

Where, What, Why: Towards Explainable Driver Attention Prediction

Yuchen Zhou^{1,2*}; Jiayu Tang^{1*}; Xiaoyan Xiao¹, Yueyao Lin¹, Linkai Liu¹, Zipeng Guo¹,
Hao Fei², Xiaobo Xia², Chao Gou^{1†}

¹Sun Yat-sen University, ²National University of Singapore

<https://github.com/yuchen2199/Explainable-Driver-Attention-Prediction>

Abstract

Modeling task-driven attention in driving is a fundamental challenge for both autonomous vehicles and cognitive science. Existing methods primarily predict where drivers look by generating spatial heatmaps, but fail to capture the cognitive motivations behind attention allocation in specific contexts, which limits deeper understanding of attention mechanisms. To bridge this gap, we introduce Explainable Driver Attention Prediction, a novel task paradigm that jointly predicts spatial attention regions (**where**), parses attended semantics (**what**), and provides cognitive reasoning for attention allocation (**why**). To support this, we present **W³DA**, the first large-scale explainable driver attention dataset. It enriches existing benchmarks with detailed semantic and causal annotations across diverse driving scenarios, including normal conditions, safety-critical situations, and traffic accidents. We further propose **LLada**, a **L**arge **L**anguage model-driven framework for driver attention prediction, which unifies pixel modeling, semantic parsing, and cognitive reasoning within an end-to-end architecture. Extensive experiments demonstrate the effectiveness of LLada, exhibiting robust generalization across datasets and driving conditions. This work serves as a key step toward a deeper understanding of driver attention mechanisms, with significant implications for autonomous driving, intelligent driver training, and human-computer interaction.

1. Introduction

As an old proverb goes, “The eyes are the window to the soul.” Human eyes serve as a crucial gateway to cognition, reflecting how human perceive, interpret, and interact with their environment [6, 54, 59, 60, 77, 79, 81]. Modeling and understanding human visual attention provide fundamental insights into cognitive abilities, experiential knowledge,

*Equal Contribution.

†Corresponding Author.

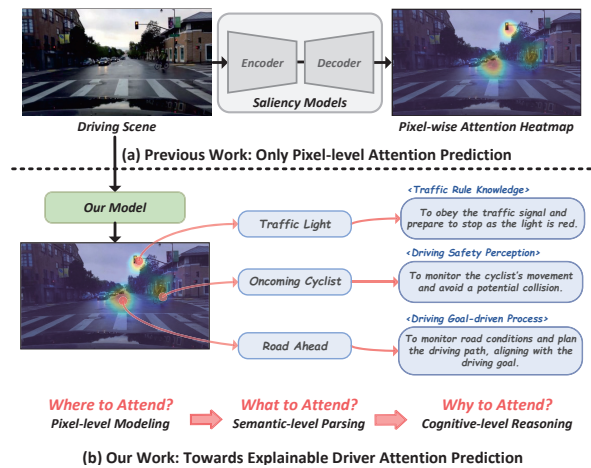


Figure 1. Prior studies focus only on predicting where drivers allocate their attention, offering an implicit and shallow representation of attention. In contrast, our work proposes a more explicit and in-depth paradigm that answers where to attend, what to attend, and why attention is allocated there, revealing underlying factors behind attention allocation such as traffic rule knowledge, driving safety perception, and goal-driven behaviors. Our work unlocks a deeper understanding of driver attention mechanisms, providing new insights to the research community.

and decision-making strategies, especially in complex, task-driven scenarios such as driving [8, 28, 32, 33, 53, 80, 82].

In recent years, driver attention prediction has emerged as a key research direction and has achieved remarkable progress [22, 27, 50, 51, 64]. Existing paradigms [10, 20, 31, 78, 84, 85] primarily focus on predicting *where* drivers look by generating spatial heatmaps, as shown in Fig. 1(a). These methods fundamentally perform pixel-space regression, offering only a shallow and implicit representation of attention mechanisms. However, these approaches fail to capture the underlying cognitive motivations behind attention allocation in specific driving contexts. For instance, in the scenario depicted in Fig. 1, a driver attends to the red traffic light to comply with traffic rules, monitors an

oncoming cyclist at an intersection to ensure driving safety, and observes the road ahead to navigate toward the intended destination. Unfortunately, no existing work has effectively modeled these semantic-level (i.e., *what* is being attended to) and cognitive-level (i.e., *why* attention is allocated) aspects of driver attention. This critical limitation hinders both a deeper understanding of driver attention mechanisms and the real-world applicability of these models.

To bridge these critical gaps, we propose a novel unified task paradigm, termed **Explainable Driver Attention Prediction**, as shown in Fig. 1(b). This paradigm requires models to simultaneously predict pixel-level attention regions (answering *where*), provide semantic-level parsing (answering *what*), and perform cognitive-level reasoning to explain the underlying causes of attention allocation (answering *why*). By integrating spatial, semantic, and cognitive knowledge, this enables a more comprehensive and explainable modeling of driver attention.

Specifically, our work introduces several key advancements over existing driver attention prediction methods: **(1)** We introduce **W³DA**, a large-scale explainable driver attention dataset, which integrates multiple mainstream driver attention benchmarks and incorporates additional annotations for semantic parsing (*what*) and cognitive reasoning (*why*). It provides spatial, semantic, and cognitive labels for attention regions, explicitly enabling *where*, *what*, and *why* reasoning in driver attention modeling. **(2)** We propose **LLada**, a **L**arge **l**anguage model-driven framework for **d**river **a**ttention prediction, which is the first unified architecture capable of jointly predicting attention regions (*where*), semantic interpretations (*what*), and cognitive reasoning (*why*) in driver attention modeling. **(3)** Unlike existing models trained and evaluated separately on individual datasets, we leverage the W³DA dataset for cross-domain collaborative training, enabling a unified modeling approach for normal driving, safety-critical situations, and traffic accidents.

Extensive experiments validate the effectiveness of LLada, demonstrating its superior performance over all state-of-the-art (SOTA) methods on W³DA in both attention map prediction and textual explanation generation. It consistently excels across various driving scenarios. Moreover, trained solely on W³DA, LLada outperforms most independently trained specialized models across the entire DR(eye)VE [51], BDDA [64], and DADA-2000 [19] datasets, achieving highly competitive performance and demonstrating strong generalization. Further qualitative analyses confirm its strong alignment with human driver attention. Our work establishes a solid foundation for comprehensive and explainable attention modeling, paving the way for advancements in cognitive science and real-world applications such as interpretable autonomous driving and intelligent driver training. Before delving into details, we

clearly emphasize our contributions as follows:

- **(Paradigm)** We introduce Explainable Driver Attention Prediction, a novel paradigm that extends traditional driver attention prediction by jointly reasoning over *where*, *what*, and *why* to achieve a more comprehensive understanding of attention in driving.
- **(Dataset)** We present W³DA, a large-scale explainable driver attention dataset, which enhances existing benchmarks with semantic (*what*) and causal (*why*) annotations, spanning normal, safety-critical, and accident scenarios.
- **(Methodology)** We propose LLada, the first large language model-driven driver attention prediction framework, enabling end-to-end training and joint prediction of attention regions (*where*), semantic parsing (*what*), and cognitive reasoning (*why*).
- **(Experiments)** Extensive experiments demonstrate that LLada consistently outperforms existing SOTA methods across multiple tasks, scenarios, metrics, and settings, demonstrating superior robustness and generalization.

