VTRAJEMBED: ALIGNING TEXT WITH VIDEO OBJECT TRAJECTORIES VIA SPATIOTEMPORAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained alignment between text and video object trajectories is a core challenge in multimodal understanding. Existing embedding models excel at global retrieval, yet their single global representation entangles multiple objects and their spatiotemporal dynamics, limiting direct object-level grounding. To address this issue, we propose **VTrajEmbed**, an embedding model that decomposes a video into object trajectories and aligns each trajectory embedding with text for object-level grounding through retrieval. Our main contributions: (1) **Slow-Fast Trajectory Encoding**, a dual-branch architecture that establishes trajectory representations by densely sampling low-resolution frames for high-frequency motion cues and sparsely selecting high-resolution frames for scene context, aggregated by a shared LLM backbone; (2) **Mask-Guided Contrastive Learning**, a training objective built upon these representations that couples an intra-group loss and a batch-level bidirectional loss under a relation mask, excluding unverified relations and preventing confirmed matches from becoming false negatives; (3) **Text2Traj-Bench**, a trajectory-level text-video benchmark with model-assisted construction and human-verified query–trajectory relations. Experimental results show that VTrajEmbed-5B improves text-to-trajectory and trajectory-to-text R@1 over Qwen3-VL-Embedding-8B by **15.48** and **18.05** percentage points, respectively, and achieves 49.48% average R@1 across five global retrieval benchmarks, compared with 42.86% for Qwen3-VL-Embedding-8B.

1 INTRODUCTION

Multimodal embedding models have become a key technique for mapping data from different modalities into a unified semantic space. The contrastive vision-language pretraining paradigm Radford et al. (2021); Zhai et al. (2023) has learned high-quality alignment on large-scale image-text pairs, and video embedding models further incorporate temporal dynamics into unified representations Zhu et al. (2024); Li et al. (2026b), achieving notable progress in video-level retrieval. However, this success relies on the premise of compressing an entire video into a single semantic unit, and when the task shifts from "does this video match this sentence" to "which object in the scene corresponds to this description," the granularity limitation of embedding models becomes apparent.

Real-world applications such as autonomous driving, robotic manipulation, and digital content moderation Zitkovich et al. (2023); Sima et al. (2024); Li et al. (2026a) demand object-level alignment. In the static-image domain, FG-CLIP Xie et al. (2026) and ObjEmbed Fu et al. (2026) achieve region-text matching through fine-grained annotation or object-level decomposition, yet these advances do not extend to video: general-purpose video embedding models typically compress multiple objects and their dynamics into a single global representation Zhou et al. (2026), and training signals remain sample-level without object-level positive or negative examples. Video grounding tasks introduce spatiotemporal supervision Gu et al. (2025); Gao et al. (2026) but output localized tubes rather than a shared text-trajectory embedding space. A concurrent work Chen et al. (2026) similarly leverages ego-trajectory similarity as training-only supervision, yet still produces single-vector representations without object-level alignment. How to equip embedding models with object-level alignment while keeping learned representations useful for broader video-level retrieval remains an open problem.

To address these challenges, we propose **VTrajEmbed**, an embedding model that uses *object trajectories* as the basic alignment unit, mapping object-level visual representations and text into a shared semantic space. Given a video and a user query, VTrajEmbed first uses a lightweight query parser to extract textual descriptions of the referenced objects, which serve as prompts for a segmentation-and-tracking model to generate candidate trajectories. Each trajectory is then encoded through a

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.

Slow-Fast dual-branch architecture, where dense low-resolution observations capture object appearance and motion, while sparse high-resolution frames provide scene context. A shared language-model backbone aggregates the trajectory features and separately encodes the original query, producing embeddings for similarity-based retrieval. For training, a mask-guided contrastive objective learns from confirmed query-trajectory matches and mismatches while excluding unverified relations. To provide relation supervision and evaluate object-level alignment, we construct **Text2Traj-Bench**, a model-assisted benchmark with human-verified query-trajectory relations. Unlike query-conditioned proposal generation in deployment, Text2Traj-Bench generates candidates from fixed category prompts and freezes each video's candidate pool before query generation, ensuring that all trajectory-level evaluations use the same query-independent candidates while the associated verified relations define relevance for object-centric video retrieval. Our main contributions:

- A Slow-Fast trajectory encoding architecture that captures high-frequency motion features via low-resolution dense sampling and scene context via high-resolution sparse sampling.
- A mask-guided contrastive learning objective that enables object-level training, coupling local discrimination and bidirectional batch retrieval under a query-trajectory relation mask.
- A model-assisted, human-verified benchmark that couples within-video evaluation with dense within-video and sparse cross-video relation annotations.
- VTrajEmbed-5B achieves 69.67% text-to-trajectory and 66.34% trajectory-to-text R@1, substantially outperforming Qwen3-VL-Embedding-8B (54.19% and 48.29%). In the global retrieval evaluation, its trajectory-derived representations average 49.48% R@1 across five widely used datasets, versus 42.86% for the same baseline.

2 Related Work

Multimodal Embedding Models map different modalities into a shared semantic space for cross-modal matching, with much of the cited work focusing on either video-level global representations or static-image object-level alignment Zhou et al. (2024); Günther et al. (2025); Zhou et al. (2026). Starting from CLIP Zhai et al. (2023), subsequent video embedding models Wang et al. (2024); Li et al. (2026b) extend contrastive pretraining to the temporal domain but produce a single global embedding per video, lacking the granularity to distinguish individual objects. A concurrent work, TraVEL Chen et al. (2026), uses ego-trajectory similarity as privileged training supervision to improve motion sensitivity of driving-video embeddings, yet still produces single-vector representations without object-level alignment. Recent efforts address fine-grained alignment in the static-image domain: FG-CLIP Xie et al. (2026) and ObjEmbed Fu et al. (2026) achieve region-level matching through fine-grained annotation or object-level decomposition, while more recent works Zhang et al. (2025); Lin et al. (2025); Huang et al. (2026) leverage multimodal LLM backbones for unified cross-modal retrieval, moving beyond the conventional separate dual-encoder paradigm. VTrajEmbed instead takes an object trajectory as its retrieval unit and explicitly aligns its appearance, motion, and scene context with text.

Open-Vocabulary Object Tracking and Grounding produce bounding boxes, masks, or spatio-temporal tubes rather than text-comparable embeddings Ren et al. (2024); Qian et al. (2025). Open-vocabulary detectors Cheng et al. (2024); Liu et al. (2024); Fu et al. (2025); Lu et al. (2025) enable text-driven localization of arbitrary categories. For video segmentation and tracking, recent methods Ravi et al. (2025); Carion et al. (2026); Li et al. (2025); Liang et al. (2025) provide complementary capabilities for video mask propagation, object association, and open-vocabulary segmentation. VTrajEmbed bridges this gap by encoding object trajectories as embedding vectors aligned with text, supporting trajectory grounding through embedding retrieval.

3 Method

VTrajEmbed follows a proposal-and-ranking pipeline (Figure 1). A lightweight query parser extracts the subject to prompt a frozen segmentation-and-tracking module for query-conditioned trajectory proposals. Slow-Fast encoding integrates appearance, motion, and scene context for similarity ranking against the original query in a shared embedding space. Mask-guided bidirectional contrastive learning uses verified query-trajectory relations to jointly optimize within-video discrimination and bidirectional retrieval. Text2Traj-Bench provides supervision and controlled evaluation: query-independent pools frozen before query generation supply common trajectory-level candidates, while verified relations establish relevance for cross-video evaluation.

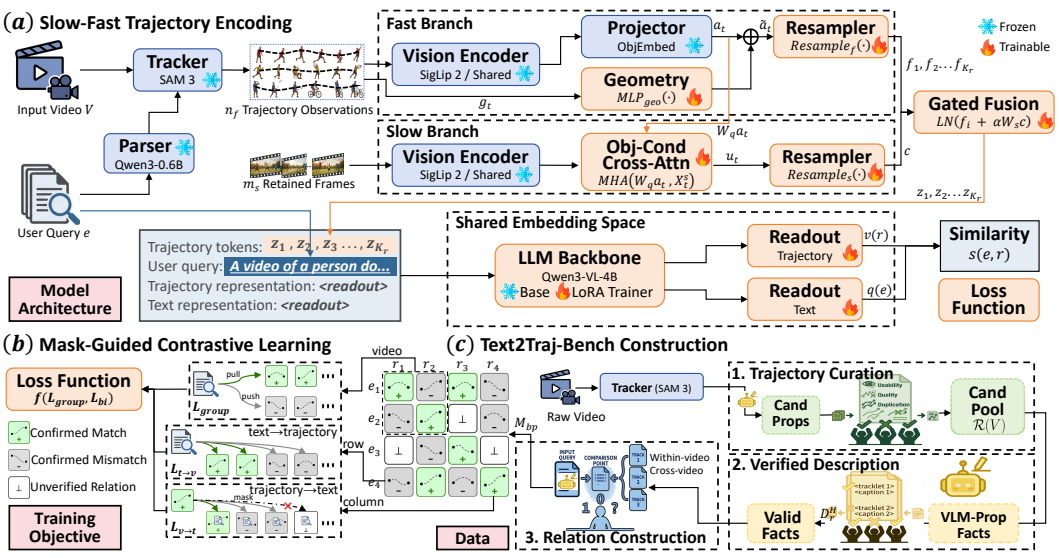


Figure 1: **Overview of VTrajEmbed. (a) Model architecture.** The Slow-Fast trajectory encoder combines dense low-resolution object observations with sparse high-resolution scene context and maps trajectories and text into a shared embedding space for similarity-based retrieval. **(b) Training objective.** Mask-guided contrastive learning jointly optimizes within-video trajectory discrimination and bidirectional trajectory-text retrieval using verified query-trajectory relations while excluding uncertain pairs. **(c) Data.** Text2Traj-Bench uses trajectory curation, verified description, and relation construction to produce fixed per-video candidate pools and human-verified annotations, including descriptions and within- and cross-video query-trajectory relations.

3.1 Slow-Fast Trajectory Encoding

Given a video V and query e, VTrajEmbed retrieves the best-matching object trajectory by ranking candidates in a shared embedding space. A lightweight parser extracts the target description from e to prompt a frozen segmentation-and-tracking model, yielding candidates $\mathcal{R}(V, e) = \{r_1, \ldots, r_K\}$. Each trajectory records the object's spatial extent and observation timestamps. VTrajEmbed independently embeds e and each candidate in a shared space and retrieves $\hat{r} = \arg\max_{r \in \mathcal{R}(V,e)} s(e, r)$ by trajectory-text similarity $s(e, r)$. Learning targets trajectory representation and cross-modal ranking; candidate generation remains frozen. For controlled evaluation, Text2Traj-Bench instead uses the query-independent pool $\mathcal{R}_{\text{bench}}(V)$, fixed before query generation and shared across models in trajectory-level evaluation; its verified relations establish relevance for object-centric video retrieval.

For each candidate, the Slow-Fast encoder combines dense low-resolution observations of fine-grained object appearance and motion with sparse high-resolution, object-conditioned scene context, as shown in Figure 1(a). Gated fusion yields an ordered feature sequence that the shared language-model backbone aggregates into a trajectory embedding. A separate text-only pass through the same backbone embeds the original query.

Fast branch: ordered object evidence. It encodes ordered object appearance and geometry from n_f uniformly sampled valid observations, processing their full frames chronologically at low pixel budget P_f. The frozen vision encoder produces spatial visual features, from which a frozen ROI projector extracts an object appearance feature a_t using the tracked box at time t. A trainable MLP embeds g_t, comprising normalized box center, size, area, inter-observation displacement, relative scale change, and original timestamp, to expose motion beyond appearance. Consecutive observations are grouped into temporal blocks B_i and pooled by a shared attention resampler:

$$\widetilde{a}_t = a_t + \text{MLP}_{\text{geo}}(g_t), \qquad f_i = \text{Resample}_f([\widetilde{a}_t]_{t \in B_i}), \quad i = 1, \ldots, K_r. \tag{1}$$

The resulting $(f_1, \ldots, f_{K_r})$ compresses local observations while retaining block order for temporal modeling in the language backbone. Timestamps expose irregular observation gaps; uniform list sampling does not imply uniform physical-time spacing.

Slow branch: object-conditioned scene context. The Slow branch complements the Fast sequence with sparse high-resolution scene context. It uniformly preselects n_s frames from ordered Fast observations and scans them chronologically using unit-normalized semantic image embeddings from a frozen auxiliary encoder. It retains the first frame and discards later frames only if cosine similarity to an already retained representative reaches τ. Representatives are never deleted and removed frames are not replenished. The retained set S contains $m_s \leq n_s$ frames and a representative for every discarded frame; only these retained frames incur high-resolution vision encoding. This pruning leaves the Fast observations and temporal blocks intact.

Object-conditioned cross-attention and gated fusion integrate the retained scene context into each valid Fast block. The branches share the frozen vision encoder's weights but run separate forward passes at pixel budgets P_f and $P_s > P_f$. At a retained time t, the matching Fast appearance feature a_t queries the high-resolution full-frame token bank X_t^s through cross-attention. A Slow attention resampler pools the resulting context features, and a learned residual gate injects their summary into each valid Fast block:

$$u_t = \mathrm{MHA}(W_q a_t, X_t^s, X_t^s), \qquad c = \mathrm{Resample}_s(\{u_t : t \in S\}), \qquad z_i = \mathrm{LN}(f_i + \alpha W_s c). \quad (2)$$

Here W_q and W_s are learned projections and α controls the context contribution. The shared c summarizes surroundings across retained frames without frame-aligned interactions; event order remains in the Fast sequence and timestamps. The design allocates temporal detail to object observations and spatial detail to contextual evidence, without assuming a strict frequency decomposition. Small objects and rapidly changing interactions may still require more spatial or temporal detail.

