SYNTHETIC ACCIDENTS AS PROBES FOR OPEN-VOCABULARY EGOCENTRIC DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained understanding of egocentric traffic accidents is essential for safe autonomous driving and requires large-scale, diverse training data. Synthetic supervision offers a scalable way to expand training coverage beyond what can be readily obtained from real accidents. To investigate whether such supervision can support generalizable accident understanding, we first introduce **Ego-view Crashes (EgoCrash)**, a large-scale synthetic video-text benchmark comprising 20,000 egocentric accident videos across 56 fine-grained traffic-event categories, with rich multimodal annotations covering accident descriptions, question-answer pairs, motion cues, event categories, and entity trajectories.

We further formulate open-vocabulary video-text retrieval for traffic accident understanding, where models retrieve accident videos and free-form descriptions bidirectionally and generalize from seen events to category-disjoint unseen events and real-world accident videos. To address this task, we propose **maneuver-aware retrieval-augmented video-language matching (MRA-VLM)**. It learns motion-sensitive video representations by combining RGB appearance and optical-flow cues, and further enhances inference with prototype-based retrieval from seen events and uncertainty-guided score fusion, improving fine-grained matching and unseen-event generalization. In extensive experiments on EgoCrash, MRA-VLM consistently outperforms strong baselines in both retrieval directions and across both seen and unseen category splits, and models trained exclusively on synthetic data also transfer to real accident videos, demonstrating the potential of synthetic video-text supervision for generalizable accident understanding.

1 INTRODUCTION

Safety-critical autonomous driving relies on the ability to perceive and reason about hazardous events from an egocentric viewpoint (Chan et al., 2016; Fang et al., 2024b). Among these events, traffic accidents pose a particularly challenging understanding problem, requiring models to identify the involved road participants, interpret their interactions, and reason about how an incident unfolds over time. Advancing such fine-grained understanding requires large-scale and diverse accident data. However, existing egocentric accident benchmarks largely rely on naturally occurring events, whose rarity and uncontrolled nature make systematic collection difficult and constrain the coverage of accident types, interactions, environments, and event dynamics (Caesar et al., 2020; Cordts et al., 2016; Malla et al., 2023; Sachdeva et al., 2024; Kang et al., 2022; Chan et al., 2016; Liu et al., 2023; You & Han, 2020). Real accident recordings can further introduce privacy concerns when individuals, vehicles, or locations are identifiable.

Synthetic generation offers a promising way of expanding accident coverage through scalable and controllable scenario construction. However, simply increasing the number of generated videos is not sufficient. Effective supervision depends on whether each video correctly depicts the actions and interactions that define the accident and whether its paired description reflects those events precisely. This requirement is especially important in egocentric videos, where accident types can be distinguished by only a short sequence of actions, such as a vehicle entering the ego lane, turning across the ego vehicle, or losing control. Existing synthetic accident resources provide only limited video-language supervision or remain too small to support a systematic study of transfer to real-world driving scenarios (Kim et al., 2019; Aliakbarian et al., 2018; Li et al., 2025). It therefore

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

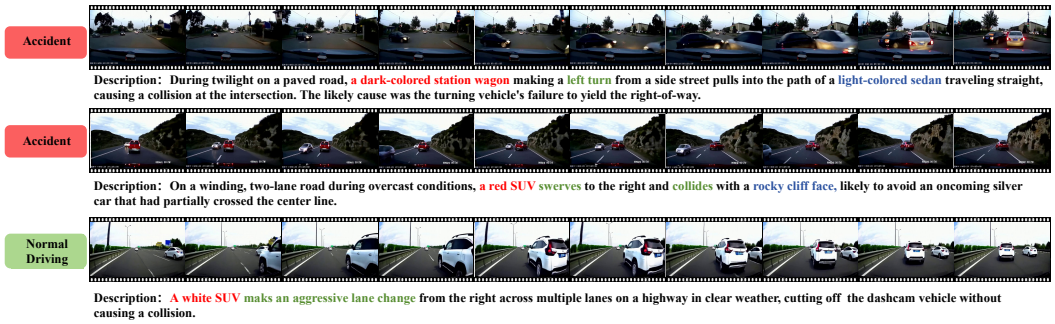


Figure 1: Traffic accidents are rare yet highly diverse, making it difficult to collect sufficient real-world training data for robust accident understanding. EgoCrash is the first large-scale synthetic egocentric video-text benchmark designed to bridge this gap by generating realistic accident videos paired with fine-grained textual accident descriptions. Two accident scenarios (top) and one normal driving scene (bottom) illustrate both the scarcity and diversity of traffic accidents. Red, green, and blue text highlight key vehicles, driving maneuvers, and road or environmental context, respectively.

remains unclear whether large-scale synthetic video-text data can support fine-grained and transferable understanding of real traffic accidents.

To investigate this question, we introduce **EgoCrash**, a synthetic benchmark containing 20,000 egocentric video-text pairs across 56 traffic-event categories, with rich multimodal annotations. Every video is accompanied by complementary supervision, including accident descriptions, coarse accident-occurrence and fine-grained event category labels, question-answer pairs (QA), optical flow, and entity trajectories. EgoCrash is constructed through a scalable generation and validation pipeline that combines language-guided scenario diversification, text-to-video synthesis, multimodal model-based annotation, and human quality control. Together, these annotations support multiple levels of accident understanding while providing controlled coverage of accident types and driving conditions without requiring the collection of identifiable real-world accident recordings. Figure 1 presents representative accident and non-accident examples from EgoCrash paired with textual descriptions.

We then formulate open-vocabulary video-text retrieval for traffic accident understanding, requiring models to associate accident videos with natural-language descriptions beyond the event categories observed during training. For this task, we propose **MRA-VLM**, which combines motion-aware representation learning with retrieval augmentation to improve fine-grained accident matching and generalization beyond the training categories. Specifically, fine-grained accidents can share similar appearance while differing mainly in participant motion and interaction dynamics. The maneuver-aware fusion module therefore enriches RGB representations with optical-flow cues while suppressing unreliable motion responses caused by ego-motion and estimation noise. In addition, although unseen accident categories have no direct training examples, their underlying maneuvers, interactions, and semantics can still overlap with patterns observed in seen events. The prototype-based retrieval-augmented matching module exploits these transferable patterns by retrieving relevant visual and textual prototypes from seen training events to enrich the query and provide additional evidence for ambiguous matches.

We evaluate MRA-VLM under the proposed open-vocabulary retrieval setting across seen, category-disjoint unseen, and real-world accident videos. On EgoCrash, MRA-VLM improves overall text-to-video R@1 by 5.44% over the strongest baseline (27.59% versus 22.15%) and yields consistent gains for both retrieval directions and unseen categories. When trained only on EgoCrash and evaluated directly on real videos, it also outperforms the compared methods on overall R@1 in both text-to-video and video-to-text retrieval. These results provide evidence that carefully constructed synthetic supervision can transfer beyond the generated domain, while also revealing a substantial remaining synthetic-to-real gap. Our main contributions are summarized as follows:

- We introduce **EgoCrash**, a large-scale synthetic benchmark for fine-grained egocentric traffic accident understanding, containing 20,000 video-text pairs across 56 event categories with multi-granularity annotations.

Table 1: Comparison of EgoCrash with other public ego-view accident understanding datasets. 'AD': Accident Description, 'AR': Accident Recognition, 'QA': Question-Answer Pairs, 'OF': Optical Flow, 'AC': Accident Category, 'Tracker': Entity Trajectory Tracking. 'N:A' is the ratio of videos with and without accidents. The N:A ratio for EgoCrash is approximate. 'R/S' indicates Realistic or Synthetic data.

Datasets	#Videos	Annotation Types						N:A	R/S
		AD	AR	QA	OF	AC	Tracker		
GTACrash (Kim et al., 2019)	7,720	✗	✗	✗	✗	✗	✗	1:2	S
VIENA^2(Aliakbarian et al., 2018)	15,000	✗	✗	✗	✗	✗	✗	7:243	S
DADA-2000 (Fang et al., 2019)	2,000	✓	✗	✗	✗	✗	✗	0:1	R
CTA (You & Han, 2020)	1,935	✓	✗	✗	✗	✗	✗	0:1	R
CCD(Bao et al., 2020)	4,500	✓	✓	✗	✗	✗	✓	2:1	R
MM-AU (Fang et al., 2024a)	11,727	✓	✓	✓	✗	✗	✗	0:1	R
EMM-AU (Li et al., 2025)	10,399	✓	✗	✗	✗	✗	✗	0:1	R+S
EgoCrash (Ours)	**20,000**	✓	✓	✓	✓	✓	✓	**1:9**	**S**

- We formulate open-vocabulary video-text retrieval for unseen and real-world accident events, and propose **MRA-VLM**, which combines maneuver-aware RGB-flow fusion with prototype-based retrieval augmentation for fine-grained matching and generalization beyond the training categories and synthetic domain.
- Extensive experiments demonstrate that synthetic video-text supervision supports generalizable accident understanding across both category and domain shifts, with consistent gains on seen and unseen EgoCrash events and effective transfer to real-world accident videos.

2 Related Work

Datasets for Egocentric Accident Understanding. Several egocentric accident datasets have been developed to support safety-critical driving perception and reasoning (Kim et al., 2019; Aliakbarian et al., 2018; You & Han, 2020; Fang et al., 2021; 2024a). DADA-2000 (Fang et al., 2019), CTA (You & Han, 2020), and the dashcam collection of Bao et al. (Bao et al., 2020) provide complementary annotations for driver attention, causal reasoning, and accident anticipation, while MM-AU (Fang et al., 2024a) extends coverage to over 10K videos with multi-granularity annotations for eight tasks. However, collecting and annotating real accidents remains challenging, especially for rare scenarios, motivating synthetic alternatives. Early simulator-based datasets, including GTACrash (Kim et al., 2019) and VIENA (Aliakbarian et al., 2018), support dangerous-vehicle classification and action anticipation, respectively. More recently, EMM-AU (Li et al., 2025) supplements MM-AU with 2,000 generated videos, but its synthetic subset remains limited in scale and is used jointly with real data. However, its modest synthetic subset and reliance on real-world data limit its scalability as a standalone benchmark. To bridge this gap, we introduce **EgoCrash**, a large-scale synthetic dataset for egocentric accident understanding. EgoCrash features realistic rare accidents and rich multimodal annotations, providing a scalable benchmark for studying safety-critical, long-tail traffic scenarios. A more detailed comparison between these datasets is shown in Table 1.

