Mind the Gap: Quantifying and Aligning Human-AI Visual Attention for Accident Anticipation

Hoe Sung Ryu[1], **Christian Wallraven**[1, 2*]

[1]Department of Artificial Intelligence, Korea University
[2]Department of Brain and Cognitive Engineering, Korea University
wallraven@korea.ac.kr

Abstract

Quantifying and understanding human-AI alignment in high-risk tasks such as traffic accident prediction is crucial for deployment of AI systems. Existing alignment studies, however, focus mostly on the static domain and neglect the importance of attentional processing. Here, we present Attention-DADA, a dataset of accident and non-accident traffic situations that contains detailed human prediction and frame-level eye gaze annotations. Using this benchmark, we evaluate open- and closed-source, state-of-the-art large vision-language-models (VLMs) in terms of their alignment in accident prediction performance and attentional processing in both zero-shot and attention-guided settings. Our results show that human prediction performance and consistency improve as the event time approaches. Similarly, human attentional patterns show dynamic updating throughout event progression. Conversely, while attention guidance improves VLM prediction performance, both performance and attentional alignment stay significantly below human levels as the event approaches, with the performance gap becoming significant 3.5 seconds (s) prior to the event. These results provide the first quantitative evidence of misalignment both in terms of performance and attentional processing during analysis of time-critical, dynamic events, highlighting the need for future improvements in this area.

Code — https://github.com/hoesungryu/Mind-the-Gap

Introduction

Traffic accidents often unfold in mere moments. Human drivers swiftly identify crucial visual cues from a vast array of information for decision-making (Wickens 2002). The advancement of autonomous driving systems (Yuan et al. 2024) promises to enhance road safety, yet what is required from these AI systems is not merely predicting accidents with high accuracy, but also providing a trustworthy and transparent processing aligned to human standards (Kuznietsov et al. 2024). However, the complex, black-box nature of modern AI (Amodei et al. 2016) means that we cannot be certain that an AI's decision is based on the same evidence as an expert human's. This uncertainty critically impedes the social acceptance and deployment of AI technology (Baheri 2022).

To bridge this trust gap, recent research has explored the alignment between human processing and AI decision-making. This includes visual attentional alignment—assessing how closely a model's attention corresponds with human attention (e.g., measured via eye gaze)—to determine if AI systems rely on similar visual cues to humans (Liang et al. 2024; Xu et al. 2024). Progress has been made in areas such as visual grounding (Kang et al. 2025), attention-based alignment (Zhou et al. 2025), and vision-language model (VLM) alignment (Jose et al. 2025; Li and Li 2025). Recent studies also demonstrate that cognitive processing time plays a crucial role in performance gaps between humans and AI systems (Ollikka et al. 2024). Efforts to apply these alignment principles to traffic safety have also emerged (Piergiovanni et al. 2024; Jiang et al. 2023).

Importantly, however, most existing methods focus solely on static environments, neglecting the continuous nature of emergency or accident situations. Similarly, existing frameworks typically do not account for the temporal dynamics of human attention, which narrows to focus on critical visual information during emergencies (Wickens 2002). The temporal alignment between state-of-the-art VLMs (Jin et al. 2024) and this adaptive human attention process remains under-explored. This gap underscores the need for a systematic framework capable of evaluating how visual alignment between humans and AI evolves in decision-making in time-critical situations.

This paper introduces a framework (Figure 1) for evaluating the dynamics of visual alignment between humans and AI in potential traffic accident scenarios. Using a combined human-AI study, we examine how accident prediction performance *and* attended regions evolve over time, benchmarking human drivers against state-of-the-art VLMs, including both open-source and commercial API-based systems. We also explore how attention guidance for models influences these temporal patterns, providing novel insights into human-AI alignment. Our main contributions are:

- We present *Attention-DADA*, a novel, human-annotated (eye-gaze, behavioral data) dataset for temporal analysis of decision-making in critical traffic situations.
- We benchmark accident prediction performance and attentional alignment of state-of-the-art zero-shot and

*Corresponding author

attention-guided VLMs against human data.

- We identify a critical window (3.5s before events) where human attention dynamically reallocates while VLMs fail to adapt—even with attention guidance—revealing fundamental temporal reasoning limitations.

Related Works

Human-AI Visual Alignment

Human-AI visual attention alignment research investigates the extent to which computational models' visual perception aligns with human perception. Early studies focused on visual saliency models that predicted where human gaze fixates (Itti, Koch, and Niebur 2002), subsequently evolving toward comparing deep neural networks' attention mechanisms with human eye gaze (Guo et al. 2022). With recent advances in vision-language models (VLMs), such visual alignment has become crucial for ensuring semantic consistency between text and images (Zeng, Zhang, and Li 2021).

These alignment efforts provide insights into models' reasoning capabilities by evaluating whether AI models attend to semantically relevant information in human-like ways (Muttenthaler et al. 2024; Ryu, Ju, and Wallraven 2025). To enable more sophisticated measurement, alignment evaluation datasets have been developed that capture diverse cultural perspectives and individualized feedback (Kirk et al. 2024) as well as various scenarios (Lee et al. 2023).

Alignment in Complex and Time-Critical Environments

Recent research has revealed that alignment reliability is not static, but varies considerably according to temporal and contextual factors (Zhang, Tseng, and Kreiman 2020). Foundational work has established that human cognitive performance and decision-making patterns significantly deteriorate in time-critical contexts (Corvelo Benz and Rodriguez 2023). In environments with high time pressure and cognitive burden, alignment patterns may become unstable or constrained by sociotechnical barriers (Mastrianni et al. 2025).