Shared embedding space. The shared language backbone independently embeds trajectories and text for similarity-based retrieval. Projected, ordered z_i fill the visual slots; a designated visual readout token yields h_v and hence $v(r) = \mathrm{L2Norm}(W_v h_v)$. In a separate text-only pass, the query fills the prompt with all trajectory slots masked, yielding $q(e) = \mathrm{L2Norm}(W_t h_e)$ from text readout h_e. Scores are $s(e, r) = \exp(\ell) q(e)^\top v(r)$, with learned logit scale ℓ. Neither pass uses the other modality's input, allowing one query encoding and reusable trajectory embeddings for a fixed pool. Appendix C gives the ROI implementation, geometry definition, and input templates.

Lemma 1 (Task-relevant preservation under Slow-Fast sampling)**.** *Let S_0 contain n_s preselected frames, and let semantic pruning retain $S \subseteq S_0$. Every frame i is assigned to a retained representative $\pi(i)$ with unit semantic image embeddings satisfying $\bar{g}_i^\top \bar{g}_{\pi(i)} \geq \tau$. Suppose the object-conditioned context features obey $\|u_i - u_{\pi(i)}\|_2 \leq L_u \|\bar{g}_i - \bar{g}_{\pi(i)}\|_2 + \eta$, and the resampler R is K_R-Lipschitz under maximum per-frame perturbations at fixed input length. Define $c_0 = R([u_i]_{i \in S_0})$, $c = R([u_j]_{j \in S})$, and*

$$\varepsilon_{\mathrm{mult}} = \left\| R([u_{\pi(i)}]_{i \in S_0}) - R([u_j]_{j \in S}) \right\|_2 . \quad (3)$$

Then $\|c - c_0\|_2 \leq K_R \rho_\tau + \varepsilon_{\mathrm{mult}}$, where $\rho_\tau = L_u \sqrt{2(1-\tau)} + \eta$. If c_D is a dense-context reference with $\|c_0 - c_D\|_2 \leq \varepsilon_{\mathrm{pre}}$, and the normalized trajectory encoder is K_c-Lipschitz in context with the Fast sequence fixed, then

$$\|v(r) - v_D(r)\|_2 \leq K_c \big(\varepsilon_{\mathrm{pre}} + K_R \rho_\tau + \varepsilon_{\mathrm{mult}} \big) \equiv \delta_r. \quad (4)$$

For a unit text embedding, a correct reference top-1 result with cosine margin $\gamma > 2 \max_r \delta_r$ is preserved.

The result explains when sparse high-resolution context can replace dense context while retaining the Fast observations. The semantic-to-context condition and resampler stability are explicit assumptions; deleting similar frames is not automatically lossless. Appendix F proves the bounds, quantifies the multiplicity term for attention pooling, and accounts for sampling and encoding cost.

3.2 Mask-Guided Contrastive Learning

Mask-guided contrastive learning jointly optimizes within-video discrimination and bidirectional batch retrieval. Query-centered trajectory groups from annotated and automatically expanded data supply these comparisons, as shown in Figure 1. Multiple query-trajectory pairs may match, while many others remain unverified. A relation mask therefore includes confirmed positives and negatives and excludes uncertain pairs. Training updates lightweight spatiotemporal modules and low-rank language adapters, keeping the proposal pipeline and pretrained encoder weights frozen. Appendix D details the training data and relation annotations.

Supervision and valid comparisons. Each training group b contains a query e_b, positives $\mathcal{P}_b$, and confirmed within-video negatives $\mathcal{N}_b$. For annotated data, positives include the annotated tube, its stochastically perturbed variant, and an identity-checked tracker output when available. These augmented positives are training-only and leave fixed benchmark candidate pools unchanged. Expanded examples pair generated expressions with verified target trajectories. Low overlap or large description distance suggests negative candidates, not confirmed semantic mismatches.

Within-video discrimination. It maximizes probability mass on each group's full positive set. Let $s_{br} = s(e_b, r)$ and $\mathcal{B}_N$ index groups with negatives:

$$L_{\text{group}} = -\frac{1}{|\mathcal{B}_N|} \sum_{b \in \mathcal{B}_N} \log \frac{\sum_{r \in \mathcal{P}_b} \exp(s_{br})}{\sum_{r \in \mathcal{P}_b \cup \mathcal{N}_b} \exp(s_{br})}. \tag{5}$$

All retained negatives share one denominator per group; set $L_{\text{group}} = 0$ when $\mathcal{B}_N = \varnothing$.

Relations across query groups. Across query groups, masked bidirectional retrieval compares queries with the batch's indexed positive-record list $\mathcal{P} = [(b, r) : b \in \mathcal{B}, r \in \mathcal{P}_b]$. For query b and record p, the relation mask M_{bp} is 1 for a confirmed match, 0 for a confirmed mismatch, and $\perp$ otherwise. Queries may match records from other groups. Define $\mathcal{J}_b^+ = \{p : M_{bp} = 1\}$, $\mathcal{J}_b = \{p : M_{bp} \neq \perp\}$, and their column counterparts $\mathcal{I}_p^+$ and $\mathcal{I}_p$. With $s_{bp} = s(e_b, r_p)$,

$$L_{t \to v} = -\frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \log \frac{\sum_{p \in \mathcal{J}_b^+} \exp(s_{bp})}{\sum_{p \in \mathcal{J}_b} \exp(s_{bp})}, \qquad L_{v \to t} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \log \frac{\sum_{b \in \mathcal{I}_p^+} \exp(s_{bp})}{\sum_{b \in \mathcal{I}_p} \exp(s_{bp})}. \tag{6}$$

Each row and column has a positive by construction. Confirmed positives enter both numerator and denominator; unconfirmed pairs enter neither. We optimize

$$L = \lambda_g L_{\text{group}} + \lambda_{\text{bi}} L_{\text{bi}}, \tag{7}$$

where $L_{\text{bi}} = (L_{t \to v} + L_{v \to t})/2$, with nonnegative weights λ_g and λ_{bi}.

The following lemma connects the joint objective to ranking-error control. Each comparison maximizes probability mass on its full positive set.

Lemma 2 (Task-aligned multi-positive ranking control)**.** *For a local, row, or column comparison A with finite scores and correct relation labels, let A^+, A^- be its positive and negative sets, $m_A = |A^+| \geq 1$, and $Z_A^{\pm} = \sum_{j \in A^{\pm}} e^{s_j}$. Let E_A indicate that a negative scores at least as high as every positive, counting ties as errors and setting $E_A = 0$ when A^- is empty. Then*

$$E_A \leq \frac{\mathcal{L}_A}{\log(1 + 1/m_A)}, \qquad \mathcal{L}_A = \log(1 + Z_A^-/Z_A^+). \tag{8}$$

For families $x \in \{g, t, v\}$ of local, row, and column comparisons, let E_x be their mean errors and $c_x = \log(1 + 1/m_x)$, where m_x is the maximum positive count. Empty families contribute zero. The joint objective satisfies

$$\lambda_g c_g E_g + \frac{\lambda_{\text{bi}}}{2}(c_t E_t + c_v E_v) \leq L. \tag{9}$$

Thus L bounds weighted errors across the three comparison families, even with overlapping score pairs. Including all confirmed matches and excluding uncertain relations aligns comparisons with the intended retrieval labels. Appendix G gives the proof, gradient consequences, and scope.

3.3 TEXT2TRAJ-BENCH: HUMAN-VERIFIED TRAJECTORY RELATIONS

Text2Traj-Bench provides human-verified query-trajectory relations for learning (Section 3.2) and controlled object-centric retrieval evaluation. Its training split supports within-video discrimination and bidirectional batch retrieval, while its held-out splits support three controlled evaluation settings: within-video text-to-trajectory retrieval, trajectory-to-text retrieval, and object-centric video retrieval. Construction follows three stages within each data split: trajectory curation, verified description, and relation construction, as illustrated in Figure 1(c). Models propose annotations; human review confirms relations. Source recordings are assigned to training, development, or test splits before construction, with duplicate clips, related clips, and derived records kept together.

Trajectory curation. It establishes and freezes the query-independent candidate pool $\mathcal{R}_{\text{bench}}(V)$ before query generation. A frozen SAM 3 tracker generates trajectory proposals from fixed category prompts. Three independent annotators assess trajectory usability and proposed duplicate pairs; retention and duplicate consolidation require at least two supporting votes. Only confirmed redundant trajectories are consolidated, while distinct objects, nonredundant fragments, and unresolved identity pairs remain separate candidates. All queries for V and all models share this pool in trajectory-level evaluation; description or query rejection leaves it unchanged.

Verified description. For curated trajectories, evidence-grounded query construction separates visual fact verification from query generation. A VLM proposes appearance and motion descriptions, structured facts, and evidence references from ordered trajectory observations, supplemented by target-marked full frames for contextual claims. Three annotators independently check the descriptions and facts against the video; the final version, including any minor evidence-supported corrections, requires at least two approvals. An LLM then generates query e solely from human-approved facts D_r^H that support its required predicates, temporal scope, and relational qualifiers.

Relation construction. It first determines query admission through human review of query faithfulness, naturalness, and every within-video query-trajectory pair against visual evidence, optionally using LLM pre-labels. Unresolved relations may temporarily receive $\perp$ during construction. A faithful, natural query with intended target trajectory r^* is admitted only when

$$\bar{Y}(e, r^*) = 1, \qquad \bar{Y}(e, r) = 0 \quad \forall r \in \mathcal{R}_{\text{bench}}(V) \setminus \{r^*\}. \tag{10}$$

Thus within-video relations are binary across all three splits. Additional confirmed matches or uncertain relations block admission without changing the candidate pool. Accepted training relations define $\mathcal{P}_b$ and $\mathcal{N}_b$ for L_{group}, whereas held-out relations provide human-verified evaluation labels.

Following within-video verification, cross-video expansion constructs within-split relations for bidirectional training and cross-video evaluation. Verified descriptions guide mining of easy- and hard-negative candidates and potential matches; these categories guide sampling, not relation labels. An LLM compares query requirements with verified candidate facts and proposes a relation with supporting or conflicting predicates. For training, humans individually verify potential matches and hard-negative candidates and review easy-negative candidates in stratified samples; unverified pairs remain $\perp$. These cross-video and binary within-video relations populate training mask M_{bp}; L_{bi} excludes $\perp$ entries. Optional calibrated automatic easy negatives are training-only and reported separately as pseudo-labels. For development and test, every cross-video relation required by trajectory-to-text and object-centric video retrieval galleries is individually confirmed as a match or mismatch.

Fixed candidate pools and human-verified relevance labels define the benchmark's controlled evaluation protocol. Verified within-video relations label text-to-trajectory retrieval; verified cross-video relations define query-trajectory and query-video relevance for trajectory-to-text and object-centric video retrieval. All relations for development and test evaluation have binary labels; only training supervision retains unresolved pairs. For trajectory-level evaluation, fixed pools isolate representation and ranking from proposal quality; associated verified relations define relevance for object-centric video retrieval. Appendix E details annotation, batching, auditing, and release protocols.

4 Evaluation

4.1 Experimental Setup

Implementation. VTrajEmbed trains only lightweight trajectory modules and LoRA adapters, keeping pretrained components frozen. It uses the SigLIP2 Tschannen et al. (2025) vision encoder from Qwen3-VL-4B-Instruct, the ObjEmbed Fu et al. (2026) object projector, and the language component of Qwen3-VL-4B Bai et al. (2025) as the shared trajectory-text backbone. Training uses the joint objective in Section 3.2 for three epochs, with $\lambda_g = 0.1$, $\lambda_{\text{bi}} = 1.0$, an initial learning rate of 5×10^{-5}, and an effective batch size of 32. Ablations share this configuration unless stated otherwise. Architecture, optimization, and runtime details appear in Appendices C and B.

Data and evaluation tasks. Text2Traj-Bench provides three retrieval settings using videos from HC-STVG Tang et al. (2021), Kinetics-700 Carreira et al. (2019), MammAlps Gabeff et al. (2025) and other datasets. Source recordings are split before construction; development data support model selection, and test data are reserved for final evaluation. External-evaluation recordings from overlapping datasets are excluded from Text2Traj-Bench training and cross-video relation expansion.

Table 1: **Object-centric retrieval results on Text2Traj-Bench.** R@1, R@3, and mAP are reported in percent (higher is better). Task definitions, model interfaces, and abbreviations follow Section 4.1. **Bold** marks column bests; parentheses show percentage-point changes from Qwen3E-8B.

Model	T2Traj			Traj2T			OCVR		
	R@1	R@3	mAP	R@1	R@3	mAP	R@1	R@3	mAP
Jina-v4 Günther et al. (2025)	26.31	77.72	53.52	43.89	67.68	59.62	47.91	73.39	63.44
BGE-VL Zhou et al. (2024)	60.66	92.73	76.79	39.56	57.31	53.93	60.85	84.08	73.88
WeMM-2B Zhou et al. (2026)	45.10	86.33	66.31	20.98	40.43	37.14	42.60	60.61	56.10
WeMM-4B Zhou et al. (2026)	46.61	87.20	67.59	21.46	38.83	37.20	44.13	62.38	57.41
WeMM-9B Zhou et al. (2026)	46.76	86.81	67.56	18.09	33.92	33.52	45.26	62.86	58.36
Qwen3E-2B Li et al. (2026b)	44.79	86.56	66.01	43.66	57.54	51.53	65.86	84.58	76.35
Qwen3E-8B Li et al. (2026b)	54.19	89.57	72.24	48.29	61.54	55.21	**69.65**	87.32	79.28
VTrajEmbed-5B	**69.67** (15.48↑)	**98.51** (8.94↑)	**83.72** (11.48↑)	**66.34** (18.05↑)	**94.65** (33.11↑)	**80.27** (25.06↑)	68.04 (1.61↓)	**95.21** (7.89↑)	**81.43** (2.15↑)

Table 2: **Global video retrieval results.** R@1 is reported in percent (higher is better); Avg. is the unweighted mean over five datasets. Evaluation settings and abbreviations follow Section 4.1. **Bold** marks column bests; parentheses show percentage-point changes from Qwen3E-8B.

