Seeing the Evidence, Not Just the Answer: Training-Free SalMask Evidence Routing for Traffic Accident Video Understanding

Hejiu Lu[1] and Sang Hun Lee[1]

Graduate School of Automobile and Mobility, Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Republic of Korea
{lhj17,shlee}@kookmin.ac.kr

Abstract. Vision-language models (VLMs) show strong potential for autonomous-driving scene understanding, yet their decisions on safety-critical accident videos may rely on language priors rather than visually grounded evidence. We propose SalMask Evidence Routing (SalMask-ER), a training-free framework that routes complementary global and localized evidence to a frozen VLM. A compact question-agnostic caption preserves scene-level context, while open-vocabulary detection and driver-oriented saliency produce object-aware crops with attenuated backgrounds. On VRU-Accident, SalMask-ER improves Qwen2.5-VL-3B from 51.12% to 67.33%, a 16.21-point gain, and reaches 71.30% with Qwen3.5-4B. In the same way intended for the dense accident captioning task, the SalMask configuration reports higher scores than the published Qwen2.5-VL-3B baseline on all six metrics and outperforms VideoPerceiver-3B, which is trained using a combination of supervised fine-tuning (SFT) and reinforcement learning (RL). Evidence ablations further show that removing or mismatching localized content weakens accident-centric and causal reasoning.

Keywords: Vision-Language Models · Traffic Accident Video Understanding · Visual Evidence Routing

1 Introduction

Vision-language models (VLMs) [1,15,17,26] are increasingly used for autonomous-driving scene understanding, including traffic-rule reasoning, risk explanation, and accident analysis. Accident videos remain challenging because vulnerable road users (VRUs) may occupy only a small region or appear for only a few frames, while correct interpretation depends on road layout, visibility, relative motion, conflict, collision, and aftermath.

Accident understanding is a focused but semantically demanding subset of traffic anomaly analysis. Whereas anomaly methods [7, 19] commonly score deviations from routine traffic dynamics or localize unusual segments, accident understanding [5,8,26] must identify interacting agents, reconstruct the pre-impact

sequence, classify the event, and infer visually supported causes, prevention measures, and outcomes. Because crashes are rare across diverse combinations of road geometry, weather, occlusion, and road-user behavior, they also exemplify the long-tail safety problem in autonomous driving [6, 27].

Chain-of-Thought (CoT) prompting [20, 22] can improve language-based reasoning, but a fluent reasoning trace does not guarantee that the model has observed the decisive visual cues. When subtle or transient hazards are missed, a VLM may still produce a plausible explanation by relying on linguistic priors. In traffic accidents, however, these subtle visual cues may determine the cause and progression of the event. Safety-critical reasoning therefore requires not only accurate predictions but also sensitivity to the visual evidence supporting them.

As illustrated in Fig. 1, Post-training with SFT/RL incurs substantial training cost, whereas multi-step reasoning increases test-time latency. We therefore investigate inference-time evidence routing for frozen VLMs. Our motivation is to study whether visual evidence can be constructed and routed at inference time without supervised fine-tuning or reinforcement learning in vision-language models. Our **SalMask Evidence Routing** (SalMask-ER) combines three inputs: uniformly sampled video frames for global context, a compact question-agnostic self-caption, and localized SalMask crops. The crops are obtained by detecting accident-related entities and applying driver-oriented saliency to suppress distracting background while retaining local context. The same routing interface supports multiple-choice VQA and dense accident captioning.

On VRU-Accident, SalMask-ER raises Qwen2.5-VL-3B VQA accuracy from 51.12% to 67.33% and reaches 71.30% with Qwen3.5-4B. SalMask also outperforms the unmasked RawCrop variant by 4.98 points. In the dense caption task comparison between papers, it reports higher scores than the published Qwen2.5-VL-3B baseline on all six metrics and exceeds VideoPerceiver-3B on five of six. Controlled evidence variations confirm that performance depends on the semantic content of the routed regions rather than merely on the presence of an additional image input.

Contributions.

- We introduce SalMask-ER, a training-free framework that routes global context, a compact self-caption, and object-saliency evidence to a frozen VLM for accident-video understanding.
- We demonstrate consistent VQA gains across three VLM backbones and competitive dense-captioning results without target-benchmark parameter updates.
- Through RawCrop ablation and evidence variations, we show that both entity localization and saliency-guided background attenuation contribute to accident-centric and causal reasoning.

