

# Reading Risk: Zero-Shot Traffic Accident Estimation via Vision-Language Models

Vinicius Avena, Rodrigo S. Couto, Luís Henrique M. K. Costa

Universidade Federal do Rio de Janeiro - GTA/Poli/COPPE/UFRJ - Rio de Janeiro, Brazil

{avena, rodrigo, luish}@gta.ufrj.br

**Abstract**—Modern vision-based advanced driver assistance systems (ADAS) are largely blind to semantic notions of risk. They can detect lanes, vehicles, and pedestrians but struggle to determine if a scene is becoming dangerous. In this work, we explore whether a pretrained vision-language model (VLM) can supply such semantic structure without any task-specific supervision. Using MobileCLIP in a strictly zero-shot setting, we design a risk-aware textual vocabulary and compute semantic accident scores through image-text similarities. We then study four minimal temporal aggregation strategies operating either in embedding or similarity space, revealing how short-range context modulates the latent risk signal. Evaluated on the DAD dataset, our approach achieves nearly 50% AP and anticipates accidents up to 2.5s ahead, approaching fully supervised early baselines. These results suggest that general-purpose VLMs already encode actionable priors for risk understanding in vehicle environments, offering a new path toward semantically grounded ADAS.

**Index Terms**—Vision-Language Models, Zero-Shot Learning, Risk Understanding, Accident Anticipation

## I. INTRODUCTION

Advanced Driver Assistance Systems (ADAS) play a key role in improving modern vehicle safety, offering functions such as collision warnings and automated emergency braking. However, most systems rely on task-specific detectors and narrowly supervised objectives [1], focusing on identifying **what** is present and **where**, but not on **whether** a scene is semantically risky or likely to lead to an accident.

This limitation is most evident in accident risk assessment. Existing pipelines estimate risk from low-level visual cues or from models trained on labeled hazardous events, which anchor their predictions to specific sensors and training distributions. They lack an explicit notion of semantic risk: recognizing that a configuration of vehicles, VRUs (Vulnerable Road Users), and road elements corresponds to a high-risk, pre-accident situation. While traditional models see bounding boxes, trajectories, and optical flow, human drivers see “a car suddenly braking”, “a pedestrian jaywalking”, or “a motorcycle lane splitting”. This semantic gap is a key obstacle to more proactive and generalizable safety systems.

Recent advances in vision-language models offer an alternative. Trained on large-scale image-text data from diverse domains, VLMs acquire broad knowledge about objects, agents, actions, and typical scene dynamics [2]. They embed visual inputs and linguistic descriptions in a shared representation space, aligning scenes with high-level semantic concepts. Crucially, these models have shown strong zero-shot performance in a wide range of downstream tasks [3]: given