Method	Breakfast	HMDB51	UCF101	MSVD	MSR-VTT	Avg. R@1
WeMM-2B Zhou et al. (2026)	23.60	75.52	95.59	55.92	47.56	59.64
WeMM-4B Zhou et al. (2026)	28.14	**77.07**	**96.01**	56.26	**47.87**	**61.07**
WeMM-9B Zhou et al. (2026)	27.99	76.68	95.49	**56.52**	47.08	60.75
Qwen3E-2B Li et al. (2026b)	22.86	73.02	90.88	3.09	10.86	40.14
Qwen3E-8B Li et al. (2026b)	22.85	76.26	94.97	5.89	14.33	42.86
VTrajEmbed-5B	**44.84** (21.99↑)	75.12 (1.14↓)	94.55 (0.42↓)	19.64 (13.75↑)	13.26 (1.07↓)	49.48 (6.62↑)

Dataset statistics appear in Appendix E. *Text-to-Trajectory Retrieval (T2Traj)* ranks a video's fixed pool $\mathcal{R}_{\text{bench}}(V)$ for a query; *Trajectory-to-Text Retrieval (Traj2T)* ranks all test queries for a trajectory; and *Object-Centric Video Retrieval (OCVR)* ranks videos for an object-centric query.

Baselines. Baselines cover four multimodal embedding families: jina-embeddings-v4 (Jina-v4) Günther et al. (2025), BGE-VL-v1.5-mmeb (BGE-VL) Zhou et al. (2024), WeMM-Embedding (WeMM) Zhou et al. (2026), and Qwen3-VL-Embedding (Qwen3E) Li et al. (2026b). Scale suffixes identify model sizes, e.g., WeMM-4B and Qwen3E-8B.

Evaluation protocol. Evaluation uses the same query-independent benchmark trajectories across models for T2Traj and Traj2T, with human-verified relevance labels for all three settings. Baselines without native object-level inputs independently encode each trajectory as a video with its boxes overlaid on the corresponding frames. For OCVR, conventional baselines use native whole-video embeddings, while VTrajEmbed scores videos by $S(e, V) = \max_{r \in \mathcal{R}_{\text{bench}}(V)} s(e, r)$. Relation labels and construction-time evidence annotations are withheld from model inputs. All development and test gallery relations have confirmed binary relevance labels; no unverified pairs are scored. All three settings report R@1 (primary), R@3, and mAP. Global retrieval evaluates whether trajectory-derived representations remain useful after benchmark-specific alignment. For each benchmark, VTrajEmbed uses multiple trajectory representations $v(r)$ per video for video-text alignment on the training split, followed by retrieval evaluation on the test split.

4.2 Overall Performance

Table 1 compares VTrajEmbed with existing multimodal embedding models on the three object-centric retrieval settings of Text2Traj-Bench. Table 2 further evaluates whether the trajectory representations learned by VTrajEmbed can be effectively adapted to standard global video-text retrieval. Overall, VTrajEmbed substantially improves fine-grained object-centric retrieval while its trajectory representations remain effective for broader global video-text retrieval after adaptation.

Compared with Qwen3E-8B, VTrajEmbed-5B improves T2Traj R@1 from 54.19% to 69.67% and Traj2T R@1 from 48.29% to 66.34%, corresponding to gains of 15.48 and 18.05 percentage points, respectively. The improvement is even more pronounced under broader ranking metrics: Traj2T

R@3 increases from 61.54% to 94.65%, and mAP from 55.21% to 80.27%. After alignment on the training split of each global retrieval benchmark, VTrajEmbed achieves an average test-set R@1 of 49.48%, compared with 42.86% for Qwen3E-8B, with similar performance on HMDB51, UCF101, and MSR-VTT and substantially higher results on Breakfast and MSVD. Taken together, these results show that VTrajEmbed substantially strengthens trajectory-level discrimination while its learned trajectory representations remain effectively adaptable to broader video-text retrieval.

Trajectory-level modeling substantially improves fine-grained object discrimination. On T2Traj, VTrajEmbed achieves 69.67% R@1, 98.51% R@3, and 83.72% mAP. Compared with Qwen3E-8B, these results correspond to improvements of 15.48, 8.94, and 11.48 percentage points, respectively. VTrajEmbed also outperforms BGE-VL, the strongest competing baseline on T2Traj, by 9.01 points in R@1, 5.78 points in R@3, and 6.93 points in mAP. The gain is particularly meaningful because all methods rank exactly the same query-independent trajectory pools; therefore, it cannot be attributed to improved proposal recall or access to different candidate objects. Instead, the results directly reflect a stronger ability to distinguish the target trajectory from visually or semantically competing objects within the same video. This behavior is consistent with the Slow-Fast trajectory encoder: the Fast branch preserves ordered object appearance, geometry, and motion evidence, whereas the Slow branch supplements these observations with sparse high-resolution, object-conditioned scene context.

Bidirectional supervision produces larger gains when retrieving text from trajectories. The advantage of VTrajEmbed becomes even more pronounced on Traj2T, where it reaches 66.34% R@1, 94.65% R@3, and 80.27% mAP. Relative to Qwen3E-8B, the improvements are 18.05, 33.11, and 25.06 percentage points, respectively. The gain is therefore not restricted to identifying a target object from a small within-video candidate set: a trajectory representation also retrieves its corresponding descriptions more reliably from the substantially broader test-text gallery. This observation provides empirical support for the bidirectional component of our training objective. In addition to the within-video group loss that explicitly trains object discrimination, L_{bi} performs both text-to-trajectory and trajectory-to-text comparisons over verified cross-group relations. The particularly large improvements in Traj2T R@3 and mAP are consistent with this design, suggesting that the learned trajectory embedding captures semantics that remain discriminative beyond its source-video candidate pool.

Object-level alignment remains effective when aggregated to video-level retrieval. In OCVR, conventional embedding baselines rank videos using their native whole-video embeddings, whereas VTrajEmbed derives each video score from its highest text-trajectory similarity. Under this trajectory-centric scoring interface, VTrajEmbed obtains 68.04% R@1, 95.21% R@3, and 81.43% mAP. Compared with Qwen3E-8B, R@3 and mAP increase by 7.89 and 2.15 percentage points, respectively, although R@1 decreases slightly by 1.61 points. The combination of substantially higher R@3 and mAP with a small R@1 decrease suggests that trajectory-level matching more reliably places relevant videos among the highest-ranked candidates, while the ordering of the first-ranked video remains less stable. This result is also consistent with the OCVR formulation $S(e, V) = \max_{r \in \mathcal{R}_{\mathrm{bench}}(V)} s(e, r)$: trajectory-level correspondence can support cross-video retrieval without introducing a separately trained video-level scoring module. At the same time, the remaining R@1 gap indicates that aggregating independently scored object trajectories by a maximum operation does not fully resolve video-level ranking.

Trajectory representations transfer effectively to global video retrieval. We further evaluate whether VTrajEmbed's learned trajectory representations support conventional global video-text retrieval after task-specific alignment. For each benchmark, multiple trajectory representations $v(r)$ from each video are used for video-text alignment on the training split, with retrieval evaluated on the test split. After benchmark-specific adaptation, VTrajEmbed achieves an average R@1 of 49.48%, compared with 42.86% for Qwen3E-8B, despite using a smaller 4B language backbone. Performance varies across datasets: VTrajEmbed substantially exceeds Qwen3E-8B on Breakfast (44.84% vs. 22.85%, $+21.99$ points) and MSVD (19.64% vs. 5.89%, $+13.75$ points), while remaining close on HMDB51 (75.12% vs. 76.26%), UCF101 (94.55% vs. 94.97%), and MSR-VTT (13.26% vs. 14.33%). The latter three results trail Qwen3E-8B by only 0.42–1.14 percentage points. These results indicate that trajectory-level semantics remain useful beyond object-centric retrieval and can be effectively adapted to global video-text alignment. Together with the substantially stronger T2Traj and Traj2T results, these findings suggest that VTrajEmbed improves fine-grained object discrimination without restricting its representations to trajectory-only retrieval.

Table 3: **Ablation study of Slow-Fast encoding and training objectives.** Other settings follow Section 4.1. Scores are percentages; parentheses show percentage-point drops from the full model.

Variant	R@1	R@3	mAP
Full model	**69.67**	**98.51**	**83.72**
w/o Slow branch	64.58(5.09↓)	95.49(3.02↓)	79.86(3.86↓)
w/o Fast branch	41.08(28.59↓)	87.01(11.50↓)	64.94(18.78↓)
Fast observations: 32 → 16	64.31(5.36↓)	95.05(3.46↓)	79.81(3.91↓)
w/o L_{bi} (group only)	49.29(20.38↓)	93.55(4.96↓)	71.10(12.62↓)
w/o L_{group} (bidirectional only)	48.06(21.61↓)	81.45(17.06↓)	66.27(17.45↓)

4.3 Ablation Study

Table 3 examines the contributions of the Slow-Fast trajectory encoder and the two training objectives on T2Traj. The full model performs best across all three metrics. Removing either architectural branch or either loss term consistently degrades performance, indicating that both the trajectory representation and the joint supervision contribute to object-level retrieval.

Dense Fast observations provide core trajectory evidence, complemented by sparse Slow context. Removing the Slow branch decreases R@1 from 69.67% to 64.58% and mAP from 83.72% to 79.86%. With the Fast sequence unchanged, this degradation supports the contribution of sparse high-resolution, object-conditioned scene context beyond dense low-resolution trajectory observations. In contrast, removing the Fast branch reduces R@1 to 41.08%, R@3 to 87.01%, and mAP to 64.94%. The substantially larger drop is consistent with the Fast branch preserving ordered object appearance, geometry, and temporal progression.

Denser Fast sampling further improves trajectory discrimination. Reducing the maximum number of Fast observations from 32 to 16 lowers R@1 to 64.31% and mAP to 79.81%. The smaller degradation relative to complete branch removal suggests that reduced sampling retains much of the Fast pathway's benefit but weakens fine-grained discrimination.

Joint supervision improves T2Traj ranking over either objective alone. Removing L_{bi}, leaving only the within-video group objective, reduces R@1 from 69.67% to 49.29% and mAP from 83.72% to 71.10%. Conversely, removing L_{group} and training only with the bidirectional objective reduces R@1 to 48.06% and mAP to 66.27%. The group-only variant retains higher R@3 than the bidirectional-only variant (93.55% versus 81.45%), consistent with L_{group} directly training within-video candidate discrimination. However, its lower R@1 and mAP than the full model support combining local discrimination with relation-masked bidirectional retrieval.

5 Conclusion and Future Work

We present VTrajEmbed, an embedding model that aligns text and video object trajectories in a shared semantic space. Its Slow-Fast encoder combines dense low-resolution observations for appearance and motion with sparse high-resolution frames for scene context, while mask-guided contrastive learning couples within-video discrimination with bidirectional trajectory-text retrieval under verified relation supervision. Text2Traj-Bench provides human-verified query-trajectory relations and fixed candidate pools for controlled trajectory-level evaluation, alongside verified cross-video relations for object-centric video retrieval. VTrajEmbed-5B improves T2Traj and Traj2T R@1 over Qwen3E-8B by 15.48 and 18.05 percentage points, respectively. After benchmark-specific adaptation, its trajectory-derived representations achieve 49.48% average R@1 across five global retrieval datasets, compared with 42.86% for Qwen3E-8B. These results indicate finer object discrimination while preserving transferability to broader video-text retrieval.

Future work will extend fixed-pool evaluation to the full query-conditioned proposal-and-ranking pipeline, jointly measuring proposal recall, trajectory quality, and retrieval accuracy. This will quantify how missing objects, fragmented trajectories, and identity switches affect downstream retrieval. We will investigate learned aggregation of trajectory-level similarities to replace maximum-similarity scoring for more robust object-centric video retrieval. Finally, we will expand Text2Traj-Bench with more object categories, video domains, naturally collected user queries, and verified cross-video relations to assess robustness and generalization across bidirectional and video-level retrieval.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Yi-Chung Chen, Philip Jacobson, Tom Lampo, Yiren Lu, Jin Yao, David I Inouye, Jing Gao, Danhua Guo, and Burhan Yaman. Travel: Trajectory-guided video embedding learning for driving-video retrieval. *arXiv preprint arXiv:2608.13495*, 2026.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911. IEEE, 2024.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14987–14997. IEEE, 2025.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing LYU, Xiaohua Xie, and Wei-Shi Zheng. Objembed: Towards universal multimodal object embeddings. *arXiv preprint arXiv:2602.01753*, 2026.

Valentin Gabeff, Haozhe Qi, Brendan Flaherty, Gencer Sumbul, Alexander Mathis, and Devis Tuia. Mammalps: A multi-view video behavior monitoring dataset of wild mammals in the swiss alps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13854–13864, 2025.

Hong Gao, Jingyu Wu, Xiangkai Xu, Kangni Xie, Yunchen Zhang, Bin Zhong, Xurui Gao, and Min-Ling Zhang. Omniground: A comprehensive spatio-temporal grounding benchmark for real-world complex scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17588–17597, 2026.

Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. In *International Conference on Learning Representations*, volume 2025, pp. 62122–62142, 2025.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Sedigheh Eslami, Scott Martens, Bo Wang, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval, 2025. URL `https://arxiv.org/abs/2506.18902`.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Usman Naseem, Chunyu Wang, Qi Dai, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer cross-modality representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5131–5139, 2026.

Anqi Li, Wenwei Jin, Jintao Tong, Pengda Qin, Jiawei Li, and Guo Lu. Towards trustworthy multimodal moderation via policy-aligned reasoning and hierarchical labeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2278–2289, 2026a.

Jinyang Li, En Yu, Sijia Chen, and Wenbing Tao. Ovtr: End-to-end open-vocabulary multiple object tracking with transformer. In *International Conference on Learning Representations*, volume 2025, pp. 49927–49947, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026b.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20009–20019. IEEE, 2025.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 44215–44234, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yehao Lu, Minghe Weng, Zekang Xiao, Rui Jiang, Wei Su, Guangcong Zheng, Ping Lu, and Xi Li. Dynamic-dino: Fine-grained mixture of experts tuning for real-time open-vocabulary object detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20847–20856. IEEE, 2025.

Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. Covtrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Ji Ao, Dawei Leng, and Yuhui Yin. Fg-clip 2: A bilingual fine-grained vision-language alignment model. In *International Conference on Machine Learning*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9274–9285. IEEE, 2025.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*, 2024.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. Wemm-embedding: Wechat multi-modal embedding technical report. 2026. URL https://arxiv.org/abs/2608.24053.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pp. 9588–9608, 2024.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

AI USE STATEMENT

Generative AI assisted with language editing, restructuring the appendix, checking mathematical derivations.

ETHICS STATEMENT

This work does not involve the recruitment of human subjects or the collection of new personally identifiable or private data.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, experimental settings, datasets, and evaluation protocols in the main paper and appendix. Additional implementation details and hyperparameters are included in the supplementary materials. Source code and configuration files will be released to support reproducibility.

A NOTATION

Table 4 summarizes the mathematical notation used in the main text, including the lemmas and experimental captions. Related symbols are grouped by role; meanings that depend on context are stated explicitly.

Table 4: Notation used in the main text.

Symbol	Meaning
Task and candidate ranking	
V, e	Input video and natural-language query, respectively.
r, $\mathcal{R}(V)$	A candidate object trajectory and the fixed candidate pool for video V. Each trajectory stores boxes and original timestamps.
$\hat{r}$, r^*	Highest-scoring retrieved trajectory and designated benchmark target, respectively.
$s(e, r)$	Retrieval logit: $\exp(\ell) q(e)^\top v(r)$.
M	Number of candidates in an evaluation pool; uniform random selection succeeds with probability $1/M$.
$Y(e, r)$, $\widehat{Y}(e, r)$	True binary semantic-match relation and the verifier's estimate, respectively; the estimate can also be uncertain ($\perp$).
Slow-Fast trajectory encoding (Section 3.1)	
t	Original observation time or frame timestamp; the subscript t in a feature identifies that observation.
i, j	Indices whose domains follow the surrounding expression: i indexes Fast blocks or preselected frames, j indexes retained frames in Lemma 1, and j indexes comparison items in Lemma 2.
n_f, n_s, m_s	Numbers of valid Fast observations, preselected Slow frames, and retained Slow frames, respectively; $m_s \le n_s \le n_f$.
P_f, P_s	Per-frame pixel budgets for the Fast and Slow branches; $P_s > P_f$.
a_t, g_t	Target-ROI feature from the frozen object projector and its additional geometry/motion input, respectively. The ROI feature combines visual appearance with normalized box-position encoding, while g_t further contains geometry, displacement, relative scale change, and timestamp information.
$\widetilde{a}_t$	Geometry-enhanced object feature: $a_t + \mathrm{MLP}_{\mathrm{geo}}(g_t)$.
B_i, K_r	The ith consecutive block of Fast observations and the number of valid blocks for trajectory r, respectively.
m, K	Fast temporal-block length and the fixed maximum number of Fast trajectory slots, respectively. The experimental configuration uses $m = 4$ and $K = 8$.
f_i	Fast feature obtained by attention-pooling the geometry-enhanced observations in block B_i.
S_0, S	Preselected Slow frame set and the subset retained after semantic pruning, respectively.
τ	Cosine-similarity threshold for removing a preselected frame that is sufficiently similar to a retained representative.
X_t^s	High-resolution full-frame visual token bank at retained time t; superscript s denotes the Slow branch.
u_t, c	Object-conditioned scene-context feature at time t and the pooled context across retained Slow frames, respectively.
W_q, W_s	Learned projections of the object appearance into the Slow attention query, and of pooled Slow context into the Fast residual, respectively.
α, z_i	Learned scalar context gate and the fused block feature $z_i = \mathrm{LN}(f_i + \alpha W_s c)$.

Continued on the next page

Table 4 (continued)

Symbol	Meaning
h_v, h_e	Visual and text readout hidden states from their separate passes through the shared language backbone.
W_v, W_t	Output projections mapping visual and text readouts into the shared embedding space, respectively. Here the subscript t denotes text.
$v(r)$, $q(e)$	Unit-normalized trajectory and text embeddings, respectively.
ℓ	Learned logit-scale parameter; $\exp(\ell)$ is the positive scale multiplying cosine similarity.
Slow-context preservation (Lemma 1)	
$\pi(i)$, $\bar{g}_i$	Retained representative assigned to preselected frame i, and the unit semantic image embedding of frame i, respectively. Thus $\bar{g}_{\pi(i)}$ is the representative's embedding; these embeddings differ from geometry inputs g_t.
u_i, $u_{\pi(i)}$	Object-conditioned context features of preselected frame i and its retained representative, using the same feature definition as u_t.
L_u, η	Nonnegative coefficient and residual in the assumed bound relating semantic-image distance to context-feature distance.
R, K_R	Slow attention resampler ($R = \mathrm{Resample}_s$) and its Lipschitz factor under maximum per-frame perturbations at fixed input length.
c_0, c_D	Context pooled from all preselected frames before pruning, and the dense-context reference, respectively. The dense reference encodes all Fast observations at high resolution.
$v_D(r)$	Normalized trajectory embedding obtained with c_D, using the same model parameters and Fast sequence as $v(r)$.
$\varepsilon_{\mathrm{pre}}$	Upper bound on the preselection error $\|c_0 - c_D\|_2$.
ρ_τ	Per-frame context-replacement bound $L_u\sqrt{2(1-\tau)} + \eta$.
$\varepsilon_{\mathrm{mult}}$	Difference between pooling representatives repeated once per assigned preselected frame and pooling each retained frame once; see Eq. (3). It accounts for the change in attention multiplicities.
K_c	Lipschitz factor of the normalized trajectory encoder with respect to pooled context, holding the Fast sequence fixed.
δ_r	Upper bound on trajectory-embedding distortion: $K_c(\varepsilon_{\mathrm{pre}} + K_R\rho_\tau + \varepsilon_{\mathrm{mult}})$.
γ	Dense-reference cosine margin of the correct target over its highest-scoring distractor. The lemma preserves this decision when $\gamma > 2\max_r \delta_r$.
Training groups, relations, and losses (Section 3.2)	
b, e_b, $\mathcal{B}$	Query-group index, the group's query, and the set of query-group indices in a training batch, respectively.
$\mathcal{P}_b$, $\mathcal{N}_b$	Positive trajectories and confirmed within-video negatives for query group b.
$\mathcal{B}_N$	Subset of batch groups containing at least one confirmed negative.
s_{br}, L_{group}	Within-group logit $s(e_b, r)$ and the multi-positive loss averaged over $\mathcal{B}_N$; the loss is zero if this set is empty.
$\mathcal{P}$, p, r_p	Indexed list of all positive records in the batch, a record index, and the trajectory input of record p, respectively. Repeated records retain their multiplicities.
M_{bp}, $\perp$	Query-record relation mask: 1 for a confirmed match, 0 for a confirmed mismatch, and $\perp$ for an unconfirmed relation excluded from the loss.
$\mathcal{J}_b^+$, $\mathcal{J}_b$	Positive record indices $\{p : M_{bp} = 1\}$ and all confirmed record indices $\{p : M_{bp} \neq \perp\}$ for query b.

Continued on the next page

Table 4 (continued)

Symbol	Meaning
$\mathcal{I}_p^+, \mathcal{I}_p$	Positive query indices $\{b : M_{bp} = 1\}$ and all confirmed query indices $\{b : M_{bp} \neq \perp\}$ for record p.
s_{bp}	Batch logit $s(e_b, r_p)$ between query b and trajectory record p.
$L_{t\to v}, L_{v\to t}$	Masked multi-positive batch losses for text-to-trajectory and trajectory-to-text retrieval, respectively.
L_{bi}, L	Mean of the two batch-retrieval losses and the joint objective $L = \lambda_g L_{\mathrm{group}} + \lambda_{\mathrm{bi}} L_{\mathrm{bi}}$, respectively.
$\lambda_g, \lambda_{\mathrm{bi}}$	Nonnegative weights of the within-video and bidirectional batch losses.
Multi-positive ranking control (Lemma 2)	
A, A^+, A^-	One local, query-row, or trajectory-column comparison, together with its positive and negative item sets.
m_A, s_j	Number of positive items in comparison A and the finite logit of comparison item j, respectively.
Z_A^+, Z_A^-	Sums of exponentiated logits over positive and negative items: $Z_A^{\pm} = \sum_{j\in A^{\pm}} \exp(s_j)$.
E_A	Conservative top-1 error indicator: one when a negative scores at least as high as every positive, and zero otherwise (including an empty negative set). Ties count as errors.
$\mathcal{L}_A$	Single-comparison multi-positive loss $\log(1 + Z_A^-/Z_A^+)$.
$x \in \{g, t, v\}$	Comparison-family label: g denotes local within-video groups, t query rows (text-to-trajectory), and v trajectory columns (trajectory-to-text).
E_x; E_g, E_t, E_v	Mean conservative ranking error for family x; the three explicit subscripts identify the corresponding local, row, and column errors.
m_x	Maximum number of positives in a comparison from family x. Empty families contribute zero to the joint bound.
c_x; c_g, c_t, c_v	Ranking-control coefficient $\log(1 + 1/m_x)$ for each nonempty comparison family, distinct from the context vector c.
Experimental diagnostics and retrieval metrics	
$\overline{m}_s$	Mean number of retained Slow frames over the evaluated trajectories.
$\overline{\Delta}_v$	Mean measured distortion $\|v(r) - v_D(r)\|_2$ over evaluated trajectories, distinct from the theoretical bound δ_r.
$\tau_-, \tau_\star, \tau_+$	Lower, reference, and higher pruning thresholds fixed on development data, with $\tau_- < \tau_\star < \tau_+$.
R@1, R@3	Recall at ranks one and three. With one relevant trajectory per query, each is the fraction of queries whose target is within that rank cutoff.
mAP	Mean average precision over queries; the relevant-item sets must follow the stated retrieval protocol.
Operators and notation conventions	
$\mathrm{MLP}_{\mathrm{geo}}$	Trainable multilayer perceptron mapping geometry/motion inputs to appearance-feature space.
$\mathrm{Resample}_f, \mathrm{Resample}_s$	Attention resamplers for Fast temporal blocks and Slow context, respectively.
MHA, LN	Multihead attention and layer normalization, respectively.
L2Norm	Normalization of a nonzero vector to unit Euclidean norm.
$(\cdot)^\top, \|\cdot\|_2, \|\cdot\|$	Transpose, Euclidean norm, and set cardinality or indexed-list length, respectively. A list length counts repeated records.
$[\cdot], \{\cdot\}$	An indexed sequence/list and a set, respectively; repeated representatives in a list retain their multiplicities.
exp, log, arg max	Natural exponential, natural logarithm, and an argument attaining the maximum. The exponential e^{s_j} in Lemma 2 is unrelated to query e.

B Experimental and Reproducibility Details

This section consolidates the concrete model, optimization, batching, and runtime configurations used for the reported experiments. Detailed encoder operations and training-data construction are described separately in Appendices C and D.

B.1 Model and Encoder Configuration

The frozen visual encoder is initialized from the vision tower of `Qwen3-VL-4B-Instruct`. This vision tower is based on SigLIP2-Large and has undergone dynamic-resolution adaptation and multimodal continued training. We extract intermediate spatial features from Transformer layers 12, 18, and 24, using one-based layer indices. The frozen target-ROI projection module is initialized from `ObjEmbed-4B`. The language backbone is initialized from the language component of `Qwen3-VL-4B-Instruct`. The pretrained visual encoder, target-ROI projector, and base language-model parameters remain frozen during VTrajEmbed training. For the Fast pathway, we sample at most $n_f^{\max} = 32$ valid trajectory observations and use a temporal block length of $m = 4$. A trajectory therefore contains at most

$$K = \left\lceil \frac{n_f^{\max}}{m} \right\rceil = \left\lceil \frac{32}{4} \right\rceil = 8 \tag{11}$$

Fast temporal blocks. The three ROI feature scales are extracted from visual Transformer layers 12, 18, and 24. Each original feature has width 1024 and is transformed by 2×2 spatial merging followed by an independent projection to a 2560-channel spatial feature map. Before ROI extraction, the first two feature maps are upsampled by transposed convolutions with spatial factors of 4 and 2, producing 640 and 1280 channels, respectively; the third feature map retains 2560 channels. A 7×7 ROI Align operation is applied independently at all three scales. The resulting ROI features are concatenated along the channel dimension to obtain $640 + 1280 + 2560 = 4480$ channels and are linearly projected to 2560 channels at each spatial location. The resulting $7 \times 7 \times 2560$ region representation is flattened to a 125,440-dimensional vector and passed through a $125440 \to 2560 \to 2560$ MLP with GELU activation to obtain the appearance vector. In parallel, the bounding-box center coordinates, width, and height are normalized by the image dimensions, encoded using sinusoidal positional encoding, and processed by an independent $2560 \to 2560 \to 2560$ MLP. The appearance and positional vectors are added to form the final 2560-dimensional target-ROI representation for each frame.

