# VLM-Driven Proactive Escalation for Tele-operations in Autonomous Vehicles

Anonymous CVPR submission

Paper ID *****

## Abstract

*Recent advances in autonomous driving enable vehicles to operate in increasingly complex environments, yet safe deployment still requires timely human intervention when the autonomous system encounters ambiguous or safety-critical situations. Existing remote driving systems rely on heuristic or confidence-based triggers to request assistance, which lack semantic understanding of driving context and often cause over- or under-escalation. We propose a vision–language reasoning-driven escalation framework that enables autonomous vehicles to request remote driving assistance based on context-aware semantic reasoning rather than numerical uncertainty. Our approach uses a vision–language model (VLM) aligned using Direct Preference Optimization (DPO) on pairwise operator preferences to calibrate escalation decisions and reduce unnecessary or missed escalations. The model interprets visual observations and driving context to generate structured explanations that help operators quickly understand and respond to situations. We evaluate the approach using NVIDIA Cosmos-2 across diverse uncertainty scenarios, including intersection negotiation and dynamic obstacle ambiguity. Results show improved intervention efficiency, reduced unnecessary operator engagement, and enhanced safety compared to heuristic triggering, highlighting a scalable pathway toward reliable human-in-the-loop autonomous driving.*

## 1. Introduction

Level 4 automated vehicles (AVs) have been progressively deployed in both commercial and research settings in recent years. However, despite significant advances in perception, planning, and control, AVs still encounter situations that cannot be reliably resolved by automation alone. These situations often lead to disengagements, where the automated driving system (ADS) must be deactivated, preventing mission completion or safe continuation of the driving task. Such failures typically arise in long-tail scenarios, complex social interactions with other road users, or context-dependent decision-making that exceed the capabilities of rule-based or confidence-driven autonomy [24, 34]. As a result, remote driving and teleoperation have emerged as critical human-in-the-loop mechanisms for ensuring safety, reliability, and operational continuity in real-world AV deployments [1, 10, 11, 21]. Beyond serving as a fallback control interface, remote operators can interpret semantically ambiguous traffic situations and provide guidance or direct control when autonomy becomes uncertain. In practical deployments such as fleet-based autonomous mobility services, logistics vehicles, and supervised AV operations, timely human intervention becomes particularly important when a limited number of operators must oversee multiple vehicles under constraints of bandwidth, latency, and attention [5, 7, 12].

In large-scale deployments of autonomous vehicles, particularly in robotaxi and autonomous mobility services, a fundamental challenge lies in determining when an automated vehicle should request remote operator assistance and ensuring that such requests are accompanied by clear and interpretable contextual information. Autonomous driving remains inherently difficult due to dynamic interactions with other road users, evolving traffic patterns, environmental uncertainty, and rare long-tail events that cannot be reliably handled by pre-programmed behaviors or statistical confidence estimates. Conventional autonomous driving stacks, typically composed of perception, prediction, and planning modules, are prone to cascading errors and limited robustness when confronted with unfamiliar or socially complex traffic situations [15, 32]. As a result, automated driving performance may degrade for multiple reasons, broadly including perception ambiguity, planning uncertainty, and control-level difficulties, each requiring different levels of urgency and types of remote assistance. In fleet-scale robotaxi operations, requesting intervention too frequently can overwhelm remote operators and reduce system efficiency, whereas delayed escalation may lead to safety risks or service disruptions. Existing remote driving and tele-assistance frameworks often rely on heuristic rules, confidence thresholds, or handcrafted triggers to determine when a vehicle should request human intervention [11, 28].

However, these mechanisms lack semantic understanding of the surrounding driving context and are therefore limited in their ability to prioritize assistance requests or communicate the underlying reason for escalation [25].

Motivated by these challenges and emerging needs, to the best of our knowledge, we are the first to introduce a VLM-based reasoning framework for proactive escalation of remote driving requests. Our key contributions are summarized as follows:

**Reasoning-Driven Escalation Framework:** We introduce a vision–language reasoning-based escalation framework that enables autonomous vehicles to proactively request remote driving assistance using context-aware semantic understanding, moving beyond heuristic and confidence-based triggers.

**Preference-Aligned Decision Calibration via DPO:** We develop a task-conditioned post-training strategy using Direct Preference Optimization (DPO) to align the VLM with operator preferences, improving escalation accuracy and reducing both over- and under-intervention.

**Diverse Scenario Generation and Evaluation:** We address dataset limitations by generating and evaluating diverse driving scenarios using NVIDIA Cosmos-2, enabling comprehensive assessment under varied uncertainty conditions.