2. Related Work

2.1. Driver Attention Prediction

Datasets. Specialized driver attention datasets have significantly advanced the field. Early works like DR(eye)VE [51] introduced a pipeline to analyze driver gaze in normal traffic, recording attention data from eight drivers. Similarly, the LBW dataset [31] provided gaze data collected from 28 drivers in real driving conditions. Cheng et al. introduced IVGaze [11], which collected in-vehicle gaze data but lacked corresponding driving footage. The BDD-A dataset by Xia et al. [64] focused on safety-critical events like emergency braking and traffic congestion. DADA-2000 [19] extended this to traffic accidents, including 2,000 videos and corresponding gaze data. While existing datasets have provided large-scale driver attention data, they primarily associate spatial heatmaps with driving scenes without capturing the cognitive reasoning behind attention allocation, which limits a deeper understanding of driver attention mechanisms. To address this, we introduce W³DA, the first large-scale dataset for explainable driver attention modeling. W³DA extends beyond heatmaps by providing fine-grained annotations that link attention heatmaps to their semantic meaning and underlying causes, offering unprecedented insights into driver attention mechanisms.

Models. Attention prediction models have evolved from early image processing techniques like ITTI [29] and GBVS [23] to deep learning-based architectures. CNNs became fundamental, enabling richer visual feature representations. Palazzi et al. [51] integrated RGB, optical flow, and segmentation, while Xia et al. [64] employed convolutional LSTMs. Deng et al. [14] proposed a fully convolutional encoder-decoder. Multi-resolution networks [25] and

heterogeneous networks [26] further incorporated multi-level visual content. ASIAFNet [37] emphasized semantic-based object-level attention. Beyond CNNs, Inverse Reinforcement Learning [3] and Generative Adversarial Networks [2, 3] have been explored to model driver attention. Amadori et al. [1] incorporated driving inputs and vehicle states in VR simulations. Unsupervised methods [85] explored scene dynamics without reliance on labeled attention data. Shi et al. [57] proposed FOD-Net with saliency guidance. Recent advances integrate Transformers [18] to enhance global context modeling. ACT-Net [21] pioneered CNN-Transformer fusion, followed by FBLNet [10], which introduces a feedback loop mechanism for experience accumulation, further bridging attention modeling with human-like learning. Despite these advancements, existing models still fail to explicitly capture the cognitive motivations behind driver attention. To address this limitation, we propose LLada, a unified end-to-end framework that jointly models spatial (*where*), semantic (*what*), and cognitive (*why*) factors, enabling a more comprehensive and explainable approach to driver attention prediction.

2.2. Multimodal Large Language Models

Recent advancements in multimodal large language models (MLLMs) [38, 39, 42, 43, 47–49, 63, 69, 71, 72, 76, 83] have driven significant progress across image-level captioning [7], object-level detection [74], and pixel-level segmentation [35, 65], with notable applications in autonomous driving [13, 58, 62, 75] and robotics [40, 45, 73]. Building on this foundation, recent studies have explored incorporating human attention as an additional modality to enhance model performance and interpretability. Voila-A [68] aligned MLLMs with the human gaze (captured via AR/VR devices), ensuring better attention correspondence between models and users. GazeXplain [9] predicted visual scanpaths and generated natural language explanations for fixations using BLIP [36]. However, its explanations remained semantic-level parsing, describing what is attended to without explicitly uncovering the underlying cognitive motivations behind attention allocations. In contrast, LLada introduces a unified framework that jointly models spatial (*where*), semantic (*what*), and cognitive (*why*) factors, enabling a more comprehensive and explainable approach to driver attention modeling. To the best of our knowledge, this is the first attempt to model underlying cognitive motivations of attention using MLLMs.

3. W³DA Dataset

Existing driving attention datasets primarily predict where drivers look via spatial heatmaps. However, they lack semantic and cognitive understanding, limiting their ability to comprehensively model task-driven visual attention in driving. To bridge this gap, we move beyond spatial attention

prediction to a deeper, more explainable understanding of driver attention. A key challenge is the absence of a dataset that provides semantic parsing (*what*) and cognitive reasoning (*why*) for driver attention.

To address this, we introduce W³DA, a semantic and reasoning-aware driving attention dataset that enhances conventional *where* annotations with high-quality *what* and *why* labels. As shown in Fig. 2(a), W³DA unifies multiple mainstream driving attention datasets, integrating DR(eye)VE [51] and LBW [31] for normal driving, BDDA [64] for critical situations, and DADA-2000 [19, 20] for traffic accidents. This diverse data composition allows for more robust cross-domain training, improving model generalization and real-world applicability. Furthermore, as depicted in Fig. 2(b-d), we introduce an efficient and highly reliable data annotation pipeline that assigns semantic labels to attention targets (*what*) and provides context-aware cognitive reasoning for attention causes (*why*), enabling a more comprehensive understanding of driver attention.

3.1. Data Annotation Pipeline

Creating a large-scale, high-quality dataset for explainable driver attention prediction presents several inherent challenges. First, while vast amounts of driver attention data exist, many driving scenarios—such as cruising, highway driving, or following another vehicle—exhibit stable attention patterns, with drivers primarily fixating on the preceding vehicle or the road’s vanishing point. In such cases, attention shifts are minimal, making frame-by-frame annotation inefficient and resource-intensive. Secondly, manual annotation by human experts incurs substantial time and financial costs. While leveraging powerful MLLMs for automated annotation is a promising alternative, MLLMs are prone to hallucinations, often generating erroneous or contextually inappropriate annotations that compromise dataset reliability. To address these challenges, we introduce an efficient and highly reliable semi-automatic pipeline for dataset annotation.

For the first challenge, we introduce an attention-aware key sample selection strategy that avoids redundant frame-by-frame annotation. This approach prioritizes identifying key moments indicative of significant attention shifts by evaluating perceptual and attention-related changes between consecutive samples, ensuring the annotations capture high-information moments. Our selection process is grounded in three key criteria: (1) **semantic similarity of driving scenes**, capturing perceptual variations in the environment; (2) **spatial divergence of driver attention**, capturing variations in the spatial distribution of attended regions; and (3) **semantic similarity of driver attention**, reflecting contextual shifts in attended regions. Specifically, criteria (1) and (3) are computed using the CLS token from the CLIP-Large image encoder [56], while (2) is evaluated

