TRAVEL: TRAJECTORY-GUIDED VIDEO EMBEDDING LEARNING FOR DRIVING-VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieving scenarios of interest from large driving-video collections is essential for data curation, model development, and safety validation. Program-based systems provide explicit control over well-specified events, but commonly depend on auxiliary structured inputs and an additional language-to-program translation step. Multimodal embedding models offer a more direct interface through single-vector text–video retrieval. However, our evaluation across several pretrained embedding models reveals weak performance on driving videos, particularly for motion-centric queries. Standard supervised fine-tuning substantially narrows the domain gap but remains insufficient for capturing fine-grained driving dynamics. Ego trajectories offer a denser description of motion than captions, and their pairwise similarity can serve as graded motion supervision. Based on this insight, we introduce TraVEL (Trajectory-Guided Video Embedding Learning), a trajectory-aware adaptation framework that uses ego trajectories as privileged training supervision. TraVEL applies Group Relative Policy Optimization with complementary rewards: a graded video-to-video reward organizes the embedding space according to ego-motion similarity, while a text-to-video reward preserves language alignment. Trajectories are used only during training, and inference remains single-vector retrieval without ego poses or auxiliary perception outputs. We construct two complementary driving-video retrieval benchmarks derived from nuReasoning and the Argoverse 2 (AV2) Sensor Dataset for in-domain and zero-shot cross-domain evaluation. Across three embedding-model families, TraVEL consistently improves motion retrieval over supervised fine-tuning while maintaining comparable instance-level performance. On Qwen3-VL-Embedding-2B, it improves lateral and longitudinal motion mAP by 7.5 and 10.3 points, respectively. These improvements also transfer to zero-shot evaluation on the AV2 benchmark, demonstrating cross-dataset generalization.

1 INTRODUCTION

Autonomous-driving fleets collect massive volumes of driving logs, supporting the development of downstream perception and planning models (Lu et al., 2026; Yao et al., 2026a). Since much of this data captures routine behaviors such as driving straight at a constant speed, constructing useful datasets for model development and evaluation requires deliberate scenario mining. For example, the Waymo Open Motion Dataset and Argoverse 2 (AV2) explicitly mine interactive scenarios, while nuScenes manually selects interesting scenes from a larger pool of raw data (Ettinger et al., 2021; Wilson et al., 2021; Caesar et al., 2020). Efficiently identifying scenarios of interest is thus essential for driving-data management and curation. Retrieved examples can be prioritized for annotation or selectively sampled for downstream model training (Liang et al., 2024; Feng et al., 2026), used to construct or condition scenarios for simulation and validation (Ding et al., 2024; Yao et al., 2026b), or serve as precedents for assessing candidate driving actions and counterfactual alternatives (Patrikar et al., 2025). These needs have motivated growing interest in automated driving-video retrieval and scenario mining (Ki et al., 2025; Davidson et al., 2026; Yao et al., 2026b).

In practice, program-based retrieval is a common and dependable approach for identifying well-specified driving events. These systems translate expert knowledge into executable functions over detected events and motion cues, providing explicit and interpretable control over the target scenario (Davidson et al., 2026; Park et al., 2026; Yao et al., 2026b). However, their execution com-

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.

monly relies on auxiliary structured data or outputs from upstream perception modules, such as object detections, tracks, HD maps, and ego poses. They also require an additional query-time translation step to support natural-language retrieval. Even when all constituent events are already represented by existing functions, each compositional query must still be converted into an executable program. For example, RefAV (Davidson et al., 2026) uses an LLM to decompose each natural-language query into predefined atomic functions for scenario mining: a query such as "the ego vehicle turns right at an intersection while a pedestrian is crossing" must be converted into a program that combines functions for the right turn, the intersection, and the crossing pedestrian. This step complicates the retrieval pipeline and may introduce errors when converting the user's intent into executable logic.

In contrast, multimodal embedding models offer a direct retrieval interface by mapping natural-language queries and videos into a shared representation space. Each video can be encoded offline as a single vector and ranked through efficient similarity search. Recent multimodal embedding models (Jiang et al., 2025; Meng et al., 2025; Li et al., 2026) have demonstrated strong performance across diverse multimodal retrieval tasks. However, our initial evaluation across several pretrained multimodal embedding models reveals consistently weak performance on driving-video retrieval, particularly for motion-centric queries. Standard domain adaptation through supervised fine-tuning (SFT) improves overall retrieval performance, but the fine-tuned models still struggle to capture motion concepts, even when those concepts are present in the training captions. For example, although a training caption such as "Gently accelerate to execute the left turn at the green light" explicitly describes a left turn, the fine-tuned model still struggles to retrieve relevant videos for the query "Turn left." This observation suggests that caption supervision alone may be insufficient for learning motion-aware video representations. It is also consistent with prior findings that multimodal embedding models can rely on contextual shortcuts rather than action dynamics (Ma et al., 2024).

Motivated by this limitation, we explore using ego trajectories, which can be obtained from onboard sensors or visual pose estimation (Zuo et al., 2026), as additional supervision to make the learned video embedding space motion-aware. Unlike captions, which only abstractly describe motion through semantic terms, ego trajectories provide a dense physical description of how the ego vehicle moves. Pairwise trajectory similarity can therefore provide graded supervision about motion relationships among videos. Based on this, we introduce TraVEL (Trajectory-Guided Video Embedding Learning), a trajectory-aware adaptation framework for learning motion-aware driving-video embeddings. TraVEL first adapts a general-purpose embedding model to the driving domain with paired caption supervision, and then applies trajectory-aware GRPO (Guo et al., 2025), which combines an ego-trajectory similarity reward that organizes the video embedding space by vehicle motion with a text-to-video reward that preserves language alignment. As illustrated in Fig. 1, TraVEL encourages higher embedding similarity between videos with similar ego motion and lower similarity between videos with different ego motion. Since trajectories are used only as training supervision, retrieval remains single-vector similarity search that requires neither ego poses nor auxiliary perception outputs at query time.

To evaluate motion-centric driving-video retrieval, we construct and release two benchmarks[1] derived from publicly available driving datasets. The first reformulates nuReasoning (Huang et al., 2026) into a retrieval benchmark for both instance-level matching with reasoning-trace queries and category-level retrieval with lateral and longitudinal motion queries. The second uses AV2 (Wilson et al., 2021) ego-motion queries to evaluate zero-shot cross-domain transfer. Across different backbones, TraVEL consistently improves motion-centric retrieval over standard supervised fine-tuning, with gains of up to 7.5 and 10.3 mAP points on lateral and longitudinal motion retrieval, respectively, while maintaining comparable instance-level performance. These improvements also transfer to the AV2 benchmark without training or tuning on AV2.

Our contributions are summarized as follows:

- We introduce TraVEL, which uses ego trajectories as privileged supervision through complementary video-to-video and text-to-video GRPO rewards to learn motion-aware video embeddings while preserving language alignment.
- We construct and release motion-centric driving-video retrieval benchmarks derived from nuReasoning and the AV2 Sensor Dataset for in-domain and cross-domain evaluation.