Importantly, these studies primarily focus on optimizing system task performance or enhancing user convenience, rather than directly exploring how human-AI alignment changes as situations unfold. Only recently has criticism emerged that evaluating alignment with fixed measures is insufficient, given that human attention patterns vary according to task context (Kadner et al. 2023). Research has demonstrated that alignment patterns maintained under controlled conditions can collapse when environmental factors introduce additional complexity or urgency (Vaccaro, Almaatouq, and Malone 2024). Building on this foundation, our research benchmarks how human-AI visual alignment changes dynamically as the critical point of a final accident occurrence approaches.

Traffic Accident Prediction with VLMs

The task of traffic accident prediction has evolved significantly with deep learning advancements. Early approaches relied on statistical methods or classical computer vision techniques combined with sequential models, while recent developments have been dominated by Transformer-based models that demonstrate exceptional capabilities in capturing long-range spatiotemporal dependencies in complex traffic scenarios (Jiang et al. 2023). Concurrently, advancements in vectorized high-definition map representations have provided models with richer environmental context, further enhancing predictive power (Yuan et al. 2024).

A more recent paradigm shift has been the integration of language, leading to the rise of Vision-Language Models (VLMs) in the traffic domain. Unlike traditional vision-only models that output predefined class labels or regression values, VLMs can process and generate natural language, enabling more nuanced and interpretable analysis of traffic events (Sima et al. 2023; Pan et al. 2024). This opens new possibilities for zero-shot reasoning and detailed situational questioning, providing enhanced flexibility in traffic prediction tasks (Piergiovanni et al. 2024; Jin et al. 2024).

Despite these significant advancements, the primary focus of existing research remains on improving predictive accuracy (Fang et al. 2024). Model evaluation is predominantly outcome-oriented, assessing what they predict rather than how they arrive at that prediction. Crucially, there is a significant lack of research investigating whether the internal decision-making process of these models aligns with that of human experts, especially concerning visual attention over time. Our work adds this perspective to the outcome-based evaluation, providing a detailed benchmark for quantifying the temporal evolution of visual alignment between humans and VLMs in traffic scenarios.

Methods

Attention-DADA Benchmark Dataset

Video data. In this study, we constructed the *Attention-DADA* dataset to enable comprehensive benchmarking of predictive decision-making processes between humans and computational models in accident anticipation scenarios. We first selected 100 base videos (50 accident, 50 non-accident) from the popular DADA-2000 dataset (Fang et al. 2019). This paired accident/non-accident design is enabling a controlled, discriminative analysis to isolate attentional patterns unique to critical events versus normal driving conditions. For accident videos, we used the annotated physical collision time as T_E; for non-accident videos (safe controls), we designated the video's endpoint as T_E. We excluded the final 0.5-second (s) interval ($[T_E - 0.5s, T_E]$) from all videos to eliminate trivial visual cues of imminent impact. To analyze how and when anticipation evolves, we implemented a fine-grained temporal segmentation. We extracted the preceding 6-second period ($[T_E - 6.5s, T_E - 0.5s]$) and partitioned it into five sequential 2-second clips (T5, T4, T3, T2, T1) overlapping a 1-second sliding window. T5 denotes the earliest segment ($[T_E - 6.5s, T_E - 4.5s]$) and T1 the latest ($[T_E - 2.5s, T_E - 0.5s]$), with each treated as an independent judgment unit (Gu et al. 2018).

Human annotation. We next recruited a total of 30 participants (18 male; mean age = 29.8 years, SD = 3.8) for

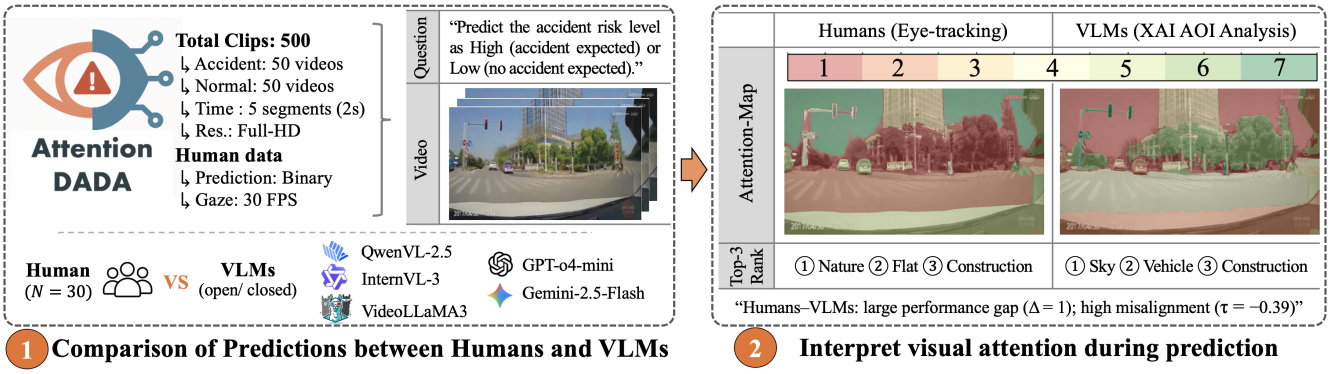


Figure 1: Framework for evaluating performance and visual attention alignment between humans and VLMs on identical driving scenarios. Performance is measured by F1 score, while alignment is quantified by comparing human eye-tracked AOI rankings against VLM XAI—derived rankings. In the illustrated failure case, humans succeed (F1=1.0) while VLMs fail (F1=0.0). This performance gap is reflected in a severe attentional misalignment (Kendall's $\tau = -0.4$)

human annotation. All participants held a valid driver's license with at least one year of driving experience, had normal or corrected-to-normal vision, and provided informed consent before the experiment. The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Korea University (IRB KUIRB-2021-0226-02).