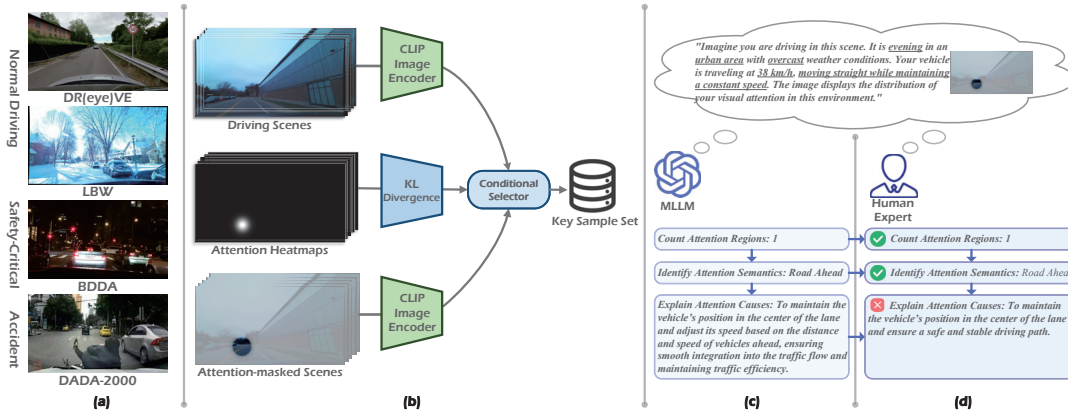


Figure 2. The W³DA annotation pipeline. (a) W³DA integrates multiple driver attention benchmarks, covering normal driving, safety-critical scenarios, and accidents. (b) To reduce annotation redundancy in scenarios with stable attention patterns (e.g., cruising, following another vehicle), we propose a key sample selection method, leveraging semantic scene similarity, spatial attention divergence, and semantic attention similarity to identify key moments of significant attention shifts. (c) We employ visual and contextual prompts with a structured chain-of-thought to guide an advanced MLLM in generating preliminary context-aware annotations. (d) Finally, human experts validate and refine these annotations to ensure accuracy, contextual relevance, and alignment with real driving scenarios.

via KL divergence between attention heatmaps. To implement this, we initialize the key sample set K with the first sample from the original driving video $V \in \mathbb{R}^{T \times H \times W \times 3}$. For each subsequent sample in V , we evaluate these three metrics relative to the most recent key sample in K . A sample is added to K if the KL divergence (criterion 2) exceeds a predefined threshold or if the semantic similarity metrics (criteria 1 and 3) fall below their respective thresholds. Otherwise, it is considered redundant and excluded from further consideration.

For the second challenge, we propose an effective approach that integrates visual and contextual prompts with a clever chain-of-thought strategy to guide an advanced MLLM in generating preliminary annotations that are both useful and contextually relevant. Specifically, we provide the MLLM with comprehensive contextual cues, including driving behavior, weather conditions, and scene categories, encouraging it to generate annotations aligned with the current scenario. Additionally, we overlay attention maps onto images using a simple grayscale mask, directing the MLLM’s focus toward the relevant regions where driver attention is concentrated. We further employ a chain-of-thought to guide the MLLM through a progressive analysis of the semantics and underlying causes of driver attention. This process prompts the MLLM first to determine the number of attention regions within the scene, then identify the content within each region (answering *what*), and finally explore the causal factors behind the attention allocation (answering *why*). In practice, we utilize the Qwen-VL-Max API as our MLLM to implement this approach.

Despite the MLLM’s strong visual reasoning capabilities, hallucinations remain a potential issue. To address this,

we involve human experts to review and refine the model’s outputs based on their domain expertise, ensuring the final annotations are accurate, reasonable, and contextually aligned with the driving scenario.

3.2. Dataset Statistics

Building on our pipeline, W³DA comprises 69,980 key samples extracted from 3,548 video scenes, covering diverse driving conditions. Specifically, it includes 22,839 samples from normal driving, 22,950 samples from critical situations, and 24,191 samples from accidents. W³DA spans a variety of weather conditions (sunny, rainy, cloudy, overcast, snowy, foggy), geographic locations (urban, rural, highway, suburban, mountain, tunnel), and time periods (morning, afternoon, night). Furthermore, W³DA provides rich textual annotations: the average length of *what* annotations is 5.40 words, while the average length of *why* annotations is 24.32 words, offering fine-grained semantic and cognitive insights into driver attention. To maintain a rigorous experimental setup, we adhere to the original training-validation-test splits of the four source datasets, ensuring no data leakage. W³DA advances driver attention modeling, offering valuable insights for both academia and industry. More details can be found in the supplementary material.

4. Explainable Driver Attention Prediction

4.1. Problem Definition

The goal of Explainable Driver Attention Prediction is to predict **where a driver is likely to attend, what regions are attended, and why attention is allocated**. This task goes beyond conventional attention prediction by unifying

spatial, semantic, and causal reasoning, enabling a more explainable and cognition-driven understanding of driver attention. Formally, given a driving scene represented as an image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ with contextual information \mathbf{C} (e.g., weather, road conditions, driving behavior), the goal is to estimate a pixel-wise attention map $\mathbf{A} \in \mathbb{R}^{H \times W}$, a set of semantic descriptions of attended regions $\mathcal{S} = \{s_i\}_{i=1}^N$, and a set of causal explanations for attention allocation $\mathcal{E} = \{e_i\}_{i=1}^M$, where s_i denotes the semantic label of the i -th attended region, and e_i represents the explanation behind the attention allocated to the i -th region.

4.2. LLada Architecture

LLada comprises four core components: 1) a pretrained visual encoder \mathcal{F}_{vis} , 2) a large language model \mathcal{F}_{LLM} , 3) a special attention token [ATTN], and 4) a cognitive-aware attention decoder \mathcal{F}_{dec} , as illustrated in Fig. 3. Given a driving scene, the visual input \mathbf{I} and contextual information \mathbf{C} are first processed by the pretrained visual encoder \mathcal{F}_{vis} and the LLM \mathcal{F}_{LLM} , respectively. Then, the attention token [ATTN] encodes high-level cognitive cues within \mathcal{F}_{LLM} and enables interaction between the language and vision modalities through the attention decoder \mathcal{F}_{dec} . Finally, \mathcal{F}_{dec} generates the spatial attention map \mathbf{A} , while the \mathcal{F}_{LLM} outputs the semantic descriptions \mathcal{S} and causal explanations \mathcal{E} . Next, we provide a detailed description of each component.

Visual Encoder. We utilize a visual encoder consisting of the pretrained vision foundation model CLIP-ViT-L [56] and a linear projector ϕ , following the setup of LLaVA [42]. Given the visual input \mathbf{I} , CLIP extracts image features, which are then projected by ϕ into the visual token space, producing embeddings $\mathbf{h}_{\text{vis}} \in \mathbb{R}^{N \times d}$ for further processing by \mathcal{F}_{LLM} .

Large Language Model. The driving scene context \mathbf{C} is tokenized into context tokens by the LLM tokenizer \mathcal{T} : $\mathbf{h}_{\text{con}} = \mathcal{T}(\mathbf{C})$. The visual tokens \mathbf{h}_{vis} and context tokens \mathbf{h}_{con} are then concatenated with a fixed prompt sequence $\mathbf{h}_{\text{prompt}}$ and fed into the LLM, denoted as \mathcal{F}_{LLM} . The output token embeddings $\tilde{\mathbf{y}}_{\text{txt}}$ are generated as:

$$\tilde{\mathbf{y}}_{\text{txt}} = \mathcal{F}_{\text{LLM}}([\mathbf{h}_{\text{prompt}}, \mathbf{h}_{\text{vis}}, \mathbf{h}_{\text{con}}]), \quad (1)$$

where $[\cdot]$ denotes concatenation. The final text responses are generated by applying a linear classifier to $\tilde{\mathbf{y}}_{\text{txt}}$ for next-token prediction. During attention prediction, these responses include semantic descriptions of the attended regions \mathcal{S} and causal explanations for attention allocation \mathcal{E} . In practice, we use Vicuna-7B [12] as the \mathcal{F}_{LLM} .