## 2. Related Work

**Vision-Language Models in Autonomous Driving.** The integration of vision-language models (VLMs) into autonomous driving has recently gained significant traction [15]. Recent works such as DriveVLM [33] and DriveLM [30] use large vision-language models to interpret traffic scenes, perform structured reasoning, and support driving-related decision making or explanation generation [30, 33, 35]. Related multimodal driving frameworks such as DriveLLaVA and Dolphins further show that language-guided models can support behavior-level driving decisions and scene understanding in complex road environments [22, 35]. More broadly, embodied multimodal systems such as PaLM-E and RT-2 demonstrate that large models can connect visual perception, language reasoning, and action-oriented outputs in real-world decision-making settings [3, 8]. Collectively, these studies suggest that VLMs can contribute not only to perception enhancement, but also to higher-level semantic interpretation and human-interpretable reasoning in autonomous systems [3, 8, 15].

This capability is particularly relevant in driving because some challenging situations arise not only from sensor noise or low confidence, but from ambiguity in scene semantics, intent, and interaction context [22, 30, 33]. Examples include unusual construction layouts, temporary traffic control, and other long-tail scenarios where purely rule-based or perception-only reasoning may be insufficient [15, 30, 33]. However, most prior VLM-based driving systems are designed to improve the autonomy stack itself, such as through scene understanding, driving assistance, decision making, or planning [18, 22, 26, 30, 33, 35]. In contrast, our work focuses specifically on the human-robot interaction layer, using VLM-based reasoning not to replace vehicle control, but to determine when autonomy should request remote human assistance. This framing shifts the role of the VLM from direct driving support to context-aware, explainable escalation for teleoperation.

**Teleoperation and Remote Driving.** Teleoperation remains an important fallback mechanism for autonomous vehicle deployment, particularly when the onboard system encounters situations it cannot safely or reliably resolve on its own [6, 20]. Prior work on remote driving and remote assistance has largely focused on communication latency, predictive displays, interface design, and operator performance, reflecting the practical difficulty of supervising or controlling a vehicle from a distance [5, 23, 32]. More recent studies also emphasize human-centered workstation design and operator situation awareness as critical requirements for safe and effective remote assistance [29]. At the system level, latency mitigation and communication robustness continue to be treated as major technical bottlenecks for scalable teleoperation [11, 16, 20].

However, the question of *when* a vehicle should request teleoperation is often treated more simply than the question of *how* teleoperation should be delivered. In many existing systems, escalation is tied to heuristic triggers such as low confidence, low speed, operational design domain boundaries, or predefined remote-assistance workflows [11, 28, 29]. While such rules are practical for deployment, they do not fully address semantically ambiguous situations in which the challenge lies not only in sensing failure, but in interpreting intent, unusual roadway structure, or nonstandard human signaling [2, 31]. Examples include temporary construction control, uncommon edge-case traffic behavior, and traffic-directing gestures from pedestrians or police officers, all of which can challenge automated systems even when no clear numerical failure signal is present [2, 31].

**Reasoning Distillation.** Aligning smaller, deployable models via teacher supervision has evolved beyond simple label imitation [14]. Recent works demonstrate the efficacy of Chain-of-Thought (CoT) distillation and reasoning supervision, where student models are trained to output the intermediate logic traces generated by large teacher models [13, 17]. Our framework relies heavily on this reasoning distillation paradigm to transfer complex driving context interpretation into a lightweight model.

**Preference Alignment and Safety.** Aligning language models with human preferences through Supervised Fine-Tuning (SFT) and methods such as Direct Preference Op-

timization (DPO) has become a common paradigm for improving model behavior in instruction-following systems [27]. DPO is particularly attractive because it optimizes preferred outputs over rejected ones without requiring an explicit reward model [27]. However, applying DPO to safety-critical autonomous driving introduces additional challenges, especially under severe class imbalance [9]. In remote driving, truly safety-critical handoff events are rare compared with routine situations, which can bias standard preference optimization toward under-escalation in high-risk cases [9]. To address this, we build on standard DPO with an asymmetric, safety-aware loss that places greater penalty on misclassified edge cases. This design is motivated by the same principle as focal loss, which emphasizes hard and underrepresented examples during training [19], but we adapt it to pairwise preference learning for escalation alignment.

## 3. Problem Statement

We formulate remote driving request escalation as a discrete three-class triage problem, moving beyond heuristic or confidence-based approaches that lack contextual understanding and typically trigger intervention only after failures occur. Given a camera observation $\mathcal{X}$, a vision–language model (VLM) predicts an escalation level $y \in$ AUTONOMY, SUPERVISION, TELEOPERATION, along with a natural-language reasoning trace $\mathcal{R}$ that explains the decision. Unlike reactive methods, our reasoning-driven approach enables proactive escalation before potential incidents arise and provides context-aware guidance, allowing operators to select appropriate actions based on the evolving driving scenario.