Open-Vocabulary Learning. Open-vocabulary learning aligns visual representations with language to recognize concepts beyond a predefined label space (Wu et al., 2024a), with applications in recognition, detection, and segmentation (Zhu & Chen, 2024; Wang et al., 2025; Fu et al., 2025; Hajimiri et al., 2025). In traffic scenes, existing work primarily addresses perception, including 2D object detection (Liu et al., 2024; Cheng et al., 2024) and language-aligned 3D segmentation, detection, and occupancy prediction (Chen et al., 2023; Peng et al., 2023; Lu et al., 2023; Vobecky et al., 2023; Tan et al., 2023). The setting most similar to ours is open-vocabulary video anomaly detection (Wu et al., 2024b), which aims to recognize anomalous events unseen during training but typically represents them with a single event label. In contrast, traffic accident understanding requires finer-grained characterization of how road participants interact and how an event develops over time. An unseen accident therefore differs from seen cases not only in its category, but also in the composition of participant behaviors and interaction dynamics. This challenge is further amplified by the egocentric viewpoint, where the collision itself could be partially observed or outside the camera view. Existing accident benchmarks, however, are still largely defined over closed event

taxonomies (Fang et al., 2024a). We study open-vocabulary video-text retrieval for traffic accidents, evaluating whether visual-language correspondences learned from seen categories generalize to category-disjoint unseen events.

3 EGOCRASH DATASET

3.1 DATASET CONSTRUCTION

EgoCrash is constructed through a multi-stage generation, annotation, and validation pipeline (Figure 4 in Appendix A.2) to provide large-scale synthetic supervision for egocentric accident understanding. We supplement a real-world source pool with publicly available videos from underrepresented accident categories and select 2,000 videos through category-aware proportional sampling. Qwen-VL (Wang et al., 2024) generates accident-focused descriptions, and the resulting video–text pairs are used to fine-tune Wan2.1-T2V-1.3B (Wan et al., 2025). To expand scenario diversity, we augment the descriptions by varying participants, maneuvers, and environmental conditions while preserving the event category, and generate videos from multiple fine-tuning checkpoints. Human screening removes samples with severe visual artifacts, temporal inconsistencies, implausible interactions, or description mismatches. The final dataset contains 20,000 synthetic egocentric traffic-event videos across 56 categories, totaling 0.82M frames. Detailed generation and quality-control procedures are provided in Appendix A.1.

3.2 MULTI-LEVEL ANNOTATION

EgoCrash complements its 20,000 videos with event descriptions, accident-occurrence and category labels, pixel and entity level motion annotations, and 80,000 QA pairs, supporting fine-grained video–language alignment, event recognition, and motion and scene understanding. Detailed annotation procedures and statistics are provided in Appendix A.2 and Appendix A.3, respectively.

Accident Descriptions. GPT-4o (OpenAI, 2024) and Gemini-2.5 Pro (Comanici et al., 2025) independently describe each generated video, after which Qwen-VL 72B selects the more visually consistent and informative candidate, with human review when needed. The resulting descriptions are annotated from the generated videos rather than copied from generation prompts and capture participants, maneuvers, interactions, outcomes, causes, and environmental context. They average 42.8 words and provide paired supervision for video–text retrieval.

Accident Labels. All 20,000 videos are annotated with accident-occurrence labels and one of 56 fine-grained event categories. Accident occurrence is determined using complementary video- and text-based predictions, with disagreement cases resolved through human verification. Fine-grained categories are assigned according to the predefined accident taxonomy. We manually audit the annotated labels to ensure the accuracy.

Motion Annotations. We provide dense optical flow and entity trajectories over approximately 0.77M annotated frames, comprising 3.18M bounding boxes and 354K tracks. Frame-wise detections are associated with persistent track IDs, while optical flow captures finer motion beyond object displacement. These annotations support traffic-interaction analysis and the maneuver-aware component of MRA-VLM.

Multi-Granularity QA Pairs. Each video has four QA pairs covering environment, road type, time of day, and weather/visibility. GPT-4o (OpenAI, 2024) generates answers, with unresolved cases manually annotated and sampled answers checked for visual consistency. These 80,000 pairs complement event descriptions with structured context and can support downstream tasks such as scene question answering and condition-aware accident analysis, although they are not used in the retrieval task studied in this work.

3.3 BENCHMARK USAGE

We partition the 56 EgoCrash categories into 41 seen and 15 category-disjoint unseen categories and the complete category lists are provided in Figure 6 of Appendix A.3. The training split contains 10,235 video–text pairs from seen categories only. The test split contains 9,765 pairs, including 5,144 seen-category pairs and 4,621 unseen-category pairs, with all unseen-category samples ex-

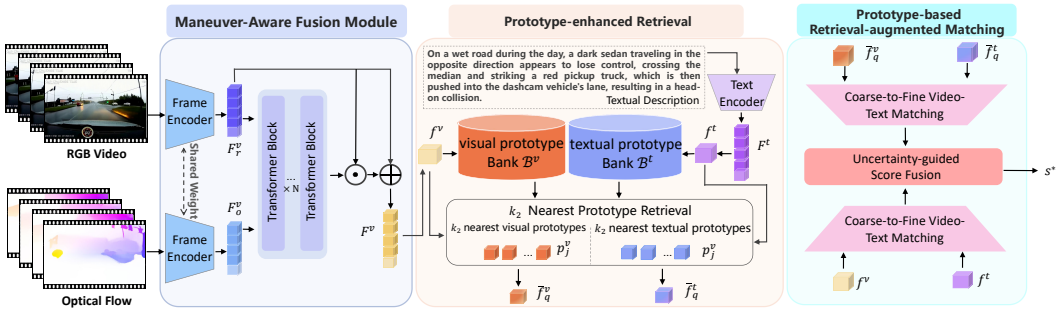


Figure 2: **The overall architecture of MRA-VLM.** We apply Maneuver-Aware Fusion Module to enhance the raw video representation by capturing the related dynamic cues from optical flow frames. In the inference process, we introduce Prototype-based Retrieval-Augmented Matching Strategy to improve the performance on unseen traffic accidents, which contains two steps: Prototype-based Query Enhancement and Retrieval-Augmented Refinement.

cluded from retrieval-model training and prototype construction. To assess generalization on domain shift, we additionally evaluate on 2,000 real-world videos from MM-AU (Fang et al., 2024a). These videos are also excluded from model training and prototype construction. We generate accident descriptions and optical-flow annotations using the corresponding EgoCrash annotation procedures while retaining the original MM-AU category labels. Further split statistics and category distributions are provided in Appendix B.1.

4 Method

We propose MRA-VLM (Figure 2) for open-vocabulary accident video-text retrieval. During training, the Maneuver-Aware Fusion Module integrates motion cues with RGB features to distinguish visually similar events with different participant dynamics. At inference, Prototype-based Retrieval-Augmented Matching reuses visual and semantic patterns from seen training events to refine retrieval of unseen accidents without unseen-category supervision.

4.1 Representation Extraction

Let $\mathcal{D}_{\text{train}} = \{(v_i, t_i, y_i)\}_{i=1}^{N},\ y_i \in \mathcal{Y}_{\text{seen}}$, denote the training set, where v_i, t_i, and y_i are an accident video, its paired textual description, and its event category, respectively. The unseen categories $\mathcal{Y}_{\text{unseen}}$ are disjoint from $\mathcal{Y}_{\text{seen}}$. Following Zeng et al. (2023), we uniformly sample frames from each video and encode them using a Vision Transformer (Dosovitskiy et al., 2020). Temporal embeddings are added to the frame-wise patch features and aggregated over time, producing RGB token features F_r^v. The same visual encoder is used to obtain optical-flow features F_o^v. The paired description is encoded using BERT (Devlin et al., 2019), producing token-level features F^t, with the corresponding `[CLS]` embedding f^t used as the global textual representation.

4.2 Training

Maneuver-Aware Fusion Module. The RGB representation captures scene appearance but can be insufficient to distinguish accidents with similar visual content but different participant motions. We therefore use optical flow to provide explicit motion cues and construct a motion-sensitive video representation. Given the RGB representation F_r^v and optical-flow representation F_o^v, we first use cross-attention to retrieve motion information relevant to each RGB token:

$$\hat{F}^v = \text{CrossAttn}\left(F_r^v, F_o^v, F_o^v\right). \tag{1}$$

Here, F_r^v serves as the query, while F_o^v provides the keys and values. Not all optical-flow responses are equally reliable. In particular, camera motion and flow-estimation noise can introduce motion patterns that are unrelated to the underlying accident dynamics. We therefore predict a feature-wise gate G from the joint RGB and motion representation, then get the final maneuver-aware representation F^v:

$$G = \sigma\Big(\phi([F_r^v, \hat{F}^v])\Big), \qquad F^v = F_r^v + G \odot \hat{F}^v, \tag{2}$$

where $[\cdot,\cdot]$ denotes feature concatenation, $\phi(\cdot)$ is an MLP projection, and $\sigma(\cdot)$ is the sigmoid function. $\odot$ denotes element-wise multiplication. The gate assigns smaller weights to unreliable motion responses and larger weights to informative motion cues, while the residual connection preserves the original RGB information. We take the `[CLS]` embedding in F^v as the final global-level fused visual representation f^v.