Attention Token. To adapt the LLM for attention prediction, we introduce a special attention token, [ATTN], extending the original LLM vocabulary. This token encodes high-level cognitive cues within the LLM and facilitates cross-modal interaction between language and vision

via \mathcal{F}_{dec} . During attention prediction, the output token sequence $\tilde{\mathbf{y}}_{\text{txt}}$ includes the [ATTN] token. We extract its corresponding embedding $\tilde{\mathbf{y}}_{\text{txt}}[\text{ATTN}]$, which is then projected into the \mathcal{F}_{dec} space via an MLP projector ψ :

$$\mathbf{h}_{\text{attn}} = \psi(\tilde{\mathbf{y}}_{\text{txt}}[\text{ATTN}]). \quad (2)$$

This enables \mathcal{F}_{dec} to decode high-level cognitive attention information from \mathbf{h}_{attn} .

Cognitive-aware Attention Decoder. The attention decoder \mathcal{F}_{dec} facilitates the interaction between the attention token embedding \mathbf{h}_{attn} and the visual features \mathbf{h}_{vis} , decoding context-aware cognitive information to generate the pixel-wise attention map $\mathbf{A} \in \mathbb{R}^{H \times W}$. Specifically, we first introduce a cross-attention mechanism to guide the decoder’s awareness of high-level cognitive cues from \mathbf{h}_{attn} and derive cognitively-driven visual features \mathbf{h}'_{vis} :

$$\mathbf{h}'_{\text{dec}} = \mathbf{h}_{\text{vis}} + \text{Repeat}(CA(\mathbf{h}_{\text{attn}}, \mathbf{h}_{\text{vis}})), \quad (3)$$

where $CA(q, kv)$ denotes the cross-attention operation and $\text{Repeat}(\cdot)$ is a replication operation, which adds $CA(\mathbf{h}_{\text{attn}}, \mathbf{h}_{\text{vis}})$ to each token in \mathbf{h}_{vis} . Next, the \mathbf{h}'_{vis} are reshaped from a 2D shape of $\frac{HW}{p^2} \times C$ to a standard 3D feature map of size $\frac{H}{p} \times \frac{W}{p} \times C$, where C is the feature channel size and p is the image patch size. Then, we apply a series of five 3×3 convolutional layers with batch normalization and ReLU activation to reduce the feature dimensionality to $\frac{H}{p} \times \frac{W}{p}$. Finally, a bilinear upsampling operation is applied to generate the attention map \mathbf{A} at the full image size of $H \times W$.

Training Objectives. The LLada model is trained end-to-end using the attention map prediction loss \mathcal{L}_{map} and the textual explanation generation loss \mathcal{L}_{txt} . The overall objective \mathcal{L} is the weighted sum of these losses:

$$\mathcal{L} = \lambda_{\text{map}} \mathcal{L}_{\text{map}} + \lambda_{\text{txt}} \mathcal{L}_{\text{txt}}, \quad (4)$$

where λ_{map} and λ_{txt} are scaling factors. Specifically, the attention map loss \mathcal{L}_{map} consists of binary cross-entropy (BCE) and Kullback-Leibler (KL) divergence, encouraging accurate pixel-wise attention maps (answering *Where?*):

$$\mathcal{L}_{\text{map}} = \lambda_{\text{bce}} \mathbf{BCE}(\hat{\mathbf{A}}, \mathbf{A}) + \lambda_{\text{kl}} \mathbf{KL}(\hat{\mathbf{A}}, \mathbf{A}), \quad (5)$$

where $\hat{\mathbf{A}}$ is the predicted attention map and \mathbf{A} is the ground truth. The textual explanation generation loss \mathcal{L}_{txt} is formulated as the autoregressive cross-entropy (CE) loss, guiding the LLM to generate correct semantic descriptions (answering *What?*) and causal explanations (answering *Why?*) of the attended regions:

$$\mathcal{L}_{\text{txt}} = \lambda_{\text{what}} \mathbf{CE}(\hat{\mathcal{S}}, \mathcal{S}) + \lambda_{\text{why}} \mathbf{CE}(\hat{\mathcal{E}}, \mathcal{E}), \quad (6)$$

where $\hat{\mathcal{S}}$ and $\hat{\mathcal{E}}$ are the predicted semantic descriptions and causal explanations, respectively. That \mathcal{S} and \mathcal{E} are the ground truth descriptions and explanations.

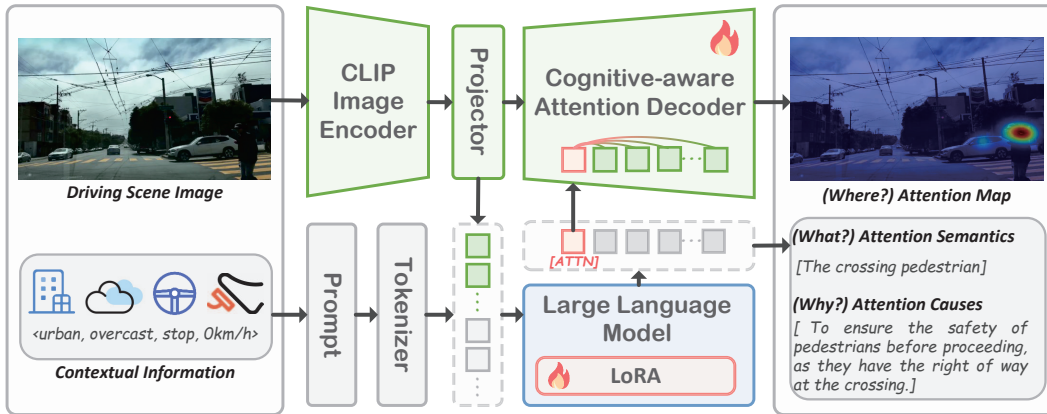


Figure 3. Overall architecture of LLada. Given a driving scene represented as an image with contextual information, the image is processed by the visual encoder (CLIP encoder and projector) to extract visual tokens, while the contextual information, along with prompt sequences, is tokenized by the LLM tokenizer. The visual and text tokens are then fed into the LLM. To adapt the LLM for explainable attention prediction tasks, a special attention token [ATTN] is introduced to encode high-level cognitive cues. Finally, [ATTN] performs cross-attention with the visual tokens, generating the attention map (answering *Where?*) through cognitive-aware attention decoder, while the LLM produces the semantic descriptions (answering *What?*) and causal explanations (answering *Why?*) of attended regions.

5. Experiment

5.1. Experimental Setting

Implementation Details. We train our model on 4 NVIDIA A100 GPUs, utilizing the DeepSpeed engine for efficient optimization. The training process employs the AdamW [46] optimizer with a learning rate of 0.0003, and WarmupDecayLR as the learning rate scheduler with 100 warmup iterations. For loss balancing, we set the scaling factors \mathcal{L}_{map} , \mathcal{L}_{txt} , \mathcal{L}_{bce} , \mathcal{L}_{kl} , $\mathcal{L}_{\text{what}}$, and \mathcal{L}_{why} to 2, 1, 1, 0.1, 1, and 1, respectively. Training is conducted with a batch size of 8 per device, using gradient accumulation with a step size of 5 to accommodate memory constraints. To retain the knowledge in the pre-trained LLM \mathcal{F}_{LLM} , we apply LoRA [24] for parameter-efficient fine-tuning while keeping the visual encoder \mathcal{F}_{vis} entirely frozen. The cognitive-aware attention decoder \mathcal{F}_{dec} is trained from scratch.