B.2 Optimization and Runtime Configuration

We train VTrajEmbed using fused AdamW (`adamw_torch_fused`) with an initial learning rate of 5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and zero weight decay. The learning rate follows a linear decay schedule from its initial value to zero, with no warm-up (`warmup_steps`=0). Each training process reads four expression-centered training samples per step. We use one process per GPU on eight NVIDIA A800 GPUs with 80 GB memory each and one gradient-accumulation step. The effective global batch size is therefore

$$4 \times 8 \times 1 = 32 \tag{12}$$

expression samples per optimizer update. Training is controlled by an epoch budget of three epochs; no positive `max_steps` value overrides this budget. The training random seed is 42. The software environment uses PyTorch 2.10.0, torchvision 0.25.0, and Transformers 5.17.0. Unless otherwise specified, architectural ablations use the same data split, optimizer, learning-rate schedule, batch configuration, training duration, and random seed as the full model. Objective ablations modify only the corresponding loss terms.

B.3 Expression-Group Construction and Noisy-GT Augmentation

Each expression-centered training sample contains four trajectory records. One clean ground-truth trajectory and one confirmed negative are always retained. The remaining two records are filled by either (i) one noisy-GT trajectory and one matched raw positive, or (ii) two matched raw positives when the noisy-GT variant is unavailable. Thus, each expression group contains one clean GT, zero

or one noisy GT, one or two matched raw positives, and one confirmed negative while maintaining a fixed group size of four trajectory records. Noisy-GT augmentation is enabled by default. The temporal augmentation first removes one contiguous segment comprising approximately 10%-20% of the trajectory observations and then independently drops each remaining observation with probability 0.1. The augmented trajectory is required to retain at least one valid observation. The clean ground-truth trajectory is maintained as a separate positive record and is not replaced by the noisy variant. For spatial perturbation, the box center is independently displaced by up to $\pm 5\%$ of the original box width and height along the horizontal and vertical axes, respectively. The box width and height are independently scaled by factors in $[0.9, 1.1]$. Perturbed boxes are clipped to the image boundaries. If clipping produces a degenerate box, the original box for that observation is restored.

C Candidate Generation and Encoder Details

C.1 Candidate generation and task scope

VTrajEmbed separates proposal generation from embedding-based ranking. A candidate $r = \{(t, b_t)\}$ stores nonempty bounding boxes b_t in original-image coordinates and their physical frame indices t. It represents the observations associated with one tracker identity. "Complete trajectory" means retaining the available identity track rather than clipping it to the action interval; it does not imply that every physical appearance of the object was successfully tracked. VTrajEmbed neither predicts new boxes nor regresses event boundaries.

The general query-conditioned proposal interface uses a frozen Qwen3-0.6B parser to extract appearance entities, retaining attributes such as clothing and color while discarding action and interaction descriptions. For example, a query about a man in a white suit and a woman in a black dress walking together yields two appearance prompts. Each prompt drives the frozen SAM3 tracker, and the union of its outputs forms the proposal pool. An appearance-matching object is retained even when its behavior does not match the query. When several queries are processed together, their prompts and proposals can be deduplicated and shared. This mode conditions proposal recall on the query, although the embedding of an already specified trajectory remains query-independent.

Text2Traj-Bench generates candidate trajectories from fixed category prompts and freezes each video's candidate pool after trajectory-quality and redundancy review, before verified description and query generation. Subsequent description or query rejection does not alter the frozen pool. The current person-oriented description examples and source setting should not be interpreted as evidence of performance on every object category. Other categories can use the same construction procedure, but require their own prompts, quality checks, and validation. The benchmark does not rerun the appearance parser to prune candidates separately for each evaluated query.

SAM3 detects objects at anchor frames and propagates masks forward and backward over the video timeline. Bounding boxes are extracted from nonempty masks and mapped to the original image coordinates. The encoder consumes rectangular bbox ROIs, not segmentation masks. Empty masks provide no observation; propagation failures and identity switches are tracking errors, not guarantees resolved by the encoder. Both the parser and tracker remain frozen, and neither receives gradients from the embedding objective. Replacing the tracker changes candidate recall and quality, so results using different proposal pools are not directly comparable as ranking results.

C.2 Sampling and temporal blocks

Sort each trajectory's $n(r) \geq 1$ observations by timestamp. The Fast branch selects $n_f(r) = \min(n(r), n_f^{\max})$ distinct observations uniformly over this ordered list. The reference maximum is $n_f^{\max} = 32$. If multiple observations are selected, evenly spaced list indices include both endpoints; a single observation uses the first available index. Sampling is performed independently for each trajectory without modifying the stored observations.

The Slow branch first preselects $n_s(r) = \min(n_s^{\text{budget}}, n_f(r))$ distinct observations, where the configured budget is one to four. Preselection uses evenly spaced indices in the ordered Fast observations, including endpoints when more than one frame is selected and using the first observation when only one is selected. These are positions in the valid-observation list, not a claim of uniform

physical-time spacing across occlusions. Let S_0 denote this initial set and $\bar{g}_i$ its unit-normalized global CLIP features.

Next, scan S_0 chronologically and maintain a retained set S. Keep the first frame. For each subsequent frame i, find $j^* = \arg\max_{j \in S} \bar{g}_i^\top \bar{g}_j$, breaking ties by timestamp. If the similarity is at least a fixed threshold $\tau \in [-1, 1]$, discard i and assign $\pi(i) = j^*$; otherwise retain i and set $\pi(i) = i$. Retained representatives are never deleted, and removed frames are not replenished. Therefore

$$1 \le m_s(r) = |S| \le n_s(r), \qquad \bar{g}_i^\top \bar{g}_{\pi(i)} \ge \tau \quad (i \in S_0). \tag{13}$$

This is semantic pruning of a preselected set, rather than selection by maximum semantic difference. The representative condition applies to the final retained set: a chain of pairwise merges that later deletes a representative would not provide the same guarantee. The CLIP comparison occurs before high-resolution encoding; only retained frames enter the Slow vision encoder and temporal resampler. The Fast observations, ROI features, and temporal blocks are unchanged by this pruning.

The assignment defines cluster sizes $\kappa_j = |\{i \in S_0 : \pi(i) = j\}|$, which may be recorded for auditing. The current attention resampler receives the retained features without cluster-size reweighting. These counts are used to analyze the resulting change in attention mass, not to introduce a new weighted pooling module. Fix the CLIP checkpoint, preprocessing, preselection budget, and threshold independently of test performance. The threshold controls the semantic covering radius $\sqrt{2(1-\tau)}$; a larger threshold tightens this guarantee but does not imply a monotone retained count for every greedy run.

With the block length fixed to $m = 4$, the valid block count is

$$K_r = \left\lceil \frac{n_f(r)}{4} \right\rceil. \tag{14}$$

Because the Fast pathway uses at most $n_f^{\max} = 32$ observations, the fixed maximum slot count is

$$K = \left\lceil \frac{32}{4} \right\rceil = 8. \tag{15}$$

Each Fast temporal block therefore contains at most four valid observations. The final block contains only its actual observations; short trajectories are never extended by copying frames. Slots $K_r + 1, \ldots, K$ are zero-filled and masked at the language-model input. The complete experimental configuration is summarized in Appendix B.

C.3 Fast branch: ROI features and geometry

The Fast pathway uses the frozen vision tower initialized from `Qwen3-VL-4B-Instruct`, which is based on SigLIP2-Large and has undergone dynamic-resolution adaptation and multimodal continued training. Low-resolution full frames are processed with a per-frame pixel budget $P_f = 65{,}536$.

For each input frame, we extract spatial Transformer features from layers 12, 18, and 24, using one-based layer indices. The original feature width at each tap is 1024. Each tapped feature is first subjected to 2×2 spatial merging and an independent projection, producing three spatial feature maps with 2560 channels while preserving their spatial organization.

The frozen target-ROI projection module is initialized from `ObjEmbed-4B`. Before ROI extraction, the first two feature maps are upsampled by transposed convolutions with spatial factors of 4 and 2, reducing their channel widths to 640 and 1280, respectively, while the third feature map remains at 2560 channels. For each tracked object box, we apply 7×7 ROI Align independently to all three scales and concatenate the resulting region features along the channel dimension, yielding

$$640 + 1280 + 2560 = 4480$$

channels at each ROI grid location. A position-wise linear projection maps this 4480-dimensional representation to 2560 dimensions.

The resulting $7 \times 7 \times 2560$ region tensor is flattened into a 125,440-dimensional vector and passed through a $125440 \to 2560 \to 2560$ MLP with GELU activation to produce an appearance vector.

In parallel, the bounding-box center coordinates, width, and height are normalized by the input-image dimensions and transformed by a sinusoidal positional encoding followed by an independent $2560 \to 2560 \to 2560$ MLP. The appearance and positional vectors are added to form the frozen 2560-dimensional target-ROI feature $a_t \in \mathbb{R}^{2560}$. The complete checkpoint and interface configuration is summarized in Appendix B.1.

For normalized box center (c_x, c_y), width w, height h, and area $A_t = wh$, the geometric input is

$$g_t = \left[c_x, c_y, w, h, A_t, d_t, \log \frac{A_t}{A_{t^-}}, \frac{t}{\max(T-1, 1)} \right], \qquad \widetilde{a}_t = a_t + \mathrm{MLP}_{\mathrm{geo}}(g_t). \tag{16}$$

Here T is the original video length, t^- is the preceding sampled observation, and d_t is the Euclidean displacement of normalized centers. Degenerate boxes are discarded, so areas are positive. The first observation has zero displacement and zero log-area change. Displacement is measured between observations and is not a time-normalized velocity; timestamps expose unequal sampling intervals to subsequent temporal modeling. The geometry MLP is $\mathrm{Linear}(8, H) \to \mathrm{GELU} \to \mathrm{Linear}(H, H)$. A shared, single-latent multi-head attention resampler with LayerNorm aggregates each observation block into f_i. The blocks retain chronological order before entering the language model.

C.4 Slow branch and fusion

The Slow branch independently runs the same frozen SigLIP2-Large-based vision tower initialized from `Qwen3-VL-4B-Instruct` at pixel budget $P_s = 200{,}704$. It produces full-frame token banks X_t^s with N_t spatial tokens. The cross-attention projections accommodate the vision-encoder and object-projector feature widths. The matching Fast appearance feature a_t, before explicit geometry is added, supplies the cross-attention query in Equation 2. ROI extraction anchors the query to the tracked object, but does not guarantee distinct attention patterns for nearby objects with similar appearance. In particular, implicit positional information in visual features is not a substitute for a guaranteed geometric correspondence.

The Slow resampler pools the $m_s(r)$ retained context features into one vector c, without cluster-size reweighting. Pruning changes the temporal attention mass as well as the number of frames; Lemma 1 explicitly includes this effect. Without temporal positional inputs this pooling is permutation-invariant. Accordingly, it represents aggregate scene context rather than temporally aligned interactions or event order. Fast features and timestamps carry the ordered evidence. Rapidly changing interactions that need high-resolution detail are a limitation of this division of labor.

Fusion uses $z_i = \mathrm{LN}(f_i + \alpha W_s c)$ for valid blocks only. The scalar α is initialized to zero, so the initial representation is Fast-only. At that initialization, gradients to the Slow residual's parameters through this gate are also zero; the gate can update first, after which the context pathway can receive gradients. Padded slots are set to zero after fusion and excluded from attention, rather than receiving the shared context residual.

C.5 Language-model inputs and adaptation

The shared language backbone is initialized from the language component of `Qwen3-VL-4B-Instruct`, with hidden width H_{LM}. The visual slot embedding is $W_{\mathrm{in}} z_i + \delta_{\mathrm{step}}$, where W_{in} denotes LayerNorm followed by a linear projection from the fused feature width to H_{LM}. Both paths use a fixed maximum of K slots and the following template:

```
Trajectory steps: <trajectory_step> ... <trajectory_step>
Read the complete temporal representation at <action_visual_readout>.
Match the expression to a trajectory: {expression}
Read the text representation at <action_text_readout>.
```

The visual pass fills valid trajectory slots, masks padding, leaves the expression empty, and reads h_v at the visual readout. The text pass fills the expression, zeros and masks all K trajectory slots, and reads h_e at the text readout. Its slot contents and mask are constant across candidates, including candidates with different K_r. Thus $q(e)$ depends only on the query and fixed prompt. Input isolation is established by these separate passes and masks, not by causal attention alone.

The output maps W_v and W_t are independent bias-free linear projections followed by L2 normalization. The logit scale starts at $\ell = \log(1/0.05)$, so the initial inverse temperature is $\exp(\ell) = 20$

and the temperature is 0.05. A positive shared scale changes the training distribution but not cosine rankings for a fixed query.

The pretrained language-model weights and language head remain frozen. LoRA adapters of rank 8, scaling parameter 16, and dropout 0.05 are inserted in the query and value projections of the last eight layers. Trainable modules are the geometry MLP; Fast and Slow resamplers; Slow cross-attention, query projection, residual projection, and gate; input and output projections; step and readout-token deltas; the LoRA parameters; and the logit scale. The vision encoder and ROI projector are frozen. Exact checkpoint identifiers, optimization settings, batch configuration, training duration, hardware, software versions, and the random seed are summarized in Appendix B.