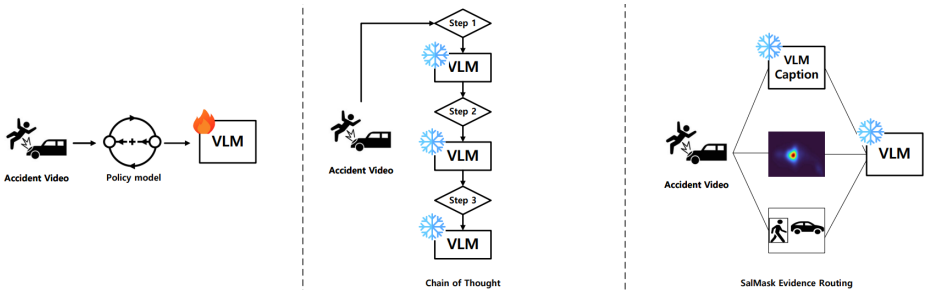

Fig. 1: Comparison of traffic accident reasoning paradigms. Policy-based methods use models trained through a combination of supervised fine-tuning (SFT) and reinforcement learning (RL), while textual Chain-of-Thought may generate reasoning steps without reliable visual grounding. SalMask-ER instead routes complementary object-centric, saliency-guided and self-generated caption evidence from the accident video to a frozen VLM before answer generation.

2 Related Work

2.1 VLMs and Accident-Scene Reasoning

VLMs [1, 11, 12] have been studied for traffic-rule reasoning, scene question answering, explanation, and planning-oriented driving analysis. NuScenes-QA [15], DriveLM [17], and LingoQA [14] address general driving scenes, whereas accident videos often depend on small, transient, or occluded cues. VRU-Accident [8] evaluates VQA and dense captioning for VRU-related crashes, and VideoPerceiver [25] focuses on fine-grained temporal perception of short accident events. Unlike parameter-updating approaches, SalMask-ER constructs a compact set of global and localized visual inputs before answer generation. Therefore, our work employs evidence routing to enhance the capability of vision–language models (VLMs) for accident videos.

2.2 Visual CoT and Grounded VLM Reasoning

Textual CoT [3, 20, 23] does not guarantee attention to the correct visual region. Visual-CoT [16] and tool-augmented methods therefore introduce object regions, spatial descriptions, or visual tools into reasoning. VTool-R1 [21], for example, trains a model to interleave language with visual tool use. SalMask-ER instead remains training-free and routes accident-relevant visual evidence before inference; controlled perturbations then test whether predictions depend on that evidence.

2.3 Object-Centric and Saliency-Guided Evidence

Open-vocabulary detectors such as Grounding DINO [13], and YOLO-World [4] are capable of localizing objects such as pedestrians, cyclists, vehicles, traffic

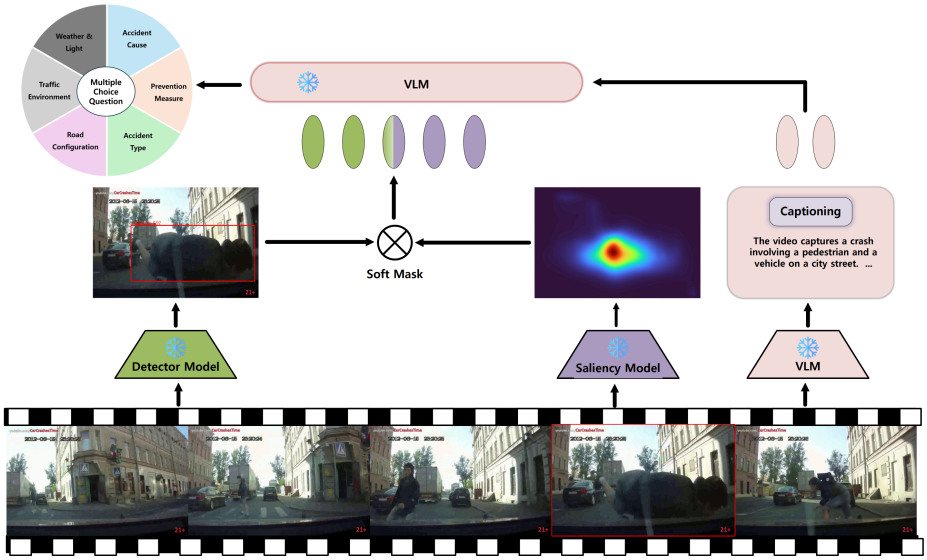

Fig. 2: Overview of SalMask-ER. A frozen detector and saliency model construct localized evidence from sampled frames. For VQA, the frozen VLM first generates a compact caption and then receives the video frames, caption, and SalMask evidence; dense captioning routes the frames and SalMask evidence directly.