[1] https://anonymous.4open.science/r/TraVEL_evaluation_benchmark-F8CF

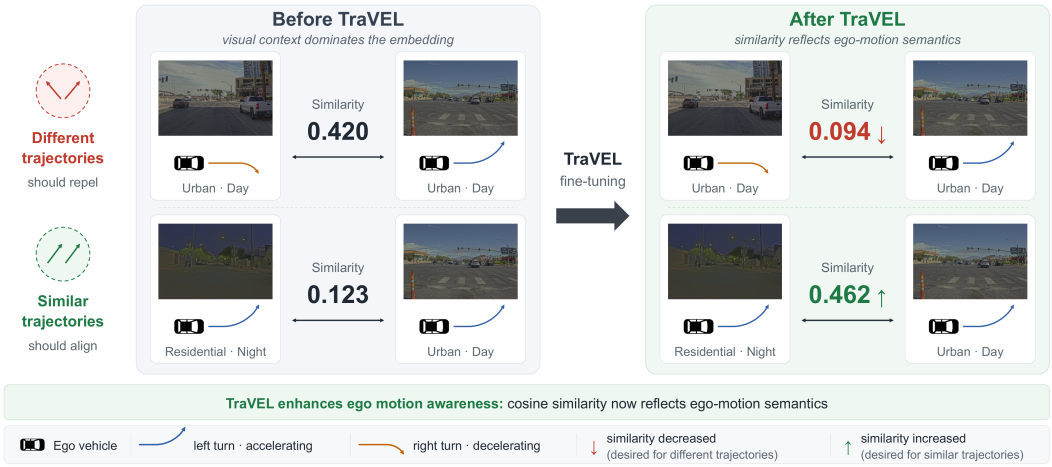

Figure 1: Effect of trajectory-aware fine-tuning on video embeddings. Before TraVEL (left), similarity is driven largely by visual context; after TraVEL (right), clips with similar ego motion become closer, while clips with different motion move farther apart despite similar appearances.

- Across multiple embedding-model families, TraVEL improves motion retrieval over supervised fine-tuning while maintaining comparable instance-level performance, and its improvements transfer to AV2 without training or tuning on it.

2 Related Work

Embedding-based video retrieval. Multimodal embedding models map text and visual inputs into a shared space, enabling direct similarity-based retrieval with natural-language queries. CLIP and SigLIP established this approach for image–text retrieval (Radford et al., 2021; Zhai et al., 2023). Video-specific work extends the interface through temporal modeling, auxiliary captions, and more fine-grained retrieval settings (Xu et al., 2021; Luo et al., 2022; Wu et al., 2023; Hur et al., 2025; Halbe et al., 2026). Recent general-purpose multimodal embedding models, including VLM2Vec, VLM2Vec-V2, Qwen3-VL-Embedding, Gemini Embedding 2, WeMM-Embedding, and Douyin Multimodal Embedding, further support video retrieval as one of many multimodal tasks (Jiang et al., 2025; Meng et al., 2025; Li et al., 2026; Shanbhogue et al., 2026; Zhou et al., 2026; Chen et al., 2026). However, these models target broad multimodal retrieval and are not specifically designed for driving-video retrieval. Prior studies also show that multimodal embedding models can rely on object or scene cues and struggle to distinguish changes in action and temporal order (Ma et al., 2024; Wang et al., 2023; Sohn et al., 2024). We therefore study how to adapt general-purpose multimodal embedding models for driving-video retrieval, with a particular focus on making their video representations sensitive to ego motion even when clips share similar visual context.

Driving-scenario retrieval and mining. Existing driving-scenario retrieval methods differ mainly in how they represent scenarios. Vision-language approaches retrieve driving data through visual and textual matching: Sohn et al. (2024) compare pretrained vision-language models for natural-language retrieval of driving images, and CARIM (Ki et al., 2025) describes driving videos with scene captions and retrieves them through object-level text matching. Program-based approaches instead execute natural-language queries over structured perception outputs: RefAV (Davidson et al., 2026) composes atomic functions over object tracks, HYU_OASIS (Park et al., 2026) extends this approach with additional atomic functions and visual-attribute modeling, and STRIVE-D (Yao et al., 2026b) combines a calibrated rule library with dense embedding retrieval and keyword matching. Other methods exploit motion signals directly: Sohn et al. (2025) learn geometric embeddings from road-user trajectories and search them through LLM-generated descriptions of trajectory clusters, while CITMR (Peng et al., 2026) encodes measured ego motion with a dedicated trajectory encoder alongside images and text. Both methods require trajectory data at retrieval time, whereas TraVEL uses ego trajectories only as training supervision and retrieves videos from visual input

alone. Closest to our setting, Cosmos-Embed1 (NVIDIA et al., 2025) trains a joint video–text embedder for physical AI and evaluates it on autonomous-driving scenario retrieval. TraVEL builds on this embedding-based direction and uses ego trajectories to impose motion-aware structure on the video embedding space.

3 PRELIMINARIES

Multimodal embedding-based retrieval. A multimodal embedding model f_θ maps text and video into a shared embedding space, and each training example (q_i, x_i) pairs a natural-language description with a video. At retrieval time, a query q scores each video x_j by $s_\theta(q, x_j) = \langle f_\theta(q), f_\theta(x_j) \rangle$, which is cosine similarity when the embeddings are unit-normalized, and videos are ranked by this score. Since corpus video embeddings can be precomputed, retrieval can use standard nearest-neighbor indexing methods.

Group relative policy optimization. Group Relative Policy Optimization (GRPO) (Guo et al., 2025) optimizes a policy using rewards normalized within a group of sampled candidates, avoiding a separately learned value function. For a group $\mathcal{G}_i$ of k candidates with rewards r_{ij}, GRPO computes the group-relative advantage

$$A_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i + \epsilon}, \tag{1}$$

where μ_i and σ_i are the mean and standard deviation of rewards within group $\mathcal{G}_i$, and ϵ is a small constant for numerical stability. Let π_θ, π_{old}, and π_{ref} denote the current, rollout, and reference policies, respectively. The clipped GRPO objective with reference-policy regularization is

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_i \left[\frac{1}{k} \sum_{j \in \mathcal{G}_i} \min(\rho_{ij} A_{ij}, \text{clip}(\rho_{ij}, 1-\varepsilon, 1+\varepsilon) A_{ij}) \right] + \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{2}$$

where $\mathbb{E}_i$ denotes the expectation over queries, $\rho_{ij} = \pi_\theta(j)/\pi_{\text{old}}(j)$ is the importance ratio, ε is the clipping threshold, and β controls an optional KL penalty toward the reference policy. Following Retrieval-GRPO (Liu et al., 2026), we treat the embedding model as a retrieval policy network. A softmax over its query–candidate similarity scores defines a probability distribution over candidate videos, allowing GRPO to update the model using group-relative rewards.

4 TRAVEL: TRAJECTORY-GUIDED VIDEO EMBEDDING LEARNING

We first analyze why general-purpose multimodal embedding models perform poorly on driving-video retrieval (Section 4.1). The analysis motivates TraVEL, a two-stage framework in which supervised contrastive fine-tuning first bridges the domain gap (Section 4.2), after which trajectory-aware GRPO uses physical motion to organize the video embedding space (Section 4.3).