We define these triage levels according to the degree of human involvement required for safe operation:

- **AUTONOMY:** The AV can safely handle the scene without operator involvement. This includes routine but non-trivial driving situations such as following vehicles, navigating normal lane traffic, and yielding to pedestrians at crosswalks.
- **SUPERVISION:** The AV can continue operating, but environmental degradation or scene complexity may reduce the reliability of perception or planning. The remote operator should monitor the situation closely, although immediate intervention is not yet necessary.
- **TELEOPERATION:** The AV has reached its operational limits and cannot safely proceed without direct human control. The remote operator must assume control immediately. Examples include road blockages, police-directed traffic overriding signals, or large animals obstructing the lane.

A teleoperation center must oversee $N$ simultaneous AV feeds. Consequently, the system must process $N$ visual streams under strict latency and compute constraints, precluding the use of VLMs on every feed. Our objective is therefore to train a lightweight VLM policy $\pi_\theta(y, \mathcal{R} \mid x)$ that reliably identifies TELEOPERATION cases while maintaining a low false-positive rate on AUTONOMY scenes.

## 4. Methodology

### 4.1. System Overview

operates as a modular monitoring layer at the teleoperation centre, entirely decoupled from the vehicle's onboard autonomy stack. Each autonomous vehicle streams its front-camera feed to the centre, where a lightweight student VLM (Qwen3-VL-2B-Instruct) processes incoming frames and produces a structured triage assessment. This architecture ensures that the monitoring system neither competes for onboard compute nor introduces a single point of failure into the driving pipeline: if the triage VLM fails, the AV continues operating under its native autonomy stack while the operator retains manual monitoring capability. The student model is trained through a two-stage pipeline comprising (i) preference data generation via a large teacher VLM, followed by (ii) safety-aware preference alignment, described in detail below.

### 4.2. Dataset Construction

Because no publicly available datasets explicitly label scenes for Level-4 teleoperation triage, we construct a custom preference alignment dataset from the nuScenes driving corpus.

**Edge-case biased sampling.** We sample approximately 9,500 camera keyframes from nuScenes [4], with an additional 1000 frames reserved for held-out evaluation. Rather than sampling uniformly, we apply a priority-biased strategy that deliberately over-represents semantically challenging scenes. Specifically, we assign higher sampling probability to frames containing pedestrians near the roadway, construction vehicles or traffic cones, complex multi-way intersections, and scenes captured under adverse weather or lighting conditions (rain, fog, heavy glare). This strategy surfaces the long-tail scenarios most likely to cause perception or planning failures, precisely the situations where correct triage is critical.

**Two-stage annotation pipeline.** Ground-truth reasoning traces and triage labels are generated through a two-stage process designed to balance annotation throughput with label quality:

1. **Machine annotation:** Cosmos Reason 2 8B, a large-scale VLM post-trained on driving data with physical reasoning capabilities, processes each of the

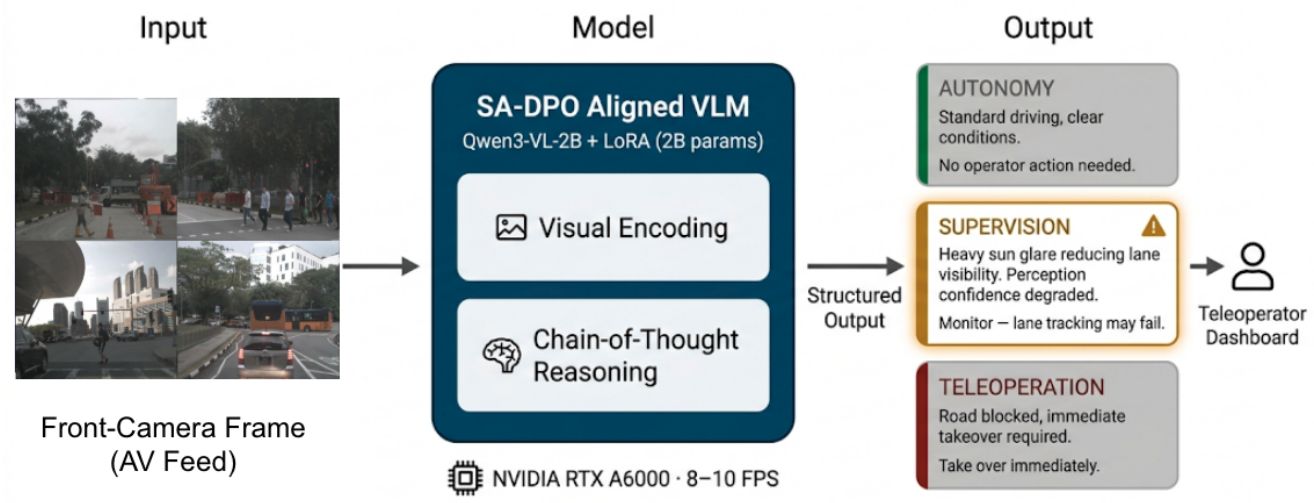


Figure 1. Architecture Diagram

9,500 frames. For every frame, the model generates a detailed Chain-of-Thought (CoT) reasoning trace analysing the driving context identifying relevant actors, environmental conditions, and potential hazards followed by a final triage label $y \in \{\text{AUTONOMY}, \text{SUPERVISION}, \text{TELEOPERATION}\}$.

