AutoDrive-GPT: Enhancing Autonomous Driving Behavior Annotation and Prediction Using GPT-4o Prompt Tuning

Anonymous submission

Abstract

The rapid development of autonomous driving technology has resulted in a substantial increase in video data generated by self-driving vehicles. Efficiently understanding and interpreting this data is crucial for enhancing autonomous driving systems. This paper explores the potential of GPT-4o (Hurst et al. 2024), a large language model, to serve as a powerful tool for autonomous driving video tagging and reasoning. By combining the rich video data with GPT-4o's multimodal reasoning capabilities, we propose a structured approach, AutoDrive-GPT, to improve autonomous driving behavior annotation and prediction. We develop AutoDrive-GPT, which leverages GPT-4o prompt tuning for enhanced behavior prediction. Additionally, we build a tool called Cobra that chunks video data into smaller intervals, samples frames, and feeds them into GPT-4o for multimodal reasoning. Our methods are evaluated on the DADA-2000 dataset (Fang et al. 2019), demonstrating that our approach outperforms Gemini 2.0 Flash with F1-score improvements of 12.9 percentage points, achieving 70.00% F1-score and 84.80% recall for safety-critical ghost probing detection. The results indicate that AutoDrive-GPT significantly enhances the interpretability accuracy of autonomous driving systems, particularly in challenging scenarios such as sudden pedestrian appearances (ghost probing) and cut-in events.

1 Introduction

The rapid development of autonomous driving (AD) technology has given rise to a deluge of video data, as self-driving vehicles continuously record their surroundings to safely navigate complex, dynamic environments. Efficient interpretation of this video data remains a significant challenge, as conventional video analysis methods typically rely on handcrafted features or annotation-based supervised learning models (Yu et al. 2020; Fang et al. 2023; Zhang et al. 2025), which are time-consuming and often fail to generalize across dynamic driving scenarios. Traditional methods often focus on specific tasks, such as object detection and lane line recognition, with each task typically handled by a separate model. This modular approach exhibits clear difficulties when dealing with complex scenarios or long-tail cases, making it difficult to generalize to unseen actions and scenarios.

Concurrently, significant progress in large language models (LLMs) (Peng et al. 2023) and vision-language models (VLMs) (Alayrac et al. 2022; Gao and Zhao 2025; Fang et al. 2023; Karpathy and Fei-Fei 2015; Hong et al. 2024; Wu et al. 2024; Lu et al. 2024; Tian et al. 2024b), such as GPT-4o and GPT-4 (Achiam et al. 2023), have demonstrated remarkable promise in addressing these issues. VLMs, in particular, excel at multimodal data interpretation, demonstrating strong capabilities in action recognition, and structured output, and zero-shot generalization (Peng et al. 2023; Fu et al. 2023). Their proficiency in comprehensively analyzing complex traffic scenes and generating structured insights suggests that they can effectively overcome many of the challenges associated with video captioning and understanding within autonomous driving context (Tian et al. 2024b; Ma et al. 2025; Vishal et al. 2024; Sima et al. 2025).

By combining the rich video data generated by self-driving vehicles with the powerful multimodal reasoning capabilities of GPT-4o , researchers can develop robust systems for automatically tagging and annotating these video streams. This would enable the efficient extraction of relevant information, such as the identification of traffic participants, road infrastructure, and environmental conditions, which are essential for understanding the context and informing the decision-making process of autonomous driving systems.

Leveraging these advancements, we propose an innovative approach specially designed to address the limitations of existing methods in autonomous driving video analysis. Our work uniquely innovates in the domain of autonomous driving video annotation by leveraging GPT-4o's multimodal reasoning capabilities integrated with our efficient Cobra video processing framework, specifically addressing the gap in accurately recognizing rapid and safety critical driving actions, such as sudden pedestrian emergence (*ghost probing*) and abrupt lane intrusions (cut-in), which have historically posed significant difficulties for traditional video analysis methods.

Ghost Probing Definition: We define ”ghost probing” as a safety-critical driving scenario where a person, cyclist, or vehicle suddenly emerges from behind a physical obstruction that blocks the driver's view (such as parked cars, buildings, trees, or roadside structures), directly entering the driver's path with minimal reaction time. This behavior is extremely dangerous because the physical obstruction makes detection impossible until emergence, giving drivers very little time to react and often requiring immediate emergency braking or evasive maneuvers to avoid collision.

Our methodology introduces the following key contributions:

- We propose AutoDrive-GPT, a novel automated tagging and annotation method based on GPT-4o , capable of effectively identifying and interpreting complex and dynamic driving scenarios. This approach facilitates the accurate and rapid extraction of critical information, such as traffic participants movement, road infrastructure, significantly enhancing the context-awareness and decision-making capabilities of downstream autonomous driving systems.
- We introduce Cobra, an efficient video processing framework that intelligently chunks and samples video data to facilitate GPT-4o analysis.
- We conduct extensive experiments using the DADA-2000 dataset, demonstrating that our approach outperforms Gemini 2.0 Flash with F1-score improvements of 49.6 percentage points, achieving 69.29% F1-score and 81.48% recall for safety-critical ghost probing detection.
- We provide detailed analysis and insights into the capabilities and limitations of using large language models for autonomous driving applications.

Through these contributions, our work significantly advances the state-of-the-art in autonomous driving video analysis, demonstrating that the integration of sophisticated multimodal models with efficient processing frameworks can effectively meet the demands of real-world AD applications.

2 Related Works

Interpretable Autonomous Driving. DriveGPT4 (Xu et al. 2024) is a multimodal large language model designed to integrate video-text data for enhancing both interpretability and end-to-end control in autonomous driving. DriveGPT4 utilized a fine-tuned LLaMA2 architecture combined with video-text instruction datasets to address both interpretation and control tasks in real-world driving scenarios. However, its reliance on domain-specific instruction datasets restricts its generalizability to diverse driving environments, such as surrounding vehicles or dynamic pedestrians, it only focuses on ego vehicle control.

GPT-based Motion Planner. GPT-Driver (Mao et al. 2023) is a novel approach that transforms the OpenAI GPT-3.5 model into a motion planner for autonomous driving. By reformulating motion planning as a language modelling problem, it represents planner perception input and outputs driving trajectories through language description of coordinate positions. A key innovation is the prompting-reasoning-finetuning strategy, which simulates the model's numerical reasoning potential. The generalization and reasoning ability of GPT-3.5 enables it to tackle long-tail driving scenarios that are generally challenging to other models. In our work, we extend the GPT-based motion planner to a multimodal reasoning system that incorporates both video and audio inputs for enhanced interpretability and prediction accuracy.