Participants completed the experiment in a quiet, distraction-free room. They were seated $\approx$ 70 cm from a display screen, subtending a visual angle of $\approx$ 43°. The experiment, controlled and presented using PsychoPy software (Peirce et al. 2019), consisted of 500 trials, divided into three blocks, and presented in randomized order.

Participants first read a standardized instruction sheet that explained the goal—predicting whether a traffic accident would occur within the next few seconds. After a brief practice block with feedback, they completed the main trials independently. Each trial began with a central fixation cross displayed for 50 milliseconds (ms), followed by a video segment presented in randomized order to minimize correlation and memory effects. Upon completion, participants viewed a centered response screen (550ms) and made a binary judgment by pressing the left (non-accident) or right arrow key (accident). This design maintained central fixation for 600ms, ensuring gaze stabilization and central bias control (Tatler 2007).

Stimuli were presented on a 52.1 cm $\times$ 29.3 cm Full HD monitor integrated with a Tobii TX300 eye tracker. While participants viewed each video, the eye tracker simultaneously recorded gaze data at a sampling rate of 120 Hz. Calibration was performed before each experimental block to ensure high-precision gaze tracking.

Vision-Language Models

Our study employed a diverse suite of VLMs to ensure a comprehensive evaluation, encompassing larger-scale, closed-source and interpretable, open-source models (see supplemental material). We applied both prompt-based and visual-based attention guidance techniques to each model to evaluate their effectiveness in improving human-AI attention alignment.

Open-source Models. To obtain fine-grained access to internal representations and enable detailed analysis of attention mechanisms, we include three representative high-performing open-source VLMs. InternVL-3 (Zhu et al. 2025) demonstrates exceptional performance in fine-grained object recognition and multi-object relational reasoning, making it particularly suitable for complex traffic scenarios. Qwen-2.5-VL (Bai et al. 2023) is explicitly designed with interpretability in mind, providing access to token-level attention weights and cross-modal fusion maps that are essential for our alignment analysis. VideoLLaMA3 (Zhang et al. 2025) maintains coherent context during conversational interactions, potentially enabling robust temporal performance.

Visual Attention Extraction. We extracted VLM visual attention using Input × Gradient (Ancona et al. 2017) to generate pixel-level importance scores quantifying cross-modal (text-video) interactions. Since gradient-based methods yield sparse, localized heatmaps (Adebayo et al. 2018), *direct* pixel-wise comparison with continuous human gaze distributions is unreliable. Therefore, we aggregated both VLM attributions and human gaze maps at the AOI level. This approach enables robust comparison of regional priorities, focusing our analysis on semantic attention allocation—which is more interpretable and relevant for human-AI alignment in safety-critical scenarios (see below).

Closed-source Models. To establish a state-of-the-art performance benchmark while maintaining practical deployment viability, we utilized Gemini-2.5-Flash (Google 2025) and GPT-o4-mini (OpenAI 2025). These models were chosen for their advanced multimodal reasoning capabilities combined with low-latency processing suitable for time-critical tasks.

Model Enhancement Strategies

Prompt-based Attention Guidance. For both open- and closed-source models, we employed additional attention-directing prompts (e.g., "Pay special attention to the vehicle") to guide model focus toward safety-critical regions. This approach demonstrated improvements in alignment scores in previous studies in other contexts without requiring architectural modifications (Yu, Yu, and Wang 2024; Gu et al. 2023).

Visual-based Attention Guidance To quantify the utility of human gaze for accident anticipation, we implemented a visual-based human attention guidance mechanism. This method assesses whether VLM performance improves when provided with explicit guidance on where humans allocate attention. We achieved this by enhancing the visual inputs with aggregated human gaze maps, applying multiplicative pixel enhancement based on human gaze data (Woo et al. 2018). This technique selectively amplifies visual features at human-attended regions while preserving the original spatial structure, enabling a controlled comparison between the baseline VLM and this human-guided model.

Human-AI Alignment Analysis Framework

We investigated human-AI alignment by administering an identical accident-anticipation task to humans and VLMs, then quantifying performance as well as the focal points each agent prioritized during prediction. To make sure that the model's contextual priors were comparable to the task given to humans, we established a concise role specification for VLMs that replicated the human instructions, including the task description and a JSON output template. Additionally, given the prompt sensitivity (Zhou et al. 2022) and ordering bias (Tian et al. 2025) documented in prior VLM studies, we implemented two bias mitigation strategies for VLMs: (1) we developed six semantically equivalent but syntactically diverse prompts (see supplemental material), and (2) we counterbalanced option presentation order to minimize positional bias.

Prediction Performance and Consistency

Prediction Performance Metrics. While the overall class distribution in our study is balanced between accident and non-accident cases, relying solely on overall accuracy can mask meaningful differences in how each class contributes to prediction outcomes. We therefore report the macro F1-score, which balances precision and recall across classes.

Prediction Consistency Metrics. We assessed inter-rater agreement using Fleiss' Kappa (Falotico and Quatto 2015) to measure the degree of inter-agent agreement beyond chance. We used this metric to quantify how consistently humans (across participants) and AI (across prompts) evaluate identical videos. This focuses not only on overall correctness, but on the relative consistency with which different humans and AI models interpret the same scenario.