Evaluation Metrics. We follow the evaluation setup of prior driver attention prediction studies [10, 20] and assess the spatial accuracy of generated attention maps using AUC_J , AUC_B , SIM, CC, KLdiv, and NSS. To evaluate the language quality of the generated *what* and *why* textual explanations, we adopt BLEU, METEOR, ROUGE, and CIDEr-R. This comprehensive evaluation allows us to effectively measure the model’s performance in spatial, semantic, and cognitive reasoning for driver attention.

5.2. Results on W³DA

Attention Map Prediction (Where). We first evaluate the attention map prediction quality on the W³DA dataset, comparing our proposed LLada with 10 state-of-the-art methods: ITTI [29], GBVS [23], DeepGaze I [34], DeepGaze

IIE [41], MLNet [17], CDNN [14], FBNet [16], ConvNeXt [44], ERFNet [70], and GazeXplain [9]. ITTI and GBVS are early saliency models based on handcrafted features, while the others are learning-based approaches. DeepGaze I and DeepGaze IIE use official pre-trained models due to unavailable training codes. MLNet, CDNN, FBNet, ConvNeXt, and ERFNet are specialized attention prediction models optimized solely for this task, whereas GazeXplain and LLada jointly optimize attention map prediction and textual explanation generation. Table 1 presents the performance comparison. LLada surpasses all SOTA methods in every scenario. In particular, LLada achieves substantial gains on the KLdiv , AUC_J , and AUC_B metrics. For KLdiv, it outperforms the second-best method (ERFNet) by 38.40%, 22.79%, and 11.65% in normal driving, safety-critical situations, and traffic accidents, respectively. Compared to GazeXplain, which jointly optimizes attention map prediction (*where*) and textual explanation generation (*what & why*), LLada consistently achieves superior performance across all metrics, demonstrating its effectiveness.

Textual Explanation Generation (What & Why).

We further evaluate the quality of textual explanations on the W³DA dataset, comparing our proposed LLada with the multi-task model GazeXplain. To ensure a comprehensive evaluation, we also establish strong baseline methods to validate LLada’s effectiveness. Unlike our end-to-end approach, these baselines adopt a two-stage paradigm: first, a conventional attention prediction model generates an attention map; second, an MLLM interprets the attention map to generate textual explanations for *what* and *why*. Specifically, we select DeepGaze I, DeepGaze IIE, MLnet, CDNN,

Table 1. Comparison of attention map prediction performance. **Bold** and underline indicate the best and second-best results. † marks multi-task models jointly optimizing Attention Map Prediction (*where*) and Textual Explanation Generation (*what/why*).

Method	Normal Driving						Safety-Critical Situation						Traffic Accident					
	KLdiv↓	CC↑	SIM↑	AUC _J ↑	AUC _B ↑	NSS↑	KLdiv↓	CC↑	SIM↑	AUC _J ↑	AUC _B ↑	NSS↑	KLdiv↓	CC↑	SIM↑	AUC _J ↑	AUC _B ↑	NSS↑
ITTI [29]	3.216	0.093	0.080	0.676	0.665	0.662	2.807	0.049	0.119	0.618	0.613	0.410	3.339	0.033	0.073	0.595	0.600	0.266
GBVS [23]	2.572	0.294	0.139	0.868	0.857	2.057	2.238	0.246	0.176	0.839	0.804	1.501	2.826	0.173	0.105	0.814	0.765	1.212
DeepGaze I [34]	3.103	0.207	0.080	0.885	0.546	1.254	2.550	0.237	0.137	0.874	0.595	1.297	3.060	0.206	0.082	0.883	0.673	1.412
DeepGaze IIE [41]	3.071	0.226	0.082	0.887	0.610	1.382	2.505	0.296	0.141	0.926	0.589	1.641	3.026	0.227	0.085	0.917	0.584	1.476
MLnet [17]	2.129	0.547	0.460	0.914	0.836	4.494	1.953	0.528	0.433	0.928	0.874	5.254	2.897	0.344	0.288	0.893	0.784	3.008
CDNN [14]	2.614	0.465	0.394	0.887	0.790	3.687	2.646	0.401	0.350	0.885	0.767	3.831	3.568	0.283	0.254	0.851	0.714	2.446
FBnet [16]	2.980	0.406	0.343	0.869	0.769	3.103	2.585	0.431	0.364	0.887	0.803	3.975	3.197	0.329	0.246	0.873	0.789	2.727
ConvNeXt [44]	2.042	<u>0.570</u>	0.412	0.916	0.848	<u>4.661</u>	1.765	<u>0.567</u>	0.413	0.938	<u>0.877</u>	5.438	3.049	0.377	0.248	0.891	0.806	3.425
ERFNet [70]	1.979	0.558	0.425	<u>0.923</u>	0.840	4.304	<u>1.593</u>	0.538	0.410	<u>0.942</u>	0.868	5.201	<u>2.181</u>	<u>0.391</u>	0.253	<u>0.930</u>	0.846	3.035
GazeXplain† [9]	2.578	0.477	0.389	0.857	<u>0.866</u>	3.945	2.769	0.383	0.321	0.848	0.743	2.299	3.109	0.371	0.236	0.902	0.804	1.598
LLada† (Ours)	1.219	0.583	<u>0.436</u>	0.952	0.908	5.376	1.230	0.579	<u>0.420</u>	0.950	0.912	<u>5.271</u>	1.927	0.396	<u>0.262</u>	0.934	0.889	<u>3.216</u>

Table 2. Comparison of textual explanation generation performance. **Bold** and underline show the best and second-best performances.

Method	Normal Driving				Safety-Critical Situation				Traffic Accident			
	BLEU↑	METEOR↑	ROUGE↑	CIDEr-R↑	BLEU↑	METEOR↑	ROUGE↑	CIDEr-R↑	BLEU↑	METEOR↑	ROUGE↑	CIDEr-R↑
DeepGaze I [34] + LLaVA [42]	0.281	0.274	0.357	0.287	0.130	0.178	0.269	0.078	0.179	0.213	0.291	0.167
DeepGaze IIE [41] + LLaVA [42]	0.281	0.275	0.361	0.301	0.253	0.283	0.346	0.289	0.253	0.266	0.344	0.272
MLnet [17] + LLaVA [42]	0.294	0.283	0.398	0.279	0.285	0.294	0.395	0.312	0.258	0.264	0.373	0.260
CDNN [14] + LLaVA [42]	0.299	0.287	<u>0.403</u>	0.283	0.285	0.294	0.397	0.312	0.263	0.270	0.377	0.265
ConvNeXt [44] + LLaVA [42]	0.291	0.282	0.396	0.268	0.117	0.173	0.268	0.078	0.225	0.247	0.345	0.209
ERFNet [70] + LLaVA [42]	0.294	0.283	0.396	0.274	<u>0.287</u>	<u>0.297</u>	<u>0.398</u>	0.307	<u>0.270</u>	<u>0.272</u>	0.375	0.258
GazeXplain [9]	<u>0.307</u>	<u>0.299</u>	0.216	<u>0.422</u>	0.185	0.290	0.367	<u>0.554</u>	0.167	0.195	<u>0.438</u>	<u>0.656</u>
LLada (Ours)	0.436	0.360	0.582	0.963	0.444	0.375	0.593	1.233	0.376	0.318	0.520	1.002

ConvNeXt, and ERFNet as first-stage predictors. For the second stage, we fine-tune LLaVA [42] (CLIP-ViT-L [56] + Vicuna-7B [12]) via LoRA [24], using the same MLLM as LLada to process grayscale attention maps for textual explanation generation. As shown in Table 2, LLada consistently outperforms all baselines across scenarios and metrics, surpassing both GazeXplain and all two-stage baselines, which demonstrates its superior cognitive reasoning capabilities in driver attention analysis. The results from Table 1 (Attention Map Prediction) and Table 2 (Textual Explanation Generation) collectively confirm that our LLada surpasses all previous methods, showcasing its capability for a deeper understanding of driver attention mechanisms.

