑
		Top-1	Top-3			
TM2T [Guo et al., 2022b]	NR	51.6	82.3	2.935	7.00	32.2
MotionGPT [Jiang et al., 2023]	770M	54.3	82.7	2.821	12.47	32.4
MG-MotionLLM [Wu et al., 2025b]	220M	59.2	86.6	2.581	8.06	36.7
MoRe-DVC (Ours)	1T / 32B active	**61.5**	**88.5**	**2.450**	**13.85**	**38.4**

Table 2: Detailed motion captioning on FineMotion at sequence level. The MG-MotionLLM parameter counts follow its official small/base/large T5 configurations.

Model	Params	BLEU@1 ↑	BLEU@2 ↑	BLEU@4 ↑	ROUGE ↑	BERTScore ↑
MG-MotionLLM (Small) [Wu et al., 2025b]	60M	79.02	63.61	51.02	64.6	50.2
MG-MotionLLM (Base) [Wu et al., 2025b]	220M	81.78	66.44	53.75	65.5	52.3
MG-MotionLLM (Large) [Wu et al., 2025b]	770M	81.73	66.29	53.56	65.8	52.4
MoRe-DVC (Ours)	1T / 32B active	**83.50**	**68.20**	**55.40**	**67.5**	**54.8**

4.2 Results on Motion-Centric Video Narration

HumanML3D. MoRe-DVC improves the reported HumanML3D metrics relative to the benchmark rows in Table 1. Because these methods differ substantially in architecture and scale, we treat this table as benchmark continuity rather than causal attribution. The stronger MotionGPT3 comparison, official-checkpoint reproduction, and 20-trial/statistical analysis are reported in Appendix B.2.1 and Appendix B.5.

FineMotion. On FineMotion, MoRe-DVC improves the sequence-level lexical and semantic metrics in Table 2. The task is motion-to-detailed-text generation under the dataset's native protocol; full protocol alignment is summarized in Appendix B.2.

BoFiT. On BoFiT, MoRe-DVC attains the strongest values among the benchmark rows in Table 3. The general-purpose Video-LLM rows are inherited reference points from the official benchmark, whereas our architectural claims rely on the controlled Kimi-K2 comparisons in the next subsection.

4.3 Ablation Study

Controlled same-backbone baselines. All variants in Table 4 use the same Kimi-K2 backbone, the same pose/proposal inputs, the same prompts, and the same decoding temperature. Relative to direct prompting, the full method improves FCE-Motion by 7.8 points and BERTScore by 6.6 points; relative to text-only RAG, the gains are 6.1 and 5.3 points, respectively. These controls are the primary evidence that the improvement is not an artifact of backbone choice. Three-seed reliability and paired bootstrap tests are reported in Appendix B.5.

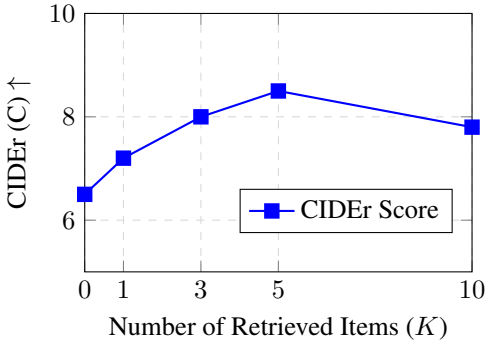

Figure 4: Impact of retrieval size K on BoFiT.

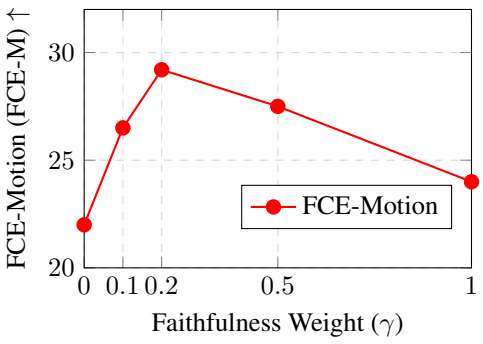

Figure 5: Impact of faithfulness γ on BoFiT.

Table 3: Fine-grained motion video narration on BoFiT. Scores for shared published baselines follow the official BoFiT benchmark; NR indicates that a comparable parameter count is not specified in that benchmark table.

Methods	Params	B@1 ↑	B@2 ↑	B@3 ↑	B@4 ↑	R ↑	M ↑	C ↑	BERT ↑	FCE-M ↑
Video-LLaMA [Zhang et al., 2023]	NR	16.7	5.2	1.7	0.6	15.6	7.9	0.3	19.8	-8.6
Video-LLaVA [Lin et al., 2024]	NR	35.6	18.1	9.9	6.1	24.0	17.7	2.3	40.7	17.3
Video-ChatGPT [Maaz et al., 2024]	NR	21.2	8.9	4.3	2.3	17.2	9.2	1.0	31.1	8.1
BoFiTGen-inc [Zhao et al., 2025]	NR	38.4	21.9	13.0	8.3	27.2	18.3	6.8	43.8	25.7
BoFiTGen-tb [Zhao et al., 2025]	NR	34.0	19.2	11.6	7.4	25.3	16.9	4.0	43.8	27.0
MoRe-DVC (Ours)	1T / 32B active	**40.5**	**23.5**	**14.5**	**9.8**	**29.5**	**20.5**	**8.5**	**46.5**	**29.2**

Table 4: **Controlled Baseline Analysis on BoFiT.** All variants use identical Kimi-K2 backbone, pose/proposal inputs, prompts, and decoding. "w/o POEM/reward" disables both symbolic prompt descriptors and $\mathcal{L}_{\text{faith}}$.