4.1 DIAGNOSING DRIVING-VIDEO EMBEDDING GAPS

We first evaluate pretrained multimodal embedding models on driving-video retrieval without adaptation. Table 1 shows weak retrieval performance across backbones in this pretrained setting. To investigate the cause of the poor performance, we visualize the caption and video embeddings of Qwen3-VL-Embedding-2B (Li et al., 2026) with t-SNE (Van der Maaten & Hinton, 2008) in Fig. 2. In the pretrained embedding space, caption and video embeddings remain far apart, consistent with the weak retrieval performance and indicating poor cross-modal alignment on driving data. Moreover, although caption embeddings show recognizable grouping by longitudinal-motion category, video embeddings from different categories are intermixed, indicating little motion-specific

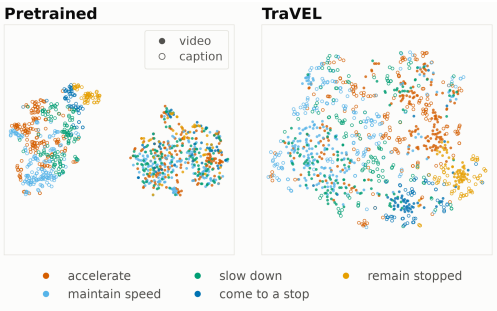

Figure 2: t-SNE visualization of Qwen3-VL-Embedding-2B embeddings.

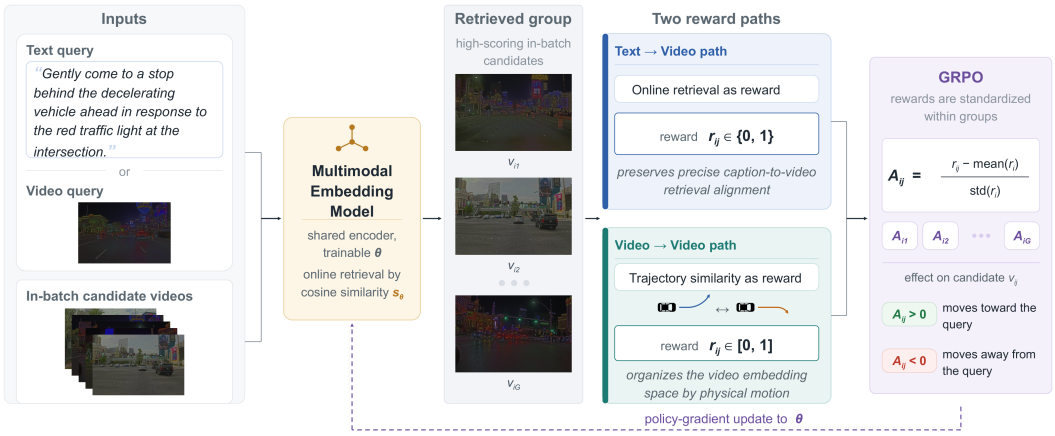

Figure 3: Overview of trajectory-aware GRPO fine-tuning. For each text or video query, the embedding model retrieves a group of high-scoring candidates from the in-batch videos. The text-to-video group contains the paired video and hard negatives with binary instance rewards, while the video-to-video group receives graded ego-trajectory similarity rewards.

structure in the video embedding space. The former motivates supervised domain adaptation; the latter motivates the trajectory-aware fine-tuning stage.

4.2 Driving-Domain Supervised Fine-Tuning

We first use supervised contrastive fine-tuning to align driving-video clips with their reasoning-trace captions. For each mini-batch $\mathcal{B}$ of paired examples, the paired video is the positive match for a caption, while other videos in the batch serve as negatives. The text-to-video InfoNCE objective (van den Oord et al., 2018) is

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(s_\theta(q_i, x_i)/\tau_{\text{SFT}})}{\sum_{j \in \mathcal{B}} \exp(s_\theta(q_i, x_j)/\tau_{\text{SFT}})}, \tag{3}$$

where τ_{SFT} is the temperature. This objective improves text-to-video retrieval, but its supervision is based on paired instances rather than relationships between video motions. Two clips with similar ego trajectories may be treated as negatives when they occur in the same batch, and the loss does not explicitly reward their motion similarity. This motivates the trajectory-derived reward in the next stage.

4.3 Trajectory-Aware GRPO Fine-Tuning

To impose motion structure on the video embedding space, we propose trajectory-aware GRPO, using ego trajectories as privileged supervision during training. Its video-to-video reward encourages embeddings of clips with similar trajectories to be closer together, making the learned space more motion-aware, while a text-to-video reward preserves language alignment. Fig. 3 illustrates this training stage.

Text-to-video alignment reward. To preserve the paired text-to-video retrieval learned during SFT, we use a binary instance reward. For caption q_i and candidate video x_j, we set $r_{ij}^{\text{t2v}} = \mathbf{1}[j = i]$, rewarding retrieval of the paired video.

Video-to-video motion reward. To impose motion-similarity structure on the video embedding space, we first define a similarity between the ego trajectories of two clips, which serves as the video-to-video reward. We represent each trajectory with two complementary components: ego-frame lateral displacement, which captures the direction and extent of sideways motion, and speed, which captures the vehicle's rate of travel. For a pose sequence of T steps, component c of clip i is denoted by

$$\mathbf{u}_i^c = \big[u_i^c(1), \ldots, u_i^c(T)\big], \qquad c \in \mathcal{C}. \tag{4}$$

However, equal differences in raw values do not correspond to equal differences in driving behavior. For example, a speed difference of 2 m/s distinguishes a stopped vehicle from a moving one, but reflects only a minor variation between two vehicles cruising at 10 and 12 m/s. To make distances more sensitive to such low-magnitude changes, we apply a signed logarithmic transform to each component, which preserves its sign while compressing large values:

$$g_c(u) = \operatorname{sign}(u) \log(1 + |u|) \,. \tag{5}$$

Given two transformed trajectories, a naive way to compare them is to measure their pointwise distance, which is sensitive to temporal offsets: two clips may both contain a right turn, but if one turn starts slightly later, their pointwise difference can be large even though the two curves have similar shapes. We therefore use dynamic time warping (DTW) (Sakoe & Chiba, 1978), which finds a low-cost temporal alignment between the two sequences. For clips i and j, the DTW distance for component c is

$$d_c(i,j) = \frac{1}{T} \min_{\gamma \in \Gamma} \sum_{(t,t') \in \gamma} \left| g_c(u_i^c(t)) - g_c(u_j^c(t')) \right|, \tag{6}$$

where Γ denotes the set of monotone alignment paths. We then normalize each component distance and combine them with non-negative weights w_c to obtain the video-to-video reward:

$$r_{ij}^{\mathrm{v2v}} = R \left(1 - \frac{\sum_{c \in \mathcal{C}} w_c \widehat{d}_c(i,j)}{\sum_{c \in \mathcal{C}} w_c} \right). \tag{7}$$

Here $\widehat{d}_c(i,j)$ denotes the normalized component distance. The resulting reward lies in $[0, R]$, with larger values assigned to clips exhibiting more similar ego motion.

On-policy retrieval groups. Both rewards are optimized with GRPO, which compares candidates within a group for each query. Since the two rewards use different queries, captions for text-to-video and videos for video-to-video, we construct a separate group for each. The candidates in a group determine which ranking errors the model learns to correct, so we build each group from the candidates that the current model scores highest, as these most directly affect its retrieval ranking. At each optimization step, given a candidate pool $\mathcal{P}$ of in-batch videos, we define the following groups for each paired caption–video example (q_i, x_i):

$$\mathcal{G}_i^{\mathrm{t2v}} = \{i\} \cup \mathrm{Top}_{k-1} \left\{ s_\theta(q_i, x_j) : j \in \mathcal{P}, j \neq i \right\}, \tag{8}$$

$$\mathcal{G}_i^{\mathrm{v2v}} = \mathrm{Top}_k \left\{ s_\theta(x_i, x_j) : j \in \mathcal{P}, j \neq i \right\}, \tag{9}$$

where Top_k returns the indices of the k highest-scoring candidates. The text-to-video group always includes the paired video; otherwise, every candidate could receive zero binary reward, leaving no learning signal. The video-to-video group excludes the query video itself and needs no designated positive, since its graded rewards already differ across candidates.