Multi-Grained Video-Text Matching. To learn representations that can transfer beyond the seen event taxonomy, we optimize global alignment, fine-grained cross-modal matching, and token-level semantic prediction using video-text contrastive learning (VTC), video-text matching (VTM), and masked language modeling (MLM), respectively.

Video-Text Contrastive Learning. VTC provides global video-text alignment and serves as the coarse matching stage. For video v_i and text t_j, their global similarity is $g_{ij} = \mathrm{cosine}(f_i^v, f_j^t)$. Given a mini-batch of B matched pairs, the video-to-text contrastive loss is

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(g_{ii}/\tau)}{\sum_{j=1}^{B}\exp(g_{ij}/\tau)}, \qquad \mathcal{L}_{\mathrm{VTC}} = \frac{1}{2}\left(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}\right), \tag{3}$$

where $\mathcal{L}_{t2v}$ is defined symmetrically as $\mathcal{L}_{v2t}$ and τ is a learnable temperature parameter.

Video-Text Matching. While VTC provides global alignment, fine-grained retrieval requires token-level interaction between visual and textual features. We therefore use the global similarities g_{ij} to select semantically similar hard negatives and feed the corresponding token-level representations into a cross-modal encoder. For video-to-text matching, the fine-grained score between video v_i and text t_j is defined as

$$s_{v2t}(i,j) = \mathrm{MatchHead}\left(\mathrm{CrossEncoder}(F_i^v, F_j^t)\right). \tag{4}$$

The text-to-video score $s_{t2v}(j,i)$ is computed symmetrically using the reversed input order. A standard binary cross-entropy loss $\mathcal{L}_{\mathrm{VTM}}$ is applied to positive and hard-negative pairs. Thus, g_{ij} provides coarse global matching for candidate selection, whereas s_{v2t} and s_{t2v} provide the fine-grained scores used for cross-modal discrimination.

Masked Language Modeling. MLM further encourages token-level semantic alignment by predicting masked words from both textual context and the associated video representation. We denote the resulting standard masked-token prediction loss by $\mathcal{L}_{\mathrm{MLM}}$ (Zeng et al., 2023). The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{VTC}} + \mathcal{L}_{\mathrm{VTM}} + \mathcal{L}_{\mathrm{MLM}}. \tag{5}$$

4.3 Open-Vocabulary Inference

Base coarse-to-fine retrieval. At inference, MRA-VLM follows the same coarse-to-fine matching process used during training. Given all test videos and texts, we first compute the global similarity matrix $G = [g_{ij}]$, where $g_{ij} = \mathrm{cosine}(f_i^v, f_j^t)$. For each query, the top-C candidates are selected according to G and re-ranked using the fine-grained matching function in Eq. 4, producing $s_{v2t}(i,j)$ and $s_{t2v}(j,i)$ for video-to-text and text-to-video retrieval, respectively. Since the two directions can select different candidate sets and use different query orders in the cross-modal encoder, s_{v2t} and s_{t2v} are not necessarily symmetric.

Although the trained model can directly retrieve unseen events, an unseen accident can still share motion, interaction, and semantic patterns with events observed during training. We therefore introduce prototype-based retrieval-augmented matching, which reuses representative patterns exclusively from seen training events. It augments the base coarse-to-fine retrieval pipeline with a prototype-enhanced branch and selectively combines the two according to retrieval uncertainty.

Multimodal prototype bank. We independently cluster the global video representations f^v and text representations f^t from $\mathcal{D}_{\mathrm{train}}$. The number of clusters is set to $K = |\mathcal{Y}_{\mathrm{seen}}|$. For each cluster, the k_1 samples closest to its centroid are retained, forming the visual prototype bank $\mathcal{B}^v = \{p_j^v\}_{j=1}^{k_1 K}$ and textual prototype bank $\mathcal{B}^t = \{p_j^t\}_{j=1}^{k_1 K}$. These prototypes summarize recurring visual, motion, interaction, and linguistic patterns observed across seen events.

Prototype-enhanced retrieval. Prototype-based retrieval-augmented matching enhances the query modality independently for each retrieval direction. Given a query representation f_q, we retrieve its

Table 2: Text-to-video (T2V) and video-to-text (V2T) retrieval results under synthetic-seen-to-unseen. "Recall@1": R@1 , "Mean Average Precision": mAP.

Methods	T2V (overall)		V2T (overall)		T2V (seen)		V2T (seen)		T2V (unseen)		V2T (unseen)	
	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
CLIP4Clip (Luo et al., 2022)	20.30	32.00	19.50	31.50	26.60	39.60	26.40	39.40	28.50	40.80	27.50	40.80
DRL (Wang et al., 2022)	20.70	32.10	20.30	32.10	26.70	39.90	27.20	39.90	28.20	41.10	27.90	40.60
X-CLIP (Ma et al., 2022)	20.99	32.57	20.89	32.69	26.40	39.03	24.75	37.56	28.31	41.09	26.80	39.69
X^2-VLM (Zeng et al., 2023)	22.15	34.07	20.15	31.50	30.30	43.71	27.10	39.71	30.47	42.23	27.63	40.15
MRA-VLM (Ours)	**27.59**	**39.82**	**24.95**	**36.43**	**35.46**	**48.15**	**31.92**	**43.72**	**34.70**	**47.35**	**32.23**	**44.05**

k_2 nearest prototypes from the corresponding prototype bank. Let $\mathcal{N}_{k_2}(q)$ denote the retrieved set and a_j the softmax-normalized cosine similarity between the query and prototype p_j. The enhanced query representation is

$$\bar{f}_q = \frac{1}{2}\left(f_q + \sum_{j \in \mathcal{N}_{k_2}(q)} a_j p_j\right). \tag{6}$$

For video-to-text retrieval, only the video query is enhanced, and the resulting representation is passed through the same coarse-to-fine pipeline to obtain $\bar{s}_{v2t}$. For text-to-video retrieval, only the text query is enhanced, producing $\bar{s}_{t2v}$.

Uncertainty-guided score fusion. The prototype-enhanced branch is used only when the base retrieval is ambiguous. For a video query v_i, let $j_i^* = \arg\max_j s_{v2t}(i,j)$ be its top-ranked text. We then use $t_{j_i^*}$ as a text query and obtain its top-ranked video $i_i^* = \arg\max_k s_{t2v}(j_i^*, k)$. If $i_i^* = i$, the match is reciprocally consistent and the original video-to-text ranking is retained. If $i_i^* \neq i$, the two retrieval directions disagree, indicating an ambiguous correspondence. We measure the disagreement using the difference between their top matching scores:

$$\alpha_i = \left| s_{v2t}(i, j_i^*) - s_{t2v}(j_i^*, i_i^*) \right|. \tag{7}$$

The prototype-enhanced score is then used to refine the base score:

$$s^*_{v2t}(i,j) = s_{v2t}(i,j) + \alpha_i \, \bar{s}_{v2t}(i,j). \tag{8}$$

For reciprocally consistent queries, we set $\alpha_i = 0$. The same procedure is applied symmetrically to text-to-video retrieval, producing s^*_{t2v}. In this way, the original coarse-to-fine matcher determines reliable retrievals, while prototype-enhanced matching contributes primarily when the two retrieval directions indicate an uncertain correspondence.

5 Experiment

Dataset and Evaluation Metrics. EgoCrash comprises 20,000 video–text pairs across 56 traffic event categories, including 41 seen and 15 unseen categories. We construct the dataset split to evaluate both retrieval on seen categories and generalization to unseen accident types. We evaluate open-vocabulary traffic event understanding using Recall@K (R@K) and mean Average Precision (mAP), reported separately for seen and unseen categories. Meanwhile, mAP measures ranking quality over the complete candidate list by averaging the Average Precision across all queries. We report R@1 and mAP. More details are shown in Appendix B.1.

Implementation Details. The frame encoder of MRA-VLM is initialized with BEiT-2 (Peng et al., 2022) and takes RGB frames of resolution 224×224 as input. We initialize the text encoder and the cross-modal fusion encoder with the first 12 layers and the last 6 layers of BERT (Devlin et al., 2019), respectively. For all experiments, we sample 15 frames per video and set the maximum number of text tokens to 56. We train MRA-VLM for 50 epochs with a mini-batch size of 32, using the AdamW (Loshchilov & Hutter, 2017) optimizer with a weight decay of 0.01 and a learning rate of 4×10^{-5}. We set the prototype-bank hyper-parameters to $k_1 = 5$ and $k_2 = 3$. All experiments are conducted on a single NVIDIA H100 GPU.

5.1 Results

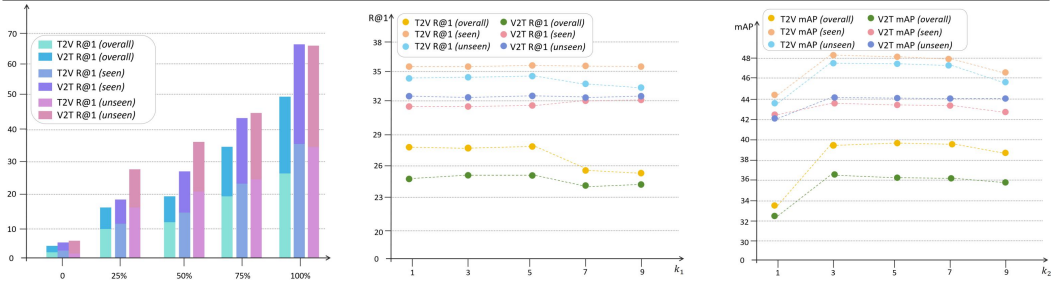
We evaluate EgoCrash and MRA-VLM under two complementary open-vocabulary settings. The first one is termed **synthetic-seen-to-unseen** that all models are trained EgoCrash seen split and

Table 3: Text-to-video (T2V) and video-to-text (V2T) retrieval results under synthetic-to-real. "Recall@1": R@1 , "Mean Average Precision": mAP.