Variant	Retrieval Type	Graph Cond.	POEM	Boundary Refine.	FCE-M (↑)	BERT (↑)
Kimi-K2 (Direct Prompt)	None	✗	✗	✗	21.4	39.9
Kimi-K2 + Text RAG	Text Only	✗	✗	✗	23.1	41.2
Kimi-K2 + Motion RAG	Motion Only	✗	✗	✗	25.8	42.8
Full w/o POEM/reward	Hybrid	✓	✗	✓	27.2	44.5
Full w/o boundary refinement	Hybrid	✓	✓	✗	27.9	45.1
MoRe-DVC (Full)	**Hybrid**	✓	✓	✓	**29.2**	**46.5**

Hyperparameters and reward interaction. Retrieval peaks at $K = 5$, and the default faithfulness weight is $\gamma = 0.2$. The apparently conflicting POEM observations in Table 5 and Fig. 5 correspond to different factorial cells: Table 5 activates POEM together with its default faithfulness objective, whereas $\gamma = 0$ in Fig. 5 keeps the full retrieval context and exposes symbolic descriptors without the reward. The latter causes *unanchored specificity*: rich retrieved language plus unregularized numeric descriptors encourages specific kinematic claims whose direction/body-part/order can be paraphrased incorrectly. The complete factorial, reward-only isolation, and extended error taxonomy are in Appendix C.5; diversity side effects are quantified in Appendix C.9.

Boundary robustness and retrieval attribution. Table 6 shows that boundary refinement substantially improves temporal localization for the default proposal source. The same conclusion holds for uniform-window and jittered-boundary sources (Appendix C.7); the original pose-corruption stress test is retained in Appendix C.8. A separate univariate retrieval experiment holds POEM, the faithfulness reward, boundary refinement, MMR, and K fixed while toggling only graph conditioning; it is reported in Appendix C.6. Computational/storage overhead, open-stack reproducibility, memory-size dependence, physical-error taxonomy, and descriptor ablations are reported in Appendix C.2–C.10.

5 Conclusion

MoRe-DVC presents a pose-conditioned retrieval-augmented framework for fine-grained motion narration. Its main contribution is graph-conditioned retrieval over refined event structure, complemented by a deterministic kinematic faithfulness reward that constrains decoder-side physical inconsistencies. Across three motion-centric benchmarks, the method shows consistent improvements, while same-backbone controls isolate gains from model scale and configuration.

The framework also has important limitations. It depends on upstream pose quality and cannot reliably recover from severe occlusion or catastrophic pose-estimation failures. Absolute performance depends on generator scale and memory-bank size, so cross-model benchmark comparisons are not intrinsically like-for-like; our causal claims therefore rely on same-backbone controls, and scale/memory sensitivity is documented in the Appendix. The flagship configuration uses a remotely hosted 1T Kimi-K2 backend, creating an accessibility barrier, although the fully local 70B open-stack reproduction preserves the conclusions on a 4×A100 node. Finally, the rule parser can over-penalize rare but correct expressions at large γ, motivating future work on more flexible yet auditable kinematic verification. Future work will also explore stronger end-to-end pose tracking and robustness under severe in-the-wild occlusion.

Table 5: Module add-back study on BoFiT. Every active POEM variant is trained with the *default objective*; therefore "+POEM" includes $\mathcal{L}_{\text{faith}}$ with $\gamma = 0.2$. This differs from the $\gamma = 0$ full-retrieval cell in Fig. 5.

Model	Components			B@4 ↑	R ↑	M ↑	C ↑	BERT ↑	FCE-M ↑
	POEM	Boundary Refine.	GCHR						
Base LLM				6.8	22.4	14.5	4.2	39.9	21.4
+ POEM	✓			7.5	24.5	16.5	5.5	41.5	24.5
+ Boundary refinement	✓	✓		8.2	27.0	18.5	6.5	43.5	26.5
Full MoRe-DVC	✓	✓	✓	**9.8**	**29.5**	**20.5**	**8.5**	**46.5**	**29.2**

Table 6: **Temporal Boundary Refinement on BoFiT** using the default pose-based proposal source. Proposal-source robustness is reported in Appendix C.7.

Stage	mIoU (↑)	tIoU@0.3 (↑)	tIoU@0.5 (↑)	Start MAE (↓)	End MAE (↓)
Coarse Proposals (Input)	42.1	55.4	31.2	18.4 frames	19.1 frames
After Refinement	**58.7**	**76.2**	**52.8**	**6.2 frames**	**7.5 frames**

References

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22680–22698, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1263.

Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Hongyang Chao, and Tao Mei. Retrieval augmented convolutional encoder-decoder networks for video captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(1s):48:1–48:24, 2023. doi: 10.1145/3539225.

D. Minola Davids, A. Arul Edwin Raj, and C. Seldev Christopher. SportSummarizer: A unified multimodal fusion transformer for context-aware sports video summarization. *Neurocomputing*, 652:131011, 2025. doi: 10.1016/j.neucom.2025.131011.

James Diebel. Representing attitude: Euler angles, unit quaternions, and rotation vectors. Technical report, Stanford University, October 2006.

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 (CVPR)*, pages 5152–5161, June 2022a. doi: 10.1109/CVPR52688.2022.00509.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *Computer Vision – ECCV 2022*, pages 580–597. Springer, 2022b. doi: 10.1007/978-3-031-19833-5_34.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video ReCap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18198–18208, June 2024. doi: 10.1109/CVPR52733.2024.01723.

MinJu Jeon, Si-Woo Kim, Ye-Chan Kim, HyunGee Kim, and Dong-Jin Kim. Sali4Vid: Saliency-aware video reweighting and adaptive caption retrieval for dense video captioning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25777–25790, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1308.

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, pages 20067–20079, 2023.

Yeonjoon Jung, Minsoo Kim, Seungtaek Choi, Jihyuk Kim, Minji Seo, and Seung-won Hwang. Retrieval-augmented video encoding for instructional captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8554–8568, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.543.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13894–13904, June 2024. doi: 10.1109/CVPR52733.2024.01318.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. HiCM2: Hierarchical compact memory modeling for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4293–4301, 2025. doi: 10.1609/aaai.v39i4.32451.

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.342.