2. **Human-in-the-loop verification:** Human annotators review and, where necessary, correct the machine-generated labels. This verification pass is particularly important for ambiguous edge cases where the boundary between SUPERVISION and TELEOPERATION is context-dependent (e.g., a construction zone with partial lane blockage).

This curated data constitutes the "chosen" preference tier $y_w$. To construct the "rejected" tier $y_l$, we run the pre-trained, unaligned student model (Qwen3-VL-2B-Instruct) on the same 9,500 frames, producing baseline zero-shot responses. Together, these form the preference dataset $\mathcal{D} = \{(x_i, y_{w,i}, y_{l,i})\}_{i=1}^{9500}$ required for DPO training.

**Observed class distribution.** After edge-case biased sampling, the resulting dataset exhibits the following class frequencies: approximately 49% AUTONOMY, 38% SUPERVISION, and 13% TELEOPERATION. While our biased sampling strategy deliberately enriches the representation of safety-critical scenes to provide sufficient training signal, in uncurated AV logs TELEOPERATION-level events occur in fewer than 1–2% of frames. This deployment-time class imbalance is precisely what motivates the asymmetric loss formulation described in Section 4.4.

### 4.3. Teacher-Student Distillation Pipeline

Our framework employs reasoning distillation to transfer domain-specific driving knowledge from a large, capable teacher into a compact student model suitable for multi-feed deployment.

**Teacher model.** We utilise as the teacher. Unlike general-purpose VLMs, has been post-trained on embodied driving scenarios and exhibits strong physical reasoning capabilities—for instance, understanding that a police officer's hand signal overrides a green traffic light, or that heavy solar glare may render lane markings invisible to the vehicle's perception system. These nuanced judgments are encoded in the teacher's CoT reasoning traces, which serve as rich supervision signals beyond simple label matching.

**Student model and LoRA fine-tuning.** The student model, Qwen3-VL-2B-Instruct, is a 2-billion parameter VLM chosen specifically for its favourable compute-to-quality trade-off in multi-feed deployment scenarios. Rather than fine-tuning all model parameters, we apply Low-Rank Adaptation (LoRA) to the attention projection matrices, adding fewer than 0.5% additional trainable parameters. LoRA enables rapid adaptation with modest GPU memory requirements while preserving the student's pre-trained visual and linguistic representations. During training, the LoRA weights are optimised via SA-DPO (described below) while the base model weights remain frozen. At inference, the LoRA adapters are merged back into the base model, incurring zero additional latency.

**Preference pair construction.** For each training frame $x_i$, the preference pair $(y_{w,i}, y_{l,i})$ captures the contrast between expert-level reasoning (teacher) and naïve baseline reasoning (student). The "chosen" response $y_{w,i}$ contains the teacher's full CoT trace followed by the correct triage

label and a structured action recommendation, while the "rejected" response $y_{l,i}$ is the student's unaligned zero-shot output on the same frame. This pairing teaches the student not only *what* the correct triage should be, but *why*—transferring the semantic reasoning process rather than merely imitating labels.

### 4.4. Safety-Aware DPO (SA-DPO)

Direct Preference Optimization (DPO) aligns language model outputs with human preferences by optimising the policy to increase the likelihood of preferred responses relative to dispreferred ones. The standard DPO loss is:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left( \beta \left( \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right) \tag{1}$$

where $\pi_\theta$ is the policy being trained, $\pi_{\text{ref}}$ is the frozen reference policy, $\beta$ controls the deviation from the reference, and $\sigma$ is the logistic function. While effective for general alignment, standard DPO assigns equal importance to all preference pairs regardless of the safety implications of misclassification.

**The under-escalation problem.** Although our edge-case biased sampling enriches TELEOPERATION representation in the training set, AUTONOMY and SUPERVISION frames still constitute the vast majority of preference pairs. Because DPO aggregates gradients across all pairs, the loss landscape is dominated by these routine examples. Consequently, a model trained with standard DPO learns to systematically under-escalate—predicting AUTONOMY or SUPERVISION even when the scene requires immediate human takeover because the rare under-escalation errors are "drowned out" by the frequent correct classifications on routine scenes. This problem is further amplified at deployment, where TELEOPERATION-level events occur in fewer than 1–2% of frames. The behaviour is analogous to the class-imbalance problem in object detection that motivated Focal Loss, but manifests in the preference-learning domain.