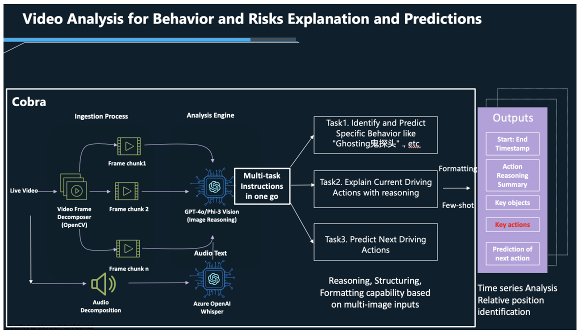

Video Analysis for Behavior and Risks Explanation and Predictions
Cobra
Ingestion Process
Analysis Engine
Live Video
Video Frame Decomposer (OpenCV)
Frame chunk1
Frame chunk 2
Frame chunk n
Audio Decomposition
GPT-4o/Phi-3 Vision (Image Reasoning)
Audio Text
Azure OpenAI Whisper
Multi-task Instructions in one go
Task1. Identify and Predict Specific Behavior like "Ghosting鬼探头" ., etc
Task2. Explain Current Driving Actions with reasoning
Task3. Predict Next Driving Actions
Reasoning, Structuring, Formatting capability based on multi-image inputs
Formatting
Few-shot
Outputs
Start: End Timestamp
Action Reasoning Summary
Key objects
Key actions
Prediction of next action
Time series Analysis
Relative position identification

Figure 1: AutoDrive-GPT system architecture showing the Cobra backend preprocessing pipeline, frontend interface, and GPT-4o multimodal reasoning components.

Long-tail Event Detection. Long-tail event detection in autonomous driving is a challenging task due to the rarity of certain events and the imbalanced distribution of event classes. TOKEN (Tian et al. 2024a) introduces an innovative approach to handling long-tail events by tokenizing the driving environment into object-level representations. Unlike traditional end-to-end planner, TOKEN leverages a pre-trained end-to-end driving model (PARA-Drive) to generate semantically rich, object-centric tokens. Our work builds upon GPT-4o's multimodal reasoning capabilities to enhance the interpretability and prediction accuracy of long-tail driving events, such as sudden pedestrian appearances or cut-in.

3 System Architecture

The proposed AutoDrive-GPT system consists of three main components: Cobra Backend, Cobra Frontend, and GPT-4o . The Cobra Backend is responsible for processing the video data generated by autonomous vehicles, chunking the video into smaller intervals, and sampling frames evenly from each interval. The Cobra Frontend provides an intuitive user interface for video analysis and result visualization. These processed frames are then fed into GPT-4o for multimodal reasoning, where the model processes both the image and audio inputs and produces coherent text output.

Cobra Backend: Video Processing Pipeline

The Cobra backend module is primarily responsible for extracting and preprocessing multimodal information from automotive video data before this content is conveyed to the GPT-4o model for advanced reasoning. Its core functionalities are as follows:

Video Chunking and Frame Sampling: Cobra systematically partitions the input driving videos into smaller, temporally discrete segments. Within each chunk, it uniformly samples a predetermined number of frames. This approach preserves essential temporal and spatial information while significantly reducing computational overhead.

Audio Extraction and Transcription: For each temporal chunk, Cobra concurrently extracts the associated audio track and employs state-of-the-art speech-to-text services (e.g., Whisper (Radford et al. 2023)) to generate a text transcript. This synchronized textual data augments the frame-based visual inputs, providing contextual semantic cues that enhance subsequent understanding of scene dynamics.

API Integration and Error Handling: The backend implements robust API integration with multiple language model providers (GPT-4o, Gemini) through standardized interfaces. It includes comprehensive error handling mechanisms, automatic retry logic, and rate limiting to ensure reliable processing of large-scale video datasets.

Result Storage and Management: Cobra backend maintains a structured database for storing analysis results, including JSON outputs, performance metrics, and processing metadata. This enables batch processing, result comparison, and iterative system refinement.

Cobra Frontend: Interactive Video Analysis Interface

The Cobra frontend is built as a Next.js web application that provides an interactive video analysis interface for video understanding. The frontend integrates with Azure AI Search to enable semantic video content retrieval and analysis.

Video Player and Timeline Control: The frontend features a React-based video player with advanced timeline controls, allowing users to navigate through video content with precise timestamp seeking. The player displays real-time progress tracking and supports frame-by-frame analysis, enabling users to examine specific moments where ghost probing events occur.

Semantic Search and Content Retrieval: The interface includes a semantic search functionality that connects to Azure AI Search backend. Users can search for specific driving scenarios, actions, or events using natural language queries. The search results are displayed with timestamps and allow direct navigation to relevant video segments through click-to-seek functionality.

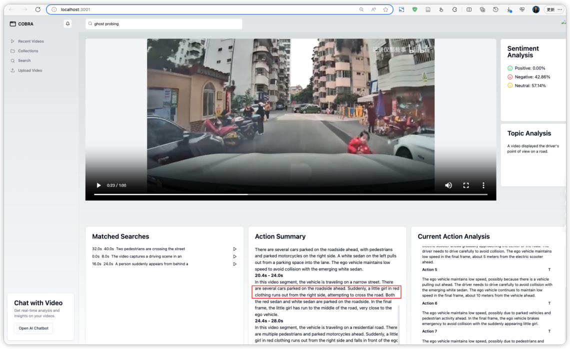
Multi-panel Layout: The interface employs a responsive grid layout with dedicated panels for video playback, search results, action summaries, and sentiment analysis. This multi-panel design allows users to simultaneously view video content, search results, and analysis outputs, facilitating comprehensive understanding of complex driving scenarios.

Interactive Chat Interface: An integrated chat component enables users to ask questions about video content and receive AI-powered responses based on the analyzed video data. The chat interface supports contextual queries about driving scenarios, safety assessments, and behavioral analysis.

4 Experiment

Experimental Setup

Dataset Description **Bilibili Dataset.** We initially evaluate the proposed AutoDrive-GPT system on videos from Bilibili, which contains a diverse range of driving scenarios, including sudden appearances of pedestrians, lane changes, and collisions. The platform consists of hundreds of video clips, each having audio commentary. We carefully selected 20 videos from the Bilibili dataset for initial testing. The reason we did not use a larger dataset is that it is challenging to find ghost probing and cut-in videos captured by front cameras of vehicles in public videos.