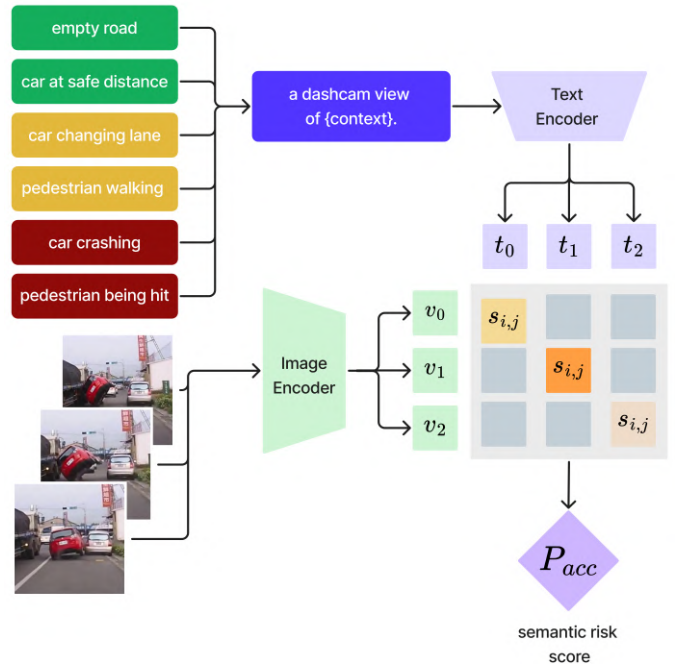


Fig. 1: Overview of the proposed zero-shot semantic risk framework. Textual contexts are grouped into three semantic categories (green: safe, yellow: low risk, red: accident) and inserted into a prompt template before being encoded into text embeddings  $\{t_i\}$ . Dashcam frames are processed by the image encoder into visual embeddings  $\{v_j\}$ . A cosine-similarity matrix  $S$  is computed between all image-text pairs; each similarity  $s_{i,j}$  is transformed via softmax, and the probability mass assigned to accident-related prompts yields the accident score  $P_{acc}$  used as our semantic risk estimate.

only a textual description, they can evaluate how well an image or video frame matches that concept, training-free. This leads to a central question for automotive perception: *can a VLM understand the semantics of risk in driving scenes without any task-specific training?*

In this paper, we address this question by treating a VLM as a semantic risk sensor. We use MobileCLIP [3], a lightweight image-text model designed for real-time applications, and employ it strictly in an inference-time, zero-shot setting. Instead of training on labeled crashes, we design a textual vocabulary spanning three categories: accident situations, low-risk behaviors, and safe driving conditions. As depicted in

Figure 1, for each sequence the model compares its visual content with these descriptions and produces a similarity distribution, from which we derive a semantic accident score that leverages contextual information for zero-shot accident risk recognition without further training.

Building on this similarity-based formulation, this study goes beyond frame-wise inference, investigating temporal aggregation strategies that integrate information across frames, such as exponential moving averages and short sliding-window pooling. Although these strategies introduce minimal computational overhead, they help reveal how short-term temporal context interacts with semantic reasoning in models not explicitly optimized for video-based risk assessment.

Our contributions can be summarized as follows:

- We propose a zero-shot framework for semantic risk estimation in driving scenes, where a VLM is queried with natural-language descriptions of accidents, risky behaviors, and safe driving, yielding a training-free risk assessment.
- We investigate lightweight temporal aggregation schemes on top of VLM-derived embeddings and similarity scores, providing insight into how short-term temporal context interacts with semantic reasoning in models not explicitly trained for video understanding.
- We perform a comprehensive quantitative and qualitative study on the widely used Dashcam Accident Dataset (DAD) for Traffic Accident Anticipation, benchmarking our zero-shot VLM-based risk estimation against supervised state-of-the-art baselines.

Our experiments on the DAD dataset [4] show that leveraging the semantic priors of a lightweight VLM brings the zero-shot performance surprisingly close to supervised accident anticipation benchmarks. Using only inference-time prompting, our approach achieves nearly 50% average precision, competitive with early supervised methods, and provides warnings up to 2.5 seconds before annotated accidents. Overall, our results indicate that off-the-shelf VLMs already provide meaningful risk-aware signals in traffic scenes. This points toward future ADAS and autonomous driving systems where semantic risk assessment is built directly on top of general-purpose multimodal models and further strengthened by fine-tuning for accident anticipation.

## II. RELATED WORK

### A. Accident Anticipation

Traffic Accident Anticipation (TAA) has traditionally been addressed by modeling visual evidence from object-centric perception and motion. Early approaches rely on object detections [5], while others incorporate motion cues via optical flow or tracking [6]. A parallel line of work uses recurrent spatio-temporal encoders (RNN/LSTM/GRU) to aggregate evidence across frames [7]. DSA-RNN [4] assigns attention to candidate objects and propagates these cues through the sequence, enabling the model to capture subtle risk indicators. DSTA [8] extends this idea with coordinated spatial-temporal attention,

jointly selecting salient regions and informative segments. Recent transformer-based accident anticipation models (e.g., AAT-DA [9]) achieve high performance by explicitly modeling traffic participants and driver attention through strong object-centric supervision.