Methods	T2V (overall)		V2T (overall)		T2V (seen)		V2T (seen)		T2V (unseen)		V2T (unseen)	
	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
CLIP4Clip (Luo et al., 2022)	10.85	19.54	9.45	17.22	11.19	20.11	9.31	17.51	10.32	19.00	9.89	18.25
X-CLIP (Ma et al., 2022)	11.10	19.96	10.10	17.39	11.03	19.79	10.17	17.77	11.17	20.38	10.03	19.46
DRL (Wang et al., 2022)	11.78	20.06	9.55	17.34	12.99	21.93	10.17	18.18	9.89	18.66	8.74	16.73
X^2-VLM (Zeng et al., 2023)	12.65	19.86	10.40	17.14	13.85	22.05	11.13	18.15	11.18	18.00	7.31	13.66
MRA-VLM (Ours)	**12.95**	**20.12**	**10.85**	**17.44**	**15.81**	**24.82**	**11.89**	**20.04**	**15.62**	**25.90**	**13.90**	**22.91**

Table 4: Ablation studies on the key components of MRA-VLM. OF, M, P, and U denote Optical Flow, Maneuver-Aware Fusion Module, Prototype-enhanced retrieval, and Uncertainty-guided score fusion, respectively.

Configuration	OF	M	P	U	T2V (Overall)		V2T (Overall)		T2V (Seen)		V2T (Seen)		T2V (Unseen)		V2T (Unseen)	
					R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
w/o P&R	✓	✓			21.52	32.96	20.54	31.60	28.48	40.78	27.53	39.25	26.87	38.80	24.25	35.43
w/o M	✓		✓	✓	22.03	32.01	20.30	32.01	28.62	40.50	27.16	39.51	28.67	41.14	28.09	40.83
w/o OF		✓	✓	✓	25.48	36.59	19.05	29.94	26.15	37.08	19.52	30.35	24.73	36.04	18.53	29.49
w/o P	✓	✓		✓	26.60	36.39	24.80	33.91	34.74	45.51	32.15	42.28	34.35	45.26	32.57	42.89
w/o R	✓	✓	✓		26.87	38.80	24.25	35.43	34.23	46.79	30.68	42.42	34.18	46.57	31.53	43.20
Full MRA-VLM	✓	✓	✓	✓	**27.59**	**39.82**	**24.95**	**36.43**	**35.46**	**48.15**	**31.92**	**43.72**	**34.70**	**47.35**	**32.23**	**44.05**



(a) Ablation studies on synthetic training data scale. (b) Ablation studies on different k_1. (c) Ablation studies on different k_2.

Figure 3: (a) We train 0%, 25%, 50%, 75%, 100% of the EgoCrash training set and then report the recall rate on the real-world test set. (b) Effect of the number of representative prototypes retained per cluster k_1 on retrieval performance. (c) Effect of the number of selected prototypes k_2 on retrieval performance.

tested on the unseen split. As shown in Table 2, MRA-VLM outperforms four video–text retrieval baselines in both directions across the overall, seen, and unseen settings. Specifically, on the overall test set, MRA-VLM achieves 27.59% R@1 and 39.82% mAP for T2V retrieval, exceeding the strongest baseline (X^2-VLM) by 5.44% and 5.75% percentage points, respectively. For V2T retrieval, our method achieves 24.95% R@1 and 36.43% mAP, surpassing the best competing method. Moreover, MRA-VLM maintains clear advantages on both seen and unseen accident types. These results demonstrate that MRA-VLM not only learns discriminative bidirectional alignments between accident videos and textual accident descriptions, but also generalizes effectively to previously unseen accident semantics, highlighting its potential for open-vocabulary accident video understanding. The higher absolute scores on the unseen split are largely attributed to its smaller gallery size, where fewer distractor videos make retrieval relatively easier.

In the second setting, termed **synthetic-to-real**, all models are trained exclusively on the generated training set and directly evaluated on a real accident video-text test set. This setting is designed to examine whether the accident semantics learned from synthetic videos can transfer to realistic and more challenging traffic scenarios. To construct the real-world accident video-text test set, we randomly sample 2,000 MM-AU videos with balanced coverage of seen and unseen categories, strictly excluding those used to fine-tune Wan2.1. This shared test set is excluded from both generator fine-tuning and retrieval-model training, with descriptions and optical flow generated following Section 3.2. As shown in Table 3, MRA-VLM achieves overall R@1/mAP scores of 12.95%/20.12% for T2V and 10.85%/17.44% for V2T. Performance is lower than in synthetic-domain evaluation, consistent with a substantial synthetic-to-real domain gap. Nevertheless, on unseen categories, MRA-VLM achieves 15.62%/25.90% for T2V and 13.90%/22.91% for V2T. These results support the transferability of EgoCrash supervision and the ability of MRA-VLM to generalize to real-world accident retrieval.

5.2 Ablation Study

Effectiveness of Key Components. We conduct ablation studies on the key components of the proposed MRA-VLM to show their effectiveness. We evaluate five variants: 1) *w/o OF*, removing optical-flow inputs; 2) *w/o M*, replacing Maneuver-Aware Fusion Module with simple addition; 3) *w/o P*, replacing Prototype-enhanced retrieval with one randomly selected prototype per modality; 4) *w/o R*, fixing $\alpha_i = 0.5$ in Eq. 8; and 5) *w/o P&R*, retaining only the original coarse-to-fine matching scores without prototype augmentation. We report the ablation results in Table 4 and obtain the following observations. 1) w/o OF performs worse than MRA-VLM, confirming the necessity of optical flow for identifying abnormal motions in accidents. 2) w/o M degrades retrieval performance, indicating that the improvement of MRA-VLM does not come merely from introducing optical flow but from the adaptive interaction between appearance and motion cues. 3) w/o P suggests that retrieving multiple highly relevant prototypes enables the model to aggregate complementary semantic cues from the prototype bank. Such multi-prototype fusion provides a more stable and diverse prior and alleviates the negative effect of unreliable individual prototypes. 4) w/o R and w/o P&R are inferior to MRA-VLM, validating that incorporating the prototype-based retrieval-augmented matching further improves the matching results.

The Impact of Training Set Scale. In vision-language retrieval, the amount of training data generally plays an important role: more data usually leads to better performance. To thoroughly study the effectiveness of EgoCrash, we further explore the impact of the data scale during training. Specifically, we train separate models on 0%, 25%, 50%, 75%, and 100% of the EgoCrash training pairs for 50 epochs each. As shown in Figure 3(a), the recall on the real-world test set improves steadily as the amount of training data increases, with a clear gain from 0% to 100%.

Effectiveness of k_1 Cluster Center Selection. We further study the impact of k_1, the number of representative prototypes retained per cluster, with results shown in Figure 3(b). We observe that $k_1 = 5$ achieves the optimal balance between computational efficiency and retrieval accuracy. Smaller values ($k_1 = 1, 3$) may exclude relevant categories, while larger values ($k_1 = 7, 9$) introduce noise and increase computation without significant performance gains. These results validate the prototype-bank design, in which a moderate k_1 retains the most informative reference patterns without excessive computation.

Effectiveness of k_2 Prototype Selection. We evaluate k_2, the number of retrieved prototypes used to enhance the query inputs, the results are shown in Figure 3(c). As k_2 increases, the model achieves better retrieval performance, suggesting that multiple retrieved prototypes provide richer semantic cues for accident video understanding. However, the performance drops when $k_2 > 3$, likely because less relevant prototypes introduce noise and dilute the discriminative query enhancement. These results indicate that a moderate k_2 is essential for balancing prototype diversity and reliability, with $k_2 = 3$ offering the best trade-off for both seen and unseen accident categories.

6 Conclusion

In this paper, we focus on the egocentric traffic accident understanding task in autonomous driving. To address this data scarcity, we introduce EgoCrash, a large-scale synthetic video-text benchmark designed for this safety-critical task. Compared with existing real-world traffic event datasets, EgoCrash highlights the distinctive advantages of synthetic data: it supports the scalable generation of diverse long-tail accident scenarios, controlled modeling of hazardous interactions, and fine-grained spatiotemporal annotations that are costly and difficult to obtain from real-world videos. To handle the open-ended, compositional nature of real-world accidents, which arise from diverse combinations of traffic participants, maneuvers, road layouts, and causal interactions, we further propose MRA-VLM. By explicitly modeling the motion dynamics of traffic participants and leveraging semantically related accident prototypes, MRA-VLM improves video–text alignment and generalizes more effectively to previously unseen accident scenarios. Extensive experiments demonstrate strong performance on the accident retrieval task, as well as effective transfer from synthetic training data to real-world driving videos. These results suggest that carefully constructed synthetic benchmarks, together with models designed to capture motion and compositional semantics, provide a promising path toward scalable and generalizable traffic accident understanding.

7 AI DISCLOSURE

Generative AI tools were used in several stages of this work. During the construction of EgoCrash, large language models (LLMs) and vision-language models (VLMs), including GPT-4o, Gemini, Qwen2.5-VL, Qwen-VL, and Qwen3, were used to assist with the generation, selection, and verification of textual annotations and accident labels. In particular, candidate accident descriptions were generated and evaluated with multimodal models, while independent visual- and text-based models were used to verify accident occurrence and assist with fine-grained accident-category annotation. Model-generated annotations were subject to consistency checks and human review for ambiguous or inconsistent cases. LLMs were also used during the research process to provide feedback on methodology and experimental design, assist with the formulation and presentation of mathematical expressions, and support the interpretation and presentation of experimental results. In addition, generative AI tools were used for Chinese–English translation, literature summarization, brainstorming, and editing portions of the manuscript for clarity and readability. All AI-assisted outputs, including annotations, methodological suggestions, code-related suggestions, and manuscript text, were reviewed and verified by the authors. The research methodology, experimental protocol, reported results, and final conclusions were determined by the authors. The authors take full responsibility for the content of the paper and for all artifacts produced with the assistance of generative AI.