5.3. Results on Conventional Attention Prediction

To evaluate the effectiveness of LLada trained on the cross-domain dataset W³DA, we compare it with domain-specific attention prediction models fully trained on DR(eye)VE [51], BDDA [64], and DADA [19, 20]. Performance is evaluated on the complete test sets of these datasets. As shown in Table 3, LLada demonstrates competitive performance, particularly in terms of KLdiv, outperforming the second-best domain-specific models by 29.81%, 20.69%, and 5.49% on the DR(eye)VE, BDDA, and DADA test sets, respectively. Additionally, GazeXplain, trained with cross-domain multi-task learning on W³DA, also achieves promising results. Notably, W³DA provides 39,642 keyframes for training, while domain-specific models use 28,632 (DR(eye)VE), 26,325 (BDDA), and 33,939

(DADA) frames. Despite the cross-domain setting, training LLada on W³DA does not significantly increase cost, highlighting its strong generalization across domains and the efficiency of key-frame selection in W³DA.

5.4. Qualitative Analyses

We quantitatively compare our LLada with the latest GazeXplain [9] on explainable driver attention prediction, as shown in Fig. 4. LLada’s pixel-wise attention map and textual explanations align more closely with the Ground Truth, capturing attention mechanisms across spatial, semantic, and cognitive levels. In complex scenarios, such as busy traffic flow (Row 2) and pedestrian interactions (Row 3), LLada outperforms GazeXplain. LLada correctly focuses on the pedestrian in front of the vehicle (Row 3) and provides context-aware, task-driven explanations like “assessing the pedestrian’s movement and speed to avoid a potential collision.” In contrast, GazeXplain misses this key focus and offers less relevant, scene-independent explanations. See supplementary materials for more examples.

5.5. Ablation Study

To validate the effectiveness of LLada in jointly reasoning over *where*, *what*, and *why*, we conduct ablation studies. First, we remove the *where* task (attention map prediction) to assess its impact on *what* and *why* (textual explanation generation). As shown in Fig. 5, removing *where* causes a significant performance drop in all textual explanation metrics. This demonstrates that *where* plays a crucial role in

Table 3. Comparison with domain-specific models for attention map prediction. **Bold** and underline denote the top two performances.

(a) DR(eye)VE test set.				(b) BDDA test set.				(c) DADA test set.					
Training	Method	KLdiv↓	CC↑	Training	Method	KLdiv↓	CC↑	SIM↑	Training	Method	KLdiv↓	CC↑	SIM↑
DR(eye)VE	MLNet [17]	2.00	0.44	BDDA	HWS [64]	2.07	0.48	0.35	DADA	HWS [64]	2.77	0.33	0.22
	RMDN [5]	1.77	0.41		U2NET [55]	1.47	0.56	0.36		U2NET [55]	<u>1.86</u>	0.47	0.30
	DR(eye)VE [51]	1.40	0.56		MINET [52]	10.50	0.49	0.36		MINET [52]	9.99	0.39	0.30
	HWS [64]	1.72	0.51		DRIVE [4]	13.83	0.32	0.26		DRIVE [4]	4.03	0.37	0.25
	SCAFNet [20]	<u>1.35</u>	<u>0.59</u>		SCAFNet [20]	1.48	0.56	<u>0.40</u>		SCAFNet [20]	2.17	<u>0.48</u>	0.34
	DVAM [15]	1.38	0.57		DBNET [61]	1.85	0.56	0.38		DBNET [61]	2.77	0.40	0.26
	TransConvNet [67]	1.37	0.58		FBLNet [10]	<u>1.40</u>	0.64	0.47		PGNET [66]	5.28	0.46	0.37
	MTSF [30]	1.72	0.56		MTSF [30]	1.61	0.51	0.36		FBLNet [10]	1.92	0.50	0.33
W ³ DA	GazeXplain	1.48	0.49	W ³ DA	GazeXplain	2.85	0.38	0.32	W ³ DA	GazeXplain	1.93	0.46	0.29
	LLada	1.04	0.67		LLada	1.16	0.60	0.47		our LLada	1.82	<u>0.48</u>	<u>0.36</u>

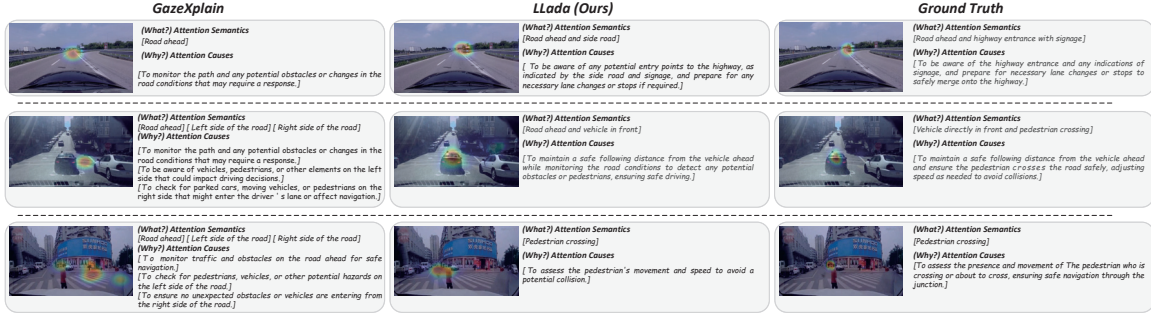


Figure 4. Quantitative examples. More in supplementary materials.

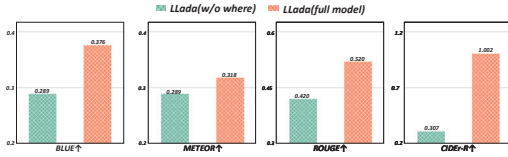


Figure 5. Impact of where on textual explanation generation.

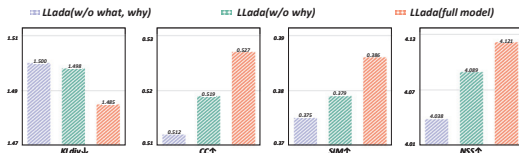


Figure 6. Impact of what and why on attention map prediction.

what and *why* reasoning by providing essential spatial-level cues for more accurate and coherent explanations.