Group-relative optimization. To optimize both groups with GRPO, we treat the embedding model as a retrieval policy over each group (Section 3). For $m \in \{\mathrm{t2v}, \mathrm{v2v}\}$, let a_i^m denote the query of group $\mathcal{G}_i^m$, that is, $a_i^{\mathrm{t2v}} = q_i$ and $a_i^{\mathrm{v2v}} = x_i$. The retrieval policy is a softmax over the retrieval scores within the group:

$$\pi_\theta^m(j \mid a_i^m) = \frac{\exp(s_\theta(a_i^m, x_j)/\tau)}{\sum_{j' \in \mathcal{G}_i^m} \exp(s_\theta(a_i^m, x_{j'})/\tau)}, \qquad j \in \mathcal{G}_i^m. \tag{10}$$

Within each group, rewards are standardized into advantages using Eq. (1), and the policy is updated with the GRPO objective in Eq. (2). This yields a text-to-video objective $\mathcal{L}^{\mathrm{t2v}}$ and a video-to-video objective $\mathcal{L}^{\mathrm{v2v}}$, which we sum to obtain the final objective:

$$\mathcal{L}_{\mathrm{TraVEL}}(\theta) = \mathcal{L}^{\mathrm{t2v}}(\theta) + \mathcal{L}^{\mathrm{v2v}}(\theta). \tag{11}$$

The first term preserves cross-modal alignment, while the second organizes the video embedding space by ego motion.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We construct two benchmarks to evaluate motion-centric driving-video retrieval within the training domain and on an unseen dataset. **(i) nuReasoning retrieval:** We use an initially released subset of nuReasoning (Huang et al., 2026) containing 2,577 source clips with human-verified reasoning annotations and synchronized ego trajectories. We split the source clips into 2,015 for training and 562 for testing before extracting video windows. Around annotated keyframes, we extract 8-second windows with 3 seconds of history and 5 seconds of future context. After discarding incomplete windows, this yields 5,360 training and 1,500 test videos. Each extracted window inherits the lateral and longitudinal motion labels provided by nuReasoning at its annotated keyframe. Each motion label also serves as a text query, such as "Turn right," and videos sharing that label are relevant to it; Appendix A.1 describes which motion queries enter the evaluation. The reasoning annotations provide instance-level natural-language queries, while the trajectories serve as training-only motion supervision. **(ii) AV2 ego-motion retrieval:** We build an evaluation-only benchmark from the AV2 Sensor Dataset (Wilson et al., 2021). Adapting the HYU_OASIS scenario-mining codebase (Park et al., 2026), we define ego-motion events using AV2's map-aligned ego poses and lane geometry, then form natural-language queries from these events. Atomic queries describe a single event, such as a left turn, while unordered, simultaneous, and temporal queries combine two events that both occur, overlap in time, or occur in a specified order, respectively. We use no AV2 data for training or tuning so that this benchmark's results test whether motion-aware retrieval transfers to a new visual domain and a different query format. Construction details are provided in Appendix A.

Evaluation metrics. We evaluate both instance-level and motion-oriented retrieval. For instance-level retrieval on nuReasoning, each reasoning-trace query has one paired video. We report recall at ranks 1, 5, and 10 (R@1/5/10), along with the median and mean rank (MdR and MnR) of the paired video. For motion-oriented retrieval, a query can match multiple videos. We compute average precision (AP) over the retrieval pool for each query and report mean AP (mAP), averaging queries equally within each lateral or longitudinal motion family on nuReasoning and within each query type in the AV2 evaluation. Higher recall and mAP, and lower ranks, indicate better retrieval.

Backbones and training stages. We evaluate TraVEL with InternVideo2-Stage2 (Wang et al., 2024), Cosmos-Embed1-224p (NVIDIA et al., 2025), and the 2B and 8B variants of Qwen3-VL-Embedding (Li et al., 2026). The first two use separate video and text encoders, while Qwen3-VL-Embedding uses a shared transformer. All models use 8-second front-view nuReasoning clips resized to 224×224; model-specific frame sampling is detailed in Appendix B. Cross-domain transfer on AV2 is evaluated only with Qwen3-VL-Embedding, which supports a larger input budget for the longer AV2 videos. For each backbone, we compare its released *Pretrained* checkpoint, *SFT* after supervised contrastive fine-tuning on nuReasoning using Eq. (3), and *TraVEL* after trajectory-aware GRPO starting from SFT.

5.2 RESULTS

In-distribution evaluation. As shown in Table 1, all pretrained backbones perform poorly on nuReasoning without domain-specific fine-tuning. Their low recall and high retrieval ranks reveal a substantial domain gap between general-purpose multimodal representations and driving data. Supervised fine-tuning largely closes this gap, producing clear improvements in instance-level retrieval across all backbones, which shows that paired reasoning annotations provide effective supervision for adapting general embedding models to the driving domain. However, strong instance-level retrieval does not imply that the embedding captures ego motion. When a single motion description such as "Turn left" is used as the query, SFT models remain limited, reaching only 28.2 lateral mAP with Qwen3-VL-Embedding-2B. TraVEL improves both motion metrics over SFT on every backbone while maintaining comparable instance-level retrieval. The gains are largest for Qwen3-VL-Embedding, where lateral and longitudinal mAP increase by 7.5 and 10.3 points at 2B and by 1.8 and 7.9 points at 8B, whereas the improvements for InternVideo2 and Cosmos-Embed1 are at most 2.4 points. We hypothesize that the shared transformer of Qwen3-VL-Embedding is easier to adapt

through fine-tuning than the separate video and text encoders of the other two backbones. Qualitative comparisons in Appendix C.3 further illustrate how supervised adaptation and trajectory-aware training progressively improve motion-sensitive retrieval.

Table 1: Retrieval results on the nuReasoning test set. *LAT* and *LON* report mAP over lateral and longitudinal motion queries, and the remaining columns report instance-level retrieval with reasoning-trace queries. Recall and mAP are in percent, and MdR and MnR are rank positions.

			Motion mAP		Instance retrieval				
Backbone	#Params	Stage	LAT ↑	LON ↑	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
Chance-level reference									
Uniform random ranking	–	–	14.7	20.1	0.1	0.3	0.7	750	750.5
InternVideo2-Stage2	6B	Pretrained	14.6	22.3	1.1	4.9	7.7	299	420.2
		SFT	19.2	32.9	**8.9**	26.7	39.5	**18**	69.2
		TraVEL	**19.7**	**33.9**	**8.9**	**27.5**	**40.2**	**18**	**66.2**
Cosmos-Embed1-224p	1.2B	Pretrained	28.4	24.2	2.7	10.3	16.2	94	205.9
		SFT	34.9	35.2	9.4	**27.6**	**40.1**	**17**	**58.5**
		TraVEL	**35.2**	**37.6**	**9.6**	27.1	38.6	19	62.0
Qwen3-VL-Embedding	2B	Pretrained	17.5	24.6	2.1	5.1	7.3	232	349.7
		SFT	28.2	45.5	**8.8**	25.3	36.5	22	**81.1**
		TraVEL	**35.7**	**55.8**	**8.8**	**25.6**	**38.3**	**20**	**81.1**
Qwen3-VL-Embedding	8B	Pretrained	18.6	26.3	2.1	5.9	8.7	232	329.4
		SFT	35.5	49.3	**13.1**	31.5	43.0	15	**63.4**
		TraVEL	**37.3**	**57.2**	12.8	**32.1**	**43.7**	**14**	64.5

Cross-domain evaluation. The AV2 benchmark differs from nuReasoning in both scenes and sensors. Besides containing unseen driving environments, AV2 uses a different front-view camera, whose portrait-oriented frames include part of the ego vehicle's hood, and its native resolution and clip duration also differ from those used during training. Despite these differences, TraVEL improves over its SFT checkpoint for every query type at both model scales (Fig. 4), including compositional queries that combine two events; for example, atomic mAP increases by 7.5 and 6.5 points at 2B and 8B, respectively. Since no AV2 data is used for training or tuning, this transfer suggests that trajectory supervision organizes the embedding space by physical motion rather than by dataset-specific appearance or caption patterns. Temporal queries remain the most challenging: even with TraVEL, their mAP is only about 5 points above random ranking (12.5 and 12.4 versus 7.5). As our objective does not explicitly model temporal relations between events, stronger temporal event understanding remains a direction for future work.