REFERENCES

Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Viena: A driving anticipation dataset. In *Asian Conference on Computer Vision*, pp. 449–466. Springer, 2018.

Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2682–2690, 2020.

Vladislav Bargatin, Egor Chistov, Alexander Yakovenko, and Dmitriy Vatolin. Memfof: High-resolution training for memory-efficient multi-frame optical flow estimation. *arXiv preprint arXiv:2506.23151*, 2025.

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.

Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Asian conference on computer vision*, pp. 136–153. Springer, 2016.

Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7020–7030. IEEE, 2023.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attention$f$ analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 4303–4309. IEEE, 2019.

Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems*, 23(6):4959–4971, 2021.

Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22030–22040, 2024a.

Jianwu Fang, Fan Wang, Jianru Xue, and Tat-Seng Chua. Behavioral intention prediction in driving scenes: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 25(8):8334–8355, 2024b.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14987–14997, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5061–5071, 2025.

Minhee Kang, Wooseop Lee, Keeyeon Hwang, and Young Yoon. Vision transformer for detecting critical situations and extracting functional scenario for automated vehicle safety assessment. *Sustainability*, 14(15):9680, 2022.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. Crash to not crash: Learn to identify dangerous vehicles using a simulator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 978–985, 2019.

Cheng Li, Keyuan Zhou, Tong Liu, Yu Wang, Mingqiao Zhuang, Huan-ang Gao, Bu Jin, and Hao Zhao. Avd2: Accident video diffusion for accident video description. *arXiv preprint arXiv:2502.14801*, 2025.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7353–7363, 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.

Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11624–11641, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1190–1199. IEEE, 2023.

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.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 638–647, 2022.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1043–1052, 2023.

ModelScope. Diffsynth-studio, 2023. URL `https://github.com/modelscope/diffsynthstudio`. GitHub repository.

OpenAI. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2024-05.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824. IEEE, 2023.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 7513–7522, 2024.

Zhiyu Tan, Zichao Dong, Cheng Zhang, Weikun Zhang, Hang Ji, and Hao Li. Ovo: Open-vocabulary occupancy. *arXiv preprint arXiv:2305.16133*, 2023.

Antonin Vobecky, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *Advances in Neural Information Processing Systems*, 36:50545–50557, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14824–14834, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022.

Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):5092–5113, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18297–18307. IEEE, 2024b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tackgeun You and Bohyung Han. Traffic accident benchmark for causality recognition. In *European Conference on Computer Vision*, pp. 540–556. Springer, 2020.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE transactions on pattern analysis and machine intelligence*, 46(5):3156–3168, 2023.

Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8954–8975, 2024.

APPENDIX CONTENTS

A EgoCrash Dataset **14**

A.1 Dataset Construction 14

A.2 Multi-task Annotation Strategy 15

A.3 Dataset Static Analysis 18

B Experiment **21**

B.1 Dataset Splits and Evaluation Protocol 21

B.2 Inference Efficiency. 23

B.3 Qualitative Results. 23

A EGOCRASH DATASET

EgoCrash is constructed through a multi-stage generation, annotation, and validation pipeline designed to provide large-scale synthetic supervision for egocentric accident understanding. We present the full construction pipeline in Figure 4.

A.1 DATASET CONSTRUCTION

In this section, we give more details about dataset Construction.

Fine-tuned Data Collection. Our data collection pipeline consists of three stages. 1) We first analyze the category distribution of the original MM-AU dataset and augment each underrepresented accident category containing fewer than 10 videos with additional accident videos collected from YouTube, increasing its size to at least 20 videos. 2) We then employ a category-aware proportional sampling strategy to select 2,000 representative videos from the resulting dataset. 3) Finally, we use Qwen-VL (Wang et al., 2024) with a predefined prompt to generate a textual description for each selected video. These descriptions focus on accident-relevant information, including the event progression, potential causes of the collision, and the motion of the key vehicles. The prompt that we used in the procession of the textual description generation is: The prompt is defined as:

> "*Summarize the traffic accident video from the ego-car's forward dashcam perspective in a paragraph under 500 characters. Focus on: 1. The immediate cause of the crash (*e.g.*, a red car sudden lane change, pedestrian violation, unexpected braking). 2. Key vehicle motion trajectories with timing/spatial cues (*e.g.*, 'ego-braked sharply 2s before impact', 'truck-2 swerved left from blind spot'), 3. Spatial/directional context: Describe all objects' positions relative to ego-car (*e.g.*, 'front', 'left-rear', 'right-blind spot') and their movement directions (*e.g.*, 'approaching head-on', 'overtaking from right'). 4. Explicitly state if ego-car is involved: 'ego-car [collided/avoided] with [object]'. Omit all non-essential details (weather, timestamps, etc.).* " .

Fine-tuning Video Generative Model. To balance the generation quality and fine-tuning efficiency, we apply three strategies specially: 1) We finetune official Wan2.1-T2V 1.3B (Wang et al., 2024) with FP8 quantization via DiffSynth-Studio ModelScope (2023); 2) TeaCache (Liu et al., 2025) (threshold=0.08) has been adopted to accelerate video generation; 3) Setting video parameters: 41 total frames, 10 fps, 480×832 resolution ($H \times W$), quality = 6. All generations use a fixed random seed (42) for reproducibility. We finetune the model for 12 epochs and save the model checkpoint after each epoch for subsequent accident video generation.

Diverse Traffic Event Video Acquisition. To broaden within-category semantic and visual coverage, we augment the descriptions associated with all 2,000 source videos selected for generator fine-tuning. For each original description, Qwen3 (Yang et al., 2025) and DeepSeek-R1 (Guo et al., 2025) each generate five variants, yielding ten augmented descriptions per source video and 20,000

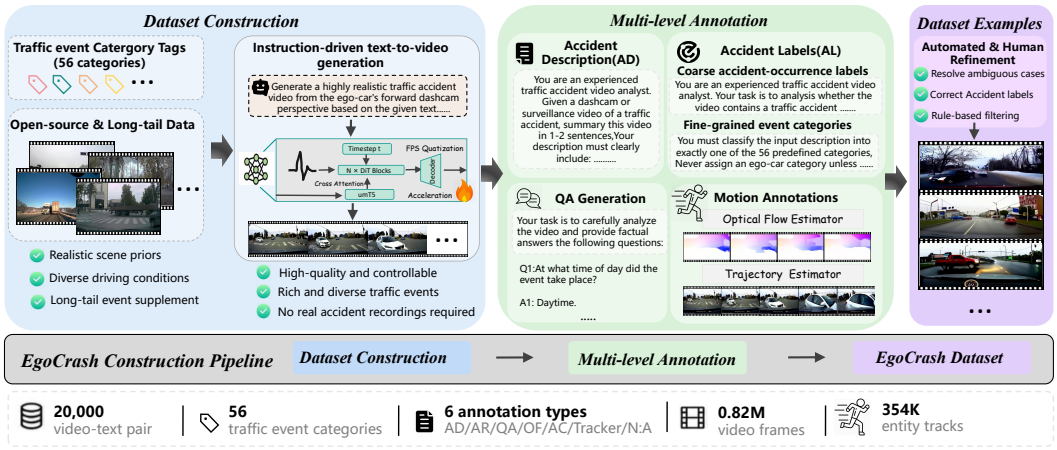


Figure 4: **Construction pipeline of EgoCrash.** It consists of two main stages: Dataset Construction and Multi-level Annotation. Representative samples from the resulting benchmark are shown.

augmented descriptions in total. These variants modify road participants, maneuvers, and environmental conditions while preserving the defining event category. The original descriptions are used only for generator fine-tuning, whereas the augmented descriptions serve as conditioning prompts for video generation.

We partition the 20,000 augmented descriptions into five groups of 4,000 and assign them cyclically to the 12 checkpoints saved after successive fine-tuning epochs. Specifically, checkpoints from epochs 1–5 use groups 1–5, respectively; checkpoints from epochs 6–10 repeat this assignment; and checkpoints from epochs 11–12 use groups 1 and 2. Each checkpoint generates one video per assigned description, producing 4,000 videos and yielding 48,000 candidate videos overall. All generations use a fixed random seed of 42. Each video contains 41 frames at a resolution of 480 × 832 and a frame rate of 10 fps. Description augmentation varies the specified scenarios, while using different checkpoints provides additional variation in appearance and motion realization.

Five annotators screen all candidate videos over one week. Each video is initially reviewed by two annotators, who assess visual quality, temporal consistency, the plausibility of traffic interactions, and alignment with the conditioning description. Videos with severe visual artifacts, temporal inconsistencies, implausible interactions, or event–description mismatches are excluded. When the two reviewers disagree, a third annotator reviews the video, and the final decision follows the majority judgment. This process removes 28,000 candidates and retains 20,000 videos, corresponding to a retention rate of approximately 41.7%. The resulting EgoCrash dataset covers 56 traffic-event categories and contains approximately 0.82 million frames. The retained videos subsequently undergo the multi-level annotation procedure described in Appendix A.2.

By separating description-level augmentation from checkpoint-induced generation variability, the proposed pipeline provides explicit control over semantic coverage while retaining substantial diversity in visual realization.