lights, and road elements; however, the confidence scores of these detectors alone do not indicate their relevance to accidents. Driver attention models and saliency models [2,9,10,26] provide complementary saliency signals. By learning from the driving experience of drivers and other experts, these models gain an understanding of which elements in traffic scenes require attention and their relationships. SalM^2 [24] is a lightweight driver saliency model, although its graph does not explicitly encode object identities. SalMask-ER fuses these two signals to generate an evidence routing mechanism that can both identify objects and preserve context, while maintaining visual selectivity.

3 Method

The task of understanding road traffic accidents is challenging; it involves the spatial configurations and dynamic interactions of vulnerable road users, as well as factors such as weather conditions. Of particular importance is the identification and characterization of the visual and cognitive focus of attention of human drivers or computational models. As illustrated in Fig. 2, SalMask-ER routes three complementary signals to a frozen VLM: sampled video frames, a compact self-generated caption, and localized object–saliency evidence.

3.1 Global Context Extraction

Given a dashcam video $V = \{f_t\}_{t=1}^{T}$, we uniformly sample N frames,

$$V_s = \text{UniformSample}(V, N), \tag{1}$$

with $N = 8$ in the main experiments. These frames preserve weather, road geometry, traffic layout, and coarse temporal progression. In the first VQA pass, the frozen VLM generates a compact question-agnostic caption,

$$c_g = \text{VLM}(V_s, p_{\text{cap}}), \tag{2}$$

where p_{cap} is the caption prompt. The caption is generated once per video without access to the question or candidate answers, truncated to 60 words, and reused across all six questions. It summarizes global semantics but may omit small road users or localized pre-collision interactions.

3.2 SalMask Evidence Construction

For each selected detector frame, an open-vocabulary detector localizes traffic entities,

$$\mathcal{B}_t = \{(b_j, l_j, s_j)\}_{j=1}^{M_t}, \tag{3}$$

where b_j, l_j, and s_j denote the bounding box, label, and detector confidence. Candidate regions are ranked using detector confidence, a traffic-category prior, and a small preference for later observations. Label diversity is retained to reduce near-duplicate crops.

SalM^2 produces a normalized saliency map $S_t \in [0,1]^{H \times W}$. For a selected context-expanded crop I_j, we extract and resize the corresponding saliency region M_j. The soft-masked evidence is

$$e_j = I_j \odot \left[\beta + (1-\beta) M_j^{\gamma}\right], \tag{4}$$

where β preserves low-intensity context, γ controls saliency contrast, and $\odot$ denotes element-wise multiplication. We use $\beta = 0.35$ and $\gamma = 1.0$. Unlike a hard mask, SalMask emphasizes salient pixels without removing the surrounding spatial context. The top-K crops form $\mathcal{E} = \{e_i\}_{i=1}^{K}$.

3.3 Evidence-Routed VQA Inference

In the second pass, the VLM receives the sampled frames, compact caption, localized evidence, question, and candidate answers:

$$\hat{y} = \text{VLM}(V_s, c_g, \mathcal{E}, q, \mathcal{A}), \tag{5}$$

where $\mathcal{A} = \{A, B, C, D\}$. We use $K = 1$ because larger evidence sets introduced distracting local content in preliminary studies. The prompt requests a single option letter rather than a free-form CoT. SalMask-ER therefore changes the evidence routed to the model, not its parameters.

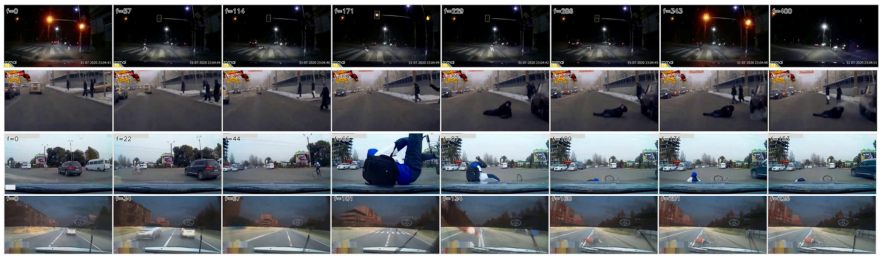

Fig. 3: Representative sampled frames from the VRU-Accident dataset.