Assessment of Visual Attention Alignment

We systematically evaluated the alignment of visual attention between humans and AI systems by analyzing their respective attention patterns during diverse traffic scenarios. To achieve this, we employed both continuous and discrete analytical approaches, leveraging high-resolution eye-tracking data for human participants and gradient-based interpretability techniques for (open-source) AI models.

Continuous Attention Analysis. For human participants, we performed continuous attention analysis using established visual saliency metrics applied directly to gaze distributions. Specifically, we quantified spatial alignment between human attention distributions using the Kullback-Leibler Divergence (KL_{div}) (Kullback and Leibler 1951), Correlation Coefficient (CC) (Borji and Itti 2012), and Normalized Scanpath Saliency (NSS) (Peters et al. 2005).

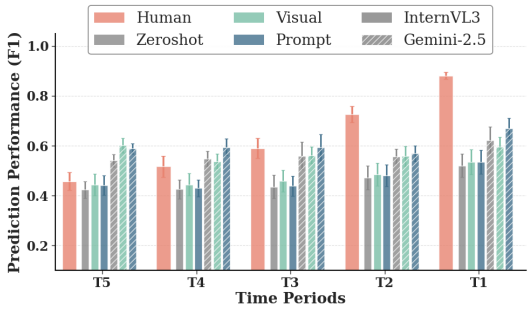
Discrete Attention Analysis. To assess the temporal alignment between human and AI visual attention, we defined discrete Areas of Interest (AOIs) using seven core semantic categories relevant to driving, adapted from the Cityscapes dataset (Cordts et al. 2016): (1) flat surfaces (road, sidewalk, parking, rail track), (2) humans (person, rider), (3) construction elements (building, wall, fence, bridge), (4) traffic-related objects (traffic light, traffic sign), (5) vehicles (car, truck, bus, motorcycle, bicycle), (6) sky, and (7) nature (vegetation, terrain). We generated high-precision AOI masks for each frame by first applying a Mask2Former model (Cheng et al. 2022) to produce coarse semantic maps, which were then refined with SAM2 (Ravi et al. 2024) to achieve precise boundaries. This process enabled detailed tracking of attention transitions, such as shifts from a vehicle to a traffic signal and then to the road.

For each frame, we quantified the attention directed at each Area of Interest (AOI) by calculating a normalized intensity score. This score was computed by summing the saliency values within the AOI mask from the model's output and dividing by the mask's area. This process yielded frame-wise attention hierarchies (i.e., AOI rankings) for both the AI model and each human driver. To establish a robust ground-truth human attention hierarchy that represents a collective consensus, we aggregated the rankings from all human participants using a leave-one-out cross-validation approach. Finally, we measured the concordance between the AI model's attention hierarchy and this aggregated human baseline using Kendall's tau (τ) coefficient (Kendall 1938). Human-to-human alignment was measured in the same way. Overall, this allowed us to benchmark the degree to which a model's assessment of object importance aligned with that of the human participants in our experiment.

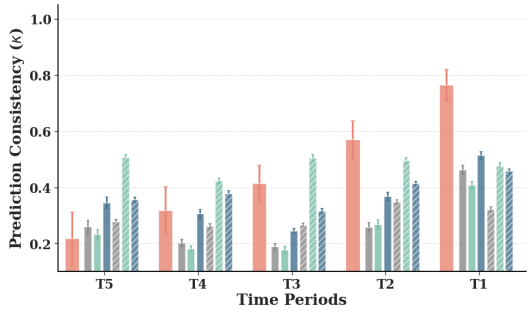
Results

Performance Gap between Human and AI Models

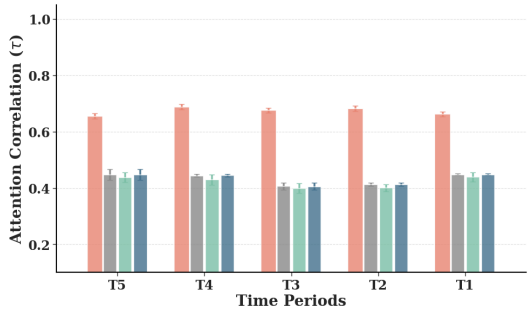
In Figure 2a and Figure 2b, contrast prediction abilities of humans and *zero-shot* AI models as a function of closeness to a potential event (see also first part of Table 1). The results clearly reveal that humans and AI systems exhibit fundamentally different response patterns to temporal urgency. Human participants demonstrated consistent improvement in prediction performance. F1 scores progressed from 0.457 (95% CI: $\pm$0.036) at T5 to 0.881 (95% CI: $\pm$0.015) at



(a) Prediction performance (F1-score). Human performance (coral) exceeds all models from T3 onward. Attention-guided models outperform zero-shot (gray) baselines, with prompt guidance (blue) showing superior results to visual guidance.



(b) AI consistency (κ) improves toward critical time T1 but remains below human-level agreement ($\kappa \geq 0.75$).



(c) Visual attention alignment (Kendall's τ) is higher among humans than among VLMs.

Figure 2: Performance comparison across time periods (T5 to T1 are time periods getting closer to a potential event T_E.) Error bars show 95% bootstrap CIs. Attention-guided model (teal, blue) outperform baselines (gray) but maintain consistent gaps with human performance (coral) across all metrics.

T1, while Fliess's Kappa coefficients rose substantially from 0.125 to 0.766-this pattern was driven by improvements in performance for the accident sequences (see supplemental material for detailed analyses). These findings suggest that humans possess the ability to effectively focus resources for enhanced decision-making towards the event.