Yunchuan Ma, Laiyun Qing, Guorong Li, Yuankai Qi, Amin Beheshti, Quan Z. Sheng, and Qingming Huang. RETTA: Retrieval-enhanced test-time adaptation for zero-shot video captioning. *Pattern Recognition*, 171(Part A):112170, 2026. doi: 10.1016/j.patcog.2025.112170.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679.

Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Exo2EgoDVC: Dense video captioning of egocentric procedural activities using web instructional videos. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 8313–8324, February 2025.

Karim Radouane, Julien Lagarde, Sylvie Ranwez, and Andon Tchechmedjiev. Guided attention for interpretable motion captioning. In *Proceedings of the 35th British Machine Vision Conference (BMVC)*. BMVA, 2024a. URL `https://papers.bmvc2024.org/0987.pdf`.

Karim Radouane, Andon Tchechmedjiev, Julien Lagarde, and Sylvie Ranwez. Motion2language, unsupervised learning of synchronized semantic motion segmentation. *Neural Computing and Applications*, 36(8):4401–4420, 2024b. doi: 10.1007/s00521-023-09227-z.

Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. doi: 10.1021/ac60214a047.

Shuzheng Si, Haozhe Zhao, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Bofei Gao, Kangyang Luo, Wenhao Li, Yufei Huang, Gang Chen, Fanchao Qi, Minjia Zhang, Baobao Chang, and Maosong Sun. Teaching large language models to maintain contextual faithfulness via synthetic tasks and reinforcement learning, 2025.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Bizhu Wu, Jinheng Xie, Meidan Ding, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. FineMotion: A dataset and benchmark with both spatial and temporal annotation for fine-grained motion generation and editing, 2025a.

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 (CVPR)*, pages 27849–27858, June 2025b. doi: 10.1109/CVPR52734.2025.02593.

Kangyi Wu, Pengna Li, Jingwen Fu, Yizhe Li, Yang Wu, Yuhan Liu, Jinjun Wang, and Sanping Zhou. Event-equalized dense video captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8417–8427, June 2025c. doi: 10.1109/CVPR52734.2025.00788.

Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. MotionLLM: Multimodal motion-language learning with large language models, 2024.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. In *International Conference on Learning Representations*, 2025d.

Zeyu Xi, Ge Shi, Xuefen Li, Junchi Yan, Zun Li, Lifang Wu, Zilin Liu, and Liang Wang. A simple yet effective knowledge guided method for entity-aware video captioning on a basketball benchmark. *Neurocomputing*, 619:129177, 2025a. doi: 10.1016/j.neucom.2024.129177.

Zeyu Xi, Ge Shi, Haoying Sun, Bowen Zhang, Shuyi Li, and Lifang Wu. EIKA: Explicit and implicit knowledge-augmented network for entity-aware sports video captioning. *Expert Systems with Applications*, 274:126906, 2025b. doi: 10.1016/j.eswa.2025.126906.

Zeyu Xi, Haoying Sun, Yaofei Wu, Junchi Yan, Haoran Zhang, Lifang Wu, Liang Wang, and Changwen Chen. Player-centric multimodal prompt generation for large language model based identity-aware basketball video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24330–24339, October 2025c.

Zeyu Xi, Ya Jing, Haoying Sun, Haoran Zhang, and Lifang Wu. DISCO: Disentangled identity-action extraction and spatiotemporal context modeling for llm-based identity-aware basketball video captioning. *Pattern Recognition*, 176:113188, 2026. doi: 10.1016/j.patcog.2026.113188.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13525–13536, June 2024.

Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. ReMoGPT: Part-level retrieval-augmented motion-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9635–9643, 2025. doi: 10.1609/aaai.v39i9.33044.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. MotionGPT: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7368–7376, 2024. doi: 10.1609/aaai.v38i7.28567.

Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9832–9841, June 2021. doi: 10.1109/CVPR46437.2021.00971.

Bingchan Zhao, Xinyi Liu, Zhuocheng Yu, Tongchen Yang, Yifan Song, Mingyu Jin, Sujian Li, and Yizhou Wang. Exploring fine-grained human motion video captioning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5247–5264, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18243–18252, June 2024. doi: 10.1109/CVPR52733.2024.01727.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *International Conference on Learning Representations*, 2026.

A Methodology Details

A.1 Notation and Scope

Table 7: Core notation used throughout the paper.

Symbol	Meaning
$\mathbf{P} = \{P_t\}_{t=1}^{T_v}$	2D/3D pose sequence with J joints
$\mathbf{B} = \{B_t\}_{t=1}^{T_v}$	orientation/Tait–Bryan features derived from pose
$\mathcal{E} = \{(s_i^p, e_i^p, r_i)\}_{i=1}^{N}$	coarse temporal proposals from the frozen pose-based upstream segmenter
$(\hat{s}_i, \hat{e}_i)$	refined start/end boundaries for event i
$\boldsymbol{\phi}_i$	deterministic symbolic kinematic descriptor for event i
$\mathcal{G} = (\mathcal{V}, \mathcal{E}_G)$	refined event graph
$\mathcal{M}$	structured motion memory bank
$\mathcal{R}_K^*$	top-K diverse evidence selected by GCHR
$\hat{\mathbf{y}}_i$	event-level generated caption
$\hat{y}^{\text{global}}$	sequence-level global motion narrative
$\mathcal{L}_{\text{faith}}$	deterministic kinematic faithfulness objective

MoRe-DVC is explicitly a *pose-conditioned* narration framework. Raw video is used only by an upstream 3D pose estimator when a dataset does not already provide poses. The default coarse temporal proposals are generated from pose alone, and no raw RGB frames are consumed by the downstream MoRe-DVC modules. This scope is important when interpreting both the comparisons and the failure modes discussed below.