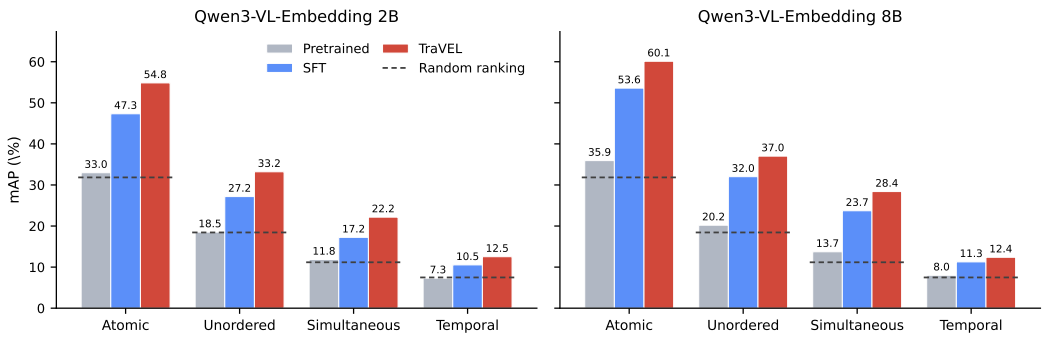

Figure 4: Zero-shot transfer to AV2 ego-motion retrieval. Bars report mAP within each query type, and dashed lines show the corresponding random-ranking baseline.

Reward ablation. To examine the contribution of each reward, we remove one reward at a time from TraVEL (Table 2). Without the video-to-video reward, both motion metrics stay within 1 point

of SFT, while instance-level retrieval remains comparable to TraVEL. Removing the text-to-video reward instead preserves motion performance but degrades instance-level retrieval below even the SFT checkpoint, reducing R@10 by 7.0 points and worsening the mean rank by 22.4 positions relative to TraVEL. These results show that the video-to-video reward introduces motion structure into the embedding space, while the text-to-video reward preserves its alignment with language.

Table 2: Reward ablation using Qwen3-VL-Embedding-2B.

	Motion mAP		Instance retrieval				
Reward	LAT ↑	LON ↑	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
SFT	28.2	45.5	8.8	25.3	36.5	22	81.1
TraVEL w/o t2v	**36.4**	55.7	7.1	21.9	31.3	27	103.5
TraVEL w/o v2v	28.7	46.5	**9.0**	**27.5**	38.1	21	**79.1**
TraVEL	35.7	**55.8**	8.8	25.6	**38.3**	**20**	81.1

Alternative uses of trajectory. Trajectory-aware GRPO is not the only way to exploit ego trajectories. We compare it with three simpler alternatives that use the same trajectories (Table 3). The most direct option serializes the trajectory as text and appends it to the video input, both during training and when encoding gallery videos at evaluation. Even with trajectories available at retrieval time, this produces no consistent motion-retrieval gain and degrades most instance-level metrics. We hypothesize that the model struggles to represent long numerical sequences through its standard text interface, as language-based time-series methods commonly rely on specialized numerical preprocessing, tokenization, quantization, or representation alignment (Gruver et al., 2023; Ansari et al., 2024; Kaur et al., 2025). A second option adds auxiliary trajectory prediction, training a linear head on the video embedding to predict the ego trajectory alongside the original SFT objective. It improves lateral mAP but not longitudinal mAP, and the motion gains quickly plateau even as the prediction loss continues to decrease. We hypothesize that this objective encourages the embedding to retain enough trajectory information for prediction, but does not organize the embedding space so that videos with similar motion lie close together. Finally, trajectory-similarity matching directly aligns video-embedding similarity with trajectory-derived similarity. It improves both motion metrics while maintaining instance-level performance, but TraVEL achieves larger gains, reaching 35.7 and 55.8 lateral and longitudinal mAP versus 32.2 and 52.0. These results show that trajectory-aware GRPO more effectively translates trajectory supervision into motion-sensitive video embeddings.

Table 3: Alternative uses of the ego trajectory with Qwen3-VL-Embedding-2B.

	Motion mAP		Instance retrieval				
Method	LAT ↑	LON ↑	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
SFT	28.2	45.5	8.8	25.3	36.5	22	81.1
Trajectory as Text Input	30.2	44.9	**9.5**	24.5	35.7	24	87.5
Trajectory Prediction	34.6	44.7	8.1	23.7	34.3	23	85.7
Trajectory Similarity Matching	32.2	52.0	8.8	25.3	37.4	21	**76.2**
TraVEL	**35.7**	**55.8**	8.8	**25.6**	**38.3**	**20**	81.1

6 Conclusion

We studied how to adapt general-purpose multimodal embedding models to motion-centric driving-video retrieval. Our results show that caption-based contrastive fine-tuning narrows the domain gap but remains limited on motion queries, and that ego trajectories, used only during training, can supply the missing motion supervision. With trajectory-aware GRPO, TraVEL improves motion retrieval over supervised fine-tuning on every tested backbone while maintaining comparable instance-level retrieval, and these improvements transfer to AV2 without training or tuning on it. These results suggest that physical motion signals can complement caption-based learning for driving-video retrieval. As an initial study of trajectory-guided embedding learning, we focus on clip-level ego motion. Temporal event understanding and interactions with surrounding agents are also central to driving scenarios, and we leave extending trajectory-guided learning to them as future work.

REPRODUCIBILITY STATEMENT

We release both benchmarks at https://anonymous.4open.science/r/TraVEL_evaluation_benchmark-F8CF, including the nuReasoning train/test split, the queries and their relevant videos for both benchmarks, evaluation code with the video preprocessing, and reference results for the pretrained backbones. The nuReasoning benchmark is derived from the publicly released nuReasoning dataset (Huang et al., 2026), and the AV2 benchmark from the public AV2 Sensor Dataset (Wilson et al., 2021) using the public HYU_OASIS codebase (Park et al., 2026). All backbones (Wang et al., 2024; NVIDIA et al., 2025; Li et al., 2026) are publicly available. Appendix A describes the benchmark construction, and Appendix B provides the training details, hyperparameters, and baseline configurations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation, experimental code, and code for dataset preprocessing and reformatting; survey related work; assist with translation; and rephrase and polish author-written manuscript text, including descriptions of the methodology and experimental results. We did not use generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for or assist with mathematical proofs; propose or refine hypotheses; design or provide substantive feedback on research methodology or experiments; support qualitative or thematic data analysis; or interpret results. All AI-assisted outputs were reviewed and revised by the authors. Retrieved references were verified against their original sources, and AI-assisted code was reviewed and tested before use. We take responsibility for the final content of this work, including text, code, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Haonan Chen, Chu Li, Zhicheng Wang, Yuanwei Liu, Yuanjiang Wang, Shaohua Jiang, and Zhicheng Dou. Douyin multimodal embedding model technical report. *arXiv preprint arXiv:2608.02148*, 2026. URL https://arxiv.org/abs/2608.02148.