In contrast, AI models tested under zero-shot conditions maintained relatively flat performance regardless of temporal conditions. The best open-source model (InternVL-3)

Model	Temporal Periods				
	T5	T4	T3	T2	T1
Human	0.46 ± 0.04	0.52 ± 0.04	0.59 ± 0.04	0.73 ± 0.03	0.88 ± 0.02
Zero-shot					
Gemini-2.5-Flash	0.54 ± 0.03	0.55 ± 0.04	0.56 ± 0.07	0.56 ± 0.04	0.62 ± 0.07
GPT-o4-mini	0.55 ± 0.04	0.56 ± 0.05	0.57 ± 0.02	0.56 ± 0.05	0.62 ± 0.08
InternVL-3	0.43 ± 0.10	0.43 ± 0.11	0.45 ± 0.14	0.48 ± 0.15	0.56 ± 0.12
VideoLLaMA3	0.46 ± 0.06	0.47 ± 0.06	0.48 ± 0.09	0.53 ± 0.07	0.56 ± 0.06
QwenVL-2.5	0.38 ± 0.05	0.38 ± 0.05	0.38 ± 0.06	0.41 ± 0.05	0.42 ± 0.06
Guidance (Prompt-based)					
Gemini-2.5-Flash	0.59 ± 0.03	0.59 ± 0.04	0.59 ± 0.07	0.57 ± 0.04	0.67 ± 0.05
GPT-o4-mini	0.62 ± 0.02	0.62 ± 0.04	0.59 ± 0.02	0.63 ± 0.051	0.63 ± 0.05
InternVL-3	0.46 ± 0.10	0.46 ± 0.11	0.48 ± 0.13	0.52 ± 0.13	0.61 ± 0.11
VideoLLaMA3	0.49 ± 0.05	0.43 ± 0.04	0.47 ± 0.04	0.52 ± 0.03	0.57 ± 0.04
QwenVL-2.5	0.37 ± 0.06	0.39 ± 0.07	0.38 ± 0.05	0.40 ± 0.05	0.43 ± 0.06
Guidance (Visual-based)					
Gemini-2.5-Flash	0.55 ± 0.03	0.56 ± 0.06	0.57 ± 0.08	0.57 ± 0.06	0.60 ± 0.04
GPT-o4-mini	0.60 ± 0.04	0.54 ± 0.04	0.56 ± 0.04	0.56 ± 0.05	0.60 ± 0.05
InternVL-3	0.45 ± 0.11	0.46 ± 0.13	0.46 ± 0.13	0.49 ± 0.14	0.59 ± 0.10
VideoLLaMA3	0.48 ± 0.08	0.49 ± 0.08	0.52 ± 0.04	0.54 ± 0.07	0.59 ± 0.07
QwenVL-2.5	0.40 ± 0.09	0.30 ± 0.08	0.40 ± 0.07	0.43 ± 0.08	0.43 ± 0.07

Table 1: Performance comparison across VLMs and attention guidance methods. Performance is measured with F1-scores across 6 prompts (mean ± 95% CI).

Condition	Metric	Temporal Periods				
		T5	**T4**	**T3**	**T2**	**T1**
Accident	CC	0.72±0.02	0.75±0.02	**0.71±0.02**	**0.74±0.02**	0.73±0.02
	NSS	1.38±0.09	1.46±0.10	**1.27±0.09**	**1.40±0.09**	1.42±0.09
	KL	0.84±0.07	0.76±0.07	**0.90±0.06**	**0.78±0.05**	0.79±0.06
Normal	CC	0.74±0.02	0.73±0.02	0.73±0.02	0.73±0.02	0.73±0.02
	NSS	1.45±0.09	1.39±0.09	1.35±0.08	1.41±0.09	1.39±0.08
	KL	0.77±0.06	0.80±0.07	0.82±0.06	0.80±0.06	0.77±0.06

Table 2: Gaze similarity by condition and period (mean ± 95% CI); **bold** indicates T3–T2 significant differences (Tukey HSD, $p < 0.05$).

had scores ranging from 0.43 to 0.564, with the best closed-source model (Gemini-2.5-Flash) scoring somewhat higher between 0.541 and 0.623. Notably, both types of models had consistently low Kappa coefficients for prediction below 0.3 throughout all temporal periods, indicating clear limitations in maintaining consistent judgment.

Effects of Attention Guidance Strategies. As Figure 2 (see also Table 1) shows, the attention guidance strategies produced small yet consistent improvements in performance. Visual guidance, implemented by blurring peripheral objects in the visual input (green bars), yielded performance improvements for open-source models of around 4% and for closed-source models of $< 1\%$. In contrast, the improvements from prompt-based guidance were larger at 7% and 6.5%, respectively.

Despite these improvements, a significant disparity compared to human performance remained during the critical time periods immediately preceding an event. This divergence highlights a "late-cue gap", suggesting that humans and VLMs employ fundamentally different mechanisms to process potential accident sequences. Whereas human performance improved closer to the critical moment, both zero-shot and guided AI models appeared unable to implement similarly adaptive responses, failing to adequately prioritize cues as the situation evolves.