A.2 Pose–Orientation Encoding (POEM)

POEM converts each refined temporal event into a dense representation and a deterministic kinematic vector

$$\boldsymbol{\phi}_i = \Phi(P_{\hat{s}_i:\hat{e}_i}, B_{\hat{s}_i:\hat{e}_i}) = [v_i, a_i, \omega_i, \text{sal}(i), \text{sym}(i), \text{phase}(i), \text{contact}(i)]. \tag{11}$$

Raw poses are pelvis-centered and normalized by average torso length. Missing or low-confidence joints are filled using confidence-weighted linear interpolation. Before differentiation, spatial trajectories are smoothed with a Savitzky–Golay filter [Savitzky and Golay, 1964]. We use a window length $w = 9$ and polynomial order $n = 3$ for all datasets. For 3D sequences, Tait–Bryan orientation follows standard roll–pitch–yaw conventions [Diebel, 2006]. When SMPL transformations are unavailable, local reference frames are built from adjacent limb vectors; for strictly 2D poses, angular velocity is approximated from temporal changes of projected inter-joint angles.

Body-part **saliency** $\text{sal}(i)$ is the temporal variance of joint angular velocity. **Symmetry** $\text{sym}(i)$ is an inverted L_2 distance between root-relative mirrored left/right extremity trajectories. **Phase** $\text{phase}(i)$ is determined from zero crossings of the most salient joint's acceleration profile. **Contact** is deterministic: a support joint is marked when its local velocity is below 0.05 m/s and it lies within the predefined ground-plane margin. These descriptors are intentionally closed-form so that the same slots can be audited by the text parser used in the faithfulness objective.

A.3 Learning Objectives and Training Pipeline

The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \alpha\mathcal{L}_{\text{ret}} + \beta\mathcal{L}_{\text{bnd}} + \gamma\mathcal{L}_{\text{faith}}^{\text{PG}} + \zeta\mathcal{L}_{\text{ord}}. \tag{12}$$

The captioning loss is

$$\mathcal{L}_{\text{cap}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{L_i}\log \pi_\theta(\hat{y}_{i,t} \mid \hat{\mathbf{y}}_{i,<t}, \mathcal{C}). \tag{13}$$

The retrieval loss aligns dense motion and caption embeddings with a temperature-scaled contrastive objective,

$$\mathcal{L}_{\text{ret}} = -\log \frac{\exp(z_i^m \cdot z_i^c/\tau)}{\sum_j \exp(z_i^m \cdot z_j^c/\tau)}. \tag{14}$$

The boundary loss is

$$\mathcal{L}_{\text{bnd}} = \frac{1}{N}\sum_{i=1}^{N}\left[\rho(\hat{s}_i - s_i^*) + \rho(\hat{e}_i - e_i^*)\right], \tag{15}$$

and the temporal-order loss is

$$\mathcal{L}_{\text{ord}} = \frac{1}{N-1}\sum_{i=1}^{N-1}\max\left\{0, \Delta - (\hat{t}_{i+1} - \hat{t}_i)\right\}, \qquad \hat{t}_i = \frac{\hat{s}_i + \hat{e}_i}{2}. \tag{16}$$

For the kinematic reward, a frozen parser maps the generated caption back to the canonical descriptor slots. With $c_i^{\text{faith}} = \|E_\psi(\hat{\mathbf{y}}_i) - \boldsymbol{\phi}_i\|_2^2$ and $r_i^{\text{faith}} = -c_i^{\text{faith}}$,

$$\mathcal{L}_{\text{faith}}^{\text{PG}} = -\frac{1}{N}\sum_{i=1}^{N}(r_i^{\text{faith}} - b_i)\sum_{t=1}^{L_i}\log \pi_\theta(\hat{y}_{i,t} \mid \hat{\mathbf{y}}_{i,<t}, x_i, p_i, \mathcal{M}). \tag{17}$$

The reward is therefore a structured consistency regularizer on the decoder output, not an independent sensor signal. Its purpose is to prevent language priors from overriding pose-conditioned evidence. The causal contribution of this term is isolated in Appendix C.5.

A.4 Inference Pipeline

At inference time, the frozen upstream stage first produces pose and coarse proposals. POEM computes dense/symbolic motion descriptors, the boundary refiner adjusts the proposal endpoints, and the refined events are linked into an event graph. GCHR then retrieves training-only exemplars, past intra-video evidence, and domain rules using semantic, kinematic, graph, and phase compatibility followed by MMR. The selected evidence is serialized into the event prompt. Event captions are generated autoregressively, after which the summary aggregator produces the global motion narrative. Ground-truth boundaries or captions are never available during inference.

B Experimental Details

B.1 Datasets

We evaluate three motion-centric datasets. **BoFiT** [Zhao et al., 2025] contains 2,360 trimmed fitness-training clips at 480×270, 30 FPS, with an average duration of 2.94 s and fine-grained exercise descriptions. We use BoFiT for fine-grained event narration and all controlled ablations. **HumanML3D** [Guo et al., 2022a] contains 14,616 motion sequences and 44,970 sequence-level descriptions and is used for coarse motion-to-text captioning. **FineMotion** [Wu et al., 2025a] augments motion-language data with dense body-part-level spatial-temporal descriptions, including more than 442K motion snippets and approximately 95K paragraph-level sequence descriptions. We follow the official dataset splits used by the corresponding benchmark protocols.

B.2 Evaluation Protocol, Task Alignment, and Baseline Reproduction

Each main-paper results table is interpreted only within the dataset's native task, as summarized in Table 8. This distinction is important because the three datasets do not define identical generation tasks.

For HumanML3D and FineMotion, we use the authors' official code/checkpoints, standard splits, and the standard motion–text evaluator, matching the MG-MotionLLM protocol. We additionally re-ran

Table 8: Task alignment and comparison protocol.