Cainan Davidson, Deva Ramanan, and Neehar Peri. RefAV: Towards planning-centric scenario mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21537–21548, 2026.

Wenhao Ding, Yulong Cao, Ding Zhao, Chaowei Xiao, and Marco Pavone. RealGen: Retrieval augmented generation for controllable traffic scenarios. In *European Conference on Computer Vision*, 2024.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9710–9719, 2021.

Shuo Feng, Haojie Zhu, Haowei Sun, Xintao Yan, Linxuan He, Jingxuan Yang, Guangzhen Su, Boqi Li, Shu Li, Ling Wang, Shengyin Shen, and Henry X. Liu. Breaking through safety performance stagnation in autonomous vehicles with dense learning. *Nature Communications*, 17:3163, 2026. doi: 10.1038/s41467-026-69761-x.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G. Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3eb7ca52e8207697361b2c0fb3926511-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shaunak Halbe, Bhagyashree Puranik, Jayakrishnan Unnikrishnan, Kushan Thakkar, Vimal Bhat, and Toufiq Parag. VeRVE: Versatile retrieval for videos via unified embeddings. *arXiv preprint arXiv:2601.12193*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Zhiyu Huang, Johnson Liu, Rui Song, Zewei Zhou, Ruining Yang, Yun Zhang, Tianhui Cai, Hanyin Zhang, Mingxuan Gao, Valeria Xu, Jiali Chen, Yishan Shen, Yiluan Guo, Tony Qi, and Jiaqi Ma. nuReasoning: A reasoning-centric dataset and benchmark for long-tail autonomous driving. *arXiv preprint arXiv:2605.31572*, 2026.

Chan Hur, Jeong-hun Hong, Dong-hun Lee, Dabin Kang, Semin Myeong, Sang-hyo Park, and Hyeyoung Park. Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24077–24086, 2025.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025.

Rachneet Kaur, Zhen Zeng, Tucker Balch, and Manuela Veloso. LETS-C: Leveraging text embedding for time series classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32365–32399. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1557. URL https://aclanthology.org/2025.acl-long.1557/.

Minjoo Ki, Daejung Kim, Kisung Kim, Seon Joo Kim, and Jinhan Lee. CARIM: Caption-based autonomous driving scene retrieval via inclusive text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22036–22045, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Mingfu Liang, Jong-Chyi Su, Samuel Schulter, Sparsh Garg, Shiyu Zhao, Ying Wu, and Manmohan Chandraker. AIDE: An automatic data engine for object detection in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14695–14706, 2024.

Xingxian Liu, Dongshuai Li, Jiahui Wan, Tao Wen, Gui Ling, Yuliang Yan, Fuyu Lv, Dan Ou, Haihong Tang, and Bo Zheng. Retrieval-GRPO: A multi-objective reinforcement learning framework for dense retrieval in taobao search. *arXiv preprint arXiv:2511.13885*, 2026. URL https://arxiv.org/abs/2511.13885.

Yiren Lu, Xin Ye, Burhaneddin Yaman, Jingru Luo, Zhexiao Xiong, Liu Ren, and Yu Yin. Reconstruction matters: Learning geometry-aligned BEV representation through 3d gaussian splatting. *arXiv preprint arXiv:2603.19193*, 2026.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end-to-end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. doi: 10.1016/j.neucom.2022.07.028.

Wufei Ma, Kai Li, Zhongshi Jiang, Moustafa Meshry, Qihao Liu, Huiyu Wang, Christian Häne, and Alan Yuille. Rethinking video-text understanding: Retrieval from counterfactually augmented data. In *European Conference on Computer Vision*, pp. 254–269. Springer, 2024. doi: 10.1007/978-3-031-72624-8_15.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

NVIDIA, Francesco Ferroni, Prithvijit Chattopadhyay, Greg Heinrich, Mike Ranzinger, Roberto Amoroso, Alice Luo, Andrew Wang, and Ming-Yu Liu. Cosmos-Embed1: A joint video-text embedder for physical ai, 2025. URL https://research.nvidia.com/labs/cosmos-lab/cosmos-embed1/.

Jeongwoo Park, Yuseung Na, Seongjae Jeong, Minwon Lee, and Kichun Jo. HYU_OASIS: Argoverse 2 scenario mining challenge entry. GitHub repository, 2026. URL https://github.com/ailab-hanyang/HYU_OASIS. CVPR 2026 Workshop on Autonomous Driving.

Jay Patrikar, Apoorva Sharma, Sushant Veer, Boyi Li, Sebastian Scherer, and Marco Pavone. The case for negative data: From crash reports to counterfactuals for reasonable driving. *arXiv preprint arXiv:2509.18626*, 2025.

Kun Peng, Shanke Li, Fei Hui, and Xiyao Liu. CITMR: Contrastive image–text–motion retrieval for autonomous driving critical scenarios extraction. *Journal of Advanced Transportation*, 2026 (1):4985228, 2026. doi: 10.1155/atr/4985228.

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.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Madhuri Shanbhogue et al. Gemini Embedding 2: A native multimodal embedding model from gemini. *arXiv preprint arXiv:2605.27295*, 2026. URL https://arxiv.org/abs/2605.27295.

Tin Sohn, Maximilian Dillitzer, Lukas Ewecker, Tim Brühl, Robin Schwager, Lena Dalke, Philip Elspas, Frank Oechsle, and Eric Sax. Towards scenario retrieval of real driving data with large vision-language models. In *Proceedings of the 10th International Conference on Vehicle Technology and Intelligent Transport Systems*, pp. 496–505, 2024. doi: 10.5220/0012738500003702.

Tin Sohn, Maximilian Dillitzer, Tim Brühl, Robin Schwager, Tim Dieter Eberhardt, Michael Auerbach, and Eric Sax. Large language model-informed geometric trajectory embedding for driving scenario retrieval. In *Proceedings of the 11th International Conference on Vehicle Technology and Intelligent Transport Systems*, pp. 66–75, 2025. doi: 10.5220/0013276500003941.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024.

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/420492060687ca7448398c4c3fa10366-Abstract-Conference.html.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4Video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10704–10713, 2023.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6787–6800. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.544.

Jin Yao, Dhruva Dixith Kurra, Tom Lampo, Zezhou Cheng, Danhua Guo, and Burhan Yaman. VLGA: Vision-language-geometry-action models for autonomous driving. *arXiv preprint arXiv:2606.12396*, 2026a.

Manyi Yao, Sparsh Garg, Christian Shelton, Amit Roy-Chowdhury, and Abhishek Aich. Driving video retrieval for complex queries with structured grounding. *arXiv preprint arXiv:2606.09109*, 2026b.

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*, pp. 11975–11986, 2023.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. WeMM-Embedding: WeChat multi-modal embedding technical report. *arXiv preprint arXiv:2608.24053*, 2026. URL https://arxiv.org/abs/2608.24053.