3.4 Evidence-Routed Dense Accident Captioning

For dense captioning, the frozen VLM receives the sampled frames, SalMask evidence, and the official prompt p_{dc}:

$$\hat{c}_{\mathrm{dc}} = \mathrm{VLM}(V_s, \mathcal{E}, p_{\mathrm{dc}}). \tag{6}$$

The evaluated dense description $\hat{c}_{\mathrm{dc}}$ is distinct from the compact caption c_g, which is used only as VQA context. The comparisons in Tab. 4 use published baselines and are therefore treated as cross-paper.

4 Experiments

4.1 Datasets and Evaluation Protocol

VRU-Accident benchmark. We evaluate SalMask-ER on VRU-Accident [8], a benchmark for accident-centric video understanding involving vulnerable road users. Representative sampled frames from the dataset are shown in Fig. 3. It contains 1,000 real-world dashcam videos, 6,000 multiple-choice VQA samples, and one dense accident description per video. The six VQA categories are weather and lighting (WL), traffic environment or location (Loc.), road configuration (Road), accident type (Type), accident reason (Reason), and prevention method (Prev.). Each question contains one correct answer and three incorrect distractors. Dense captioning requires a temporally coherent and fine-grained description of the entire accident scene, covering the surrounding environment, relevant road users, vehicle motions, spatial interactions, collision progression, and final impact. The generated caption should remain faithful to the visual evidence and the information reflected in the VQA annotations, while integrating these cues into a more complete and detailed narrative.

VQA evaluation metrics. For a question category c, accuracy is

$$\mathrm{Acc}_c = \frac{N^c_{\mathrm{correct}}}{N^c_{\mathrm{total}}}, \tag{7}$$

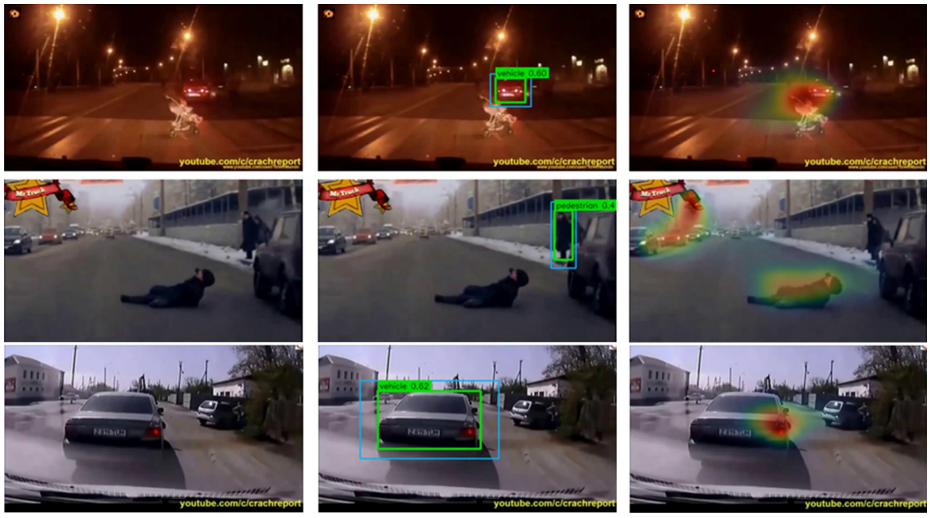

Fig. 4: Qualitative VRU-Accident examples. Each row shows the original frame, detector overlay, and saliency overlay. Object identity and driver-oriented saliency are combined to construct localized evidence for the VLM.

and overall accuracy is

$$\mathrm{Acc}_{\mathrm{overall}} = \frac{\sum_c N^c_{\mathrm{correct}}}{\sum_c N^c_{\mathrm{total}}}. \tag{8}$$

We additionally report

$$\mathrm{Critical} = \frac{\mathrm{Acc}_{Road} + \mathrm{Acc}_{Type} + \mathrm{Acc}_{Reason} + \mathrm{Acc}_{Prev}}{4}, \tag{9}$$

and

$$\mathrm{Causal} = \frac{\mathrm{Acc}_{Reason} + \mathrm{Acc}_{Prev}}{2}. \tag{10}$$

Critical emphasizes road context, accident type, cause, and prevention, while Causal isolates the two categories most directly related to defensive reasoning.