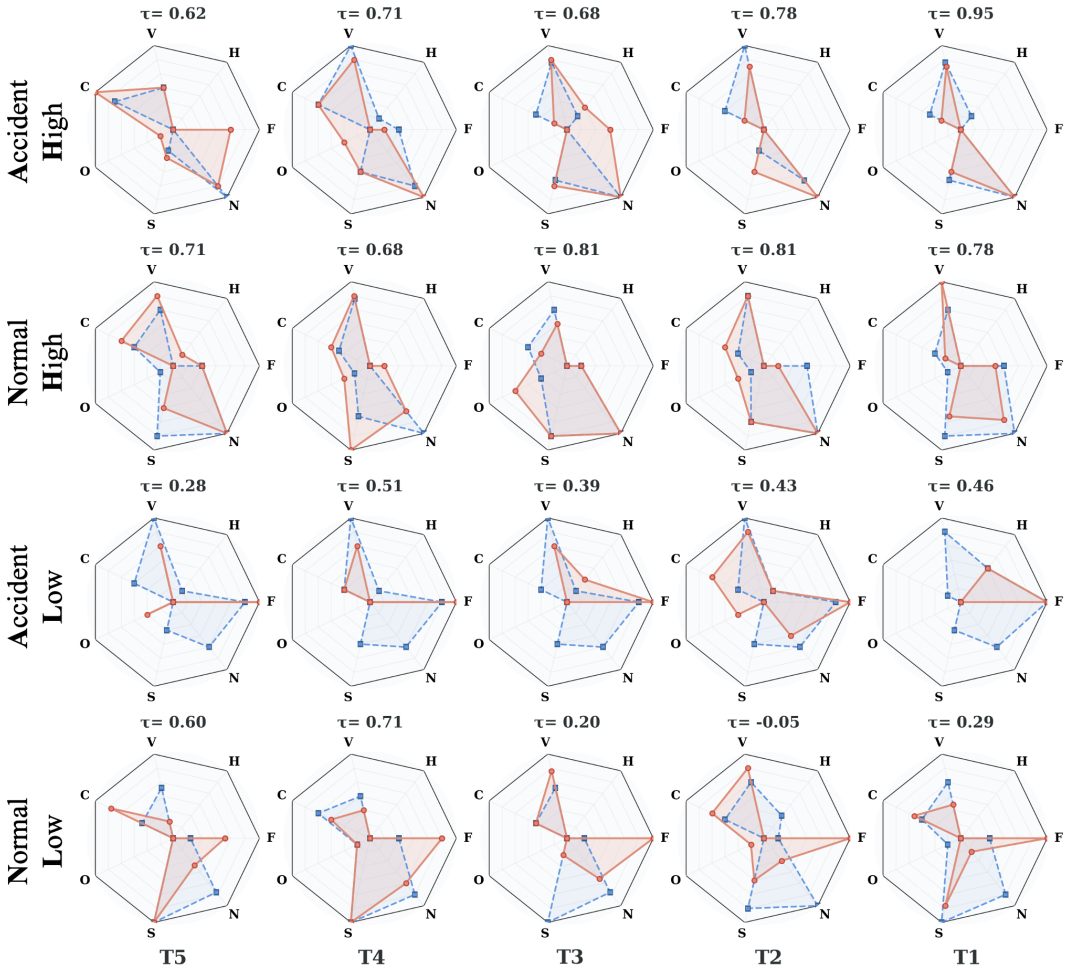


Figure 3: Radar charts compare visual-attention ranks between humans (red filled) and prompt-guided InternVL-3 (blue dashed) across seven AOIs—Vehicle (V), Human (H), Flat (F), Nature (N), Sky (S), Object (O), and Construction (C). Ranks run from 1, outer bold contour and highest attention, to 7, innermost and lowest. Columns indicate temporal segments T5–T1, and rows are ordered by average human–model τ with accident–high (AH), normal–high (NH), accident–low (AL), and normal–low (NL) from top to bottom. Videos in each row correspond to those in Figure 4.

Human-AI Attentional Alignment

Gaze Distribution Analysis. To understand the "late-cue gap" further, we next examined the attentional shifts in the human gaze data in more detail (see supplemental material for further analyses). Table 2 shows descriptive statistics comparing the gaze distributions in human participants across three different metrics. Overall, the level of similarity across participants is strong in each metric (Papageorgiou 2022).

A statistical comparison was conducted using one-way repeated-measures ANOVAs for each video condition to examine temporal patterns in the gaze distributions across time periods. Results revealed a clear divergence between accident and non-accident conditions in terms of temporal stability: accident videos showed a main effect of time period across all gaze comparison metrics (CC, $F(4, 245) = 2.742, p < 0.05, \eta^2 = 0.043$; NSS, $F(4, 245) = 2.452, p < 0.05, \eta^2 = 0.038$; and KL_{div}, $F(4, 245) = 3.196, p < 0.05, \eta^2 = 0.050$). In contrast, non-accident videos showed no significant temporal effects for any metric (all $p > 0.05$), suggesting that routine driving situations maintain consistent visual attention patterns across time periods.

Tukey HSD post-hoc comparisons identified significant pairwise differences in visual attention patterns towards later time points (underlined values in Table 2). The consistent pattern across multiple metrics suggests systematic attentional shifts closer to the event time T_E, contrasting with the temporal reliability observed in routine driving scenarios.

AOI ranking analysis. As Figure 2c shows, the agreement in AOI ranking between our best-scoring open-source model (InternVL-3) and humans falls below human-human consistency as measured by Kendall's τ. The difference here is constant throughout all time periods, and model alignment is seemingly not influenced by any attentional guidance (see supplemental materials for further analyses).

To provide further insight into this attentional misalign-

ment, we analyzed four representative video clips chosen based on InternVL-3's attentional alignment performance (see Figure 3). As the radar plots in Figure 3 show, humans employ strategic adaptation in their attention allocation, shifting focus in response to different scenarios. We identified two distinct human attention strategies dictated by environmental demands: (1) Consistent Allocation, observed in sequences similar to rows 1 and 2 (AH and NH), where attention priorities remain stable despite evolving information or threat levels; and (2) Dynamic Updating, prevalent in sequences similar to rows 3 and 4 (AL and NL), where attention is flexibly reallocated in real-time.