Although effective, those methods rely on low-level visual signals, i.e., detections, motion fields, and task-trained sequence models for reasoning. Consequently, risk cues are learned largely from annotated accident data, with limited access to broader semantic priors. In contrast, we test whether a pretrained VLM can provide risk-relevant semantics in a purely zero-shot setting, using only embedding-level temporal aggregation and no learned temporal modules.

### B. Zero-Shot Video Understanding

Zero-shot video understanding aims to recognize or localize actions without class-specific training data, motivated by the high cost of video annotation. A common approach is to repurpose vision-language models pretrained on large image-text datasets and apply them directly to video without fine-tuning.

Several methods improve text-video alignment through prompt engineering and lightweight inference-time design choices. TEAR [10], for instance, expands short action labels into richer natural-language prompts generated by LLMs, strengthening semantic matching to action categories.

Overall, videos are treated as sequences of frozen frame embeddings, with reasoning handled via prompts and simple aggregation rather than learned video modules. We adopt this training-free paradigm for driving by testing whether a pretrained VLM, guided only by prompts, can provide meaningful risk cues without any video-specific supervision.

### C. Vision-Language Models for Driving

Recent work explores vision-language models (VLMs) in driving scenarios, leveraging their ability to combine visual perception with broad world knowledge and natural-language reasoning. Many approaches integrate VLMs and LLMs into perception, prediction, or planning by formulating these tasks as language-conditioned queries, enabling scene interpretation and decision making through prompts.

GPT-Driver [11] formulates motion planning as language modeling problem, using a LLM to generate trajectories and explanations. DriveGPT4 [12] extends this idea by jointly processing video and text to output both control commands and human-readable explanations. Beyond planning, EMMA [2] represents sensor inputs and intermediate driving states as language tokens to support end-to-end reasoning across perception, prediction, and planning within a unified multimodal framework.

Despite their promise, these systems typically rely on large models and video-specific retraining, making them costly and complex. In contrast, we take a simpler direction: we evaluate whether an off-the-shelf, image-pretrained VLM already encodes sufficient semantic priors to produce useful risk signals in a purely zero-shot setting, without supervision or learned temporal modules.

### III. METHODOLOGY

We investigate whether an off-the-shelf vision-language model, used without any video-specific tuning, can generate meaningful semantic risk signals from driving footage. To isolate the model’s inherent semantics, we build a single inference pipeline and test four lightweight temporal strategies that differ only in how they aggregate information over time. All variants share the same core elements: (i) frame-level visual embeddings, (ii) textual embeddings defining a structured risk vocabulary, (iii) cosine-based vision-text matching, and (iv) risk scores derived from these similarities.

Figure 1 illustrates the general pipeline. A dashcam video is split into frames, each passed through the VLM’s image encoder to produce a frame embedding. In parallel, natural-language descriptions grouped into three risk-related categories are computed once by the text encoder. Since Mobile-CLIP maps images and text into a shared semantic space, we compare them using cosine similarity. This similarity module forms the backbone of all temporal strategies; they differ only in how the visual embeddings are temporally aggregated before matching.

#### A. Visual Embeddings

The visual pathway uses the image encoder of a CLIP-like model [13], pretrained with large-scale contrastive learning to align images and natural-language descriptions. This pretraining yields broad semantic correspondences beyond the supervised labels and enables zero-shot recognition of nuanced traffic elements and behaviors. Each dashcam frame from the forward-facing view is independently mapped to a normalized embedding in the VLM’s visual space, without any learned temporal structure. Consequently, any temporal reasoning in our system arises exclusively from the inference-time strategies described later.

#### B. Textual Space and Semantic Prompt Design

A key component of our approach is the construction of a driving-specific textual space. Following common practices in zero-shot video understanding (Sec. II-B), we adopt grounded prompts templates to anchor textual descriptions in the same visual perspective as the input videos, such as “a dashcam view of ...”. This style of prompt reduces semantic mismatch between textual concepts and the visual distribution of dashcam footage, helping the model align abstract risk descriptions with concrete driving scenes.