Next, we construct a baseline without *what* and *why* reasoning. We then progressively incorporate these tasks to assess their impact on the attention map prediction, as shown in Fig. 6. Introducing the *what* task improves *where* performance, indicating that textual descriptions provide valuable semantic cues for attention localization. Adding the *why* task further boosts *where* performance, confirming that causal reasoning strengthens attention map prediction. Overall, these findings highlight the mutual reinforcement between *where* and *what/why* reasoning, particularly

the strong facilitative effect of attention map prediction on textual explanation generation, demonstrating the effectiveness of LLada’s joint optimization framework.

6. Conclusion

We introduce Explainable Driver Attention Prediction, a novel paradigm that extends beyond traditional pixel-wise attention modeling by jointly predicting *where* drivers look, *what* they attend to, and *why* their attention is allocated. This enables a cognitive-level understanding of driver attention, bridging perception and reasoning in autonomous driving. To support this paradigm, we present W³DA, the first large-scale explainable driver attention dataset, which significantly enriches existing benchmarks with detailed semantic and causal annotations across diverse real-world driving scenarios. Furthermore, we propose LLada, a Large Language Model-driven framework that unifies attention prediction, semantic parsing, and cognitive reasoning in an end-to-end architecture, offering human-like interpretability beyond conventional models. Extensive quantitative and qualitative analyses demonstrate LLada’s strong generalization across diverse driving conditions and datasets, highlighting the impact of cognitive cues in driver attention modeling. By systematically decoding the latent cognitive factors underlying attention allocation, our work pushes the frontier of task-driven attention modeling in driving, providing new insights for cognitive science, transparent autonomous systems, and human-computer interaction.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China under Grant 62373387, in part by Natural Science Foundation of Guangdong Province (No. 2023A1515030264), in part by Beijing Natural Science Foundation (4252048), and in part by the Shenzhen Fundamental Research Program (Grant No. JCYJ20240813151301003). Xiaobo Xia is supported by MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China (Grant No. 2421002).

References

- [1] Pierluigi Vito Amadori, Tobias Fischer, and Yiannis Demiris. Hammerdrive: A task-aware driving visual attention model. *IEEE T-ITS*, 23(6):5573–5585, 2021. 3
- [2] Javier Araluce, Luis M Bergasa, Manuel Ocaña, Rafael Barea, Elena López-Guillén, and Pedro Revenga. Aragan: A driver attention estimation model based on conditional generative adversarial network. In *IV*, pages 1066–1072. IEEE, 2022. 3
- [3] Sonia Bae, Erfan Pakdamanian, Inki Kim, Lu Feng, Vicente Ordonez, and Laura Barnes. Medirl: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning. In *ICCV*, pages 13178–13188, 2021. 3
- [4] Wentao Bao, Qi Yu, and Yu Kong. Drive: Deep reinforced accident anticipation with visual explanation. In *ICCV*, pages 7619–7628, 2021. 8
- [5] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani. Recurrent mixture density network for spatiotemporal visual attention. In *ICLR*, 2017. 8
- [6] Giuseppe Cartella, Marcella Cornia, Vittorio Cuculo, Alessandro D’Amelio, Dario Zanca, Giuseppe Boccignone, and Rita Cucchiara. Trends, applications, and challenges in human attention modelling. In *IJCAI*, pages 7971–7979, 2024. Survey Track. 1
- [7] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *NeurIPS*, 37:19472–19495, 2025. 3
- [8] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *ECCV*, pages 91–107. Springer, 2020. 1
- [9] Xianyu Chen, Ming Jiang, and Qi Zhao. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *ECCV*, pages 314–333. Springer, 2024. 3, 6, 7
- [10] Yilong Chen, Zhixiong Nan, and Tao Xiang. Fbnet: Feedback loop network for driver attention prediction. In *ICCV*, pages 13371–13380, 2023. 1, 3, 6, 8
- [11] Yihua Cheng, Yaning Zhu, Zongji Wang, Hongquan Hao, Yongwei Liu, Shiqing Cheng, Xi Wang, and Hyung Jin Chang. What do you see in vehicle? comprehensive vision solution for in-vehicle gaze estimation. In *CVPR*, pages 1556–1565, 2024. 2
- [12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 5, 7
- [13] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *WACV*, pages 958–979, 2024. 3
- [14] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and B. S. Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE T-ITS*, 21(5):2146–2154, 2020. 2, 6, 7
- [15] Tao Deng, Lianfang Jiang, Yi Shi, Jiang Wu, Zhangbi Wu, Shun Yan, Xianshi Zhang, and Hongmei Yan. Driving visual saliency prediction of dynamic night scenes via a spatio-temporal dual-encoder network. *IEEE T-ITS*, 25(3):2413–2423, 2023. 8
- [16] Guanqun Ding, Nevrez İmamoglu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. Fbnet: Feedback-recursive cnn for saliency detection. In *MVA*, pages 1–5. IEEE, 2021. 6, 7
- [17] Samuel F Dodge and Lina J Karam. Visual saliency prediction using a mixture of deep neural networks. *IEEE T-IP*, 27(8):4080–4090, 2018. 6, 7, 8
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 3
- [19] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attention? analyzed by a benchmark. In *ITSC*, pages 4303–4309. IEEE, 2019. 2, 3, 7
- [20] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE T-ITS*, 23(6):4959–4971, 2022. 1, 3, 6, 7, 8
- [21] Chao Gou, Yuchen Zhou, and Dan Li. Driver attention prediction based on convolution and transformers. *The Journal of Supercomputing*, 78(6):8268–8284, 2022. 3
- [22] Chao Gou, Yuchen Zhou, Yao Xiao, Xiao Wang, and Hui Yu. Cascade learning for driver facial monitoring. *IEEE T-IV*, 8(1):404–412, 2023. 1
- [23] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *NeurIPS*, 19, 2006. 2, 6, 7
- [24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6, 7
- [25] Zhongxu Hu, Chen Lv, Peng Hang, Chao Huang, and Yang Xing. Data-driven estimation of driver attention using calibration-free eye gaze and scene features. *IEEE T-IE*, 69(2):1800–1808, 2021. 2
- [26] Zhongxu Hu, Yiran Zhang, Qinghua Li, and Chen Lv. A novel heterogeneous network for modeling driver attention