**Asymmetric safety weighting.** To address this, we propose Safety-Aware DPO (SA-DPO), which introduces a pair-dependent weighting function $w(y_w, y_l)$:

$$\mathcal{L}_{\text{SA-DPO}} = -w(y_w, y_l) \log \sigma \left( \beta \cdot \Delta_\theta(x, y_w, y_l) \right) \tag{2}$$

where the preference margin is:

$$\Delta_\theta(x, y_w, y_l) = \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \tag{3}$$

The weighting function encodes the asymmetric safety cost of different misclassification types:

$$w(y_w, y_l) = \begin{cases} \alpha_{\text{high}} & \text{if } y_w = \text{TELEOP.},\ y_l = \text{AUTONOMY} \\ \alpha_{\text{med}} & \text{if } y_w = \text{TELEOP.},\ y_l = \text{SUPERVISION} \\ 1.0 & \text{otherwise} \end{cases} \tag{4}$$

The most dangerous error—classifying a mandatory TELEOPERATION scene as AUTONOMY—receives the highest penalty ($\alpha_{\text{high}} = 5.0$). A less severe but still problematic error—downgrading TELEOPERATION to SUPERVISION—receives a moderate penalty ($\alpha_{\text{med}} = 2.0$). All other pairs, including over-escalation (e.g., AUTONOMY classified as TELEOPERATION), are weighted at 1.0, reflecting the operational reality that false alarms are inconvenient but not safety-critical.

**Key advantages over alternative approaches.** Unlike data-level balancing strategies (e.g., oversampling TELEOPERATION frames), SA-DPO preserves the natural class distribution during training, ensuring the model does not develop an artificially inflated prior for rare events. Unlike standard class-weighted classification losses that operate on individual predictions, SA-DPO operates on *pairwise preference comparisons between full reasoning trajectories*, allowing the model to learn the subtle semantic differences between scenes that require escalation and those that do not. For instance, the model learns that "a police officer standing on the sidewalk" (AUTONOMY) and "a police officer stepping into the roadway with a raised hand" (TELEOPERATION) differ by a contextual detail that a simple label-weighted classifier would struggle to capture.

### 4.5. Structured Output for Teleoperator Interface

The model produces a structured natural-language explanation alongside the triage decision, designed for rapid comprehension by teleoperators monitoring multiple feeds simultaneously. Each output consists of three fields: (i) the triage level (AUTONOMY, SUPERVISION, or TELEOPERATION), (ii) a concise reasoning trace (1–2 sentences) explaining the contextual factors driving the decision, and (iii) a recommended operator action. This structured format enables operators to assess an alert in under two seconds without interpreting abstract confidence scores or bounding-box overlays, significantly reducing cognitive load during multi-feed monitoring.

## 5. Evaluation and Discussion

### 5.1. Experimental Setup

We evaluate on a curated benchmark of 1,000 front-camera frames sampled from the nuScenes dataset. Ground-truth triage labels are generated by and verified through human-in-the-loop review, yielding an imbalanced but realistic dis-