We organize the prompts into three semantic categories:

- Safe: stable trajectories, orderly traffic flow, proper following distances, and visually safe scenes.
- Low-risk: denser traffic, lane changes, close interactions with vulnerable road users, and early signs of possible danger.
- Accident: collisions, loss of control, and other high-severity events involving vehicles or VRUs.

This three-level structure enforces a smooth semantic progression from safety to danger. The low-risk tier is key: it captures subtle deviations from safe driving features without

collapsing them into full accidents. Fewer categories would merge early warnings with collisions; more would add noise. Thus, three levels give enough detail while remaining stable.

We encode all prompts once with the CLIP text encoder and cache the resulting concept embeddings, obtaining concept embeddings that define a shared semantic space for all temporal strategies. Final predictions come from cosine similarity between visual and textual vectors, so each score directly maps to a human-readable risk description, making the model’s decisions easy to interpret.

#### C. Semantic Matching

Given a visual embedding (obtained from a single frame or a temporally aggregated representation) and all textual concept embeddings, we compute cosine similarity between the scene and each description. For CLIP models, this metric is natural, since contrastive training arranges semantically related image-text pairs to be close under cosine similarity. The resulting similarity vector forms the semantic interface between visual content and the risk vocabulary and is the basis for all scoring functions in our evaluation.

To convert similarities into interpretable estimates, we apply a temperature-scaled softmax over all textual concepts:

$$p_{t,i} = \frac{\exp(s_{t,i}/T)}{\sum_j \exp(s_{t,j}/T)},$$

where  $s_{t,i}$  is the cosine similarity between the visual embedding at time  $t$ ,  $v_t$ , and the text embedding  $t_i$  (see Fig. 1). We then pool these probabilities over the accident ( $\mathcal{A}$ ), low-risk ( $\mathcal{R}$ ), and safe ( $\mathcal{S}$ ) concept sets:

$$P_{\text{acc}}(t) = \sum_{i \in \mathcal{A}} p_{t,i}, \quad P_{\text{risky}}(t) = \sum_{i \in \mathcal{R}} p_{t,i}, \quad P_{\text{safe}}(t) = \sum_{i \in \mathcal{S}} p_{t,i}.$$

This yields a three-way semantic distribution that captures the model’s belief about which type of situation best matches the current frame. In complex scenes, multiple concepts may be simultaneously activated; therefore, softmax normalization and category-level pooling capture it as probability mass shared across concept groups, rather than enforcing a single hard decision.

## IV. TEMPORAL METHODS

The goal of the temporal methods is not to approximate a trained video model, but to probe how minimal forms of temporal context affect the behavior of a purely zero-shot semantic signal. Each method is intentionally lightweight, allowing us to evaluate whether the VLM’s semantic knowledge alone can benefit from structured temporal evidence.

#### A. Pure Frame-wise Inference

The frame-wise baseline treats each frame independently. For each  $t$ , we compute  $v_t$ , obtain similarities  $s_t$ , and derive the risk score directly. This method offers maximal responsiveness and isolates the intrinsic semantic capabilities of the model without temporal bias. However, because driving videos frequently contain motion blur, glare, transient occlusions and

rapid lighting changes, the resulting scores may fluctuate sharply. Despite this instability, frame-wise inference serves as a clean baseline that reveals exactly what the model understands from a single image.

### B. Embeddings Exponential Moving Average (EMA)

The second method introduces temporal context directly in the visual embedding space through a causal exponential moving average:

$$\tilde{v}_t = \lambda v_t + (1 - \lambda)\tilde{v}_{t-1},$$

followed by normalization. This process acts as a short-term memory, attenuating high-frequency variations while preserving responsiveness to genuine scene changes. Because smoothing occurs before the semantic projection, the cosine similarities are computed from a stabilized visual descriptor rather than a raw instantaneous embedding.