- with multi-level visual content. *IEEE T-ITS*, 23(12):24343–24354, 2022. 3
- [27] Zhixin Huang, Yuchen Zhou, and Chao Gou. Task-oriented scanpath prediction with spatial-temporal information in driving scenarios. In *PRCV*, pages 177–191. Springer, 2024. 1
- [28] Zhixin Huang, Yuchen Zhou, Jie Zhu, and Chao Gou. Driver scanpath prediction based on inverse reinforcement learning. In *ICASSP*, pages 8306–8310, 2024. 1
- [29] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE T-PAMI*, 20(11):1254–1259, 1998. 2, 6, 7
- [30] Lisheng Jin, Bingdong Ji, Baicang Guo, Huanhuan Wang, Zhuotong Han, and Xingchen Liu. Mtsf: Multi-scale temporal–spatial fusion network for driver attention prediction. *IEEE T-ITS*, 2024. 8
- [31] Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *ECCV*. Springer Nature Switzerland, 2022. 1, 2, 3
- [32] Iuliia Kotseruba and John K Tsotsos. Attention for vision-based assistive and automated driving: A review of algorithms and datasets. *IEEE T-ITS*, 23(11):19907–19928, 2022. 1
- [33] Iuliia Kotseruba and John K. Tsotsos. Data limitations for modeling top-down effects on drivers’ attention. In *IV*, 2024. 1
- [34] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. In *ICLRW*, 2015. 6, 7
- [35] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024. 3
- [36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3
- [37] Qiang Li, Chunsheng Liu, Faliang Chang, Shuang Li, Hui Liu, and Zehao Liu. Adaptive short-temporal induced aware fusion network for predicting attention regions like a driver. *IEEE T-ITS*, 23(10):18695–18706, 2022. 3
- [38] Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. Lmeye: An interactive perception network for large language models. *IEEE T-MM*, 26:10952–10964, 2024. 3
- [39] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shen Yuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025. 3
- [40] Fei Lin, Yonglin Tian, Yunzhe Wang, Tengchao Zhang, Xinyuan Zhang, and Fei-Yue Wang. Airvista: Empowering uavs with 3d spatial reasoning abilities through a multimodal large language model agent. In *ITSC*, pages 476–481. IEEE, 2024. 3
- [41] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *ICCV*, pages 12919–12928, 2021. 6, 7
- [42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023. 3, 5, 7
- [43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 3
- [44] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 6, 7
- [45] Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon-Camarasa. Robollm: Robotic vision tasks grounded on multimodal large language models. In *ICRA*, pages 12428–12435. IEEE, 2024. 3
- [46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [47] Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, et al. Deem: Diffusion models serve as the eyes of large language models for image perception. *arXiv preprint arXiv:2405.15232*, 2024. 3
- [48] Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*, 2024.
- [49] Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025. 3
- [50] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. ” looking at the right stuff”-guided semantic-gaze for autonomous driving. In *CVPR*, pages 11883–11892, 2020. 1
- [51] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver’s focus of attention: The dr(eye)ve project. *IEEE T-PAMI*, 41(7):1720–1733, 2019. 1, 2, 3, 7, 8
- [52] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020. 8
- [53] Warren Woodruff Pettine, Dhruva Venkita Raman, A David Redish, and John D Murray. Human generalization of internal representations through prototype learning with goal-directed attention. *Nature Human Behaviour*, 7(3):442–463, 2023. 1
- [54] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual Review of Neuroscience*, 13(1):25–42, 1990. 1
- [55] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 8
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Askill, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3, 5, 7
- [57] Yi Shi, Shixuan Zhao, Jiang Wu, Zhangbi Wu, and Hongmei Yan. Fixated object detection based on saliency prior in traffic scenes. *IEEE T-CSVT*, 34(3):1413–1426, 2024. 3
- [58] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *ECCV*, pages 256–274. Springer, 2024. 3
- [59] Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2(1):1–15, 2024. 1
- [60] Lisa J Stephenson, S Gareth Edwards, and Andrew P Bayliss. From gaze perception to social cognition: The shared-attention system. *Perspectives on Psychological Science*, 16(3):553–576, 2021. 1
- [61] Han Tian, Tao Deng, and Hongmei Yan. Driving as well as on a sunny day? predicting driver’s fixation in rainy weather conditions via a dual-branch visual model. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1335–1338, 2022. 8
- [62] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivelm: The convergence of autonomous driving and large vision-language models. In *CoRL*, 2024. 3
- [63] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *ICML*, 2024. 3
- [64] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *ACCV*, pages 658–674. Springer, 2019. 1, 2, 3, 7, 8
- [65] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, pages 3858–3869, 2024. 3
- [66] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *CVPR*, pages 11717–11726, 2022. 8
- [67] Chuan Xu, Bo Jiang, and Yan Su. Transconvnet: Perform perceptually relevant driver’s visual attention predictions. *Computers and Electrical Engineering*, 115:109104, 2024. 8
- [68] Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user’s gaze attention. *NeurIPS*, 37:1890–1918, 2025. 3
- [69] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3
- [70] Kaihui Yang, Junwei Han, Guangyu Guo, Chaowei Fang, Yingzi Fan, Lechao Cheng, and Dingwen Zhang. Progressive adapting and pruning: Domain-incremental learning for saliency prediction. *ACM T-MCCA*, 2024. 6, 7
- [71] Shuo Yang, Zijian Yu, Zhenzhe Ying, Yuqin Dai, Guoqing Wang, Jun Lan, Jinfeng Xu, Jinze Li, and Edith C. H. Ngai. Rama: Retrieval-augmented multi-agent framework for misinformation detection in multimodal fact-checking, 2025. 3
- [72] Yang Yang, Zhiying Cui, Junjie Xu, Changhong Zhong, Wei-Shi Zheng, and Ruixuan Wang. Continual learning with bayesian model based on a fixed pre-trained feature extractor. *Visual Intelligence*, 1(1):5, 2023. 3
- [73] Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. *arXiv preprint arXiv:2505.08548*, 2025. 3
- [74] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *IJCV*, pages 1–19, 2024. 3
- [75] Hao Zhang, Hongyang Li, Xingyu Liao, Feng Li, Shilong Liu, Lionel M Ni, and Lei Zhang. Da-bev: Depth aware bev transformer for 3d object detection. *CoRR*, 2023. 3
- [76] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*, pages 19–35. Springer, 2024. 3
- [77] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJ-CAI*, page 4951. NIH Public Access, 2020. 1
- [78] Chunyu Zhao, Wentao Mu, Xian Zhou, Wenbo Liu, Fei Yan, and Tao Deng. Salm²: An extremely lightweight saliency mamba model for real-time cognitive awareness of driver attention. In *AAAI*, pages 1647–1655, 2025. 1
- [79] Yuchen Zhou, Guang Tan, Mengtang Li, and Chao Gou. Learning from easy to hard pairs: Multi-step reasoning network for human-object interaction detection. In *ACM MM*, pages 4368–4377, 2023. 1
- [80] Yuchen Zhou, Guang Tan, Rui Zhong, Yaokun Li, and Chao Gou. Pit: Progressive interaction transformer for pedestrian crossing intention prediction. *IEEE T-ITS*, 24(12):14213–14225, 2023. 1
- [81] Yuchen Zhou, Linkai Liu, and Chao Gou. Learning from observer gaze: Zero-shot attention prediction oriented by human-object interaction recognition. In *CVPR*, pages 28390–28400, 2024. 1
- [82] Yuchen Zhou, Xinxin Liu, Zipeng Guo, Ming Cai, and Chao Gou. Hkts: A hierarchical knowledge-guided traffic scene graph representation learning framework for intelligent vehicles. *IEEE T-IV*, 2024. 1
- [83] Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. *NIPS*, 37:3122–3156, 2024. 3
- [84] Yuchen Zhou, Chao Gou, Zipeng Guo, Yihua Cheng, and Hyung Jin Chang. Behavior-aware knowledge-embedded model for driver attention prediction. *IEEE T-CSVT*, 2025. 1
- [85] Pengfei Zhu, Mengshi Qi, Xia Li, Weijian Li, and Huadong Ma. Unsupervised self-driving attention prediction via uncertainty mining and knowledge embedding. In *ICCV*, pages 8524–8534, 2023. 1, 3