Dense-caption evaluation metrics. Following the official protocol [8], we evaluate 1,000 generated descriptions using SPICE, METEOR, COMET, and the F-measures of ROUGE-1, ROUGE-2, and ROUGE-L. Higher values indicate better performance. The official detailed-caption prompt and evaluation implementation are used for the reported SalMask configuration.

Models and implementation. We evaluate Qwen2.5-VL-3B, Qwen2.5-VL-7B, and Qwen3.5-4B in a zero-shot setting without target-task parameter updates. Unless stated otherwise, VQA follows the two-pass protocol in Sec. 3 with eight

Table 1: VRU-Accident VQA results (%). SalMask-ER routes the self-caption and local SalMask evidence together with the sampled video frames. Best results within each backbone group are bold.

Method	WL	Loc.	Road	Type	Reason	Prev.	Avg	Critical	Causal
Qwen2.5-VL-3B Video	67.80	80.00	40.30	44.00	25.10	49.50	51.12	39.73	37.30
Qwen2.5-VL-3B SalMask-ER	**68.70**	**86.30**	**61.00**	**74.20**	**55.80**	**58.00**	**67.33**	**62.25**	**56.90**
Qwen3.5-4B Video	61.80	78.60	54.20	68.70	52.90	57.60	62.30	58.35	55.25
Qwen3.5-4B SalMask-ER	**69.60**	**86.40**	**60.10**	**77.00**	**61.30**	**73.40**	**71.30**	**67.95**	**67.35**
Qwen2.5-VL-7B Video	64.80	79.50	40.00	58.80	31.40	48.90	53.90	44.77	40.15
Qwen2.5-VL-7B SalMask-ER	**66.00**	**89.60**	**54.70**	**74.70**	**59.70**	**64.40**	**68.18**	**63.38**	**62.05**

uniformly sampled frames and one SalMask crop; the answer pass is constrained to return one option letter. Dense captioning uses Qwen2.5-VL-3B with the official prompt. The published Qwen2.5-VL-3B baseline in Tab. 4 was not reproduced under our exact checkpoint, sampling, prompting, decoding, or software configuration; the comparison is therefore cross-paper.

Grounding DINO uses fixed prompts for pedestrians, cyclists, vehicles, crosswalks, and traffic lights on four uniformly sampled frames, with box and text thresholds of 0.25 and 0.20. SalM^2 generates saliency maps from eight frames resized to 256×256 and scaled to $[0, 1]$. Each detected entity is aligned with the temporally closest saliency map. Video frames and evidence crops are resized with maximum side lengths of 448 and 224 pixels, respectively, while preserving aspect ratio. Experiments are implemented in PyTorch on 4 NVIDIA RTX 3090 GPUs without gradient updates; sampling, answer extraction, and generation settings are fixed across evidence configurations.

5 Results and Analysis

Main results on VRU-Accident. Table 1 shows consistent gains across all three backbones. For Qwen2.5-VL-3B, SalMask-ER improves average accuracy from 51.12% to 67.33%. The largest category gains occur in Accident Type (44.00% to 74.20%) and Accident Reason (25.10% to 55.80%), where small interacting agents and causal cues are especially important. Qwen2.5-VL-7B improves from 53.90% to 68.18%, while Qwen3.5-4B achieves the best overall, Critical, and Causal scores of 71.30%, 67.95%, and 67.35%, respectively.

Comparison with video-specialized VideoPerceiver. In Table 4, Qwen2.5-VL-3B + SalMask reported higher scores than the published Qwen2.5-VL-3B baseline model across all six metrics; compared with VideoPerceiver-3B, SalMask scored 0.001 lower on the SPICE metric but performed better on the remaining five metrics. These results suggest that training-free evidence routing can remain competitive with specialized video models, while the cross-paper setting prevents a strictly controlled comparison.Gemini 1.5 Flash remains the strongest overall among closed-source vision-language large models. The residual gap on

Table 2: Compact comparison with VideoPerceiver on VRU-Accident. "Specialized training" indicates additional SFT and RL for fine-grained action and transient-event perception.

Method	Specialized training	Avg Acc. (%)
VideoPerceiver-3B [25]	Yes	66.40
Qwen2.5-VL-3B SalMask-ER (Ours)	No	67.33
Qwen3.5-4B SalMask-ER (Ours)	**No**	**71.30**
VideoPerceiver-7B [25]	Yes	73.70

Table 3: Ablation study of global and localized visual evidence on VRU-Accident using Qwen2.5-VL-3B. Δ Avg. is measured relative to Method A.