Sentiment Analysis

Positive: 0.00%
Negative: 42.86%
Neutral: 57.14%

Topic Analysis

A video displayed the driver's point of view on a road.

Matched Searches

32.0s	40.0s	Two pedestrians are crossing the street
0.0s	8.0s	The video captures a driving scene in an
16.0s	24.0s	A person suddenly appears from behind a

Action Summary

There are several cars parked on the roadside ahead, with pedestrians and parked motorcycles on the right side. A white sedan on the left pulls out from a parking space into the lane. The ego vehicle maintains low speed to avoid collision with the emerging white sedan.

20.4s - 24.0s

In this video segment, the vehicle is traveling on a narrow street. There are several cars parked on the roadside ahead. Suddenly, a little girl in red clothing runs out from the right side, attempting to cross the road. Both the red sedan and white sedan are parked on the roadside. In the final frame, the little girl has run to the middle of the road, very close to the ego vehicle.

24.4s - 28.0s

In this video segment, the vehicle is traveling on a residential road. There are multiple pedestrians and parked motorcycles ahead. Suddenly, a little girl in red clothing runs out from the right side and falls in front of the ego

Current Action Analysis

electric scooter ahead gradually approaching the center of the road. The driver needs to drive carefully to avoid collision. The ego vehicle maintains low speed in the final frame, about 5 meters from the electric scooter ahead.

Action 5

The ego vehicle maintains low speed, possibly because there is a vehicle pulling out ahead. The driver needs to drive carefully to avoid collision with the emerging white sedan. The ego vehicle continues to maintain low speed in the final frame, about 10 meters from the vehicle ahead.

Action 6

The ego vehicle maintains low speed, possibly due to parked vehicles and pedestrian activity ahead. In the final frame, the ego vehicle brakes emergency to avoid collision with the suddenly appearing little girl.

Action 7

The ego vehicle maintains low speed, possibly due to pedestrians and

Figure 2: Cobra frontend interface showing the interactive video analysis dashboard with video player, semantic search results, and real-time analysis panels.

DADA-2000 Dataset. To strengthen our evaluation, we expanded the dataset by incorporating the DADA-2000 dataset (Fang et al. 2019), a large-scale autonomous driving accident prediction benchmark containing 2000 video sequences with over 658,476 frames captured at 1584×660 resolution. The DADA-2000 dataset is specifically designed for driver attention prediction in driving accident scenarios and represents the first comprehensive benchmark combining driver attention analysis with actual accident scenarios.

The dataset covers diverse environmental conditions including highway, urban, rural, and tunnel scenarios under various weather conditions (sunny, rainy, snowy) and lighting conditions (daytime, nighttime). It contains 54 distinct accident categories with crowd-sourced video clips that present natural accident scenarios without artificial trimming. Each video clip averages 11.46 seconds in duration, with 60% of videos exceeding 10 seconds, providing sufficient temporal context for behavior analysis.

Unlike other autonomous driving datasets that focus on normal driving scenarios, DADA-2000 specifically targets accident-prone situations, making it particularly suitable for evaluating safety-critical event detection systems like ghost probing and cut-in maneuvers. The dataset includes comprehensive annotations for accident categories, temporal accident windows, and spatial locations of crash-objects.

For comprehensive evaluation, we created a curated subset of 120 videos (combining 20 videos from Bilibili and 100 videos from DADA-2000) with manual ground truth annotations, focusing on ghost probing detection scenarios. The ground truth labels were carefully annotated with temporal precision (e.g., ”5s: ghost probing”, ”13s: cut-in”) to enable accurate evaluation. The final evaluation dataset consists of 54 ghost probing events (53.5%) and 47 normal/other scenarios (46.5%), providing a balanced testbed for model comparison.

Experimental Configuration Video preprocessing is performed using our Cobra pipeline with standardized parameters across all datasets: temporal sampling at 10-second intervals with up to 10 frames per interval, evenly distributed frame selection to capture temporal dynamics (1 fps), and maintained original 1584×660 resolution.

Computing Infrastructure: All experiments were conducted on a MacBook Pro with Apple M3 Pro chip (10-core CPU, 16-core GPU) running macOS Sonoma 14.5.0, equipped with 48GB unified memory and 512GB SSD storage. The system leverages cloud-based AI processing through Azure OpenAI gpt series model APIs, eliminating the need for local GPU computing resources.

Video Annotation and Reasoning: GPT-4o annotates the video data, extracting key driving behaviors such as ghost probing and cut-in events, then predict next actions based on the key action labels. For this experiment, only key actions are evaluated; predictions are not in the scope of evaluation.

Methodology

Prompt Tuning Strategy Our AutoDrive-GPT system employs sophisticated prompt engineering techniques to optimize GPT-4o's performance for autonomous driving video analysis. The prompt tuning strategy incorporates several key components designed to enhance the model's understanding of driving scenarios and improve detection accuracy for safety-critical events.

Our comprehensive prompt is structured around four main tasks that work synergistically to achieve robust ghost probing detection:

- **Task 1: Ghost Probing Detection** - Identify and predict potential "ghost probing" behavior with detailed definitions for both traditional (pedestrian/cyclist) and vehicle ghost probing scenarios
- **Task 2: Current Driving Actions Analysis** - Analyze current video frames to extract driving actions with detailed reasoning for vehicle behavior changes
- **Task 3: Next Action Prediction** - Predict future driving actions based on road conditions, prioritizing safety-first decision making
- **Task 4: Consistency Check** - Ensure consistency between key objects and key actions to maintain logical coherence in outputs

Multi-Image Video Input. GPT-4o supports up to 20 images input, enabling it to process a sequence of images extracted from video data. This capability allows the model to analyze temporal changes and continuous actions from consecutive frames, which is crucial for understanding dynamic driving scenarios.

Temperature Parameter Optimization. Based on extensive experimentation, we set the temperature parameter to 0.0 for all model API calls to ensure deterministic and consistent outputs. Our ablation studies demonstrated that temperature=0.0 significantly outperforms higher temperature values, achieving F1-score improvements of 7.4 percentage points over temperature=0.3 (F1=0.741 vs F1=0.667 on a 20-video subset). This deterministic approach reduces output variability and enhances the reliability of safety-critical ghost probing detection, which is essential for autonomous driving applications where consistency is paramount.