### C. Windowed Similarity Averaging

Rather than smoothing embeddings, this variant aggregates information in semantic space. After computing all frame-wise similarities  $s_t$ , we form a temporal window of the past  $R$  frames  $W_t = [t - R, \dots, t]$  and compute:

$$\bar{s}_t = \frac{1}{|W_t|} \sum_{j \in W_t} s_j.$$

This approach preserves the immediate visual representation of each frame but accumulates semantic evidence over neighboring frames. It is particularly useful for mitigating noisy semantic spikes caused by single-frame artifacts. Conceptually, it tests whether temporal smoothing applied directly to semantic predictions improves robustness without altering the visual representations.

### D. Sliding-Window Embedding Aggregation

The last method treats a short segment of the video as the fundamental semantic unit. For each timestep  $t$ , we construct a window  $W_t$  of  $R$  past embeddings and compute a weighted aggregate:

$$\bar{v}_t = \frac{1}{Z} \sum_{j \in W_t} w_j \cdot v_j, \quad Z = \sum_{j \in W_t} w_j.$$

where the weights  $w_j$  follow a fixed temporal Gaussian profile centered at the most recent frames. The aggregated vector  $\bar{v}_t$  is then normalized and used for similarity computation.

This approach exposes the VLM to a temporally integrated representation before any semantic alignment occurs, allowing textual concepts to align with a local spatio-temporal state rather than an instantaneous snapshot. In spirit, this resembles classical feature pooling used in action recognition while remaining entirely training-free. Thus, the sliding-window method preserves short-term dynamics, avoiding the inertia inherent to EMA.

## V. EXPERIMENTS

We evaluate whether zero-shot semantic signals from a VLM can handle a standard traffic safety task: Traffic Accident Anticipation (TAA). Unlike prior work that depends on temporal models, detectors, or supervised fine-tuning, we isolate the model’s intrinsic semantic understanding and study how light temporal aggregation impacts anticipation. The experiments focus on measuring how early it can detect accident semantics, how stable the risk signal is over a video, and how close this purely zero-shot setup comes to supervised TAA baselines.

We conduct all evaluations on the Dashcam Accident Dataset (DAD) [4], a widely used benchmark for Traffic Accident Anticipation (TAA). The dataset consists of 5 s dashcam clips sampled at 20 fps (100 frames per video), with temporal annotations marking the accident moment: positive clips contain a collision within the final 10 frames, while negative clips contain no high-risk event. DAD comprises 455/829 (positive/negative) clips for training and 165/301 for testing. Since our approach is fully zero-shot, we use only the test split and do not rely on supervision or dataset-specific adaptation.

All experiments use MobileCLIP-S2 [3], a lightweight VLM trained on large-scale image-text pairs and used strictly off-the-shelf, with no fine-tuning, no temporal training, and no exposure to driving data. Following standard TAA practice, we report Average Precision (AP), which measures how well the risk score separates accident vs. non-accident clips, and mean Time-to-Accident (mTTA, in seconds), which measures how early the model raises an accident warning, averaged over true positives. For context, we compare our zero-shot results to representative supervised TAA models that use explicit temporal modeling and full supervision; while not directly comparable, they indicate how far a training-free VLM semantic signal can go relative to task-specific architectures.

Additional implementation details, including the source code, vocabulary and prompt templates, and the hyperparameter configuration used in our experiments, are available in our public repository: [https://github.com/viniavena/IV\\_reading\\_risks](https://github.com/viniavena/IV_reading_risks).

## VI. RESULTS

We report accident anticipation performance on DAD for the four zero-shot variants and supervised TAA baselines. Table I summarizes Average Precision (AP) and mean Time-to-Accident (mTTA).