Dataset	Native task used here	Comparison set	Protocol
HumanML3D	Motion → text	TM2T, MotionGPT, MG-MotionLLM; MotionGPT3 in App. B.2.1	official splits + standard motion–text evaluator
FineMotion	Motion → detailed text	MG-MotionLLM small/base/large	official FineMotion/MG-MotionLLM protocol
BoFiT	Fine-grained motion-video narration	BoFiTGen; inherited Video-LLM references	official BoFiT splits + metric implementation
BoFiT controls	Same task/backbone/input	Kimi-K2 direct/RAG/module variants	identical prompts, poses, proposals, decoding

the released MG-MotionLLM checkpoint inside our evaluation harness and reproduced its published metrics within ± 0.3 on all reported metrics. For BoFiT, we follow the official benchmark of Zhao et al. [Zhao et al., 2025]; numbers for shared baselines are taken from that benchmark using identical split and metric definitions. The general Video-LLM rows in the BoFiT table are therefore reference points inherited from the benchmark rather than the basis of our claimed architectural margin.

The decisive causal comparisons are the same-backbone controls: all BoFiT control variants use the identical Kimi-K2 generator, identical pose/proposal inputs, identical prompt templates, and temperature 0.2 decoding. Only the proposed components are changed. All MoRe-DVC hyperparameters and graph coefficients are held fixed across the three datasets; no per-dataset tuning is used.

B.2.1 Additional HumanML3D Baseline: MotionGPT3

Following the review request, we evaluated the official MotionGPT3 checkpoint [Zhu et al., 2026] under the identical HumanML3D evaluation harness. Table 9 provides the comparison without expanding the main-paper table.

Table 9: Additional HumanML3D comparison under the identical evaluation protocol.

Method	R-P@1 ↑	R-P@3 ↑	MM-Dist ↓	BLEU-4 ↑	BERT ↑
MotionGPT3 [Zhu et al., 2026]	60.3	87.2	2.517	12.9	37.6
MoRe-DVC	**61.5**	**88.5**	**2.450**	**13.85**	**38.4**

B.3 Evaluation Metrics

Text quality is measured using BLEU, ROUGE, METEOR, CIDEr, and BERTScore. Motion-language alignment on HumanML3D uses R-Precision and MM-Dist following the standard motion–text evaluator. BoFiT additionally uses FCE-Motion [Zhao et al., 2025], which is sensitive to fine-grained motion content and verb/order consistency. Except for MM-Dist and FCE-Motion, the main tables display metrics on a $[0, 100]$ scale. We use the official metric implementations associated with each benchmark and do not retune evaluation thresholds per method.

B.4 Implementation Details and Frozen Stage 0

Stage 0: pose and coarse proposal extraction. The revised pipeline makes the frozen upstream stage explicit. When 3D pose is not supplied by the dataset, an off-the-shelf pose estimator first converts the source clip to a 3D skeleton sequence. The default temporal proposal generator is *pose-only*: we smooth angular-velocity and motion-energy profiles, detect kinematic change points, and assign the coarse label r_i using a lightweight pose-based action classifier. No RGB frame is consumed by MoRe-DVC after pose extraction. Coarse proposal extraction is a one-time preprocessing step and takes approximately 12 ms/clip in our profiling. The source clips are sampled at 30 FPS; the phrase "pose sequences and videos" in the review manuscript referred to sampling the source clips before this upstream estimator, not to raw-frame input within MoRe-DVC.

Fixed preprocessing and optimization. POEM uses Savitzky–Golay smoothing with $w = 9$ and polynomial order $n = 3$. The boundary search buffer is $\delta = 15$ frames. Each event window is capped at 300 frames. GCHR uses $K = 5$ retrieved items and fixed retrieval coefficients $\lambda_1 = 0.4$ (semantic), $\lambda_2 = 0.3$ (kinematic), $\lambda_3 = 0.15$ (graph), $\lambda_4 = 0.15$ (phase), with redundancy weight $\lambda_5 = 0.2$. Training uses AdamW, learning rate 10^{-4}, cosine annealing, 1,000 warmup steps, weight decay 0.01, per-GPU batch size 8, and an effective global batch size 64. The loss coefficients are $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 0.2$, and $\zeta = 0.1$. Training runs for 20 epochs. These values and generation temperature 0.2 are fixed across all datasets.

Flagship and open configurations. The flagship generator is `kimi-k2-instruct-0905` (1T total parameters, 32B active) accessed through the hosted deployment used in our experiments. The flagship configuration uses the same generator-side optimization path and faithfulness objective described in Sec. 3.3, while the surrounding motion-processing, event-refinement, retrieval, and verification components are trained or executed within our pipeline. Frozen embedding/re-ranking services are used for the flagship retrieval stack. The community-reproducible configuration removes the remote-scale and constrained-service dependencies entirely: its generator is a locally hosted Llama-3-70B adapted with LoRA, and its embedding and re-ranking components are also locally hosted open-weight models. Exact checkpoint identifiers, revisions, and loading commands are provided in the released repository so that the paper is not tied to transient model-hub aliases. The local open-stack experiment runs on one academic node with 4×A100-40GB and takes 2.1 days. Table 13 reports its complete three-dataset results.

The retrieval memory contains 100K training-only exemplars. Embeddings are stored as 4096-d fp16 vectors (≈ 0.78 GB), while symbolic descriptors and graph metadata require approximately 0.11 GB, for a total below 0.9 GB. All FAISS retrieval is exact inner-product search in the reported configuration.

B.5 Multi-Run Reliability and Statistical Significance

We repeated the full model and controlled BoFiT baselines over three random seeds. Mean±standard deviation is reported in Table 10. Retrieval itself is deterministic (exact FAISS search with fixed coefficients), and generation uses temperature 0.2, which limits run-to-run variance.

Table 10: Three-seed reliability on BoFiT. Values are mean±standard deviation across complete training runs.

Variant	FCE-M ↑	BERT ↑
Kimi-K2 Direct Prompt	21.4 ± 0.5	39.9 ± 0.4
Kimi-K2 + Text RAG	23.1 ± 0.4	41.2 ± 0.3
MoRe-DVC (Full)	$\mathbf{29.2 \pm 0.3}$	$\mathbf{46.5 \pm 0.2}$

We additionally use paired bootstrap tests with 10,000 resamples over test clips. The full model's gains over both controlled baselines are significant at $p < 0.001$. For HumanML3D, we follow the customary 20-trial evaluation protocol; MoRe-DVC obtains R-P@1 61.5 ± 0.3 and MM-Dist 2.450 ± 0.008 (95% confidence intervals). Paired bootstrap comparisons against MotionGPT3 and MG-MotionLLM are significant at $p < 0.01$. The same test-set bootstrap procedure is used for FineMotion.