These human attention strategies translate to an AI model's alignment with human visual patterns. We found that model alignment, particularly for InternVL-3, is highest when human attention is temporally stable. In scenarios characterized by consistent human attention (AH and NH), the model achieves strong and stable alignment, with correlation coefficients (τ) ranging from 0.62 to 0.95 across all temporal periods. This suggests that AI models can replicate human attention when it follows predictable patterns.

Conversely, this alignment deteriorates significantly when human attention becomes dynamic. In AL and NL scenarios, where participants must adapt to changing environments, the model struggles to keep pace. For instance, as shown in the AOI visualizations in Figure 4, when participants divert attention to novel environmental features like overpasses or overhead signs, a measurable decrease in human-AI visual alignment occurs. This indicates that as the environment becomes more complex and human attention more adaptive, the model's ability to mirror human focus diminishes.

Accuracy-tau Relationship Analysis

We next examined the relationship between model guidance approaches and human-AI attention alignment across different conditions. We analyzed three types of model training methods (zeroshot, visual guidance, prompt guidance) across two video types (accident, normal scenarios). For each condition, we classified videos into high and low accuracy groups based on each model's performance using median thresholds, then compared the distribution of Kendall's τ values between these groups using Mann-Whitney U tests.

Mann-Whitney U tests revealed significant differences in τ distributions in only one of the six conditions examined. The prompt guidance model in normal driving scenarios showed significantly higher τ values for high accuracy samples (M = 0.481, n = 208) compared to low accuracy samples (M = 0.355, n = 42), with $p < 0.01$. This indicates that the prompt guidance training approach, when performing well on normal driving videos, demonstrates significantly better alignment with human attention patterns compared to when it performs poorly. All other training approaches and conditions, including all accident video scenarios, showed no significant relationships ($p > 0.05$).

These findings demonstrate that the relationship between training method, attention alignment, and performance is highly specific. Only the prompt guidance training approach shows a correlation between performance accuracy and human-AI attention alignment, and only in routine driving situations. Complex accident scenarios show no such relationship across guidance methods.

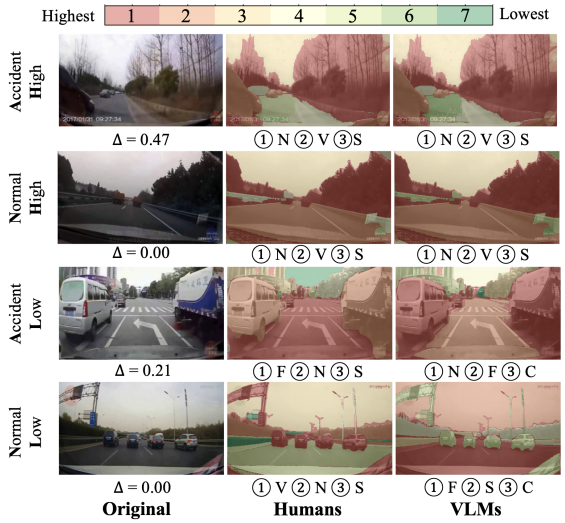


Figure 4: Visualization of attention maps. Shown from left to right are the original video frame (Δ indicates the human-VLM performance gap), the human attention map, and the VLM attention maps for the three highest-ranked regions—Vehicle (V), Human (H), Flat (F), Nature (N), Sky (S), Object (O), and Construction (C). Attentional priority is encoded with a seven-level colormap, where red (rank 1) indicates the highest priority and green (rank 7) the lowest.

Conclusion

This study demonstrates that humans and VLM systems employ fundamentally different visual attention strategies during critical situations. Humans often enhance their predictive accuracy through dynamic reallocation of attention toward core threat elements immediately preceding accidents. Conversely, VLM systems fail to improve performance to human levels, despite attention guidance. Additionally, attentional alignment remains weak across time periods, suggesting much room for improvement.

Future work will need to dive further into different guidance and fine-tuning methods as well as other types of VLMs to test how much prediction and consistency performance can be enhanced. Likewise, since explainable AI methods for closed-source VLMs are highly limited, further work is needed to obtain better black-box estimates of attentional processing in these models. Finally, our work only annotated and sampled a subset of traffic scenarios—future research will need to expand our Attention-DADA dataset to encompass a wider range of traffic-related contexts.

Nonetheless, we believe that our findings suggest that evaluation paradigms for safety-critical VLM systems shift focus from outcome-centric to also include process-centric approaches. Trust in unpredictable, real-world environments can only be established through VLMs that not only generate accurate predictions but also demonstrate reasoning and attention allocation patterns comparable to human experts.

Acknowledgments

Supported by the NRF of Korea (BK21 FOUR, RS-2025-00555141, RS-2022-NR074628) and IITP grants funded by the Korea government (MSIT) (RS-2019-II190079, Dept. of AI, Korea Univ.; RS-2021-II212068, AI Innovation Hub).

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv:1606.06565*.

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2017. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv:1711.06104*.

Baheri, A. 2022. Safe reinforcement learning with mixture density network, with application to autonomous driving. *Results in Control and Optimization*, 6: 100095.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv:2309.16609*.

Borji, A.; and Itti, L. 2012. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 185–207.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Corvelo Benz, N.; and Rodriguez, M. 2023. Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems*, 36: 14609–14636.

Falotico, R.; and Quatto, P. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2): 463–470.

Fang, J.; Li, L.-l.; Zhou, J.; Xiao, J.; Yu, H.; Lv, C.; Xue, J.; and Chua, T.-S. 2024. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22030–22040.