Even without task-specific adaptation, all four zero-shot variants operate close to 50% AP, surpassing the classical DSA baseline and narrowing the gap to early supervised methods such as AdaLEA. In terms of mTTA, the similarities-window and sliding-window variants trigger warnings about 2.4–2.5 s before impact, outperforming DSA and approaching the behavior of stronger baselines like RARE.

Fully supervised methods still have a clear advantage in accident anticipation performance. While more recent approaches report AP values above 70% on DAD [9], such performance is obtained by selecting the best-performing

TABLE I: Comparison between supervised TAA baselines and our zero-shot VLM methods on the DAD test set.

Method	AP (%) $\uparrow$	mTTA (s) $\uparrow$
DSA [4]	48.1	1.34
AdaLEA [5]	52.3	3.43
DSTA [8]	56.1	3.66
RARE [7]	62.2	2.51
Frame-wise	48.6	2.40
EMA	50.9	2.31
Similarities-Window	49.7	2.52
Sliding-Window	49.9	2.49

epoch specifically for AP, which severely degrades anticipation time (lower mTTA). Such results are not directly comparable to traditional TAA baselines, as they prioritize offline precision over early warning viability

Among our temporal variants, EMA achieves the highest AP. By smoothing the embedding space before semantic projection, it suppresses frame-level noise while preserving sharp semantic transitions, leading to cleaner high-confidence activations for accident-related prompts. In contrast, pure frame-wise inference is more sensitive to artifacts, which explains its slightly lower AP despite competitive mTTA.

These results indicate that an off-the-shelf pretrained VLM already encodes meaningful semantic cues for hazardous driving. Lightweight temporal integration provides modest but consistent gains, stabilizing the zero-shot signal with minimal temporal structure.

To better understand how each temporal strategy behaves over time, we compute the mean accident probability ( $P_{acc}$ ) across all accident clips in the DAD test set for each temporal method. Figure 2 shows the resulting trajectories for the four variants: frame-wise, EMA, similarities-window, and sliding-window.

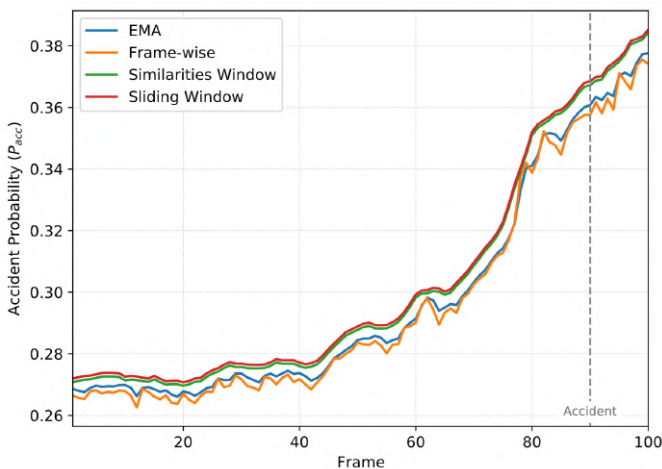


Fig. 2: Mean accident probability trajectories over time for the four temporal strategies, computed by averaging the predicted accident probability curves across all accident clips in the DAD test set at each frame index. The vertical dashed line marks the annotated accident position in the dataset.

A first observation is that the similarities-window and sliding-window curves are almost indistinguishable. Both show smoother trajectories with reduced high-frequency fluctuations, as expected from short-range temporal averaging in similarity or embedding space. This aggregation suppresses frame-level noise, yielding stable profiles and naturally explaining their nearly identical TAA performance in Table I.

By contrast, the EMA curve is closer in shape to the frame-wise baseline. Although EMA introduces causal memory, it still assigns substantial weight to the current frame, so rapid visual changes are only partially smoothed. As a result, EMA retains more local variability than the window methods, while remaining less noisy than pure frame-wise inference.

All curves exhibit a gradual rise in accident probability as clips approach the annotated accident frames, showing that the zero-shot accident semantic score captures the build-up of hazardous conditions on average. The window-based methods produce the most stable aggregate signal, while EMA strikes a compromise between responsiveness and robustness.