D Training Data and Relation Annotations

D.1 Annotated positives and overlap-based screening

A query group is indexed by $(\text{video_id}, \text{expression_id})$. For each expression-centered training sample, the batch constructor targets four trajectory records. One clean ground-truth trajectory and one confirmed negative are always retained. The remaining two records are filled by either one noisy-GT trajectory and one identity-checked matched raw trajectory, or by two matched raw positives when the noisy-GT variant is unavailable. Thus, a training group contains one clean GT, zero or one noisy GT, one or two matched raw positives, and one confirmed negative.

An annotation may cover only the query event: its observed bbox sequence is used as supervision without claiming boxes beyond its annotated interval. A matched tracker trajectory retains its full available extent and is not cropped to that interval. This distinction reconciles event-limited supervision with identity-track retrieval.

Let $G_b \neq \varnothing$ be the target's annotated visible-frame set and $a_t(r)$ its box IoU with candidate r, taking missing observations as zero. Compute metrics on original observations, before encoder sampling:

$$\text{maxIoU}_b(r) = \max_{t \in G_b} a_t(r), \quad C_\tau(r) = \frac{\sum_{t \in G_b} \mathbf{1}[a_t(r) \geq \tau]}{|G_b|}, \quad \text{meanIoU}_b(r) = \frac{\sum_{t \in G_b} a_t(r)}{|G_b|}. \tag{17}$$

Candidates with $C_{0.5} \geq 0.8$ are ordered by descending coverage, descending mean IoU, and then track ID. Include the highest-ranked candidate that retains the target identity throughout and preserves the query's visual evidence, if one exists. Identity checking outside G_b is necessary because overlap inside the annotated interval cannot rule out later identity switches.

Noisy-GT trajectory augmentation. Noisy-GT augmentation is enabled by default. Starting from the clean ground-truth trajectory, we first remove one contiguous segment comprising approximately 10%-20% of its observations and then independently drop each remaining observation with probability 0.1. The resulting noisy trajectory is required to retain at least one valid observation. The clean ground-truth trajectory is preserved independently as an unperturbed positive record.

For each retained box, the horizontal and vertical center coordinates are independently displaced by up to $\pm 5\%$ of the original box width and height, respectively. The box width and height are independently multiplied by scale factors in $[0.9, 1.1]$. Perturbed boxes are clipped to the image boundaries. If clipping produces a degenerate box, the original box for that observation is restored. The exact augmentation configuration is summarized in Appendix B.3.

Negative verification. The threshold $\text{maxIoU}_b(r) < \tau_{\text{safe}} = 0.2$ proposes a low-overlap negative. It is not a semantic certificate: a same-person fragment outside G_b can satisfy this inequality, and another person can satisfy an underspecified query. We therefore define $\mathcal{N}_b$ to contain low-overlap proposals confirmed not to match e_b. Exclude known same-identity fragments and unresolved cases. Tracks with intermediate overlap, uncertain identity, or inadequate evidence are ignored. Positive and negative memberships are query-relative and must never conflict for the same encoded input.

D.2 Model-assisted, human-verified training data

For accepted training examples without manually annotated target tubes, an LLM verbalizes the human-approved trajectory facts D_r^H into a natural query e_b, following Appendix **??**. The query's

required predicates must be supported by the approved facts, including their temporal scope. Human verification against the complete within-video candidate pool determines query admission and supplies the target positive and confirmed within-video negatives, as described in Appendix E.3.1. Cross-video candidates are mined as easy-negative candidates, hard-negative candidates, or potential matches according to Appendix E.3.2. These mining categories and text similarities propose comparisons rather than determine relation labels. Only individually human-verified matches and mismatches populate the core batch mask; unverified or unresolved pairs receive $M_{bp} = \perp$. Calibrated automatic easy negatives, when used, are separately identified as training-only pseudo-labels and excluded from the human-verified core and benchmark ground truth.

Annotated and expanded examples can be mixed in training. A source contributes to the annotated branch only when it actually supplies object tubes or when such tubes have been separately annotated. Video captions or action-class labels alone do not provide target boxes. Thus videos sourced from UCF-101, MSR-VTT, or MSVD do not automatically become human tube-annotated examples; source-specific supervision and any added annotation must be documented. HC-STVG-style event annotations can provide tube supervision, but ranking complete proposed tracks is different from the original joint spatial and temporal grounding task.

D.3 Batch mask, duplicate records, and empty comparisons

The list $\mathcal{P}$ in Section 3.2 indexes records, not physical identities. An identical encoded trajectory may occur in several source groups; separately perturbed inputs remain distinct records. Every source query matches its own positive records. Other entries are assigned by the query-trajectory relation: propagate a positive label only when the other query is also supported by that trajectory's evidence, assign zero only to established mismatches, and otherwise abstain. Shared identity alone need not establish that an event description is visible in every partial trajectory.

Within a video, existing positive/negative annotations and verified query judgments provide these relations. Across different videos, box IoU is undefined as an identity test, and different video IDs do not prove semantic mismatch. An independently verified relation may be used; otherwise the entry remains $\perp$. Consequently the batch term supplies cross-video comparisons only where those labels exist. This is a limitation of conservative supervision, not an assumption that all off-diagonal pairs are negatives.

The numerator of $L_{t\to v}$ must include *all records* satisfying $M_{bp} = 1$, including matching records from other groups. Using only the source set $\mathcal{P}_b$ in that numerator while admitting additional matching records into its denominator would incorrectly penalize those records. Repeated records act as multiplicity weights in the log-sum-exp. For identical encoded inputs their scores are tied, and their derivative contributions add; their positions in different loss expressions do not make the underlying similarities independent.

Rows and columns contain a positive by construction, avoiding undefined logarithms. A row or column with positives but no confirmed negatives contributes zero. Likewise $L_{\text{group}} = 0$ for an empty $\mathcal{B}_N$. Ignored pairs have no direct score derivative in the corresponding masked objective, though their shared encoder parameters may still change through other pairs. The multi-positive objective rewards total positive probability and does not require uniform probability across positive variants.

E Text2Traj-Bench Construction Details

This appendix details Text2Traj-Bench's three-stage construction for object-level video grounding and relation supervision for within-video discrimination and bidirectional batch retrieval. The stages follow Section 3.3: Trajectory Curation, Verified Description, and Relation Construction. Models propose candidate trajectories, descriptions, text queries, and query-trajectory relations, while human review determines the core annotations. Source assignments are fixed before construction: duplicate clips and clips from the same recording remain together, including all derived text queries and trajectory variants. Construction settings are selected on training and development sources. Cross-video training relations never connect to held-out sources, including reserved sources of overlapping external evaluation datasets.

E.1 TRAJECTORY CURATION

Proposal and quality review. Candidate trajectory proposals are generated independently of text queries by a frozen SAM 3 tracker using a shared category-prompt inventory rather than query-specific appearance prompts. Each proposal retains its available bounding boxes and original timestamps, as defined in Appendix C.1. The prompt inventory, checkpoint, anchor-frame policy, and tracking settings are versioned; "default prompts" alone do not specify a reproducible configuration. Proposal recall is limited by this process.

Three independent annotators determine proposal usability from temporal playback and target-marked frames. Observation count, temporal coverage, contiguous support, box area, and abrupt position or scale changes assist review rather than automatically determine acceptance. Reviewers label each proposal usable, unusable, or uncertain based on object recognizability, identity continuity, tracking drift, and visual support. At least two usable votes are required for retention. Cases without a majority undergo adjudication and receive another recorded decision; a proposal without two approvals is not retained. Small size, short visibility, rapid motion, and occlusion are not rejection criteria by themselves. Individual votes and exclusion reasons are retained.

Duplicate and fragment review. Duplicate and fragment review consolidates confirmed redundant trajectories while preserving distinct objects and nonredundant fragments as separate candidates. Spatiotemporal overlap and visual similarity propose duplicate pairs. All retained candidate trajectories are also displayed together to identify missed pairs. Three annotators independently judge each pair as having the same identity, different identities, or uncertain identity. At least two supporting votes are required to establish either identity decision. Same identity alone does not make two trajectory records redundant: consolidation removes only confirmed redundant observations and retains the highest-quality representative with deterministic tie-breaking. Nonredundant fragments remain separate without interpolating missing observations. Inconsistent pairwise decisions are adjudicated before duplicate groups are formed; transitive merging must not override a confirmed different-identity judgment.

The candidate pool $\mathcal{R}(V)$ is frozen after curation and remains unchanged by subsequent description or query rejection. Unresolved identity pairs remain separate candidates and are flagged for relation review rather than deleted to make text queries unique. A query matching multiple fragments is rejected, whereas a query supported by only a particular fragment is assessed against that fragment's actual observations. A video requires at least two candidate trajectories and one accepted text query for benchmark inclusion. Subsequent description failure removes only a query proposal, not its candidate trajectory. Correcting a subsequently discovered tracking defect requires a new candidate-pool version and renewed verification of all affected queries.

E.2 VERIFIED DESCRIPTION

Structured visual evidence. For each retained candidate trajectory, a VLM proposes descriptions and structured facts grounded in its ordered visual observations. Its inputs contain chronologically ordered crops and selected temporal segments. Target-marked full frames are additionally provided for spatial relations, interactions, and scene context. The VLM returns separate appearance and motion descriptions together with structured facts recording the attribute or event, value, evidence frames or interval, temporal scope, and known or unknown status. Appearance fields describe visible properties rather than inferred demographics. Action entries distinguish event intervals and participants, while direction entries record their image-relative convention. Unobserved intervals and occlusions are recorded rather than treated as evidence of absence. Sampling and targeted evidence expansion are recorded together with their budgets.

Three-person verification. Three independent annotators verify the proposed descriptions and structured facts against the video to establish the human-approved facts D_r^H. They label each description usable, usable after minor correction, unusable, or uncertain. Minor corrections may remove unsupported details, simplify an overstated action, or correct a clearly visible attribute; extensive rewriting requires a new proposal. All edits are consolidated into a single final version, which must itself receive at least two approvals. Agreement that different edits are individually acceptable does not establish approval of their union. Unresolved fields remain unknown and cannot supply facts for text query generation. The resulting D_r^H retains approved statements and their support-

ing evidence rather than providing an exhaustive account of everything that occurred. A rejected description does not remove its candidate trajectory from $\mathcal{R}(V)$.

Text query generation. An LLM generates a natural text query e and its required predicates $P(e)$ from D_r^H alone, without access to competing candidate trajectories. Each predicate must be entailed by the human-approved facts, including its temporal and relational qualifiers; lexical copying is unnecessary. Facts from different intervals cannot be combined into an unsupported simultaneous event. Text query generation uses a fixed, recorded proposal budget, and minor revisions are versioned. The next stage checks query faithfulness and naturalness; any changed wording is rechecked against the complete candidate pool $\mathcal{R}(V)$. These are model-generated text queries rather than independently collected user utterances. More fluent paraphrasing alone does not establish coverage of the distribution of natural user queries.

E.3 Relation Construction

E.3.1 Within-video Verification

Complete human verification. Within-video relation annotation verifies every proposed text query against the complete candidate pool $\mathcal{R}(V)$, including candidate trajectories without approved descriptions. With Q_V proposed queries and K_V candidate trajectories, the relation matrix has shape $Q_V \times K_V$, not necessarily $K_V \times K_V$. LLM pre-labels may highlight relevant facts, but a human annotator verifies every query-trajectory pair using the actual visual observations and available context. Disputed decisions receive an independent second review, and relations that remain uncertain retain $\perp$. Query faithfulness and naturalness are checked in the same pass.

The final relation annotation distinguishes confirmed matches, confirmed mismatches, and uncertain relations, consistent with the relation mask M_{bp} in Section 3.2. We distinguish the LLM proposal $\widehat{Y}(e, r)$, the final annotation $\bar{Y}(e, r)$, and the unobserved true semantic relation $Y(e, r)$. The labels 1, 0, and $\perp$ respectively denote a confirmed match, a confirmed mismatch, and an uncertain relation under the following policy:

$$\bar{Y}(e,r) = \begin{cases} 1, & \text{all required predicates are jointly supported,} \\ 0, & \text{a necessary condition is demonstrably violated,} \\ \perp, & \text{otherwise.} \end{cases} \tag{18}$$

Both confirmed matches and confirmed mismatches must respect the text query's temporal scope and participant bindings. Observing running in one interval does not refute "walks at some point." A confirmed mismatch based on an absent event requires adequate coverage of the relevant interval; omitted or uncertain attributes do not establish a mismatch. Human reviewers may inspect additional existing observations when the description is insufficient. Joint support additionally requires consistent event timing and participant binding.

Admission and within-video discrimination. Benchmark admission establishes a unique target trajectory and provides confirmed relations for within-video discrimination. Equation 10 requires one confirmed target match and confirmed mismatches for every alternative candidate trajectory in $\mathcal{R}(V)$. An additional confirmed match or any uncertain relation blocks admission without altering the candidate pool or other queries. This establishes an annotation-based decision within the candidate pool rather than infallible uniqueness in the physical scene. Accepted training queries define the positives and confirmed within-video negatives of each query-centered trajectory group:

$$\mathcal{P}_b = \{r \in \mathcal{R}(V_b) : \bar{Y}(e_b, r) = 1\}, \qquad \mathcal{N}_b = \{r \in \mathcal{R}(V_b) : \bar{Y}(e_b, r) = 0\}. \tag{19}$$

These sets supply the within-video discrimination objective L_{group}. Training-time trajectory augmentations are handled separately in Appendix D. The core benchmark uses only accepted single-target text queries; rejected proposals and their reasons remain in construction logs rather than providing training supervision. Every released held-out query has a fully human-reviewed row of within-video relations, not merely a sampled check of automatic labels.