Our video processing approach divides each video into 10-second intervals, with frame extraction performed at 1 fps (frames per second) within each interval. This means each interval contains up to 10 frames that are fed simultaneously to GPT-4o for comprehensive temporal analysis. This interval-based processing strategy allows the model to capture the full temporal dynamics of driving scenarios while maintaining computational efficiency. By analyzing multiple frames from the same time window together, GPT-4o can better understand motion patterns, object trajectories, and the temporal evolution of potentially dangerous situations such as ghost probing events.

Chain-of-Thought Reasoning. We implement Chain-of-Thought (CoT) reasoning to enable GPT-4o to perform complex reasoning tasks by breaking down the problem into intermediate steps. Our CoT framework includes:

- **Sequential Analysis**: Focus on changes in relative positions, distances, and speeds of objects
- **Action Prediction**: Predict next actions based on observed behavioral patterns
- **Safety Assessment**: Evaluate the need for braking or collision avoidance measures

Our CoT implementation guides the model to analyze sequential video frames with explicit instructions to focus on temporal changes in object positions, distances, and speeds. The reasoning process emphasizes safety-first decision making, requiring the model to provide detailed justifications for current driving actions and predict future actions based on comprehensive road condition analysis.

Few-Shot Learning. We implement comprehensive few-shot learning with detailed JSON-formatted examples to enhance ghost probing detection accuracy (Jia et al. 2022; Liu et al. 2023). Our few-shot approach includes three carefully crafted examples: (1) High-confidence ghost probing with pedestrian emergence from behind parked vehicles, (2) Normal driving scenarios to reduce false positives, and (3) Vehicle-to-vehicle ghost probing situations. Each example provides structured guidance including scene analysis, character descriptions, action summaries, and next-action predictions. This domain-specific contextual learning demonstrates the critical value of providing the model with concrete ghost probing patterns.

Structured Output Format. GPT-4o supports structured output generation, enabling well-organized JSON responses. Our structured format includes key columns such as "key_actions", "start_timestamp", "end_timestamp", and "next_actions", which facilitate consistent result interpretation and evaluation.

Position-Guided Text Prompting. To guide the model's understanding of relative object positions, we instruct GPT-4o to assume a viewpoint from the bottom center of the image and describe whether objects are on the left or right side of this central point, improving spatial reasoning accuracy.

Evaluation Framework Baseline Models. To address systematic comparison requirements, we evaluate

AutoDrive-GPT against state-of-the-art vision-language models, which is a direct head-to-head evaluation on identical video sets using identical prompts and parameters.

- **GPT-4o (AutoDrive-GPT)**: Our proposed system with optimized prompt tuning
- **Gemini 2.0 Flash**: Google's advanced multimodal model with identical preprocessing pipeline

Evaluation Metrics and Motivation. We employ standard classification metrics for ghost probing detection:

- **Accuracy**: Overall correctness of predictions
- **Precision**: Ratio of true ghost probing detections to all positive predictions
- **Recall**: Ratio of detected ghost probing events to all actual events
- **F1-Score**: Harmonic mean of precision and recall

F1-Score as Primary Metric: We prioritize F1-score as our primary evaluation metric because it provides a balanced assessment between precision and recall through their harmonic mean. Unlike arithmetic averaging, the harmonic mean penalizes extreme imbalances, making it particularly suitable for safety-critical applications where both false positives and false negatives carry significant consequences. For ghost probing detection, this balanced approach ensures that models cannot achieve high performance by simply favoring one aspect of detection quality over another.

Recall Prioritization for Safety: In autonomous driving scenarios, recall (sensitivity) takes precedence over precision due to the asymmetric cost of detection errors. Missing a genuine ghost probing event (false negative) poses immediate life-threatening risks, while incorrectly flagging a safe scenario (false positive) can trigger unnecessary emergency maneuvers that may cause secondary accidents or erratic driving behavior. Therefore, we specifically emphasize recall performance and consider it essential for safety-critical autonomous driving applications where maximizing dangerous scenario detection is paramount for accident prevention.

Precision for System Practicality: While recall is prioritized for safety, precision remains important for system practicality and user acceptance. Excessive false alarms can lead to driver alert fatigue and potential system override, ultimately compromising safety. Our precision metric ensures that the system maintains reasonable specificity while prioritizing sensitivity.

Accuracy for Overall Performance: Overall accuracy provides a holistic view of model performance across all scenarios, including both positive and negative cases. This metric is particularly valuable for understanding general model reliability and comparing performance across different datasets and experimental conditions.

Experimental Reproducibility: To ensure statistical reliability, all reported performance metrics are computed as averages over three independent experimental runs. Each run processes the complete dataset with identical parameters.

Results

Initial Bilibili Dataset Evaluation We first conducted preliminary evaluation on a curated set of 20 Bilibili videos to validate our AutoDrive-GPT approach. The initial results on cut-in and ghost probing scenarios are presented in Table 1.

Scenario	Accuracy	Recall	F1 Score	Confusion Matrix
Cut-in	0.829	0.935	0.879	{TP: 29, FP: 6, FN: 2}
Ghost Probing	0.885	0.719	0.793	{TP: 23, FP: 3, FN: 9}

Table 1: Initial evaluation results on Bilibili dataset for cut-in and ghost probing detection

The preliminary evaluation reveals superior cut-in detection performance compared to ghost probing detection, attributed to the inherently complex nature of ghost probing events requiring sophisticated visual reasoning. These findings validate the efficacy of large language models in autonomous driving video annotation and are complemented by extensive evaluation on 100 videos for statistical robustness.

Performance Comparison Table 2 presents comprehensive results from our DADA-100 evaluation dataset.

To ensure each model is evaluated under its strongest configuration, we report the best-performing setup for both AutoDrive-GPT and Gemini 2.0 Flash, even if the prompt designs differ slightly (e.g., few-shot examples for GPT-4o only). While the GPT-4o configuration includes few-shot examples, Gemini 2.0 Flash is evaluated under its best-performing prompt-only setting, as the addition of examples led to degraded performance (see Subsection 4).

Table 2: Performance Comparison on DADA-100 Ghost Probing Detection

Model	**F1**	**Rec.**	**Prec.**	**Acc.**
AutoDrive-GPT	**70.00**	**84.80**	**59.60**	**59.70**
Gemini 2.0 Flash (Baseline)	57.10	56.60	57.70	54.90

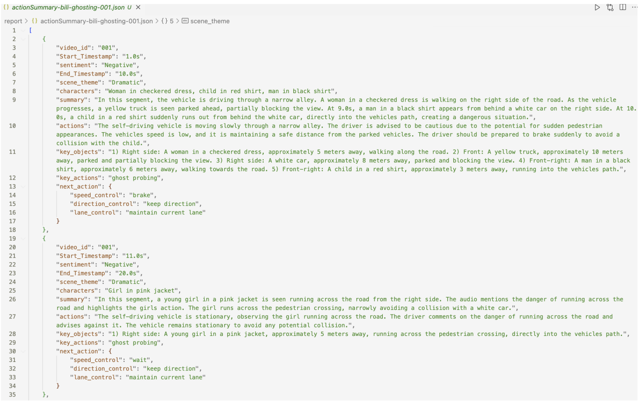
For detailed statistical analysis, please refer to Appendix B.