To complement the aggregate analysis, we also examine the behavior of the risk signal on individual video sequences. Figures 3 and 4 show one negative clip with no accident and one positive clip with a collision from the DAD test set. These qualitative visualizations illustrate how the four temporal strategies evolve over time under real driving conditions.

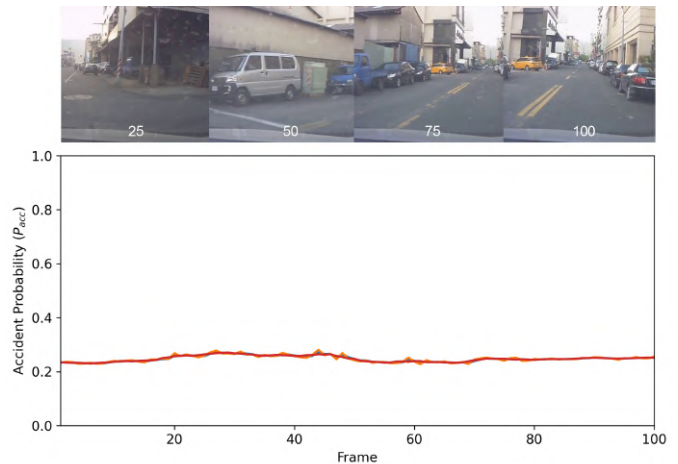


Fig. 3: Negative sequence without accident. The upper row shows representative frames (indices 25, 50, 75, and 100); the lower plot depicts the accident probability curves for the four temporal methods. The vertical dashed line marks the canonical accident position used in positive sequences (frame 90).

In the negative example, all temporal methods maintain a low and nearly flat accident probability around 0.2 throughout the sequence, with no systematic increase or abrupt peaks, indicating a low false-alarm tendency when the scene remains visually safe. In contrast, in the positive clip (Fig. 4), where the red car hits a parked car, loses stability, and flips near the end of the sequence, all temporal variants start from low risk levels and then exhibit a clear rise in probability as hazardous

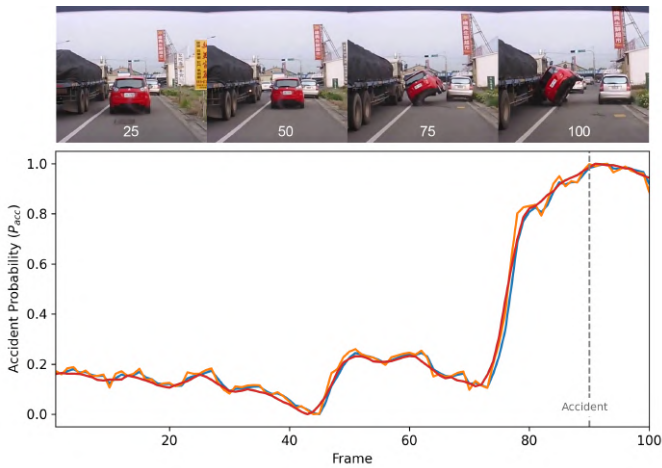


Fig. 4: Positive sequence with an actual collision. Representative frames are shown on top, with the accident probability trajectories for all methods below.

cues accumulate, with the signal approaching values close to 1.0 before the annotated accident frame.

Overall, these examples show that the zero-shot semantic signal is not only sensitive to the presence of hazardous driving events, but also tracks their temporal evolution in a coherent and interpretable manner. Despite having no exposure to driving data or temporal supervision, the VLM consistently produces risk trajectories that reflect the underlying scene dynamics, reinforcing the conclusion that meaningful accident anticipation cues can emerge purely from pretrained semantic knowledge.

## VII. CONCLUSION

In this work we investigated whether a pretrained vision-language model, used strictly in a zero-shot setting and without any form of temporal training, can provide meaningful accident anticipation signals in driving videos. By decomposing the problem into a unified CLIP-based framework and evaluating four lightweight temporal strategies, we were able to isolate the contribution of the model’s semantic knowledge without relying on any driving-specific training.