E.3.2 Cross-video Expansion

Stratified candidate mining. For bidirectional training, cross-video expansion mines additional query-trajectory comparisons exclusively from training queries and trajectories. Each text query is

compared with the candidate trajectory's facts rather than merely with another target's caption. Mining combines structured compatibility checks with appearance and motion similarity to form three strata. Easy-negative candidates exhibit obvious category or basic-property conflicts. Hard-negative candidates have otherwise similar descriptions but a possible conflict in a necessary attribute or event. Potential matches have descriptions that may support the text query. These strata guide sampling rather than determine final relation labels, and similarity ranks candidate pairs only. Category hierarchies, multicolored clothing, temporal changes, and concurrent actions prevent different field values from automatically establishing a confirmed mismatch. Random cross-video query-trajectory pairs supplement similarity-based mining to audit its selection bias. Mining budgets, random seeds, encoders, and stratum assignments are logged; duplicate pairs retain their provenance but are reviewed once.

LLM proposal and human confirmation. Mined query-trajectory pairs receive LLM relation proposals, while individual human verification determines the confirmed relations in the core relation graph. For each mined pair, the LLM receives e, $P(e)$, and D_r^H, and predicts $\widehat{Y}(e, r) \in \{1, 0, \perp\}$ together with the decisive predicates and evidence references. It applies Equation 18 and assigns $\perp$ when the descriptions leave the relation unresolved, including unresolved event scope or joint support. A textual omission does not establish a confirmed mismatch, and high similarity does not establish a confirmed match. Potential matches and hard-negative candidates receive pairwise human review with access to visual evidence; disagreements prompt independent adjudication. Easy-negative candidates are reviewed in stratified samples. Individually confirmed pairs enter the core relation graph, while unverified pairs remain $\perp$. Mining strata and final relation labels are stored separately, allowing a mined hard-negative candidate to become a confirmed match or an uncertain relation.

An optional training-only expansion uses calibrated automatic easy negatives as pseudo-labels separately from the human-verified core. Automatic easy negatives are accepted only after independent human calibration on development pairs selected by the same rule. Acceptance uses a predeclared upper bound on error with a confidence interval rather than LLM self-confidence. Uncertain audit outcomes count against acceptance. Rule-specific sample sizes, errors, coverage, and drift checks must be reported, and the rule is frozen before application. These relations are tagged as pseudo-labels, excluded from benchmark ground truth, and evaluated separately from the human-verified core. No fixed numerical accuracy or cost saving is assumed.

From verified relations to bidirectional batch retrieval. Confirmed within-video and cross-video relations populate the relation mask M_{bp} for bidirectional batch retrieval when the corresponding text queries and trajectory records occur in the same batch. To match Section 3.2, mining for this objective targets candidate trajectories with an accepted source query, making them eligible for the indexed list of positive records $\mathcal{P}$. Relation-aware batching co-selects linked query-centered trajectory groups so that their verified relations are available for the row and column comparisons of L_{bi}. A confirmed mismatch involving a trajectory absent from $\mathcal{P}$ supplies no comparison for L_{bi}. Missing relations are assigned $\perp$ rather than treated as confirmed mismatches, and each row and column retains its source positive. Reporting must include realized cross-video positive and negative counts and zero-negative row and column rates, not only the size of the offline relation graph. A positive-only comparison contributes zero loss. Transferring a relation label to an augmented trajectory record requires checking that the evidence deciding that particular relation is preserved; identity agreement alone is insufficient.

Shared auditing and release. Independent auditing and traceable release records document annotation reliability and construction provenance. Independent auditors are blinded to model suggestions and previous votes and inspect a random sample of accepted records together with a separately reported sample of difficult cases. The audit records residual errors, pre-adjudication agreement, three-vote distributions, corrections, query rejection reasons, and reviewer time per accepted query or relation. Text query and candidate-trajectory manifests store the queries, candidate and target trajectory IDs, original trajectory observations, source and split assignments, and candidate-pool versions. Linked fact, vote, and relation records retain evidence, correction history, mining strata, provenance, and pseudo-label status. Prompts, checkpoints, manifests, budgets, and audit statistics must be supplied from actual construction runs. Majority approval is an annotation rule rather than a guarantee of correctness.

Evaluation scope. Evaluation measures object-level video grounding through similarity-based ranking of candidate trajectories within each video's fixed candidate pool $\mathcal{R}(V)$. Models receive the text query e, video V, and candidate trajectories in $\mathcal{R}(V)$; target trajectory IDs, verified descriptions, relation labels, and annotation evidence tags are not model inputs. Every method ranks the same complete candidate pool for each text query. Reporting includes within-video R@1, candidate-count strata 2, 3-5, 6-10, and > 10, and a random-ranking baseline obtained by averaging $1/|\mathcal{R}(V)|$ over text queries. Single-frame and motion-dependent analyses examine whether temporal evidence is necessary, separately from whether a unique target trajectory is established.

Beyond ranking within the candidate pool, an independently sampled source subset is required to estimate proposal recall, post-curation retention, and end-to-end success. The admission and selection events satisfy

$$\Pr(S \cap A) = \Pr(A) \Pr(S \mid A), \tag{20}$$

where A denotes benchmark admission and S denotes correct target-trajectory selection. Both factors must be measured on the same source subset; within-video R@1 on the screened benchmark cannot replace the conditional accuracy measured on that subset. Reverse, cross-video, and global retrieval require separate galleries and relevance protocols and are not established by within-video ranking scores.

F Lemma 1: Task-Relevant Preservation under Slow-Fast Sampling

F.1 Reference representations and semantic coverage

Fix a trajectory, its Fast observation set $\mathcal{F}$, and all trained parameters. The analysis concerns deterministic encoding, with dropout disabled. The Slow branch first obtains $S_0 \subseteq \mathcal{F}$ by uniform index preselection and then greedily prunes it to S, as specified in Equation 13. Write $n_s = |S_0|$ and $m_s = |S|$. A representative $\pi(i) \in S$ is assigned to every $i \in S_0$, including $\pi(j) = j$ for retained frames. Since retained representatives are never deleted, all assignments satisfy the threshold in the final set.

For each potential Slow frame, u_i denotes the object-conditioned context from Equation 2, evaluated using the same Fast ROI query and the same high-resolution encoder. The features of frames excluded at runtime are reference quantities for analysis, not additional computation in the deployed model. Let

$$U_0 = [u_i]_{i \in S_0}, \quad \widetilde{U}_0 = [u_{\pi(i)}]_{i \in S_0}, \quad U_S = [u_j]_{j \in S}, \quad c_0 = R(U_0), \quad c = R(U_S). \tag{21}$$

The proxy list $\widetilde{U}_0$ has the same length as U_0 and repeats each retained representative according to its cluster size. It is not the actual pruned input. The dense-context reference is $c_D = R([u_t]_{t \in \mathcal{F}})$, which would encode high-resolution context at every Fast observation. The temporal resampler summarizes a global context vector; the reference is not a reconstruction of temporally ordered scene interactions. The Fast temporal sequence is fixed throughout this comparison.

Conditions. The lemma uses three explicit conditions. First, on assigned pairs, CLIP similarity controls object-conditioned context difference up to a residual:

$$\|u_i - u_{\pi(i)}\|_2 \le L_u \|\bar{g}_i - \bar{g}_{\pi(i)}\|_2 + \eta, \qquad L_u \ge 0, \quad \eta \ge 0. \tag{22}$$

Second, at the preselected input length, the resampler satisfies

$$\|R(U) - R(U')\|_2 \le K_R \max_i \|u_i - u'_i\|_2 \tag{23}$$

on the feature domain of interest. Third, preselection approximates the dense summary with $\|c_0 - c_D\|_2 \le \varepsilon_{\mathrm{pre}}$. This last quantity includes the effects of finite coverage and changing the number of inputs to attention. The conditions do not assert that CLIP features are sufficient for every object interaction or that sparse pooling is automatically accurate.

The first condition is weaker than equating CLIP similarity with context equality: the residual η allows discrepancies between global scene similarity and the object's contextual state. A large η makes the resulting guarantee weak, as it should. The second condition holds with suitable finite constants for standard fixed-query attention on bounded feature domains, with finite projection

norms and stable postprocessing. Output normalization additionally requires a positive lower bound on the pre-normalization embedding norm. These are local stability conditions to justify or measure; the proof does not assume that arbitrary learned representations have small sensitivity constants.

F.2 Proof of semantic-pruning and representation bounds

Restatement of Lemma 1. Under the conditions above, define $\rho_\tau = L_u\sqrt{2(1-\tau)}+\eta$ and $\varepsilon_{\text{mult}} = \|R(\widetilde{U}_0) - R(U_S)\|_2$. The context errors satisfy

$$\|c - c_0\|_2 \le K_R\rho_\tau + \varepsilon_{\text{mult}}, \qquad \|c - c_D\|_2 \le \varepsilon_{\text{pre}} + K_R\rho_\tau + \varepsilon_{\text{mult}}. \tag{24}$$

Proof. Unit normalization and the representative condition imply

$$\|\bar{g}_i - \bar{g}_{\pi(i)}\|_2^2 = 2 - 2\bar{g}_i^\top \bar{g}_{\pi(i)} \le 2(1-\tau).$$

Equation 22 gives $\max_i \|u_i - u_{\pi(i)}\|_2 \le \rho_\tau$. The original and proxy lists have equal length, so Equation 23 yields $\|R(U_0) - R(\widetilde{U}_0)\|_2 \le K_R\rho_\tau$. A triangle inequality through $R(\widetilde{U}_0)$ proves the first bound. Another triangle inequality through c_0, using the preselection condition, proves the second. The multiplicity change is handled separately because fixed-length stability does not apply directly to lists of different lengths. □

Let $F_r = (f_1, \dots, f_{K_r})$ be the unchanged Fast sequence, and let $\Phi(F_r, c)$ denote fusion, language-model aggregation, projection, and L2 normalization. For the fixed candidate assume

$$\|\Phi(F_r, c) - \Phi(F_r, c')\|_2 \le K_c\|c - c'\|_2.$$

Applying this to the second context bound proves Equation 4, with $v_D(r) = \Phi(F_r, c_D)$. The constants and errors may depend on the candidate; the ranking argument uses their maximum over the fixed candidate pool.

Retrieval consequence. For a unit text embedding q, Cauchy-Schwarz gives $|q^\top v(r) - q^\top v_D(r)| \le \delta_r$. Suppose r^* is correctly ranked first by the reference, with cosine margin

$$\gamma = q^\top v_D(r^*) - \max_{r \ne r^*} q^\top v_D(r) > 0. \tag{25}$$

Writing $\delta = \max_r \delta_r$, every competitor satisfies

$$q^\top v(r^*) - q^\top v(r) \ge q^\top v_D(r^*) - q^\top v_D(r) - \delta_{r^*} - \delta_r \ge \gamma - 2\delta.$$

Thus $\gamma > 2\delta$ preserves the correct top-1 decision. The positive logit scale does not change this ranking statement. The reference must itself be correct; no correctness guarantee is inferred merely from using more high-resolution frames.

F.3 Attention mass after deleting repeated representatives

Semantic closeness alone does not control the full effect of pruning. The proxy input repeats representative j exactly $\kappa_j = |\{i : \pi(i) = j\}|$ times, while the actual input includes it once. For a fixed-query attention head h, let z_{hj} be its query-key logit and v_{hj} its projected value for retained feature u_j. The two outputs can be written using distributions on the same representative set:

$$p_{hj} = \frac{\kappa_j e^{z_{hj}}}{\sum_{k\in S}\kappa_k e^{z_{hk}}}, \qquad q_{hj} = \frac{e^{z_{hj}}}{\sum_{k\in S} e^{z_{hk}}}, \qquad y_h^{\text{proxy}} = \sum_j p_{hj}v_{hj}, \quad y_h^{\text{pruned}} = \sum_j q_{hj}v_{hj}. \tag{26}$$

These formulas apply when the resampler has no per-frame positional or temporal bias that distinguishes repeated identical inputs, as in the order-invariant Slow pooling described in Appendix C. If $\|v_{hj}\|_2 \le V_h$, then

$$\|y_h^{\text{proxy}} - y_h^{\text{pruned}}\|_2 \le \sum_j |p_{hj} - q_{hj}|\|v_{hj}\|_2 \le V_h\|p_h - q_h\|_1. \tag{27}$$

If the output projection and subsequent postprocessing together are K_{post}-Lipschitz on concatenated head outputs, a computable upper bound is

$$\varepsilon_{\mathrm{mult}} \le K_{\mathrm{post}} \left(\sum_h V_h^2 \|p_h - q_h\|_1^2 \right)^{1/2}. \tag{28}$$

For multiple fixed latent queries, apply the same calculation to each head-query pair and to the final aggregation. The bound separates perturbing a feature from changing the frequency with which its representative enters attention. Cluster counts and retained-frame attention logits determine the mass term without requiring high-resolution features for the deleted frames.

Why this term cannot be omitted. Consider scalar values consisting of three copies of 0 and one copy of 1, with equal attention logits. Choose CLIP features identical within each value class and different between classes. Pruning exact duplicates leaves values 0 and 1. Even with zero within-cluster feature distortion, the unpruned output is 0.25 and the pruned output is 0.5. A bound depending only on the CLIP covering radius could incorrectly predict zero error. In contrast, equal cluster sizes give $p_h = q_h$ and zero multiplicity error. Adding $\log \kappa_j$ to the retained attention logits would also reproduce proxy multiplicities, but that would be a modified pooling rule; it is not assumed or introduced in the current encoder.