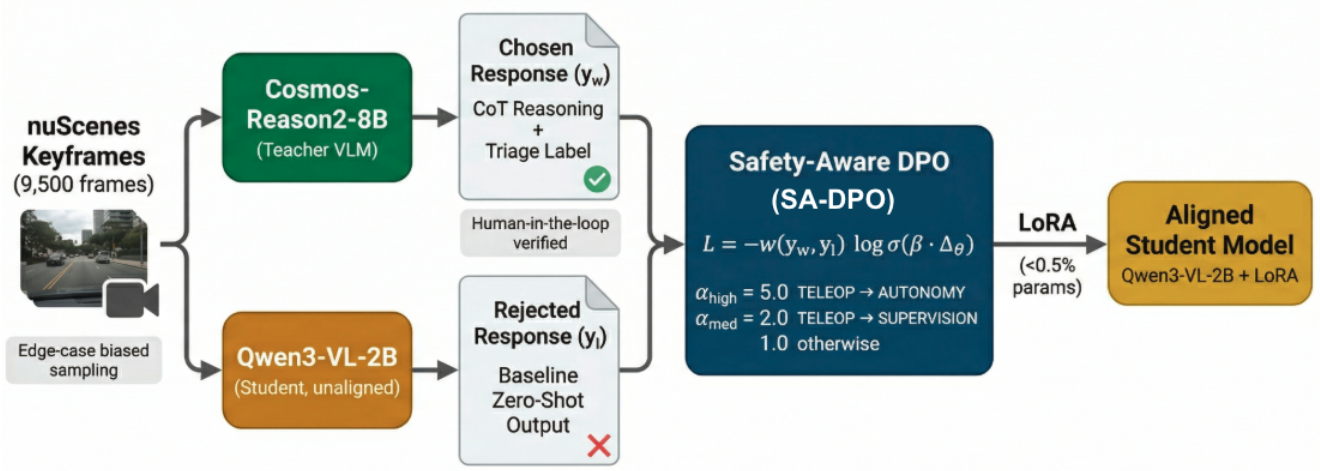


Figure 2. Data Generation and Training Pipeline

Table 1. In-domain nuScenes evaluation ($N$=1,000). Under-Escalation Rate (UE↓) is lower-is-better.

| Method | Acc. | F1 | UE↓ |
|---|---|---|---|
| Baseline (Zero-Shot) | 0.475 | 0.292 | 71.0% |
| Standard DPO | 0.730 | 0.624 | 37.9% |
| SA-DPO (Ours) | **0.921** | **0.907** | **2.1%** |

tribution: 467 AUTONOMY, 388 SUPERVISION, and 145 TELEOPERATION frames. We compare three systems:

1. **Baseline (Zero-Shot)**: The pre-trained without any preference alignment, representing a raw VLM applied directly to the triage task.
2. **Standard DPO**: aligned with the conventional DPO loss using equal weighting for all preference pairs.
3. **SA-DPO** (): Our proposed model aligned with the safety-aware asymmetric loss that up-weights under-escalation penalties.

All inference is performed using vLLM on a single NVIDIA RTX A6000 (48 GB).

### 5.2. Triage Classification Results

We report Overall Accuracy, Macro-F1, and the safety-critical *Under-Escalation Rate* (UE↓), defined as the fraction of true TELEOPERATION events misclassified as AUTONOMY.

As shown in Table 1, SA-DPO achieves 92.1% overall accuracy and a Macro-F1 of 0.907, substantially outperforming both the zero-shot baseline (47.5%, 0.292) and standard DPO (73.0%, 0.624). Most critically, SA-DPO reduces the under-escalation rate to **2.1%**—missing only 3 out of 145 true TELEOPERATION events—whereas Standard DPO still misses 37.9% and the unaligned baseline misses 71.0%. The 42.4 percentage-point accuracy gap between SA-DPO and the zero-shot baseline underscores the necessity of preference alignment: without it, the VLM defaults to predicting AUTONOMY for nearly all inputs, rendering it unsafe for deployment. Notably, the jump from Standard DPO to SA-DPO (+19.1% accuracy, +0.283 F1) is driven almost entirely by the asymmetric loss weighting, confirming that safety-aware reweighting is essential for learning from imbalanced triage distributions.

**Per-class recall analysis.** Figure 4 reports recall broken down by triage level. The zero-shot baseline achieves moderate AUTONOMY recall (90.6%) by predominantly predicting the majority class, but catastrophically fails on SUPERVISION (12.4%) and TELEOPERATION (2.8%). Standard DPO improves SUPERVISION recall to 79.1% but still captures only 24.8% of TELEOPERATION events. SA-DPO achieves the highest recall across all three classes—94.4% AUTONOMY, 89.2% SUPERVISION, and 92.4% TELEOPERATION—demonstrating that the asymmetric loss successfully shifts the model's attention toward safety-critical scenes without sacrificing performance on routine ones.

Figure 3 presents confusion matrices for all three models. The escalation-error trade-off is shown in Figure 5: SA-DPO reduces UE to just 2.1% with only 1.5% over-escalation, a favourable safety profile where the residual false alarms are far less costly than missed interventions.

### 5.3. Multi-Feed Throughput Analysis

A key operational requirement of is its ability to monitor multiple AV feeds concurrently without competing for on-board vehicle compute. We deploy the SA-DPO–aligned via vLLM on a single NVIDIA RTX A6000 GPU (48 GB VRAM). In our benchmarking, the 2B-parameter model