Sicheng Zuo, Zixun Xie, Wenzhao Zheng, Shaoqing Xu, Fang Li, Shengyin Jiang, Long Chen, Zhi-Xin Yang, and Jiwen Lu. DVGT: Driving visual geometry transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14658–14668, 2026.

A BENCHMARK CONSTRUCTION

A.1 NUREASONING RETRIEVAL

The benchmark uses the split and video windows described in Section 5.1 and defines two kinds of queries over the 1,500-video test pool. Each reasoning-trace annotation serves as the query for its paired video, the only relevant video for that query. Motion queries come from the lateral and longitudinal motion labels that nuReasoning provides at each annotated keyframe: each label is used verbatim as the text of a motion query, such as "Turn right," and the relevant videos are those whose keyframe carries that label.

Rare categories are excluded from the motion mAP. The pool contains seven lateral and eight longitudinal categories, and a motion query enters the mAP only if at least 50 videos in the pool are relevant to it, since AP over a handful of relevant videos is unstable. This removes three longitudinal categories, "Quickly come to a stop" (10 relevant videos), "Slow down quickly" (8), and "Quickly accelerate" (3), and retains all seven lateral categories. Lateral mAP therefore averages seven motion queries and longitudinal mAP averages five, as listed in Table 4.

A.2 AV2 EGO-MOTION RETRIEVAL

We use all 850 annotated training and validation logs from the AV2 Sensor Dataset, since the released test split lacks the ground-truth ego trajectories needed to assign motion-event labels. Each video in the pool consists of the complete $\sim$15.5-second `ring_front_center` video, encoded using 32 frames sampled at 2 Hz.

We build the benchmark by adapting the HYU_OASIS scenario-mining codebase (Park et al., 2026), which applies deterministic motion predicates to AV2's map-aligned ego poses and lane geometry. These predicates produce timestamp masks, which we compile into motion events. From the supported motions, we select nine ego-motion primitives: turning left or right, changing lanes left or right, continuing in lane, accelerating, braking, maintaining speed, and temporarily stopping. An event must persist for at least one second to suppress transient detections.

We compose these events into four query types: *atomic* (A); *unordered* ($\{A, B\}$), where both events occur regardless of order; *simultaneous* ($A \cap B$), where their intervals overlap by at least 0.5 seconds; and *temporal* ($A \rightarrow B$), where the events occur in the specified order with no more than a 2-second gap. Representative queries are "A video in which the ego vehicle turns left" (atomic), "A video in which the ego vehicle both turns left and accelerates" (unordered), "A video in which the ego vehicle turns left and accelerates at the same time" (simultaneous), and "A video in which the ego vehicle accelerates and then brakes" (temporal). The same executable definitions determine the relevant videos for each query, and the query text is generated from templates such as those above.

Continuing in lane and maintaining speed hold in over 97% of the videos, so a ranking cannot distinguish their positives; we therefore exclude any query satisfied by more than 90% of the pool, which removes these two atomic queries and their pairwise combinations. Both events remain in other composite queries, where the second event restricts the relevant set. The resulting benchmark contains 125 queries: 7 atomic, 35 unordered, 18 simultaneous, and 65 temporal.

B ADDITIONAL IMPLEMENTATION DETAILS

Model and input setup. All experiments use bfloat16 precision with gradient checkpointing. Qwen3-VL-Embedding-2B, Cosmos-Embed1-224p, and InternVideo2-Stage2 are trained on four NVIDIA H100 GPUs, while Qwen3-VL-Embedding-8B uses eight. For InternVideo2-Stage2 and Cosmos-Embed1-224p, we follow the frame-sampling configurations of their released models, using 4 and 8 frames per 8-second clip, respectively. For Qwen3-VL-Embedding, we sample 16 frames at 2 Hz. In the AV2 evaluation, Qwen3-VL-Embedding uses 32 frames sampled over each $\sim$15.5-second video, with frames resized from their native 1550×2048 resolution to 448×608 (width $\times$ height). We freeze the visual feature extractor of each backbone and train LoRA adapters together with model-specific pooling and projection parameters, where applicable. Both stages use LoRA (Hu et al., 2022) with rank 16, scaling factor 64, and dropout 0.1.

Supervised fine-tuning. We train each backbone for 100 steps using the text-to-video InfoNCE objective with temperature 0.02 and global in-batch negatives. Optimization uses AdamW with a constant learning rate of 10^{-4} after 10 warm-up steps, zero weight decay, and gradient clipping at 1.0. The global batch contains 1,024 paired examples for every backbone. Each reasoning trace is paired with its video, and no hard-negative mining is used.

Trajectory-aware fine-tuning. Each 8-second clip is paired with a 10 Hz ego-trajectory sequence. We normalize each component distance by its 95th-percentile corpus value and clip the normalized distance at one. We set the component weights to $w_{\text{lat}} = 1$ for lateral displacement and $w_{\text{speed}} = 2$ for speed, and $R = 10$. Stage 2 initializes from the SFT checkpoint with its LoRA weights merged into the backbone and trains a new LoRA adapter, with the SFT checkpoint serving as the reference policy. Trajectory-aware GRPO runs for 30 steps, about four epochs over the 5,360 training videos, using AdamW with a constant learning rate of 10^{-5}, no warm-up, and gradient clipping at 1.0. Each step uses a global candidate pool of 768 examples and constructs text-to-video and video-to-video groups with $k = 24$ candidates. We set $\tau = 0.05$ and $\varepsilon = 0.2$, with $\beta = 0.2$ for text-to-video optimization and $\beta = 0$ for video-to-video optimization.

Alternative uses of trajectory. For the trajectory-as-text baseline in Table 3, we serialize the same 8-second, 10 Hz trajectory used by TraVEL, comprising 80 lateral-displacement values and 80 speed values in their raw physical units. Values are formatted to three decimal places using the following serialization format:

```
The following describes the ego motion over the same
8.0-second driving clip.  Each channel contains 80 samples
in chronological order at 10 Hz.
Lateral displacement in meters in the first-frame ego
coordinate system, positive to the left:  [d_1, d_2, ...,
d_80]
Speed in meters per second:  [v_1, v_2, ..., v_80]
```

We append this text after the video input and fine-tune the SFT checkpoint for 30 steps, matching the training budget of TraVEL, using the original InfoNCE objective with temperature 0.02, a constant learning rate of 10^{-4} without warm-up, and a global batch of 768. At evaluation, the same serialization is appended when encoding each video in the pool; text queries remain unchanged.

The remaining alternatives combine the original contrastive objective with an auxiliary trajectory loss:

$$\mathcal{L}_{\text{alt}} = \mathcal{L}_{\text{SFT}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}. \tag{12}$$

We set $\lambda_{\text{aux}} = 10$. Trajectory prediction uses a linear head to predict the z-normalized lateral and speed components with a Huber loss. Trajectory-similarity matching regresses the video-embedding similarity onto the normalized motion reward over all video pairs in the global batch:

$$\mathcal{L}_{\text{aux}} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \left(\frac{1 + s_\theta(x_i, x_j)}{2} - \frac{r_{ij}^{\text{v2v}}}{R} \right)^2, \tag{13}$$

where Ω contains all pairs $i < j$ in the batch, and the cosine similarity is mapped to $[0, 1]$ to match the range of the target. Both alternatives use the same training budget as TraVEL.

C Additional Results

C.1 Per-Class Motion Retrieval

To identify which motion categories benefit from trajectory supervision, Table 4 reports AP for each motion query; the queries entering each mAP are defined in Appendix A.1. TraVEL improves every longitudinal category at both model scales, with the largest gains on stopping-related queries. For lateral motion, the largest gains occur on turns, while lane changes and subtle within-lane shifts remain challenging at every stage, with AP below 20.