B.6 Qualitative Motion Narration

The example is intended to illustrate how event-level retrieval and symbolic motion features support localized narration. In the final camera-ready figure, all highlighted physical statements should be traceable either to the reference annotation or to the deterministic pose-derived descriptors; unsupported embellishment should not be presented as evidence of improved fidelity.

C Additional Ablations and Analyses

C.1 Physical Hallucination Taxonomy

The manual error taxonomy directly tests the claim that pose conditioning alone does not prevent decoder-side physical errors. Table 11 extends the original direct-vs-full comparison with the full-retrieval model at $\gamma = 0$, i.e., the configuration that still sees POEM descriptors but receives no faithfulness reward.

The direct prompt already receives the same pose input but still exhibits substantial directional, temporal-order, and body-part errors. Thus the failure being targeted is not invalid input motion; it is the decoder's failure to preserve the conditioning signal in generated language. The $\gamma = 0$

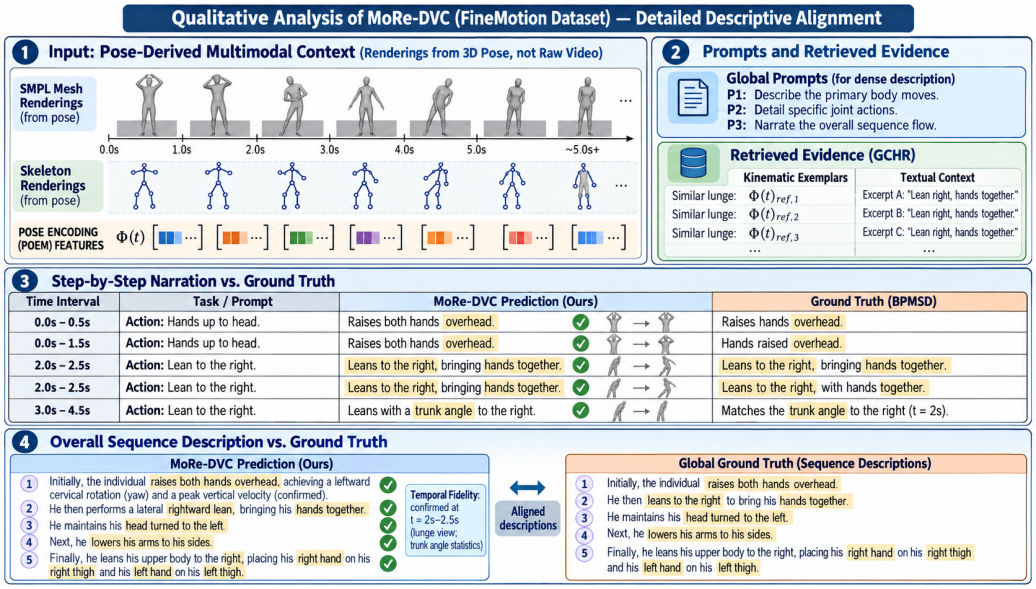

Figure 6: Qualitative FineMotion example showing pose-derived motion context, retrieved evidence, event-level narration, and sequence-level description. The rendered meshes/keyframes are visualization of pose, not an additional RGB input to MoRe-DVC.

Table 11: Physical hallucination error taxonomy (% of captions containing at least one error).

Error Type	**Direct Prompt**	**Full, $\gamma = 0$**	**MoRe-DVC Full**
Wrong Body Part	14.2%	12.6%	**3.8%**
Wrong Direction	18.5%	15.8%	**1.2%**
Wrong Temporal Order	22.1%	19.6%	**4.5%**
Unsupported Biomechanical Claim	11.0%	9.4%	**2.1%**

column shows that retrieval and pose conditioning alone leave many of these errors intact, while the faithfulness objective sharply reduces them.

C.2 Inference Latency, Parameters, and Storage

Table 12 reports the per-event cost measured on one A100. The generator dominates latency and is required by both the full method and the direct-prompt LLM baseline; the retrieval/graph components add a comparatively small overhead.

Table 12: Per-event computational budget. Latency is measured on one A100.

Component	**Trainable Params**	**Latency**
POEM	3.1M	4 ms
Boundary refinement + event graph	1.9M	5 ms
GCHR (embed/rerank/MMR)	0 (frozen)	38 ms
Retrieval-conditioned generation	LoRA 82M	$\sim$ 0.9 s

POEM plus GCHR was previously summarized as approximately 45 ms/event; the decomposed profile above explains that aggregate. Vectorized MMR and graph traversal keep retrieval below 80 ms per sequence in the evaluated setting. Relative to the generation-dominated baseline, the proposed modules add approximately 7% end-to-end latency while improving FCE-M. The one-time pose-based proposal extraction adds approximately 12 ms/clip. Storage is also modest: 100K 4096-d fp16 exemplar embeddings require approximately 0.78 GB, and symbolic/graph metadata require approximately 0.11 GB, totaling less than 0.9 GB.

C.3 Complete Open-Weight Single-Node Reproduction

To separate architecture from flagship backend scale and proprietary service access, we evaluate a fully local stack on one academic node (4×A100-40GB) with a locally hosted Llama-3-70B generator and locally hosted open-weight embedding/re-ranking components. No remote API is used anywhere in this configuration. Exact checkpoint identifiers and revisions are included in the released repository. Training takes 2.1 days. Table 13 reports all three datasets.

Table 13: Full open-stack reproduction versus the flagship configuration and the strongest prior result available for each metric.