F.4 Preselection coverage, stability, and Fast features

The preselection error is distinct from the pruning error. To make this explicit, normalize original frame indices and assign each Fast observation $t \in \mathcal{F}$ to a nearest preselected observation $\sigma(t) \in S_0$. Define the actual temporal covering radius $h_0 = \max_{t \in \mathcal{F}} |t - \sigma(t)|$. Uniform selection in the valid-observation list does not imply a fixed physical-time radius, particularly across occlusions.

If object-conditioned contexts obey $\|u_t - u_{\sigma(t)}\|_2 \le L_t h_0$, use the repeated preselection proxy $[u_{\sigma(t)}]_{t \in \mathcal{F}}$. The same proof gives

$$\|c_0 - c_D\|_2 \le K_R^D L_t h_0 + \varepsilon_{\mathrm{pre,mult}}, \tag{29}$$

where K_R^D is the fixed-length stability constant at $|\mathcal{F}|$ inputs, and $\varepsilon_{\mathrm{pre,mult}}$ is the change from the repeated proxy to the unweighted preselected input. Equation 28 applies with the preselection assignment counts. This provides one sufficient choice of $\varepsilon_{\mathrm{pre}}$ and shows why slow variation helps, while not silently assuming that changing attention multiplicities is harmless. A small time covering radius alone does not eliminate the mass term.

For completeness, a fixed-query attention head has a finite feature-perturbation constant on bounded inputs. Suppose corresponding keys and values differ by at most $L_k d$ and $L_v d$, value norms are bounded by V, and the fixed query has norm at most Q. For key dimension d_k, logits differ by at most $QL_k d/\sqrt{d_k}$. The softmax Jacobian gives $\|\operatorname{softmax}(z) - \operatorname{softmax}(z')\|_1 \le 2\|z - z'\|_\infty$, so the head-output difference is at most

$$\left(L_v + \frac{2VQL_k}{\sqrt{d_k}} \right) d.$$

Combining heads and the Lipschitz postprocessing map yields a valid K_R. This argument needs bounded projections and a stable feature domain, and does not claim that the constants are small in every trained model.

Fusion also exposes the context sensitivity explicitly. If LayerNorm is K_{LN}-Lipschitz on valid block inputs, the stack of fused blocks changes by at most

$$\sqrt{K_r}\, K_{\mathrm{LN}} |\alpha| \|W_s\|_2 \, \|c - c'\|_2$$

in Frobenius norm. Subsequent input projection, backbone aggregation, and normalized readout contribute their own stability factors to K_c. A positive lower bound on the unnormalized readout norm is needed for the last normalization. At $\alpha = 0$, the representation is independent of context, consistent with the initialization. The lemma concerns fixed parameters and does not imply that the gate or the representation error changes monotonically during training.

The reference in Lemma 1 uses the same low-resolution Fast sequence and a denser Slow context. Thus the theorem specifically justifies context compression while keeping the motion observations

unchanged. If a higher-quality Fast reference F_r^* is also considered, define $\varepsilon_F = \|F_r - F_r^*\|$ and assume joint stability

$$\|\Phi(F_r, c) - \Phi(F_r^*, c_D)\|_2 \le K_F \varepsilon_F + K_c \|c - c_D\|_2.$$

The total representation bound then adds $K_F \varepsilon_F$ to Equation 4. This isolates rather than assumes away the effect of low Fast resolution. The reference $\Phi(F_r^*, c_D)$ fixes the same dense context; comparison to a system that also changes the ROI queries or context encoder needs to include those differences in its context error. Tiny objects and fast, detailed interactions can violate the small-error regime.

F.5 Encoding cost and scope of the design argument

Let C_s be the cost of one high-resolution full-frame encoder pass under a common frame budget. Preselection followed by pruning reduces the number of Slow passes from n_s to m_s, saving $(n_s - m_s)C_s$ before accounting for pruning overhead. If C_{CLIP} and C_{prune} are additional costs relative to the unpruned baseline, a sufficient condition for a saving in this portion of the pipeline is

$$(n_s - m_s)C_s > C_{\mathrm{CLIP}} + C_{\mathrm{prune}}. \tag{30}$$

If CLIP features are shared or cached, only their incremental cost belongs on the right. Actual frame sizes, resampling overhead, ROI reuse, and caching across candidates must be included in measured end-to-end cost. No positive saving is guaranteed when no frame is removed or when the additional work dominates.

Ignoring auxiliary modules, the visual-encoder cost proxy is $n_f P_f + m_s P_s$, with $P_f = 65{,}536$ and $P_s = 200{,}704$. The Fast frame limit remains 32; the Slow preselection budget is one to four, and the actual retained count can be smaller. These are implementation budgets, not information saturation points. The design allocates temporal coverage to the Fast branch and uses the pruning tolerance to bound one source of error in sparse high-resolution context. It does not establish strict dominance over every uniform-resolution encoder, lossless compression of all scene semantics, or a generally optimal threshold. The bridge residual, attention mass discrepancy, preselection error, reference margins, and actual runtime require validation on representative data. Uniform preselection followed by CLIP pruning does not by itself guarantee correct object tracking or recall of an event absent from the observations.

G Lemma 2: Task-Aligned Multi-Positive Ranking Control

G.1 A common retrieval objective

Each local, text-to-trajectory, or trajectory-to-text term in Section 3.2 is a comparison A with a nonempty positive index set A^+ and a disjoint negative index set A^-. Indices refer to the records actually included in that comparison; repeated records retain their multiplicities. All scores are finite. Write

$$Z_A^+ = \sum_{j \in A^+} e^{s_j}, \qquad Z_A^- = \sum_{j \in A^-} e^{s_j}, \qquad \mathcal{L}_A = -\log \frac{Z_A^+}{Z_A^+ + Z_A^-} = \log\left(1 + \frac{Z_A^-}{Z_A^+}\right). \tag{31}$$

The fraction is the total softmax probability of the correct answer set. This objective marginalizes over valid answers instead of arbitrarily selecting a single positive and penalizing the others. Its retrieval interpretation is that the highest-ranked answer should belong to that set; it does not require a uniform distribution over positives or that every positive outrank every negative.

For $m_A = |A^+| \ge 1$, define the conservative top-1 error

$$E_A = \begin{cases} \mathbf{1}\left[\max_{j \in A^-} s_j \ge \max_{j \in A^+} s_j\right], & A^- \ne \varnothing, \\ 0, & A^- = \varnothing. \end{cases} \tag{32}$$

Ties between a highest-scoring positive and negative count as errors, so the bound also applies to implementations that sometimes break such ties in favor of a positive. Its semantic interpretation assumes that the included labels correctly identify matching and non-matching inputs. No claim is made about excluded, unconfirmed pairs.

G.2 Proof of the per-comparison bound

Restatement of Lemma 2, first part. For every comparison above, $\log(1 + 1/m_A)E_A \le \mathcal{L}_A$. In particular, $\mathcal{L}_A < \log(1 + 1/m_A)$ guarantees that the best positive strictly outranks every included negative.

Proof. If A^- is empty, both E_A and $\mathcal{L}_A$ are zero. If $E_A = 0$, the claim follows from $\mathcal{L}_A \ge 0$. Otherwise, set $s^+ = \max_{j \in A^+} s_j$. A negative has score at least s^+, while each of the m_A positives has score at most s^+. Hence

$$Z_A^- \ge e^{s^+}, \qquad Z_A^+ \le m_A e^{s^+}, \qquad \mathcal{L}_A \ge \log\left(1 + \frac{1}{m_A}\right). \tag{33}$$

This is the desired bound when $E_A = 1$, completing all cases. The strict-loss threshold follows by contraposition. □

Sharpness and multiple positives. The constant cannot be increased without additional assumptions: take m_A positives and one negative, all with score zero. Then $E_A = 1$ and $\mathcal{L}_A = \log(1 + 1/m_A)$. For a single positive, the threshold is $\log 2$. With multiple positives the constant becomes smaller, reflecting that aggregated positive probability can be large even when one negative ties or exceeds each individual positive. Exact duplicate records must be counted in m_A because they also appear repeatedly in the implemented sums; counting only distinct identities would not justify this bound for the current objective. The guarantee concerns the best valid answer, not alignment of every positive variant.

G.3 Joint control of the three retrieval families

Let $\mathcal{F}_g = \mathcal{B}_N$ index local comparisons, $\mathcal{F}_t = \mathcal{B}$ index batch rows, and $\mathcal{F}_v = \mathcal{P}$ index batch columns. Their positive and negative sets are

$$\begin{aligned} A_{g,b}^+ &= \mathcal{P}_b, & A_{g,b}^- &= \mathcal{N}_b, \\ A_{t,b}^+ &= \mathcal{J}_b^+, & A_{t,b}^- &= \{p \in \mathcal{P} : M_{bp} = 0\}, \\ A_{v,p}^+ &= \mathcal{I}_p^+, & A_{v,p}^- &= \{b \in \mathcal{B} : M_{bp} = 0\}. \end{aligned} \tag{34}$$

For a nonempty family $\mathcal{F}_x$, define

$$E_x = \frac{1}{|\mathcal{F}_x|} \sum_{A \in \mathcal{F}_x} E_A, \qquad m_x = \max_{A \in \mathcal{F}_x} |A^+|, \qquad c_x = \log(1 + 1/m_x). \tag{35}$$

If a family is empty, set its error and loss to zero and use $m_x = 1$. For example, this covers a batch with no local negatives. A row or column with no confirmed negatives likewise has zero loss and zero included-comparison error; it imposes no discrimination constraint.

Restatement of Lemma 2, second part. With $\lambda_g, \lambda_{\mathrm{bi}} > 0$, the objective

$$L = \lambda_g L_{\mathrm{group}} + \frac{\lambda_{\mathrm{bi}}}{2}(L_{t \to v} + L_{v \to t})$$

satisfies Equation 9. This is a deterministic bound on the empirical errors for the included comparisons; it does not require independent comparisons or disjoint score pairs.

Proof. For any $A \in \mathcal{F}_x$, $m_A \le m_x$, so $c_x \le \log(1 + 1/m_A)$. The first part of the lemma gives $c_x E_A \le \mathcal{L}_A$. Averaging over each family yields

$$c_g E_g \le L_{\mathrm{group}}, \qquad c_t E_t \le L_{t \to v}, \qquad c_v E_v \le L_{v \to t}. \tag{36}$$

The definitions make these inequalities valid for empty families as well. Multiplying by λ_g, $\lambda_{\mathrm{bi}}/2$, and $\lambda_{\mathrm{bi}}/2$, respectively, and adding proves the joint bound. Score sharing between comparisons does not affect any inequality. □

Implication for loss design. The local term controls object selection among confirmed candidates from the same video. The batch row term controls text-to-trajectory matching over the verified batch relations, while the column term controls the reverse selection of a valid description for each trajectory. A local negative absent from the batch's positive-record list receives direct supervision only through the local term. Conversely, a verified cross-video mismatch absent from a query's local pool can enter the batch objective. These are conditional sources of additional comparisons, not a claim that every batch supplies both kinds. The reverse direction also changes the conditional normalization even when the underlying score pairs overlap. Thus the joint objective accounts for the intended retrieval decisions without requiring disjoint negative sets.

G.4 Relation masking and direct supervision

The positive numerator must contain every included record known to match the query. A match erroneously included only in the denominator is treated as a negative, breaking the intended semantic interpretation of the error bound. Unknown relations are excluded from both sets; different video IDs alone do not certify a mismatch. Algebraically the lemma applies to the supplied labels, while its interpretation as control of correct semantic retrieval depends on those labels being accurate.

The derivative calculation provides an additional consistency check. With $Z_A = Z_A^+ + Z_A^-$,

$$\frac{\partial \mathcal{L}_A}{\partial s_j} = \begin{cases} e^{s_j}/Z_A > 0, & j \in A^-, \\ e^{s_j}(1/Z_A - 1/Z_A^+) \le 0, & j \in A^+, \\ 0, & j \notin A^+ \cup A^-. \end{cases} \tag{37}$$

The positive-pair derivative is strictly negative exactly when that comparison has a negative. For a batch mismatch $M_{bp} = 0$, let $D_b = \sum_{j \in \mathcal{J}_b} e^{s_{bj}}$ and $C_p = \sum_{a \in \mathcal{I}_p} e^{s_{ap}}$. Its derivative in the bidirectional term is

$$\frac{\partial L_{\mathrm{bi}}}{\partial s_{bp}} = \frac{e^{s_{bp}}}{2} \left(\frac{1}{|\mathcal{B}| D_b} + \frac{1}{|\mathcal{P}| C_p} \right) > 0. \tag{38}$$

Matching pairs have nonpositive contributions and ignored pairs have zero direct contributions. Identical query-input pairs have tied scores: the derivative with respect to that shared score sums the occurrence derivatives. Consistent labels therefore give sign-consistent contributions even when a score appears in both losses. Distinct loss indices do not make the underlying similarities independent.

Scope of the justification. The bound supports the form of a multi-positive objective and the joint treatment of local and bidirectional retrieval. It does not establish an optimal ratio of λ_g to λ_{bi}; the adopted $0.1 : 1$ weighting requires empirical validation. A small average loss controls average included-comparison error, not necessarily every individual comparison. The result neither bounds errors on unconfirmed or unseen candidates nor accounts for tracker misses, annotation mistakes, or train-test distribution shifts. It also does not guarantee that the three empirical errors decrease monotonically during training. Shared encoder parameters couple score changes: for a small gradient step,

$$\Delta s_i \approx -\eta \sum_j \frac{\partial L}{\partial s_j} \langle \nabla_\theta s_i, \nabla_\theta s_j \rangle.$$

Consequently the direct derivative signs are not a guarantee on every score update. Loss/mask ablations, relation-quality audits, and held-out retrieval evaluation remain necessary to establish the empirical benefit.