A.2 Multi-task Annotation Strategy

Textual Description. For each generated video, we uniformly sample 15 frames over its temporal duration and provide them in chronological order to GPT-4o and Gemini. Both models use the same prompt to independently generate a description of one or two sentences, covering the scene context, weather conditions, involved entities, and likely accident cause where supported by the visual evidence. This process produces two candidate descriptions per video. We use the predefined prompt to generate the descriminative description of the synestic videos. The textual prompt template is:

> *You are an experienced traffic accident video analyst. Given a dashcam video:*
> *1. Determine whether the video shows a traffic accident.*

> *2. If yes, describe the accident in 1–2 sentences, including environment, weather, entities involved, and the likely cause.*
>
> *3. If no, describe the scene factually in 1–2 sentences without inventing an accident.*

Qwen2.5-VL-72B (Wang et al., 2024) then receives the sampled frames together with both candidates and selects the description that better reflects the video content. Videos for which neither candidate is acceptable are referred for human inspection. We additionally conduct sample-based human checks of the final descriptions to assess their consistency with the corresponding videos. The final retrieval descriptions are annotated from the generated videos rather than directly inherited from the prompts used for video generation. The prompt is:

> *You are an experienced traffic video annotation reviewer.*
>
> *You will receive 15 chronologically ordered frames sampled from a video and two candidate descriptions. Select the description that most accurately represents the visible video content.*
>
> *Evaluate the candidates using the following criteria:*
>
> *1. Visual consistency: the description should correctly identify the visible entities, their actions, spatial relationships, and the event outcome.*
>
> *2. Event coverage: prioritize the main actions and interactions over incidental scene details.*
>
> *3. Scene accuracy: descriptions of the location, weather, and road conditions should be supported by the frames.*
>
> *4. Cautious interpretation: distinguish visible evidence from inference. Penalize unsupported claims about accident causes, driver intentions, responsibility, or events outside the camera view.*
>
> *5. Accident occurrence: do not assume that an accident occurs. For a non-accident video, prefer a factual description that does not invent a collision. The absence of a visible impact frame alone is not sufficient to rule out an accident; consider the full frame sequence.*
>
> *6. Clarity: prefer a concise, coherent description without unnecessary repetition.*
>
> *Select A or B if it adequately describes the main event without substantial factual errors. Do not prefer a candidate merely because it is longer or includes more details. If both candidates are acceptable, choose the more accurate and informative one; if they are equally suitable, select A. If neither candidate is acceptable, select NEITHER for human review.*
>
> *Candidate A:* $\{caption_a\}$
>
> *Candidate B:* $\{caption_b\}$
>
> *Output only A, B, or NEITHER. Do not rewrite the descriptions or provide an explanation.*

Accident Labels. We annotate both accident occurrence and fine-grained event categories. For accident occurrence, we employ two branches with different input modalities. The video branch uses Qwen2.5-VL-72B to determine whether an accident occurs from the video, using the following prompt:

> *You are an experienced traffic accident video analyst. Determine whether the given video contains a traffic accident, such as a collision, crash, or rollover. Dangerous driving, near misses, or successful avoidance alone do not count as an accident. Output only 'yes' if an accident occurs, or 'no' otherwise. Do not provide any explanation or additional text.*

The text branch uses Qwen3-8B to make the same binary prediction from the video's final textual description, using the following prompt:

> *Determine whether the following video description indicates that a traffic accident occurred, such as a collision, crash, or rollover. Dangerous driving, near misses,*

> *or successful avoidance alone do not count as an accident. Do not assume an accident unless the description indicates one.*
> *Description:* {caption}
> *Output only 'yes' or 'no'.*

Each branch outputs a binary decision. Consistent predictions are retained, whereas disagreements are resolved through human inspection of the video to determine the final accident-occurrence label.

The event taxonomy is adapted from MM-AU by merging categories with overlapping semantics and adding a non-accident category, resulting in 56 categories in total. Videos labeled as non-accident are assigned directly to the non-accident category. For the remaining videos, Qwen2.5-VL-72B assigns a fine-grained accident label using the final textual description and the predefined category list. The model is instructed to return only the corresponding category ID. The prompt is:

> *You are an assistant for autonomous driving scene understanding. Classify the given video description into exactly one of the num_accident_categories accident categories listed below. Choose the category that best matches the main accident described, considering the involved participants, their roles, and the collision or interaction type. Do not infer details that are not supported by the description. Accident categories:* {acc_list_text} Video description: {caption} *Output only the category ID from the list. Do not output any explanation or additional text.*

Motion Annotations. EgoCrash provides complementary motion annotations at the pixel and entity levels, comprising dense optical flow and temporally associated bounding boxes. These annotations are estimated from the generated videos rather than obtained from simulator states. **Dense optical flow.** We estimate optical flow using MEMFOF (Bargatin et al., 2025) with the `MEMFOF-Tartan-T-TSKH` checkpoint. The estimator processes the original 41-frame videos, preserving their temporal resolution rather than using the 15-frame sampling adopted for language annotation. The estimated flow fields are converted into RGB-encoded flow images for storage and use by the motion branch of MRA-VLM. These representations capture apparent image-plane motion, including both road-user motion and motion induced by the egocentric camera. They therefore provide complementary temporal cues to RGB appearance, but should not be interpreted as ego-motion-compensated object velocities. **Entity-level trajectories.** We use YOLO11-L (Khanam & Hussain, 2024) to detect road participants, including pedestrians, cars, motorcycles, buses, and trucks. Detections are associated across frames to form trajectories represented by track IDs and frame-wise bounding boxes. These annotations describe the spatial evolution of individual participants and complement the pixel-level motion captured by optical flow. For qualitative inspection, we additionally visualize local flow directions within detected regions; these visualizations are distinct from the dense flow images used by the retrieval model. **Usage and limitations.** MRA-VLM uses the RGB-encoded optical-flow images as motion inputs, while entity trajectories are provided as additional annotations for participant-level analysis. Both optical flow and trajectories are automatically estimated and may contain errors caused by occlusion, rapid motion, missed detections, or visual artifacts in the generated videos. Accordingly, they constitute auxiliary motion cues rather than ground-truth physical measurements. The gated fusion in MAFM is designed to regulate the contribution of flow features while preserving the original RGB representation.

Multi-Granularity QA pairs. We provide four question–answer pairs for each video, yielding 80,000 pairs across the 20,000 videos in EgoCrash. The questions follow fixed templates covering four aspects of scene context: the surrounding environment, road or area type, time of day, and weather and visibility conditions. The environment question concerns the broader setting, whereas the road-type question focuses on the visible roadway or traffic area; their answers may overlap in some scenes. The four templates are:

> *1. Identify the type of background or environment visible in the video (e.g., urban street, rural road, highway, residential area, parking lot, etc.).*
> *2. What type of road or area is shown (e.g., highway, city street, rural road, intersection, parking lot, tunnel, etc.)?*
> *3. At what time of day did the road event take place (e.g., daytime, nighttime, dawn, dusk)?*

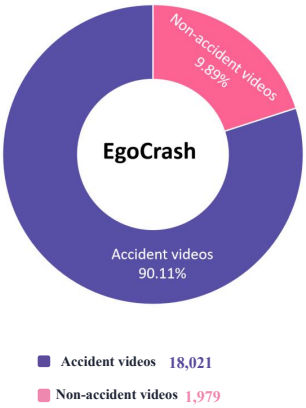


Figure 5: **Distribution of accident and non-accident videos in EgoCrash.**

> *4. What were the weather and visibility conditions (e.g., clear, rainy, foggy, snowy, sunny, overcast, etc.)?*

GPT-4o answers each question from the video using a shared instruction that requires concise, visually grounded responses and explicitly permits abstention when the evidence is insufficient:

> *You are an experienced traffic video researcher. You will be given a dashcam or surveillance video. Your task is to carefully analyze the video and provide factual answers without making assumptions beyond the visible evidence.*
> *Answer the following question:* `{question}`
> *Important:*
> - *If the information is not clearly visible in the video, answer with 'unknown'.*
> - *Do not fabricate specific names of cities, towns, or dates unless they are explicitly visible in the video.*
> - *Answer concisely, using only short sentences. Do not provide reasoning, only the final answer.*

Responses marked as `unknown` are referred for manual annotation. We additionally conduct sample-based human checks to assess whether the answers are supported by the corresponding videos, following the human spot-checking procedure used for the other semantic annotations. These QA pairs provide structured descriptions of scene context that complement the accident-focused captions. They are released as auxiliary annotations and are not used in the retrieval training objective of MRA-VLM.

A.3 Dataset Static Analysis

EgoCrash contains 20,000 video-text pairs, comprising approximately 0.82M frames across 56 traffic-event categories. In addition to accident scenarios, the dataset retains normal driving videos at a normal-to-accident ratio of 1:9. This accident-oriented distribution provides dense coverage of safety-critical events while preserving normal samples for learning the distinction between routine driving and anomalous traffic behavior. Table 1 provides a comparison with existing egocentric accident-understanding benchmarks in terms of dataset scale, annotation types, sample distribution, and data source. We present the fine-grained categories and category-disjoint seen and unseen categories in Figure 6.