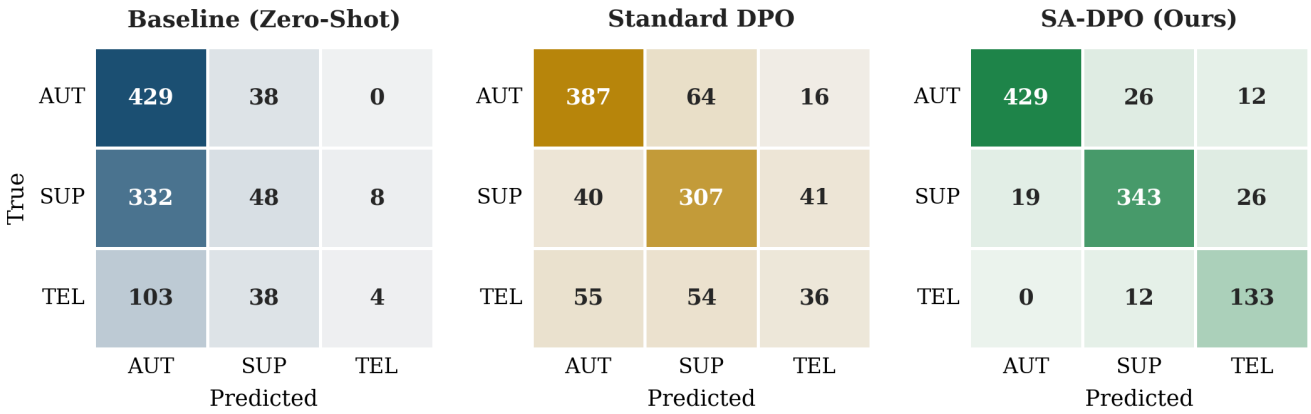


Figure 3. Confusion matrices for the three evaluated models on the 1,000-frame nuScenes benchmark

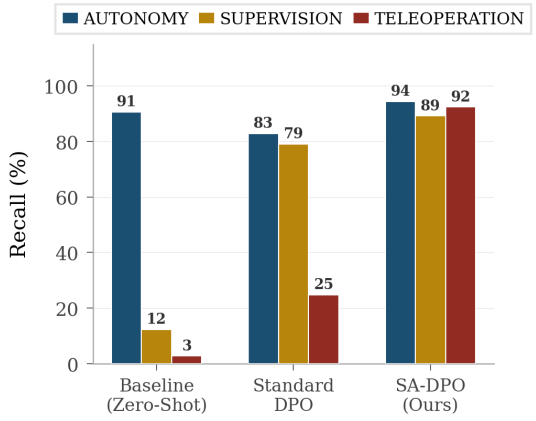


Figure 4. Per-class recall by triage level. SA-DPO achieves balanced, high recall across all three classes

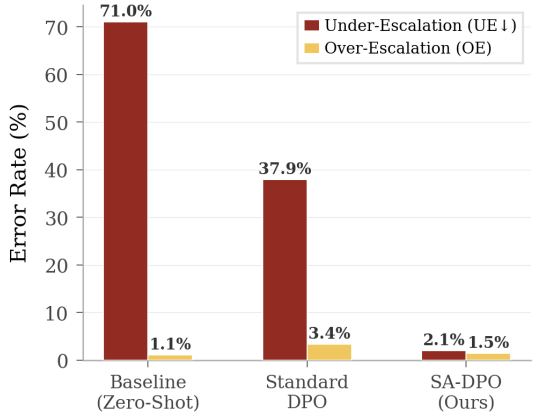


Figure 5. Under-escalation (UE, safety-critical misses) vs. over-escalation (OE, false alarms). SA-DPO reduces dangerous under-escalation to 2.1% while keeping false alarms at 1.5%

processes front-camera frames at **8–10 FPS** per stream, including full structured-output generation. The compact model size also enables deployment on more cost-effective GPUs (e.g., RTX 4090) with only modest throughput reduction, making fleet-scale triage economically viable.

## 6. Conclusion

In this work, we presented , a vision-language reasoning framework engineered to optimize the human-robot interaction bottleneck in autonomous vehicle teleoperation. By reformulating teleoperation decision from a heuristic confidence threshold to a structured, semantic reasoning problem, we enable early escalation based on contextual understanding.

Our primary contribution, Safety-Aware DPO (SA-DPO), addresses the critical class imbalance inherent in real-world driving data by applying an asymmetric loss that specifically penalizes missed intervention requests. Our results demonstrate that SA-DPO substantially mitigates under-escalation errors compared to standard alignment techniques, without resorting to artificial data resampling. As AV fleets scale, deploying lightweight, semantically aligned triage VLMs like at the teleoperation center level offers a practical, scalable mechanism for ensuring safe and highly interpretable human-in-the-loop oversight.