Our results highlight three key takeaways. First, a lightweight CLIP model alone already captures rich notions of hazardous driving, reaching about 50% AP on DAD without heavy detectors, tracking, or TAA-specific supervision. Second, adding minimal temporal structure, especially window-based aggregation and exponential smoothing, yields modest but consistent gains in AP and mTTA, indicating that simple inference-time mechanisms can amplify the model’s latent semantics. Third, qualitative analysis shows that the zero-shot risk signal evolves coherently with scene dynamics, generating interpretable trajectories that separate safe traffic from the gradual build-up leading to an accident.

Taken together, these findings indicate that pretrained VLMs already encode a latent notion of risk that can effectively support traffic accident anticipation without task-specific training. Building on this, future work will move beyond zero-shot

inference toward trainable, temporally aware architectures that retain this semantic grounding while adding explicit temporal reasoning. This includes exploring lightweight temporal networks (e.g., GRUs), lightweight adapters, TAA-tailored prompt designs, and supervision via exponential anticipation losses (e.g., AdaLEA), which have been effective for learning early-warning signals. The goal of these next steps is to turn the promising zero-shot behavior shown here into a competitive, efficient, fully trained accident anticipation system that enhances transportation safety and reliability.

## ACKNOWLEDGMENT

This work was supported by CAPES (Finance Code 001), CNPq, FAPERJ (E-26/204.562/2024 and E-26/204.122/2024), FAPESP (grants 23/00673-7 and 23/00811-0), and Development and Research Foundation - Fundep - Rota 2030, and our partners Stellantis and Mobway.

## REFERENCES

- [1] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, “DriveVLM: The convergence of autonomous driving and large vision-language models,” in *Conference on Robot Learning*. PMLR, 2025, pp. 4698–4726.
- [2] J.-H. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, Y. Zhou, J. Guo, D. Anguelov, and M. Tan, “Emma: End-to-end multimodal model for autonomous driving,” *arXiv preprint arXiv:2410.23262*, 2024.
- [3] P. K. A. Vasu\*, H. P. Ansari\*, F. Faghri\*, R. Vemulapalli, and O. Tuzel, “Mobileclip: Fast image-text models through multimodal reinforced training,” in *CVPR*, 2024. [Online]. Available: <https://arxiv.org/abs/2311.17049>
- [4] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, “Anticipating accidents in dashcam videos,” in *Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part IV 13*. Springer, 2017, pp. 136–153.
- [5] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, “Anticipating traffic accidents with adaptive loss and large-scale incident db,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3521–3529.
- [6] M. M. Karim, Z. Yin, and R. Qin, “An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1792–1803, 2023.
- [7] I. Song and J. Lee, “Real-time traffic accident anticipation with feature reuse,” in *2025 IEEE International Conference on Image Processing (ICIP)*, 2025, pp. 2312–2317.
- [8] M. M. Karim, Y. Li, R. Qin, and Z. Yin, “A dynamic spatial-temporal attention network for early anticipation of traffic accidents,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9590–9600, 2022.
- [9] Y. Kumamoto, K. Ohtani, D. Suzuki, M. Yamataka, and K. Takeda, “Aat-da: Accident anticipation transformer with driver attention,” in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 1142–1151.
- [10] M. Bosetti, S. Zhang, B. Liberatori, G. Zara, E. Ricci, and P. Rota, “Text-enhanced zero-shot action recognition: A training-free approach,” in *International Conference on Pattern Recognition*. Springer, 2024, pp. 327–342.
- [11] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, “Gpt-driver: Learning to drive with gpt,” in *NeurIPS 2023 Foundation Models for Decision Making Workshop*.
- [12] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, “Drivegpt4: Interpretable end-to-end autonomous driving via large language model,” *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8186–8193, 2024.
- [13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.