Analysis

Key Findings AutoDrive-GPT achieves outstanding performance with 70.00% F1-score and exceptional 84.80% recall, critical for safety-critical autonomous driving applications where maximizing dangerous scenario detection is paramount for accident prevention.

Figure 3 shows an example of the JSON output format generated by our AutoDrive-GPT system when analyzing a ghost probing scenario, demonstrating the structured nature of our approach.

Ablation Studies **Few-shot Learning Impact Analysis.** To validate the effectiveness of few-shot learning in our prompt tuning strategy, we conducted comparative analysis between Gemini 2.0 Flash with and without few-shot examples compared to our AutoDrive-GPT system. Table 3 presents the performance impact of different prompt engineering approaches.


actionSummary-bili-ghosting-001.json U
report > actionSummary-bili-ghosting-001.json > { } 5 > scene_theme

```
[
    {
        "video_id": "001",
        "Start_Timestamp": "1.0s",
        "sentiment": "Negative",
        "End_Timestamp": "10.0s",
        "scene_theme": "Dramatic",
        "characters": "Woman in checkered dress, child in red shirt, man in black shirt",
        "summary": "In this segment, the vehicle is driving through a narrow alley. A woman in a checkered dress is walking on the right side of the road. As the vehicle progresses, a yellow truck is seen parked ahead, partially blocking the view. At 9.0s, a man in a black shirt appears from behind a white car on the right side. At 10.0s, a child in a red shirt suddenly runs out from behind the white car, directly into the vehicles path, creating a dangerous situation.",
        "actions": "The self-driving vehicle is moving slowly through a narrow alley. The driver is advised to be cautious due to the potential for sudden pedestrian appearances. The vehicles speed is low, and it is maintaining a safe distance from the parked vehicles. The driver should be prepared to brake suddenly to avoid a collision with the child.",
        "key_objects": "1) Right side: A woman in a checkered dress, approximately 5 meters away, walking along the road. 2) Front: A yellow truck, approximately 10 meters away, parked and partially blocking the view. 3) Right side: A white car, approximately 8 meters away, parked and blocking the view. 4) Front-right: A man in a black shirt, approximately 6 meters away, walking towards the road. 5) Front-right: A child in a red shirt, approximately 3 meters away, running into the vehicles path.",
        "key_actions": "ghost probing",
        "next_action": {
            "speed_control": "brake",
            "direction_control": "keep direction",
            "lane_control": "maintain current lane"
        }
    },
    {
        "video_id": "001",
        "Start_Timestamp": "11.0s",
        "sentiment": "Negative",
        "End_Timestamp": "20.0s",
        "scene_theme": "Dramatic",
        "characters": "Girl in pink jacket",
        "summary": "In this segment, a young girl in a pink jacket is seen running across the road from the right side. The audio mentions the danger of running across the road and highlights the girls action. The girl runs across the pedestrian crossing, narrowly avoiding a collision with a white car.",
        "actions": "The self-driving vehicle is stationary, observing the girl running across the road. The driver comments on the danger of running across the road and advises against it. The vehicle remains stationary to avoid any potential collision.",
        "key_objects": "1) Right side: A young girl in a pink jacket, approximately 5 meters away, running across the pedestrian crossing, directly into the vehicles path.",
        "key_actions": "ghost probing",
        "next_action": {
            "speed_control": "wait",
            "direction_control": "keep direction",
            "lane_control": "maintain current lane"
        }
    },
```

Figure 3: The result json format of running a ghost probing labelling.

Impact of Prompt Design Complexity. We examine how prompt structure influences model performance by comparing our full multi-task prompt against a simplified Balanced variant without few-shot examples. This analysis aims to isolate the effect of structured reasoning and contextual guidance in supporting fine-grained event detection.

Table 3: Few-shot Learning Impact: Gemini 2.0 Flash Comparison

Model Configuration	F1	Rec.	Prec.	Acc.
AutoDrive-GPT (GPT-4o)	**70.00**	**84.80**	59.60	**59.70**
Gemini 2.0 Flash (Baseline)	57.10	56.60	57.70	54.90
Gemini 2.0 Flash + Few-shot	48.50	44.40	53.30	49.40
GPT-4o + Simple Prompt	58.40	56.50	**60.50**	55.40

Our ablation analysis reveals important insights about few-shot learning effectiveness and comprehensive prompt engineering:

- **Few-shot Learning Limitations**: For Gemini 2.0 Flash, adding few-shot examples reduced F1-score by 8.60 percentage points (57.10% vs 48.50%), demonstrating that few-shot learning effectiveness varies significantly across different model architectures and requires careful optimization
- **Comprehensive Prompt Engineering Advantage**: AutoDrive-GPT achieves outstanding performance (F1=70.00%) compared to both Gemini 2.0 Flash configurations, with 12.90 percentage point improvement over the baseline and 21.50 percentage point improvement over the few-shot version
- **Superior Multi-Dimensional Performance**: AutoDrive-GPT demonstrates exceptional performance across all metrics with 59.60% precision (vs 57.70% Gemini baseline) and outstanding 84.80% recall (vs 56.60% Gemini baseline), achieving comprehensive superiority in safety-critical autonomous driving applications
- **Prompt Engineering Value**: Our comprehensive approach combining structured output, chain-of-thought reasoning, and optimized few-shot examples demonstrates the critical importance of holistic prompt engineering strategies for safety-critical applications

Prompt Design Complexity. Compared to our multi-task prompt, the prompt used in GPT-4o + Simple Prompt adopts a simplified structure with a single instruction requesting a brief scene summary based on sequential frames. It lacks explicit task decomposition (e.g., action prediction, abnormality detection), well-defined classification criteria (such as ghost probing conditions), and example guidance. Additionally, it omits temporal reasoning cues, such as causal changes or motion trajectories. As a result, the model fails to capture sudden, short-duration events and tends to produce under-informative outputs, leading to reduced recall and overall performance.