Dataset / metrics	Open stack (local 70B)	Flagship (Kimi-K2)	Best prior (per metric)
HumanML3D: R-P@1 / MM-Dist / B-4 / BERT	60.8 / 2.489 / 13.1 / 37.8	**61.5 / 2.450 / 13.85 / 38.4**	60.3 / 2.517 / 12.9 / 37.6
FineMotion: B@1 / B@4 / ROUGE / BERT	82.7 / 54.6 / 66.7 / 54.0	**83.50 / 55.40 / 67.5 / 54.8**	81.78 / 53.75 / 65.8 / 52.4
BoFiT: B@4 / CIDEr / BERT / FCE-M	9.4 / 8.0 / 45.8 / 28.5	**9.8 / 8.5 / 46.5 / 29.2**	8.3 / 6.8 / 43.8 / 27.0

For HumanML3D the prior row is MotionGPT3 under the identical protocol; for FineMotion it takes the strongest MG-MotionLLM value per metric; for BoFiT it takes the strongest BoFiTGen value per metric. The local stack trails the flagship by at most 0.9 absolute points on the displayed higher-is-better metrics (and 0.039 MM-Dist), while preserving the same qualitative conclusions. This experiment supports portability to academic compute, but it does not eliminate the limitation that absolute performance depends on generator scale.

C.4 Memory-Bank Size Dependence

To quantify dependence on retrieval database size, we sweep the number of exemplar entries while holding all other components fixed.

Table 14: Effect of memory-bank size on BoFiT.

Exemplars M	10K	25K	50K	100K
FCE-M ↑	27.8	28.4	28.9	**29.2**

The gain exhibits diminishing returns. Even the 10K bank remains substantially above the direct-prompt operating point, while the 100K bank provides the strongest result. We therefore treat memory-bank size as a performance dependency and a deployment trade-off rather than a fixed property of the architecture.

C.5 Faithfulness Reward Factorial and Causal Isolation

Table 15 resolves the apparent discrepancy between the main-paper POEM add-back and the $\gamma = 0$ sweep. These are different cells of a factorial design.

Table 15: Factorial isolation of retrieval, prompt-side symbolic descriptors, and the faithfulness reward on BoFiT. All variants use identical Kimi-K2 prompts/decoding except for the indicated factors.

Variant	**Retrieval**	ϕ_i **in prompt**	$\mathcal{L}_{\text{faith}}$	**FCE-M ↑**
Direct prompt	✗	✗	✗	21.4
Direct + ϕ_i	✗	✓	✗	23.3
Direct + ϕ_i + reward	✗	✓	✓	24.5
Hybrid w/o POEM	✓	✗	✗	27.2
Full, $\gamma = 0$	✓	✓	✗	22.0
Full, reward-only (ϕ_i hidden)	✓	✗	✓	28.3
Full	✓	✓	✓	**29.2**

Rows 4 and 6 hold graph-conditioned retrieval fixed and toggle only the reward, yielding a +1.1 FCE-M gain from the reward alone. Rows 5 and 7 show the stronger interaction: when rich retrieval context

and numeric descriptors are both present, the reward converts descriptor injection from a liability into a gain. In the sparse direct-prompt regime, descriptors alone provide a modest improvement and the reward adds a further increment. This pattern is more precise than the earlier "modality distraction" explanation: the failure at $\gamma = 0$ is *unanchored specificity*, where the decoder emits more specific kinematic statements but misstates direction, body part, or order.

C.6 Univariate Isolation of Graph-Conditioned Retrieval

To isolate graph conditioning, we hold POEM, $\mathcal{L}_{\text{faith}}$, boundary refinement, MMR, and $K = 5$ fixed and vary only the retrieval conditioning.

Table 16: Univariate graph-conditioning ablation on BoFiT.

Retrieval Configuration	**FCE-M ↑**	**BERT ↑**
Motion-only retrieval, no graph	27.0	44.9
Hybrid semantic+kinematic, no graph ($\lambda_3 = 0$)	27.6	45.4
Hybrid + graph conditioning	**29.2**	**46.5**

Graph conditioning contributes +1.6 FCE-M and +1.1 BERT over the hybrid no-graph condition, larger than the +0.6 FCE-M gain from semantic–kinematic hybridization alone. At the w/o-POEM operating point, the same toggle improves FCE-M from 26.1 to 27.2. The effect is concentrated in chronological consistency: removing graph conditioning raises the temporal-order error rate from 4.5% to 9.8% in the manual taxonomy.

C.7 Proposal-Source Robustness

The default proposal source is pose-only change-point detection. To test whether boundary refinement depends on that source, we evaluate uniform windows and ground-truth boundaries perturbed by ± 20 frames.

Table 17: Boundary refinement and downstream FCE-M under three proposal sources.

Proposal Source	**Input mIoU**	**Refined mIoU**	**Final FCE-M**
Uniform sliding windows	35.2	54.6	28.3
Pose change-point (default)	42.1	58.7	29.2
GT boundaries jittered ± 20 frames	44.8	59.3	29.4

The refiner improves mIoU by 15–19 points across all three sources, while final FCE-M varies only from 28.3 to 29.4. This supports the claim that the downstream architecture is not tied to a single proposal generator.

C.8 Pose-Noise Robustness

We retain the original stress test of the downstream pose-conditioned pipeline, but move it out of the main paper to keep the camera-ready results section focused. Gaussian perturbations $\mathcal{N}(0, 0.05)$ together with 15% random joint dropout are injected into test poses. Under this corruption, MoReDVC attains FCE-M $= 26.8$. This result should be interpreted as robustness to moderate pose corruption, not as evidence that the framework can recover from catastrophic upstream tracking failures or severe multi-person occlusion.

C.9 Expression Diversity and Over-Constraint

A deterministic parser can incorrectly penalize valid but novel expressions. We quantify this effect using Distinct-1, Distinct-2, and the proportion of generated bigrams absent from the training references.