## References

[1] Association for Advancing Automation. Benefits of remote control capabilities for industrial robots. https://www.automate.org/industry-insights/benefits-remote-control-capabilities-industrial-robots, 2024. 1

[2] Tonko EW Bossen, Andreas Møgelmose, and Ross Greer. Can vision-language models understand and interpret dynamic gestures from pedestrians? pilot datasets and exploration towards instructive nonverbal commands for cooperative autonomous vehicles. In *Proceedings of the Computer*

*Vision and Pattern Recognition Conference*, pages 4779–4788, 2025. 2

[3] Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning (CoRL)*, pages 2165–2183, 2023. 2

[4] 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*, pages 11621–11631, 2020. 3

[5] J. Y. C. Chen, E. C. Haas, and M. J. Barnes. Human performance issues and user interface design for teleoperated robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1231–1245, 2007. 1, 2

[6] Cogniteam. Cloud-based teleoperation in robotics. https://www.cogniteam.com/resources/blogs/cloud-based-teleoperation-in-robotics-2/, 2024. [Online; accessed 1-Mar-2026]. 2

[7] Jacob W Crandall, Michael A Goodrich, Dan R Olsen, and Curtis W Nielsen. Validating human-robot interaction schemes in multitasking environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 35(4):438–449, 2005. 1

[8] Danny Driess, Fei Xia, Mehdi S.M. Sajjadi, et al. Palm-e: An embodied multimodal language model. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2

[9] Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024. 3

[10] Fernride. Teleoperation. https://www.fernride.com/teleoperation, 2024. 1

[11] M. Gontscharow et al. Scalable remote operation for autonomous vehicles: Integration of cooperative perception and open source communication. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 43–50, Jeju Island, Korea, 2024. 1, 2

[12] Michael A. Goodrich and Dan R. Olsen Jr. Seven principles of efficient human robot interaction. In *Proc. IEEE International Conference on Systems, Man and Cybernetics*, pages 3942–3948, 2003. 1

[13] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 14852–14882, 2023. 2

[14] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023. 2

[15] Yifan Jiang, Lei Zhang, Hao Chen, et al. A survey on vision–language–action models for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2025. Also available as arXiv:2506.24044. 1, 2

[16] Sidharth Bhanu Kamtam, Qian Lu, Faouzi Bouali, Olivier CL Haas, and Stewart Birrell. Network latency in teleoperation of connected and autonomous vehicles: A review of trends, challenges, and mitigation strategies. *Sensors*, 24(12):3957, 2024. 2

[17] Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, 2023. 2

[18] Qifeng Li, Zhiwei Peng, Lan Feng, and Quanyi Zhang. Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in CARLA-v2), 2024. Also published at ECCV 2024. 2

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[20] Mingjun Lu et al. Teleoperation technologies for enhancing connected and autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2022. Available: https://weisongshi.org/papers/lu22-teleoperation.pdf. 2

[21] Sidi Lu, Ren Zhong, and Weisong Shi. Teleoperation technologies for enhancing connected and autonomous vehicles. In *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pages 435–443. IEEE, 2022. 1

[22] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pages 403–420. Springer, 2024. 2

[23] A. Matheson, B. Donmez, F. Rehmatullah, P. Jasiobedzki, H.-K. Ng, V. Panwar, and M. Li. The effects of predictive displays on performance in driving tasks with multi-second latency: Aiding tele-operation of lunar rovers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 21–25, 2013. 2

[24] Ronja Möller, Antonino Furnari, Sebastiano Battiato, Aki Härmä, and Giovanni Maria Farinella. A survey on human-aware robot navigation. *Robotics and Autonomous Systems*, 145:103837, 2021. 1

[25] Motional. A path forward: Using AI to improve remote vehicle assistance for AVs. https://motional.com/news/path-forward-using-ai-improve-remote-vehicle-assistance-avs, 2024. 2

[26] Seonghoon Park, Jun Lee, YoungJae Kim, et al. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2024. 2

[27] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct

preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3

[28] Daniel J. Rea and Stela H. Seo. Still not solved: A call for renewed focus on user-centered teleoperation interfaces. *Frontiers in Robotics and AI*, 9:704225, 2022. 1, 2

[29] Andreas Schrank, Fabian Walocha, Stefan Brandenburg, and Michael Oehl. Human-centered design and evaluation of a workplace for the remote assistance of highly automated vehicles. *Cognition, Technology & Work*, 26(2):183–206, 2024. 2

[30] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024. 2

[31] Qunying Song, Emelie Engström, and Per Runeson. Industry practices for challenging autonomous driving systems with critical scenarios. *ACM Transactions on Software Engineering and Methodology*, 33(4):1–35, 2024. 2

[32] Felix Tener and Joel Lanir. Driving from a distance: Challenges and guidelines for autonomous vehicle teleoperation interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, pages 1–13, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2

[33] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2

[34] Adrien Wantiez, Tianming Qiu, Stefan Matthes, and Hao Shen. Scene understanding for autonomous driving using visual question answering. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2023. 1

[35] Rui Zhao, Qirui Yuan, Jinyu Li, Yuze Fan, Yun Li, and Fei Gao. DriveLLaVA: Human-level behavior decisions via vision language model. *Sensors*, 24(13):4113, 2024. 2