Table 4: Per-class motion retrieval on nuReasoning with Qwen3-VL-Embedding.

Motion query	2B			8B		
	Pretrained	SFT	TraVEL	Pretrained	SFT	TraVEL
Longitudinal motion						
Gently accelerate	34.6	59.1	**60.2**	30.4	53.1	**58.6**
Maintain speed	36.9	38.6	**45.3**	29.8	35.2	**44.5**
Slow down gently	26.2	32.4	**33.0**	33.9	37.1	**38.4**
Gently come to a stop	14.6	27.8	**50.4**	16.7	43.9	**54.5**
Remain stopped	10.8	69.6	**90.3**	20.7	77.4	**90.3**
mAP	24.6	45.5	**55.8**	26.3	49.3	**57.2**
Lateral motion						
No lateral action	65.5	72.6	**77.5**	62.0	69.1	**73.1**
Right lane change	9.3	9.3	**12.0**	11.3	12.5	**17.3**
Turn right	13.2	47.7	**65.4**	20.3	**73.6**	70.8
Left lane change	5.8	10.6	**10.9**	7.7	17.0	**17.2**
Slightly move left in the lane	9.0	**12.2**	10.3	5.1	**9.9**	9.6
Turn left	14.6	37.7	**65.2**	15.3	61.0	**64.9**
Slightly move right in the lane	5.4	7.3	**8.6**	**8.3**	5.1	8.0
mAP	17.5	28.2	**35.7**	18.6	35.5	**37.3**

C.2 Cross-Domain Transfer by Query Type

To examine how transfer varies with query type, Table 5 reports the numerical results corresponding to Fig. 4. Each query is ranked against all 850 AV2 videos. TraVEL improves over SFT for every query type at both model scales. Temporal queries remain close to random ranking for all models, indicating that improved motion sensitivity does not by itself resolve temporal event understanding.

Table 5: Zero-shot AV2 retrieval by query type.

Model	Atomic	Unordered	Simultaneous	Temporal
Random ranking	31.8	18.4	11.2	7.5
Qwen3-VL-Embedding 2B				
Pretrained	33.0	18.5	11.8	7.3
SFT	47.3	27.2	17.2	10.5
TraVEL	**54.8**	**33.2**	**22.2**	**12.5**
Qwen3-VL-Embedding 8B				
Pretrained	35.9	20.2	13.7	8.0
SFT	53.6	32.0	23.7	11.3
TraVEL	**60.1**	**37.0**	**28.4**	**12.4**

C.3 Qualitative Retrieval

Figures 5–7 compare the top-5 results across the three training stages for a query describing a gentle acceleration through a left turn while monitoring a pedestrian. The pretrained model retrieves visually relevant intersections but fails to recover either the paired video or a video matching both requested motions in its top five. SFT improves cross-modal alignment: the paired video rises from rank 92 to rank 9, and a motion-matched video appears at rank 2. TraVEL further ranks the paired video first, and three of its top five videos, including the paired one, match both the left turn and gentle acceleration. This progression illustrates how supervised adaptation improves driving-domain retrieval, while trajectory-aware training further increases sensitivity to the requested ego motion.



Figure 5: Top-5 nuReasoning retrieval results with pretrained Qwen3-VL-Embedding-2B.

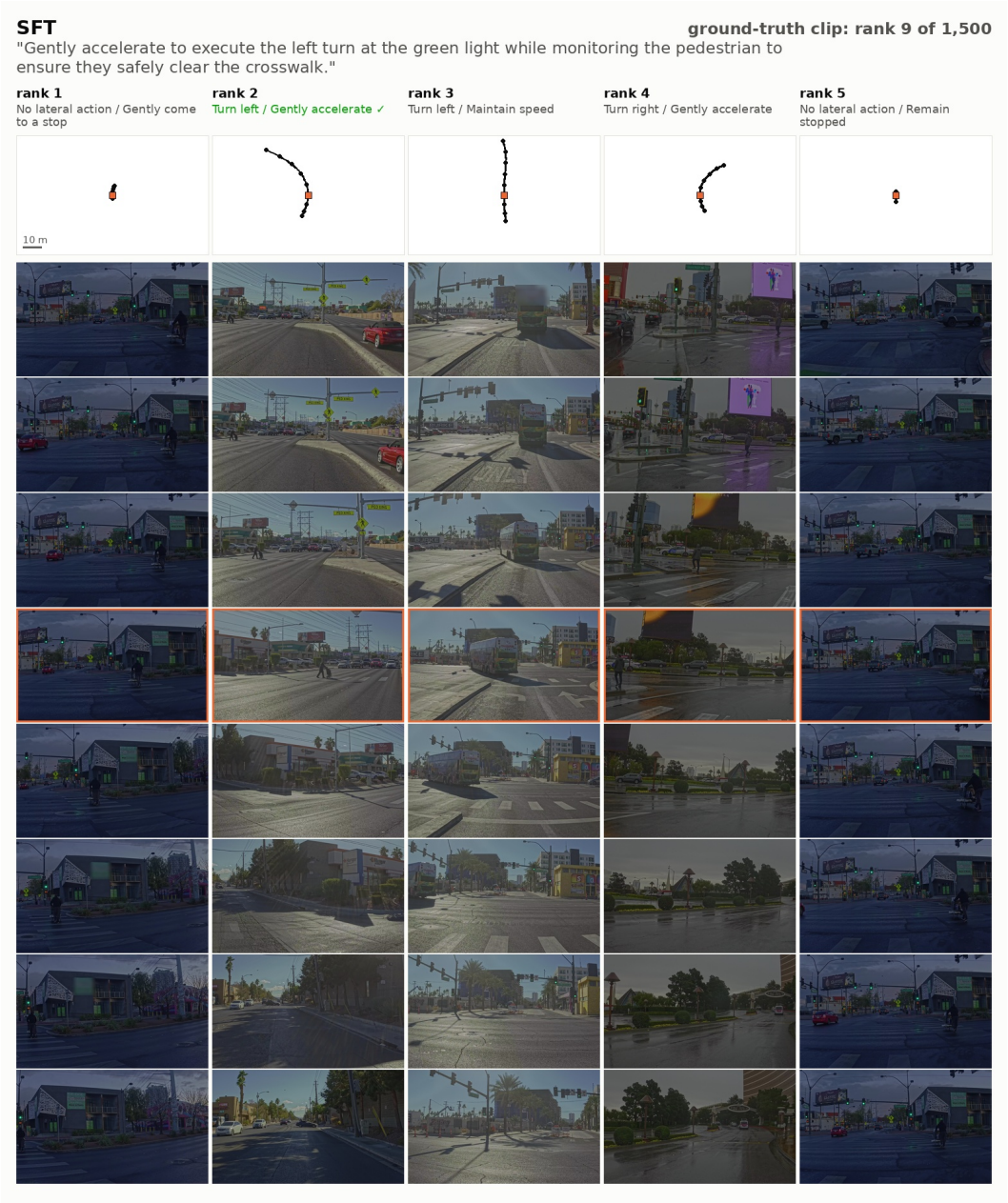

Figure 6: Top-5 nuReasoning retrieval results with Qwen3-VL-Embedding-2B after supervised fine-tuning.



Figure 7: Top-5 nuReasoning retrieval results with Qwen3-VL-Embedding-2B after TraVEL fine-tuning.