Method	Video	SelfCap	RawCrop	SalMask	Avg.	Critical	Causal	Δ Avg.
A	✓				51.12	39.73	37.30	–
B	✓	✓			59.27	48.42	45.75	+8.15
C	✓	✓	✓		62.35	56.98	53.55	+11.23
D	✓	✓		✓	67.33	62.25	56.90	+16.21

the SPICE metric suggests that detailed captions still omit some fine-grained objects, attributes or relationships.

RawCrop versus SalMask. Table 3 separates global caption context, entity localization, and saliency-guided attenuation. SelfCap improves the video-only baseline by 8.15 points, and RawCrop adds 3.08 points to reach 62.35%. SalMask reaches 67.33%, a further 4.98-point gain over RawCrop. Across 6,000 questions, the two variants produce different labels on 660 samples; SalMask corrects 405 cases that RawCrop misses, while the reverse occurs in 106 cases, yielding a net gain of 299 correct predictions. Thus, the benefit is not explained by cropping alone: attenuating distracting content while preserving local context materially improves the routed evidence.

Dense-caption results. In Table 4, Qwen2.5-VL-3B + SalMask reports higher scores than the published Qwen2.5-VL-3B baseline on all six metrics, but the difference should not be interpreted as an isolated SalMask effect because the baseline was not reproduced under the same inference configuration. Relative to VideoPerceiver-3B, SalMask is 0.001 lower in SPICE and higher on the other five metrics. Gemini 1.5 Flash remains strongest overall. The residual SPICE gap suggests that detailed captions still miss some fine-grained objects, attributes, or relations.

Sensitivity to visual evidence. Table 5 tests whether gains arise from the semantic content of SalMask rather than from an additional image slot. Replacing the

Table 4: Dense-captioning results on VRU-Accident. R1-F, R2-F, and RL-F denote the F-measures of ROUGE-1, ROUGE-2, and ROUGE-L. † denotes additional video-specific SFT/RL and ‡ denotes a closed-source model. Best and second-best results are bold and underlined, respectively.

Model	SPICE ↑	METEOR ↑	COMET ↑	R1-F ↑	R2-F ↑	RL-F ↑
Qwen2.5-VL-3B [8]	0.149	0.261	0.682	0.416	0.111	0.218
VideoPerceiver-3B† [25]	<u>0.164</u>	0.267	0.707	0.449	0.129	0.249
Qwen2.5-VL-3B + SalMask (Ours)	0.163	<u>0.277</u>	<u>0.725</u>	<u>0.455</u>	<u>0.135</u>	<u>0.253</u>
Gemini 1.5 Flash‡ [18]	**0.194**	**0.285**	**0.741**	**0.481**	**0.148**	**0.266**

Table 5: Evidence perturbation with Qwen2.5-VL-7B on VRU-Accident. Δ is relative to Original SalMask; Changed is the percentage of predictions changed by each perturbation.

Evidence	Avg	Critical	Causal	Δ Avg	Changed
SalMask	68.18	63.38	62.05	–	–
Empty	65.70	60.80	58.15	−2.48	8.30
Black	66.50	61.20	59.90	−1.68	5.33
Random	64.28	59.35	57.85	−3.90	5.82

original evidence with Empty, Black, or Random inputs lowers average accuracy by 2.48, 1.68, and 3.90 points, respectively; the corresponding Causal drops are 3.90, 2.15, and 4.20 points. The perturbations change 8.30%, 5.33%, and 5.82% of predictions. Random evidence changes fewer answers than Empty evidence but causes the largest accuracy loss, indicating that mismatched visual content can be more harmful than missing content. These results support the claim that SalMask-ER benefits from the specific evidence it routes.

6 Conclusion

We presented SalMask-ER, a training-free framework that routes global video context, a compact self-caption, and localized object–saliency evidence to frozen VLMs for accident understanding. On VRU-Accident VQA, it improves Qwen2.5-VL-3B from 51.12% to 67.33% and reaches 71.30% with Qwen3.5-4B. SalMask also exceeds RawCrop by 4.98 points, showing that saliency-guided background attenuation contributes beyond entity localization. Cross-paper dense-caption results are competitive with VideoPerceiver-3B, and evidence perturbations suggest that predictions are sensitive to the semantic content of the localized input.