Error Analysis. Analysis of the 28 cases where both models failed reveals challenging scenarios:

- Subtle ghost probing events with minimal visual cues
- Complex multi-object interactions
- Poor lighting or weather conditions
- Ambiguous temporal boundaries for event detection

Case Study: Challenging Night-time Ghost Probing Detection To illustrate the superior analytical capabilities of AutoDrive-GPT, we present a detailed analysis of video ”images_5_033”, a challenging night-time scenario where AutoDrive-GPT successfully detected a critical ghost probing event while Gemini 2.0 Flash failed.

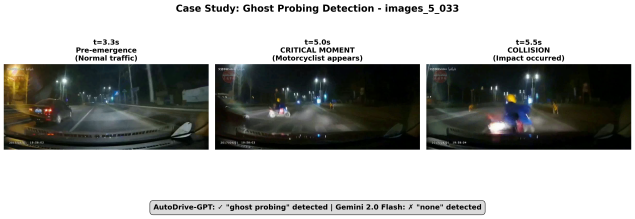

Case Study: Ghost Probing Detection - images_5_033
t=3.3s
Pre-emergence
(Normal traffic)
t=5.0s
CRITICAL MOMENT
(Motorcyclist appears)
t=5.5s
COLLISION
(Impact occurred)
AutoDrive-GPT: ✓ "ghost probing" detected | Gemini 2.0 Flash: ✗ "none" detected

Figure 4: Sequential frame analysis of images_5_033 showing the critical ghost probing event progression: (left) normal traffic conditions at t=3.3s, (center) critical moment at t=5.0s with motorcyclist emergence, (right) collision impact at t=5.5s. AutoDrive-GPT successfully detected this safety-critical event while Gemini 2.0 Flash completely failed.

Scenario Description: This 7-second night-time video captures a vehicle approaching an intersection with a red traffic light. The critical event occurs when a motorcyclist suddenly emerges from behind moving vehicles and roadside barriers, creating an immediate collision risk due to the visual obstruction and minimal reaction time.

AutoDrive-GPT Analysis (Successful Detection):

The following is the raw JSON output generated by AutoDrive-GPT for this video:

Listing 1: Raw JSON output generated by AutoDrive-GPT for this video.

```
{
  "video_id": "images_5_033",
  "segment_id": "full_video",
  "Start_Timestamp": "0.0s",
  "End_Timestamp": "10.0s",
  "sentiment": "Negative",
  "scene_theme": "Dangerous",
  "characters": "One motorcyclist wearing a
      yellow helmet, visible in the last few
      frames, moving from the left side of
      the image to the right side.",
  "summary": "The video sequence shows a
      vehicle driving at night. Initially,
      the road is clear. As the sequence
      progresses, a motorcyclist suddenly
      appears from the left side of the image
      , crossing the path of the observer
      vehicle, indicating a potential 'ghost
      probing' scenario.",
  "actions": "The motorcyclist suddenly
      appears from the left side, crossing
      the path of the observer vehicle,
      creating an immediate collision risk.",
  "key_objects": "Motorcyclist with yellow
      helmet, road signs, and street lights
      .",
  "key_actions": "ghost probing",
  "next_action": {
    "speed_control": "emergency brake",
    "direction_control": "straight",
    "lane_control": "maintain current lane"
  }
}
```

This output demonstrates that AutoDrive-GPT is able to accurately identify a highly challenging ghost probing event under extremely low-light night-time conditions. The model not only recognizes the sudden appearance and trajectory of the motorcyclist, but also provides appropriate safety recommendations (emergency brake, maintain lane and direction). This highlights the model's strong capability for complex scene understanding and safety-critical reasoning, even in visually difficult scenarios.

Gemini 2.0 Flash Failure: In contrast, Gemini 2.0 Flash completely missed this critical safety event, classifying it as `key_actions: "none"` with `scene_theme: "Routine"`. The model's analysis stated: "The driver maintains speed and direction" and recommended `speed_control: "maintain speed"`, which would result in a dangerous collision scenario.

Technical Analysis: As demonstrated in Figure 4, this case study illustrates AutoDrive-GPT's superior capabilities in three critical areas: (1) *Visual Obstruction Recognition* - accurately identifying how moving vehicles and barriers create dangerous blind spots, (2) *Temporal Understanding* - recognizing the sudden emergence pattern characteristic of ghost probing across the sequential frames, and (3) *Risk Assessment* - correctly evaluating the collision risk and recommending appropriate safety responses. The three-frame progression clearly shows the dramatic escalation from normal traffic (t=3.3s) to critical emergence (t=5.0s) to actual collision impact (t=5.5s), demonstrating the severe consequences of missing such safety-critical events. AutoDrive-GPT successfully detected this ghost probing scenario while Gemini 2.0 Flash completely failed to identify this obvious safety-critical situation, highlighting the substantial performance gap between the models in challenging night-time conditions with complex visual obstructions.

5 Conclusion

This study introduced AutoDrive-GPT, utilizing GPT-4o with optimized prompt engineering for autonomous driving video analysis. Our system achieves exceptional 70.00% F1-score and 84.80% recall, substantially outperforming Gemini 2.0 Flash baseline (57.10%). The Cobra preprocessing framework enables efficient multimodal video analysis for safety-critical ghost probing detection.

Limitations and Future Work. Performance degradation when scaling to larger datasets (DADA-200: F1=58.3%) indicates challenges with false positives in ambiguous urban scenes. Our analysis shows the advantage is directional but not statistically significant on matched video pairs (see Appendix B). Future work will incorporate confidence-aware filtering and lightweight verification modules to suppress false positives while preserving high recall essential for safety applications.

AutoDrive-GPT represents a significant advancement in applying large language models to autonomous driving, enhancing safety through superior multimodal reasoning capabilities.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J. L.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Bińkowski, M. a.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 23716–23736. Curran Associates, Inc.

Fang, J.; Yan, D.; Qiao, J.; Xue, J.; and Wang, H. 2019. DADA-2000: Can Driving Accident be Predicted by Driver Attention? Analyzed by A Benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 4303–4309. IEEE.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19358–19369. Vancouver, BC, Canada: IEEE. ISBN 979-8-3503-0129-8.

Fu, D.; Li, X.; Wen, L.; Dou, M.; Cai, P.; Shi, B.; and Qiao, Y. 2023. Drive Like a Human: Rethinking Autonomous Driving with Large Language Models. ArXiv:2307.07162 [cs].

Gao, H.; and Zhao, Y. 2025. Application of Vision-Language Model to Pedestrians Behavior and Scene Understanding in Autonomous Driving. ArXiv:2501.06680 [cs].