Dataset composition. EgoCrash contains 20,000 synthetic egocentric video–text pairs across 56 traffic-event categories. Each video contains 41 frames, yielding 820,000 video frames in total. The dataset comprises 18,021 accident videos (90.11%) and 1,979 non-accident videos (9.90%), corresponding to an approximately 1:9 non-accident-to-accident ratio. Non-accident videos provide examples in which the depicted driving event does not result in an accident. Figure 5 and Figure 6

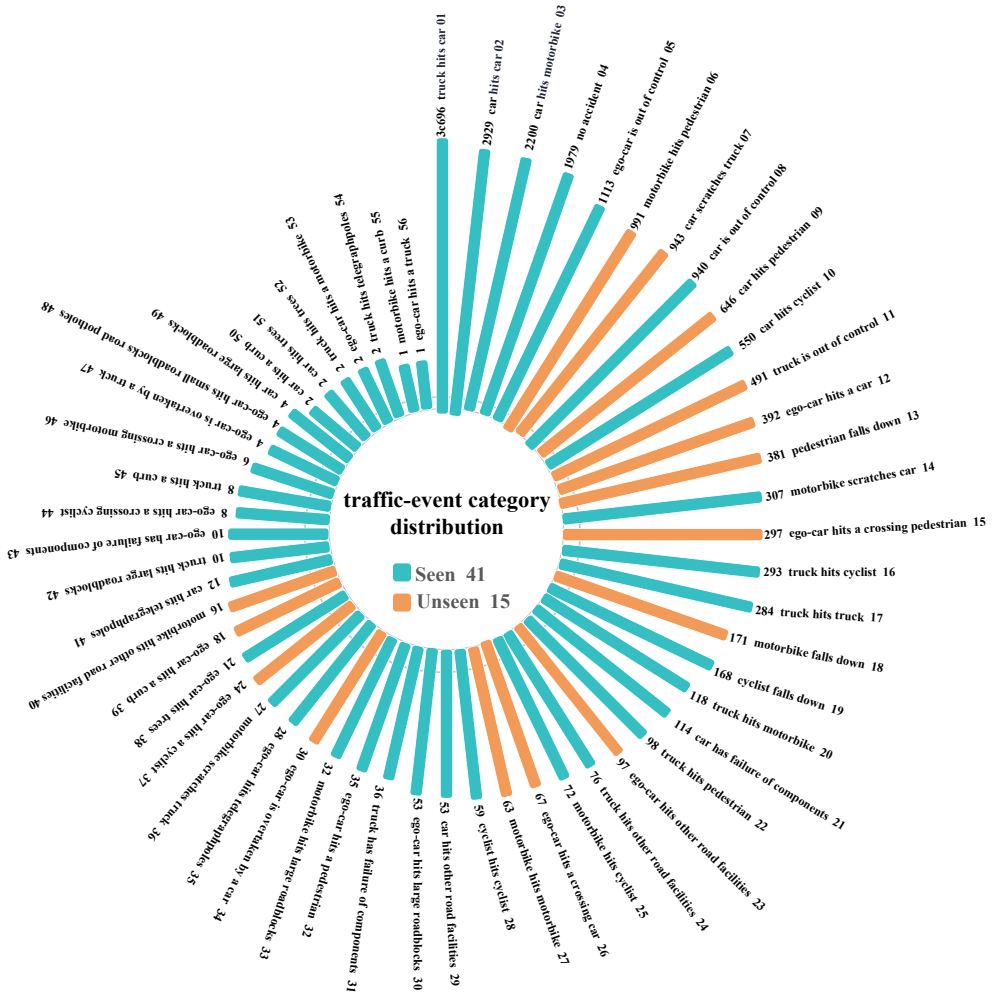


Figure 6: **Distribution of traffic-event categories in EgoCrash.** video counts across 56 traffic-event categories, with seen and unseen categories distinguished by color.

present the accident-occurrence distribution and the category frequencies, with the 41 seen and 15 unseen categories distinguished by color. The category distribution is substantially imbalanced. For example, the largest category, *truck hits car*, contains 3,696 videos, accounting for 18.48% of the dataset, whereas several categories contain only a few examples. The benchmark therefore provides uneven supervision across event types, which should be considered when interpreting aggregate retrieval performance.

Description length and semantic coverage. The 20,000 descriptions contain 855,670 words in total. Their mean and median lengths are 42.8 and 42 words, respectively, with individual descriptions ranging from 19 to 82 words. The average description contains 1.51 sentences. These statistics characterize the textual supervision provided for video–text retrieval.

To examine the content of the descriptions, we use rule-based lexicons and regular expressions to identify six groups of semantic elements: participants, actions or maneuvers, outcomes or collisions, interactions, causes, and environmental context. Table 5 summarizes their detected coverage. Participants and environmental context are detected in 99.98% of descriptions, while actions or maneuvers and interactions occur in 95.1% and 82.4%, respectively. All six groups are detected together in 60.4% of descriptions. These measurements describe lexical coverage rather than annotation accuracy or visual grounding. For example, detecting a cause-related expression does not necessarily

Table 5: Description statistics and rule-based semantic coverage in EgoCrash. Semantic groups are non-exclusive. Coverage measures the proportion of descriptions containing a detected element from each group, rather than its factual correctness.

Statistic	Value
Number of descriptions	20,000
Total word count	855,670
Mean / median length (words)	42.8 / 42
Length range (words)	19–82
Mean number of sentences	1.51
Semantic group	**Detected coverage**
Participants	99.98%
Actions / maneuvers	95.1%
Outcomes / collisions	91.0%
Interactions	82.4%
Causes	76.8%
Environmental context	99.98%
All six groups	60.4%

establish that the description explicitly attributes the accident to that cause. Under a stricter rule requiring a cause expression to occur within a causal clause, 40.1% of descriptions contain an explicit attribution. Such statements remain textual interpretations rather than verified causal explanations of the depicted events.

Environmental context in descriptions. The same lexical analysis identifies references to weather, lighting, road type, and road conditions in 75.9%, 55.5%, 92.8%, and 46.4% of descriptions, respectively. The descriptions mention varied settings, including urban roads, highways, intersections, and rural roads, as well as different weather and lighting conditions. These attributes are not mutually exclusive: a description may refer to both an urban setting and an intersection, for example. The frequencies reflect what is mentioned in the descriptions, not an exhaustive measurement of the visual scene distribution. In particular, wet roads are mentioned in 40.4% of descriptions, whereas dry roads are explicitly mentioned in only 0.31%. This asymmetry may reflect selective reporting of salient conditions and should not be interpreted as the actual wet-to-dry road ratio in the videos. The four scene-context QA pairs associated with each video provide a separate, structured annotation of the environment, road or area type, time of day, and weather and visibility.

Motion annotation statistics. Motion annotation records are available for all 20,000 videos. They comprise 3,176,063 bounding-box annotations and 353,632 tracks across 767,049 annotated frames. On average, each video contains 17.68 tracks, 158.80 bounding-box annotations, and 38.35 annotated frames. Each track contains an average of 8.98 bounding-box observations. Here, a track refers to an identity assigned by the tracking pipeline, rather than a manually verified unique physical entity. The annotated-frame count summarizes the motion annotation records and is distinct from the total number of video frames.

Comparison with existing datasets. Table 1 compares EgoCrash with existing egocentric accident datasets in terms of video count, annotation types, accident composition, and data source. EgoCrash combines synthetic video–text pairs with accident-occurrence labels, event categories, scene-context QA pairs, optical flow, and entity trajectories. These annotations provide complementary views of the same traffic events. The retrieval experiments use paired descriptions and RGB–flow inputs, while category labels define the evaluation splits and the remaining annotations support additional analysis.

EgoCrash has the following key characteristics:

- **Large-scale egocentric traffic-event coverage.** EgoCrash contains 20,000 videos, providing a larger collection than commonly used egocentric accident benchmarks such as DADA-2000 (Fang et al., 2019), CTA (You & Han, 2020), CCD (Bao et al., 2020), MM-AU (Fang et al., 2024a), and EMM-AU (Li et al., 2025). Its scale enables more systematic evaluation of fine-grained and open-vocabulary traffic-event understanding.
- **Multi-granularity annotations.** EgoCrash provides accident descriptions, recognition labels, question-answer pairs, optical flow, accident-category labels, and entity trajectories.

These annotations support a range of tasks, including accident recognition, video-text retrieval, causal reasoning, motion modeling, and trajectory-aware understanding.

- **Diverse accident scenarios.** The initial event descriptions are collected from existing real-world benchmarks and supplemented with long-tail traffic events from open-web sources. We further diversify traffic participants, road environments, weather conditions, motion patterns, and camera viewpoints through controlled language-model-based augmentation. This process produces varied video-text pairs while preserving the defining semantics of each traffic event.
- **Accident-oriented sample distribution.** EgoCrash adopts a 1:9 normal-to-accident ratio. In contrast to accident-only datasets, the retained normal-driving samples allow models to learn the boundary between routine and safety-critical events. Compared with normal-dominated driving datasets, the higher density of accident cases provides substantially greater coverage of rare and hazardous interactions.
- **Privacy-preserving construction.** Because the videos are synthetically generated, EgoCrash does not directly expose identifiable drivers, pedestrians, vehicle license plates, or sensitive real-world accident footage. Synthetic construction therefore reduces several privacy and ethical concerns associated with collecting and distributing real accident videos.

Discussion: Key Advantages of EgoCrash. Existing ego-view accident datasets are often limited by the scarcity of real-world accident data, resulting in restricted video scale, incomplete annotations, and insufficient diversity of accident scenarios. For example, several real-world datasets (Bao et al., 2020; You & Han, 2020; Fang et al., 2019; 2024a) mainly provide accident occurrence labels or coarse accident descriptions, making them less suitable for fine-grained reasoning and open-vocabulary accident understanding. Although Li et al. (2025) alleviates the data scarcity issue, it still lack comprehensive video-text annotations and trajectory-level supervision. In contrast, EgoCrash integrates large-scale synthetic video generation with rich multimodal annotations, covering accident description, recognition, question answering, optical flow, category labels, and entity trajectory tracking in a unified benchmark. EgoCrash expands accident types via LLM-driven annotation and synthesizes high-quality videos using generative models, enabling theoretically unbounded generation of diverse traffic accident samples. In this work, we generate 20,000 video-text pairs (about 0.82M frames) covering 56 traffic events, significantly surpassing EMM-AU (Li et al., 2025) and MM-AU (Fang et al., 2024a).

B Experiment

B.1 Dataset Splits and Evaluation Protocol

Real-world Dataset. To evaluate retrieval performance on real-world traffic videos, we additionally curate a real-world test set comprising 2,000 videos from MM-AU. The real-world test set contains 2,000 videos, including 1,302 videos from 41 seen categories and 698 videos from 15 unseen categories, accounting for 65.10% and 34.90% of the set, respectively We retain the original category labels and apply the corresponding components of the EgoCrash annotation pipeline to obtain optical flow and paired textual descriptions. Table B.1 reports the number of videos in each category. The resulting distribution is highly imbalanced: the three most frequent categories account for 30.25% of the videos, whereas 17 categories contain fewer than 10 videos each. This evaluation complements the synthetic benchmark by assessing retrieval on real-world visual content across diverse traffic events, although aggregate results should be interpreted in light of the uneven category representation. All videos in this real-world set are used exclusively for evaluation and are excluded from model training, model selection, and hyperparameter tuning.