The framework remains limited by detector recall, saliency quality, temporal sampling, and caption fidelity. Future work should route multiple temporally linked entities, evaluate pre-collision prediction, incorporate uncertainty-aware grounding checks, and connect accident understanding more directly to defensive action selection.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bao, W., Yu, Q., Kong, Y.: Deep reinforced accident anticipation with visual explanation. In: International Conference on Computer Vision (ICCV) (2021)
3. Chen, X., Zhang, R., Jiang, D., Zhou, A., Yan, S., Lin, W., Li, H.: Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. Advances in neural information processing systems **38**, 69110–69139 (2026)
4. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16901–16911 (2024)
5. Fang, J., Li, L.l., Zhou, J., Xiao, J., Yu, H., Lv, C., Xue, J., Chua, T.S.: Abductive ego-view accident video understanding for safe driving perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22030–22040 (June 2024)
6. Fang, J., Qiao, J., Xue, J., Li, Z.: Vision-based traffic accident detection and anticipation: A survey. IEEE Transactions on Circuits and Systems for Video Technology **34**(4), 1983–1999 (2023)
7. Khan, S.W., Hafeez, Q., Khalid, M.I., Alroobaea, R., Hussain, S., Iqbal, J., Almotiri, J., Ullah, S.S.: Anomaly detection in traffic surveillance videos using deep learning. Sensors **22**(17), 6563 (2022)
8. Kim, Y., Abdelrahman, A.S., Abdel-Aty, M.: Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 761–771 (2025)
9. Kotseruba, I., Tsotsos, J.K.: SCOUT+: Towards practical task-driven drivers' gaze prediction. In: IV (2024)
10. Kotseruba, I., Tsotsos, J.K.: Understanding and modeling the effects of task and context on drivers' gaze allocation. In: IV (2024)
11. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
12. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
13. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., 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)
14. Marcu, A.M., Chen, L., Hünermann, J., Karnsund, A., Hanotte, B., Chidananda, P., Nair, S., Badrinarayanan, V., Kendall, A., Shotton, J., Sinavski, O.: Lingoqa: Visual question answering for autonomous driving. arXiv preprint arXiv:2312.14115 (2023)
15. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. arXiv preprint arXiv:2305.14836 (2023)
16. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024)

17. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. arXiv preprint arXiv:2312.14150 (2023)
18. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
19. Tran, T.M., Bui, D.C., Nguyen, T.V., Nguyen, K.: Transformer-based spatio-temporal unsupervised traffic anomaly detection in aerial videos. IEEE Transactions on Circuits and Systems for Video Technology **34**(9), 8292–8309 (2024)
20. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
21. Wu, M., Yang, J., Jiang, J., Li, M., Yan, K., Yu, H., Zhang, M., Zhai, C., Nahrstedt, K.: Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. arXiv preprint arXiv:2505.19255 (2025)
22. Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., Yang, Y., Pang, R., Yang, Y.: Improve vision language model chain-of-thought reasoning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1631–1662 (2025)
23. Zhang, Y., Gao, S., Huang, Y., Yu, Z., Tan, K.: 3a-cot: an attend-arrange-abstract chain-of-thought for multi-document summarization. International Journal of Machine Learning and Cybernetics **16**(12), 9753–9771 (2025)
24. Zhao, C., Mu, W., Zhou, X., Liu, W., Yan, F., Deng, T.: Salm^2: An extremely lightweight saliency mamba model for real-time cognitive awareness of driver attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1647–1655 (2025)
25. Zhao, F., Zhang, L., Shi, D., Gao, Y., Ye, C., Cai, Y., Gao, J., Yan, D.: Videoperceiver: Enhancing fine-grained temporal perception in video multimodal large language models. arXiv preprint arXiv:2511.18823 (2025)
26. Zhou, Y., Tang, J., Xiao, X., Lin, Y., Liu, L., Guo, Z., Fei, H., Xia, X., Gou, C.: Where, what, why: Towards explainable driver attention prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2675–2685 (2025)
27. Zimmer, W., Greer, R., Zhou, X., Song, R., Pavel, M., Lehmberg, D., Ghita, A., Gopalkrishnan, A., Trivedi, M., Knoll, A.: Safety-critical learning for long-tail events: The tum traffic accident dataset. arXiv preprint arXiv:2508.14567 (2025)