Fang, J.; Yan, D.; Qiao, J.; Xue, J.; Wang, H.; and Li, S. 2019. Dada-2000: Can driving accident be predicted by driver attention$f$ analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 4303–4309. IEEE.

Google. 2025. Gemini 2.5 Flash: A Comprehensive Reference. Google AI Announcement. Accessed: 2025-07-04.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6047–6056.

Gu, J.; Han, Z.; Chen, S.; Beirami, A.; He, B.; Zhang, G.; Liao, R.; Qin, Y.; Tresp, V.; and Torr, P. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv:2307.12980*.

Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R. R.; Cheng, M.-M.; and Hu, S.-M. 2022. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3): 331–368.

Itti, L.; Koch, C.; and Niebur, E. 2002. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11): 1254–1259.

Jiang, C.; Cornman, A.; Park, C.; Sapp, B.; Zhou, Y.; Anguelov, D.; et al. 2023. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9644–9653.

Jin, Y.; Sun, Z.; Xu, K.; Chen, L.; Jiang, H.; Huang, Q.; Song, C.; Liu, Y.; Zhang, D.; Song, Y.; et al. 2024. Videolavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv:2402.03161*.

Jose, C.; Moutakanni, T.; Kang, D.; Baldassarre, F.; Darcet, T.; Xu, H.; Li, D.; Szafraniec, M.; Ramamonjisoa, M.; Oquab, M.; et al. 2025. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24905–24916.

Kadner, F.; Thomas, T.; Hoppe, D.; and Rothkopf, C. A. 2023. Improving saliency models' predictions of the next fixation with humans' intrinsic cost of gaze shifts. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, 2104–2114.

Kang, S.; Kim, J.; Kim, J.; and Hwang, S. J. 2025. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9339–9350.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2): 81–93.

Kirk, H. R.; Whitefield, A.; Rottger, P.; Bean, A. M.; Margatina, K.; Mosquera-Gomez, R.; Ciro, J.; Bartolo, M.; Williams, A.; He, H.; et al. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37: 105236–105344.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Kuznietsov, A.; Gyevnar, B.; Wang, C.; Peters, S.; and Albrecht, S. V. 2024. Explainable AI for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems*.

Lee, J.; Kim, S.; Won, S.; Lee, J.; Ghassemi, M.; Thorne, J.; Choi, J.; Kwon, O.-K.; and Choi, E. 2023. Visalign: Dataset for measuring the alignment between ai and humans in visual perception. *Advances in Neural Information Processing Systems*, 36: 77119–77148.

Li, H.; and Li, B. 2025. Enhancing vision-language compositional understanding with multimodal synthetic data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24849–24861.

Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19401–19411.

Mastrianni, A.; Kim, M. S.; Sullivan, T. M.; Sippel, G. J.; Burd, R. S.; Gajos, K. Z.; and Sarcevic, A. 2025. To Recommend or Not to Recommend: Designing and Evaluating AI-Enabled Decision Support for Time-Critical Medical Events. *arXiv:2505.11996*.

Muttenthaler, L.; Greff, K.; Born, F.; Spitzer, B.; Kornblith, S.; Mozer, M. C.; Müller, K.-R.; Unterthiner, T.; and Lampinen, A. K. 2024. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*.

Ollikka, N.; Abbas, A.; Perin, A.; Kilpeläinen, M.; and Deny, S. 2024. Humans beat deep networks at recognizing objects in unusual poses, given enough time. *arXiv:2402.03973*.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. OpenAI blog post.

Pan, C.; Yaman, B.; Nesti, T.; Mallik, A.; Allievi, A. G.; Velipasalar, S.; and Ren, L. 2024. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14760–14769.

Papageorgiou, S. N. 2022. On correlation coefficients and their interpretation. *Journal of orthodontics*, 49(3): 359–361.

Peirce, J.; Gray, J. R.; Simpson, S.; MacAskill, M.; Höchenberger, R.; Sogo, H.; Kastman, E.; and Lindeløv, J. K. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, 51(1): 195–203.

Peters, R. J.; Iyer, A.; Itti, L.; and Koch, C. 2005. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18): 2397–2416.

Piergiovanni, A.; Noble, I.; Kim, D.; Ryoo, M. S.; Gomes, V.; and Angelova, A. 2024. Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26804–26814.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*.

Ryu, H. S.; Ju, U.; and Wallraven, C. 2025. Seeing What Matters: Attentional (Mis-) Alignment Between Humans and AI in VR-Simulated Prediction of Driving Accidents. *IEEE Transactions on Visualization and Computer Graphics*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; and Li, H. D. 2023. Driving with graph visual question answering. *arXiv:2312.14150*.

Tatler, B. W. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7(14): 4–4.

Tian, X.; Zou, S.; Yang, Z.; and Zhang, J. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10599–10609.

Vaccaro, M.; Almaatouq, A.; and Malone, T. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1–11.

Wickens, C. D. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2): 159–177.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xu, J.; Zhou, X.; Yan, S.; Gu, X.; Arnab, A.; Sun, C.; Wang, X.; and Schmid, C. 2024. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13030–13039.

Yu, R.; Yu, W.; and Wang, X. 2024. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, 251–268. Springer.

Yuan, T.; Liu, Y.; Wang, Y.; Wang, Y.; and Zhao, H. 2024. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7356–7365.

Zeng, Y.; Zhang, X.; and Li, H. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv:2111.08276*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv:2501.13106*.

Zhang, M.; Tseng, C.; and Kreiman, G. 2020. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12985–12994.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Y.; Gao, X.; Chen, Z.; and Huang, H. 2025. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18270–18280.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv:2504.10479*.