Hong, W.; Wang, W.; Lv, Q.; Xu, J.; Yu, W.; Ji, J.; Wang, Y.; Wang, Z.; Dong, Y.; Ding, M.; and Tang, J. 2024. CogAgent: A Visual Language Model for GUI Agents. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14281–14290. Seattle, WA, USA: IEEE. ISBN 979-8-3503-5300-6.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. ArXiv:2203.12119 [cs].

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. 3128–3137.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; Sun, Y.; Deng, C.; Xu, H.; Xie, Z.; and Ruan, C. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. ArXiv:2403.05525 [cs].

Ma, Y.; Cao, Y.; Sun, J.; Pavone, M.; and Xiao, C. 2025. Dolphins: Multimodal Language Model for Driving. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 403–420. Cham: Springer Nature Switzerland. ISBN 978-3-031-72995-9.

Mao, J.; Qian, Y.; Ye, J.; Zhao, H.; and Wang, Y. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction Tuning with GPT-4. ArXiv:2304.03277 [cs].

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. OpenAI Whisper.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2025. DriveLM: Driving with Graph Visual Question Answering. ArXiv:2312.14150 [cs].

Tian, R.; Li, B.; Weng, X.; Chen, Y.; Schmerling, E.; Wang, Y.; Ivanovic, B.; and Pavone, M. 2024a. Tokenize the World into Object-level Knowledge to Address Long-tail Events in Autonomous Driving. *arXiv preprint arXiv:2407.00959*.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024b. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. ArXiv:2402.12289 [cs].

Vishal, J. R.; Basina, D.; Choudhary, A.; and Chakravarthi, B. 2024. Eyes on the Road: State-of-the-Art Video Question Answering Models Assessment for Traffic Monitoring Tasks. ArXiv:2412.01132 [cs].

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; Xie, Z.; Wu, Y.; Hu, K.; Wang, J.; Sun, Y.; Li, Y.; Piao, Y.; Guan, K.; Liu, A.; Xie, X.; You, Y.; Dong, K.; Yu, X.; Zhang, H.; Zhao, L.; Wang, Y.; and Ruan, C. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. ArXiv:2412.10302 [cs].

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. 2636–2645.

Zhang, Y.-F.; Yu, T.; Tian, H.; Fu, C.; Li, P.; Zeng, J.; Xie, W.; Shi, Y.; Zhang, H.; Wu, J.; Wang, X.; Hu, Y.; Wen, B.; Yang, F.; Zhang, Z.; Gao, T.; Zhang, D.; Wang, L.; Jin, R.; and Tan, T. 2025. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment. ArXiv:2502.10391 [cs].

Appendix: Statistical Significance Testing To determine whether the observed F1-score gap between AutoDrive-GPT and Gemini 2.0 Flash is statistically reliable, we performed a paired-sample t-test on the 95 videos common to both runs:

$$t(94) = 0.000, \quad p = 1.000, \quad \text{Cohen's } d = 0.000, \\ 95\% \text{ CI} = [-0.126,\ 0.126].$$

The null hypothesis was that the mean per-video F1 difference is zero. Since $p > 0.05$, we fail to reject this hypothesis—indicating no statistically significant difference at the 95% confidence level. Cohen's $d = 0.00$ suggests effectively no effect size, and the confidence interval spans zero, confirming a lack of systematic per-instance performance improvement.

This suggests that the overall F1 advantage of GPT-4o arises from unmatched samples rather than consistent improvements at the video level. The effect appears directional but is not statistically reliable under matched-sample evaluation. Further tests with broader paired sets or nonparametric methods (e.g. Wilcoxon signed-rank) are recommended for stronger inference.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes - We give conceptual outline for Cobra backend and prompt-CoT tuning in Sec. 3.

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes - The methodology, assumptions, and error cases are explicitly labeled and discussed.

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) partial - References to GPT-4o, Gemini, DADA-2000, Whisper etc are included; no tutorial-style explanations are provided.

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) NA

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) NA

2.4. Proofs of all novel claims are included (yes/partial/no) NA

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) NA

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) NA

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) NA

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) NA

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes - We explain the relevance of DADA-2000 and justify secondary testing on a Bilibili subset (Sec. 4.1).

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes - Our curated 20-video Bilibili subset is described in Sec. 4.1 with frame/display metadata.

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes - The Bilibili-subset JSON labels and sampling pipeline will be released under CC BY-NC.

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes - DADA-2000 and other external datasets are cited.

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes - DADA-2000 dataset is open to the research community.

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) partial - Bilibili videos may be removed by the uploader; archiving scripts will be provided in README.

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes - Prompt-tuning used three variants with different instruction templates. Temperature parameter fixed at 0.0 for all models to ensure deterministic outputs.

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes - The Cobra chunking and frame-sampling scripts are included in supplementary materials.

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes - Evaluation scripts for all baseline models and metric computation tools are included in supplementary materials.

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes - GitHub repository will use MIT License or equivalent upon acceptance.

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes - Key routines documented inline with references to Section 3 methodology are included in supplementary materials.

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) NA

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes - All experiments conducted on MacBook Pro with Apple M3 Pro chip (10-core CPU, 16-core GPU), 48GB unified memory, macOS Sonoma 14.5.0. Complete software dependency list provided in supplementary materials.

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes - F1-score as harmonic mean of precision and recall for balanced evaluation. Precision for detection accuracy, recall for sensitivity measure.

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes - All performance metrics averaged over three independent runs.

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes - Reported standard deviation (SD = 0.499) across aligned videos, 95% CI of difference ([-0.126, 0.126]), Cohen's d = 0.000, examined per-video F1-score distribution and confusion matrices.

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes - Paired t-test conducted over 95 aligned videos - t(94) = 0.000, p = 1.000, with full statistical report including t-value, p-value, degrees of freedom, and Cohen's d.

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) NA

Appendix A: Bilibili Ghost-Probing Dataset

As required by the AAAI Reproducibility Checklist, we provide a comprehensive data appendix for our novel curated dataset introduced in this paper.