At the operating point $\gamma = 0.2$, Distinct-2 decreases by 3.5% relative to the $\gamma = 0$ full model. A manual audit of 200 sampled captions by two annotators ($\kappa = 0.74$) found that the reward penalized a

Table 18: Lexical diversity and novelty under different faithfulness weights on BoFiT.

Configuration	Distinct-1	Distinct-2	Novel Bigrams
Direct prompt	0.118	0.463	14.9%
Full, $\gamma = 0$	0.116	0.458	14.5%
Full, $\gamma = 0.2$	0.112	0.447	13.8%
Full, $\gamma = 1.0$	0.093	0.381	9.6%

correct but novel phrasing in 3.5% of cases at $\gamma = 0.2$, compared with 11.2% at $\gamma = 1.0$. The parser canonicalizes common synonyms such as "raises," "lifts," and "elevates," so the default weight mainly penalizes content deviations rather than harmless paraphrase. Over-constraint becomes pronounced only at larger γ, consistent with the downslope of the main-paper sweep. Importantly, this effect cannot explain the 22.0 FCE-M cell because that cell has $\gamma = 0$ and therefore no active faithfulness constraint.

C.10 Per-Descriptor Ablation

To assess whether the hand-designed POEM features contribute beyond a single dominant statistic, we remove each descriptor while holding the remainder fixed. Table 19 reports the decrease in FCE-M relative to the full model.

Table 19: Per-descriptor FCE-M decrease when the indicated POEM descriptor is removed.

Removed Descriptor	ΔFCE-M
Angular velocity ω	-1.4
Kinematic phase	-1.1
Saliency	-0.6
Symmetry	-0.4
Contact	-0.3

Angular velocity and phase contribute most strongly, while saliency, symmetry, and contact provide smaller complementary gains. The descriptors are deliberately deterministic: replacing them with a learned latent feature would weaken the auditability of the text-to-kinematics faithfulness check.

D Limitations and Reproducibility Notes

Backend scale. Absolute performance depends on generator capacity. The main paper therefore does not interpret comparisons against smaller published models as like-for-like causal evidence. Causal claims instead rely on the same-backbone controls and the 70B open-stack reproduction in Appendix C.3.

Memory-bank size. Retrieval performance increases with memory size but exhibits diminishing returns (Appendix C.4). This creates a storage/latency trade-off and should be considered when deploying on constrained hardware.

Accessibility. The flagship configuration depends on access to a remotely hosted 1T Kimi-K2 backend. The fully local 70B stack removes this dependency and runs on a 4×A100-40GB academic node, but the two configurations are not computationally identical. The camera-ready repository releases the implementation, LoRA weights, prebuilt FAISS index, locally hosted open-weight checkpoint identifiers/revisions, and the complete open-stack reproduction recipe.

Pose and parser failures. MoRe-DVC cannot recover reliably from catastrophic upstream pose failures caused by severe occlusion or tracking errors. Separately, the rule parser may penalize rare but physically correct expressions, especially at large γ. The diversity audit in Appendix C.9 quantifies this trade-off and motivates future work on more flexible but still auditable parsers.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction define the framework as pose-conditioned, event-grounded narration downstream of frozen upstream estimators. They foreground graph-conditioned retrieval, position the symbolic faithfulness term as a domain-instantiated structured reward rather than a new reinforcement-learning mechanism, and bound the claims using same-backbone controls and open-stack validation.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The conclusion and Appendix discuss dependence on upstream pose quality, proposal quality, generator scale, memory-bank size, rule-parser coverage, and the accessibility cost of the flagship remote backbone. They also report proposal-source robustness, memory-size scaling, computational overhead, and an entirely local open-weights configuration.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper proposes an empirical framework and architecture (MoRe-DVC) for multimodal retrieval-augmented generation and does not introduce formal theoretical proofs or theorems.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 4.1 and the Appendix specify the task, benchmark protocols, proposal source, architecture, hyperparameters, retrieval coefficients, decoding settings, statistical protocol, and computational profile. The Appendix additionally reports an entirely local Llama-3-70B open-weights stack evaluated on all three datasets on a single $4\times$A100-40GB academic node.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The camera-ready release provides the implementation, LoRA weights, the prebuilt FAISS index, exact locally hosted open-weight checkpoint identifiers/revisions, and the complete open-stack reproduction recipe. These materials reproduce the single-node Llama-3-70B configuration reported in the Appendix without requiring proprietary retrieval services.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental setup, including datasets (BoFiT, HumanML3D, FineMotion), metrics, backbone LLMs, and training/loss details are specified in Section 4.1, Section 4.3, and the Appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The Appendix reports three-seed training mean±standard deviation for the full model and controlled BoFiT baselines, paired bootstrap significance tests with 10,000 test-clip resamples, and the customary 20-trial HumanML3D evaluation with 95% confidence intervals. The variability source and resampling protocol are stated explicitly.

Guidelines:

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

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: The Appendix separates the flagship remote-backbone configuration from the fully local open-weights stack. The local 70B configuration is reported on one $4\times$A100-40GB node with a 2.1-day training time, and the per-event parameter, retrieval, preprocessing, storage, and latency costs are itemized.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research relies on standard, publicly available motion and video captioning datasets and strictly conforms to the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: Broader impacts are not explicitly detailed in the main text due to space constraints; however, the core focus of the paper—reducing physical hallucinations—directly mitigates the risk of generating inaccurate or misleading physical assertions.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The framework processes 3D skeleton sequences to narrate human kinematic motion and does not inherently pose a high risk for societal misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets utilized (BoFiT, HumanML3D, FineMotion) and base foundation models are properly cited and referenced within the Experimental Setup and Related Work sections.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper utilizes existing public datasets for evaluation and does not release new datasets or data assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The research utilizes pre-existing datasets and does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: Not applicable, as the study does not involve new human subjects data collection.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The usage of LLMs is brought up for editing and formatting purposes. Secondly, The core methodology of the paper, which fundamentally relies on large language models (e.g., Kimi-k2-instruct, Llama-based NIMs) for the proposed MoRe-DVC architecture.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.