Query and gallery construction. We evaluate both text-to-video (T2V) and video-to-text (V2T) retrieval under three settings. For overall evaluation, all 9,765 test pairs are used to construct the query set and the gallery. For seen-category evaluation, both queries and gallery candidates are restricted to the 5,144 seen-category test pairs. For unseen-category evaluation, both are restricted to the 4,621 unseen-category test pairs. In T2V retrieval, descriptions serve as queries and videos as gallery candidates; V2T retrieval reverses these roles. Table 7 summarizes the evaluation sizes.

Table 6: Category distribution of the real-world test set curated from MM-AU. Category IDs and names follow the original annotations. Categories are sorted by decreasing video count, proceeding down the left block and then down the right block.

ID	Category	Videos	ID	Category	Videos
11	ego-car hits a car	203	45	truck hits motorbike	18
43	car hits car	202	9	ego-car is overtaken by a truck	15
48	truck hits car	200	18	ego-car hits other road facilities	14
50	car hits motorbike	106	51	motorbike scratches car	13
10	ego-car hits a crossing car	103	38	motorbike hits motorbike	12
8	ego-car hits a truck	103	3	ego-car hits a crossing cyclist	12
14	ego-car hits a curb	86	30	car hits trees	11
24	car hits a curb	84	44	car has failure of components	11
12	ego-car is overtaken by a car	76	33	car hits telegraphpoles	11
21	car hits large roadblocks	53	4	ego-car hits a cyclist	11
56	car hits pedestrian	52	2	ego-car hits a pedestrian	10
6	ego-car hits a motorbike	46	17	ego-car hits telegraphpoles	10
5	ego-car hits a crossing motorbike	41	55	truck hits cyclist	8
49	car scratches truck	39	59	cyclist falls down	8
40	truck hits truck	39	58	pedestrian falls down	6
13	ego-car hits large roadblocks	39	52	motorbike hits pedestrian	6
42	car is out of control	37	54	truck hits pedestrian	5
39	truck is out of control	35	0	no accident	5
15	ego-car hits small roadblocks road potholes	30	32	truck hits telegraphpoles	3
1	ego-car hits a crossing pedestrian	28	29	truck hits trees	2
37	motorbike falls down	26	26	truck hits small roadblocks road potholes	2
57	car hits cyclist	25	35	truck hits other road facilities	2
7	ego-car hits a crossing truck	22	19	motorbike hits large roadblocks	2
16	ego-car hits trees	22	34	motorbike hits other road facilities	1
20	truck hits large roadblocks	21	47	motorbike scratches truck	1
41	truck has failure of components	21	22	motorbike hits a curb	1
61	ego-car is out of control	20	62	ego-car has failure of components	1
36	car hits other road facilities	19	60	cyclist hits cyclist	1
23	truck hits a curb	19	53	motorbike hits cyclist	1
Total: 58 categories					**2,000**

Table 7: Query and gallery sizes for synthetic-domain evaluation. The same sizes apply to both retrieval directions.

Setting	Queries	Gallery candidates
Overall	9,765	9,765
Seen	5,144	5,144
Unseen	4,621	4,621

Each query has exactly one positive candidate: its originally paired video or description. Other candidates, including those sharing the same event category, are treated as negatives. The protocol therefore evaluates instance-level video–text matching rather than category-level retrieval.

Evaluation metrics. Let r_q denote the one-based rank of the unique positive candidate for query q, and let Q be the number of queries. Recall at rank K is computed as $\mathrm{R@}K = \frac{1}{Q}\sum_{q=1}^{Q} \mathbf{1}[r_q \le K]$. With one positive candidate per query, average precision reduces to $\mathrm{AP}_q = 1/r_q$. Consequently, mean average precision over the complete ranking is $\mathrm{mAP} = \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{r_q}$, which is equivalent to mean reciprocal rank under this single-positive protocol. We report R@1 and mAP as percentages for each retrieval direction.

Interpretation across evaluation settings. The overall setting includes distractors from both seen and unseen categories, whereas the subset settings use smaller, separate galleries. Overall scores are therefore not a weighted average of the seen and unseen scores. Likewise, differences between seen and unseen performance reflect both category generalization and differences in gallery size and composition.

B.2 Inference Efficiency.

We further evaluate the additional inference latency introduced by the proposed prototype-based retrieval-augmented matching. Following the standard two-stage retrieval pipeline, we separately measure the latency of the first-stage similarity-based retrieval and the second-stage cross-modal refinement for a single-video query. Without retrieval augmentation, the two stages require 0.123 ms and 12.345 ms, respectively, resulting in a total retrieval latency of 12.468 ms per query. Incorporating the prototype memory bank increases the corresponding latency to 0.135 ms and 13.456 ms. The complete retrieval pipeline therefore requires 13.591 ms per query, introducing only 1.123 ms of additional latency, or a 9.01% relative overhead. Most of this overhead arises from the prototype-enhanced second-stage refinement, whereas the additional cost in the first-stage retrieval is only 0.012 ms. These results demonstrate that the proposed strategy improves retrieval performance while maintaining practical inference efficiency.

B.3 Qualitative Results.

In Figure 7, we show the ranking list of MRA-VLM on our EgoCrash dataset. We could observe that our dataset is still challenging to prevailing cross-modality models on seen and unseen video-text pairs. Given a textual accident description, our method retrieves videos that match the main accident process described in the query. For seen categories, the top-ranked results capture key cues such as abrupt lane changes, rear-end collisions, wet-road skidding, and interactions between vehicles. For unseen categories, the retrieved results are still relevant to the query, including parking-lot turning collisions and spin-out accidents on residential streets. This suggests that the model can handle accident scenarios beyond the training categories by aligning scene context, vehicle motion, and accident-related events between text and video.

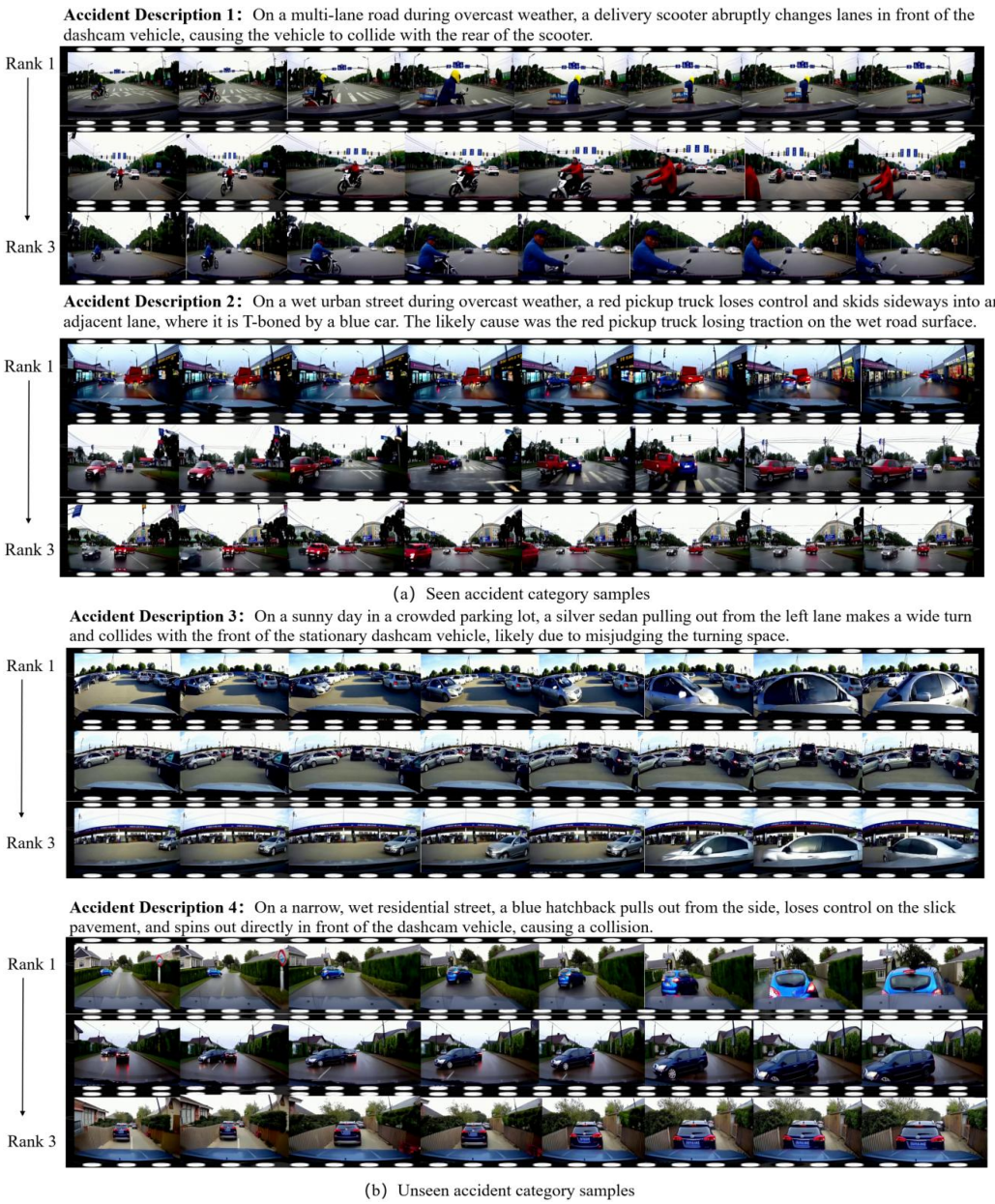


Figure 7: We visualize the top-ranked video retrieval results for representative accident descriptions from both seen and unseen accident categories. Each case presents the textual accident query and the corresponding retrieved videos ranked from Rank 1 to Rank 3.