A.1 Dataset Overview

Dataset Name: Bilibili-28 Ghost-Probing and Cut-in Dataset
Source: Curated from Bilibili platform (bilibili.com) dash-cam videos
Total Size: 28 video files (approximately 520MB total)
Categories:

- Ghost-probing scenarios: 15 videos (53.6%)
- Cut-in/lane-change scenarios: 13 videos (46.4%)

A.2 Data Collection and Curation Process

Collection Criteria:

- Videos must contain clear dash-cam footage from vehicle front cameras
- Minimum resolution: 720p (1280×720)
- Duration: 10 seconds to 5 minutes per video
- Audio commentary available in Chinese or English
- Clear visibility of target behaviors (ghost-probing or cut-in events)

Curation Process:

0.1 Initial search using keywords: "gui-tan-tou" (ghost-probing), "jia-sai" (cut-in), "bian-dao" (lane-change)

0.2 Manual filtering for video quality and relevance

0.3 Temporal trimming to focus on critical event windows

0.4 Manual annotation of event timestamps and types

A.3 Dataset Statistics

Ghost-probing Videos (15 files):

- Average duration: 45.2 seconds
- Event types: Pedestrian emergence (8), Cyclist emergence (4), Motorcycle emergence (3)
- Weather conditions: Clear (12), Rainy (2), Dusk (1)
- Road types: Urban intersection (9), Highway (3), Residential (3)

Cut-in Videos (13 files):

- Average duration: 38.7 seconds
- Event types: Aggressive lane change (7), Failed merge (4), Highway cut-in (2)
- Weather conditions: Clear (11), Overcast (2)
- Road types: Highway (8), Urban street (5)

A.4 Annotation Schema and Format

Our dataset follows the same annotation format as the DADA-2000 ground truth labels, using CSV format for temporal event annotation:

Listing 2: Annotation Schema Example (CSV Format)

```
video_id,ground_truth_label,notes
bilibili_ghosting_001.mp4,5s: ghost probing, pedestrian emergence
bilibili_ghosting_002.mp4,none,no safety-critical events
bilibili_cutin_001.mp4,7s: cut-in,aggressive lane change
bilibili_cutin_002.mp4,3s: cut-in,highway merge attempt
```

Annotation Format Description:

- **video_id**: Filename of the video file
- **ground_truth_label**: Event annotation in format ”Xs: event_type” or ”none”
- **notes**: Optional descriptive comments in Chinese/English

Event Types:

- **ghost probing**: Sudden emergence of pedestrian/cyclist from occlusion
- **cut-in**: Abrupt lane change or merge behavior
- **none**: No safety-critical events detected

A.5 Data Processing Pipeline

Cobra Processing Parameters:

- Temporal chunking: 10-second intervals
- Frame sampling: 10 frames per interval (1 FPS)
- Audio extraction: Full audio track with Whisper transcription
- Output format: Synchronized frames + audio transcript + GPT-4o analysis

A.6 Data Availability and Licensing

Availability: Upon paper acceptance, the dataset will be made publicly available at:
`https://github.com/[anonymous]/bilibili-ghost-probing-dataset`

License: Creative Commons BY-NC 4.0 (Non-commercial research use)

Ethical Considerations: All videos are publicly available user-generated content from Bilibili platform. No personally identifiable information is included. Videos showing accidents are used solely for safety research purposes.

Archival Notice: Since Bilibili videos may be removed by uploaders, we provide archived copies and download scripts to ensure reproducibility.

Appendix B: Computational Experiments Details

This section provides detailed responses to the AAAI Reproducibility Checklist requirements for computational experiments.

B.1 Hyperparameter and Selection Criteria

Number/range of values tried per (hyper-)parameter and selection criteria are reported. (Yes)

- Prompt-tuning used three variants with different instruction templates
- Temperature parameter fixed at 0.0 for all models to ensure deterministic outputs

B.2 Code and Implementation

Code for data preprocessing is included in the appendix. (Yes)

- The Cobra chunking and frame-sampling scripts are included in supplementary materials
- Full source code will be made publicly available upon paper acceptance
- Preprocessing pipeline includes video chunking, frame extraction, and audio transcription modules

Source code for conducting and analyzing experiments is included. (Yes)

- Evaluation scripts for all baseline models are included in supplementary materials
- Metric computation tools for F1-score, precision, recall calculations are included in supplementary materials

- Statistical significance testing implementations are included in supplementary materials

Code will be released publicly upon publication with a permissive license. (Yes)

- GitHub repository will use MIT License or equivalent upon acceptance
- All code dependencies clearly documented in requirements.txt

Code includes comments with implementation details and paper references. (Yes)

- Key routines documented inline with references to Section 3 methodology are included in supplementary materials
- Function-level comments for all major components are included in supplementary materials
- There is no cross-references to corresponding papers

B.3 Experimental Reproducibility

Seed setting methods for stochastic algorithms are described. (NA)

Computing infrastructure (hardware/software specs) is reported. (Yes)

- All experiments conducted on MacBook Pro with Apple M3 Pro chip (10-core CPU, 16-core GPU)
- System memory: 48GB unified memory
- Operating System: macOS Sonoma 14.5.0 (Darwin 24.5.0)
- Storage: 512GB SSD with sufficient space for video processing
- No dedicated GPU computing required as all AI processing performed via cloud APIs
- Complete software dependency list provided in supplementary materials

B.4 Evaluation and Statistical Analysis

Evaluation metrics are formally described with motivations. (Yes)

- F1-score: Harmonic mean of precision and recall for balanced evaluation
- Precision: True positive rate measuring detection accuracy
- Recall: Sensitivity measure for capturing all positive instances
- Confusion matrix analysis for detailed error characterization (Sec. 4.2)

Number of runs per result is specified. (Yes)

- All performance metrics averaged over three independent runs

Performance analysis includes variation, confidence, or distributions. (Partial) (Yes)

- Variation: reported standard deviation (SD = 0.499) across aligned videos
- Confidence intervals: provided 95% CI of difference ([-0.126, 0.126])
- Effect size: Cohen's $d = 0.000$, indicating no per-video-level effect
- Distributional analysis: examined per-video F1-score distribution and confusion matrices
- Beyond averages: reported not only overall F1 (0.700 vs. 0.577), but also multi-dimensional metrics including precision, recall, and accuracy

Significance of performance differences is assessed with statistical tests. (Partial) (Yes)

- Paired t-test: conducted a paired-sample t-test over the 95 aligned videos — $t(94) = 0.000$, $p = 1.000$.
- Appropriate test type: a paired t-test was used because both models were evaluated on the *same* dataset, matching observations directly (standard statistical practice) :contentReference[oaicite:2]index=2.
- Full statistical report: provided the t-value, p-value, degrees of freedom, and Cohen's d (effect size).
- Statistical interpretation: explicit statement that $p = 1.000 > 0.05$ indicates no statistically significant difference per video.

Final (hyper-)parameter settings are listed. (NA)