FSDAM: Few-Shot Driving Attention Modeling via Vision-Language Coupling

Kaiser Hamid[1], Can Cui[2], Khandakar Ashrafi Akbar[3], Ziran Wang[2], and Nade Liang[1]

[1] Texas Tech University, Lubbock, TX, USA
mdmunna@ttu.edu, nade.liang@ttu.edu
[2] Purdue University, West Lafayette, IN, USA
cancui@purdue.edu, ziran@purdue.edu
[3] Towson University, Towson, MD, USA
kakbar@towson.edu

Abstract. Understanding not only where drivers look but also why their attention shifts is essential for interpretable human–AI collaboration in autonomous driving. Driver attention is not purely perceptual but semantically structured. Thus attention shifts can be learned through minimal semantic supervision rather than dense large-scale annotation. We present **FSDAM** (**F**ew-**S**hot **D**river **A**ttention **M**odeling), a framework that achieves joint spatial attention prediction and structured explanation generation using 90 annotated examples. Our key insight is to decompose attention into an explicit reasoning representation, including scene context, current focus, anticipated next focus, and causal explanation, and to learn next-focus anticipation through minimal-pair supervision. To address task conflict and large sample requirements of existing models, and to mitigate task interference under limited data, we introduce a novel dual-pathway architecture in which separate modules handle spatial prediction and caption generation. In addition, we use a training-only vision–language alignment mechanism that injects semantic priors into spatial learning without increasing inference complexity, mitigating task interference under few-shot training. Despite extreme data scarcity, FSDAM achieves competitive performance in gaze prediction, and generates coherent, context-aware structural reasoning for improved interpretability. The model further demonstrates strong zero-shot generalization across multiple driving benchmarks. These results suggest that semantically grounded attention modeling enables data-efficient learning and provides a scalable path toward explainable driver attention systems in data-constrained environments.

Keywords: Vision-language coupling · Driver attention · Few-shot learning

1 Introduction

Driving is a fundamentally visual and anticipatory task. Drivers continuously allocate attention across traffic signals, pedestrians, vehicles, and road geome-

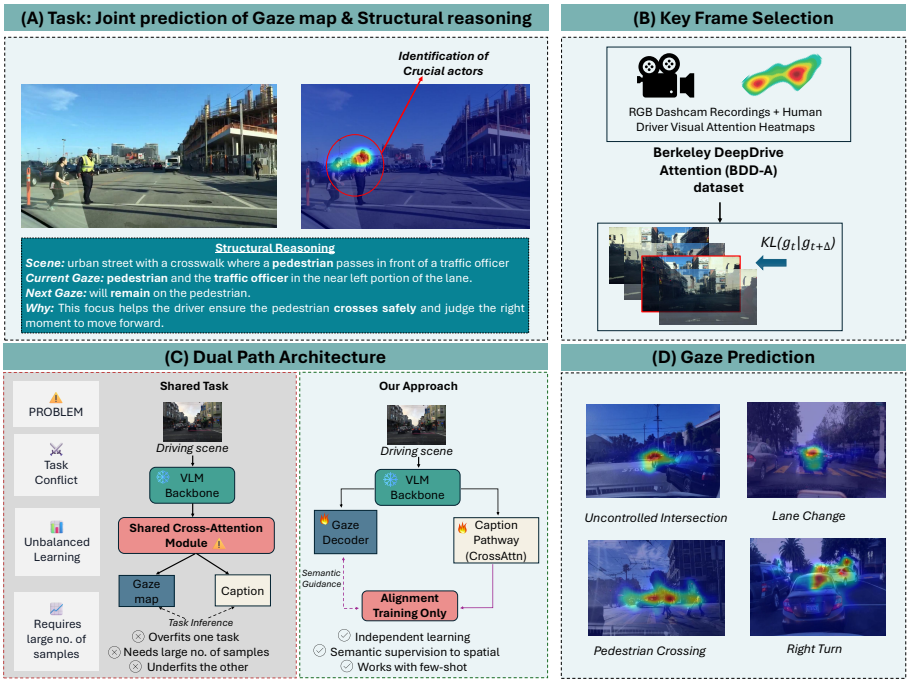

Fig. 1: Few-Shot Driving Attention Modeling (FSDAM). (A) Joint prediction of a gaze map and a structured reasoning (*Scene, Current Gaze, Next Gaze, Why*). **(B)** Key frame selection on BDD-A via KL-divergence mining to select high-change gaze-transition moments between frame pairs. **(C)** Dual-pathway design with separate gaze and caption pathways, coupled through a *training-only* vision–language alignment loss. **(D)** Qualitative gaze predictions across four representative driving scenarios.

try, not only passively reacting to static observation but actively anticipating potential hazards and right-of-way conflicts. This selective attention critically shapes situation awareness (SA) and downstream decision-making. Suboptimal attention allocation with critical actors overlooked remains a major contributor to traffic accidents, with high crash rates at intersections and during lane changes [14,25]. Understanding where and why drivers allocate attention is therefore essential not only for safety analysis but also for autonomous vehicles (AVs) operating in mixed-traffic environments. Furthermore, modeling driver attention enables cross-verification between human intent and machine decisions in human-AI collaboration, supporting interpretable and trustworthy autonomy.

Despite its importance, modeling driver attention remains challenging. Most existing approaches formulate attention prediction as a supervised saliency estimation problem, learning spatial gaze distributions from large-scale eye-tracking datasets. These methods depend heavily on densely annotated data collected under well-represented conditions to stabilize low-level perception [33]. However, large-scale gaze data collection is costly, privacy-sensitive, and often lacks di-

versity in safety-critical or rare scenarios [18, 26]. Moreover, the combinatorial growth of traffic configurations makes exhaustive dataset expansion increasingly impractical [18, 33]. This reliance on extensive supervision limits scalability and generalization to new domains.

Beyond data scarcity, a more fundamental limitation lies in how attention is conceptualized. Treating driver attention as static perceptual saliency overlooks its inherently forward-looking and semantically structured nature. Human drivers do not simply fixate on visually prominent regions; they anticipate future events based on traffic rules, agent behaviors, and scene context. Recent vision–language models (VLMs) offer a promising direction by injecting semantic reasoning into attention modeling. Works such as DriveLM [51], LLada [62], and GazeXplain [8] demonstrate that coupling spatial attention with language reasoning improves interpretability and performance. However, these approaches typically require tens of thousands of annotated samples to jointly align gaze and language representations, limiting their applicability in data-constrained scenarios.

To address these limitations, we develop **FSDAM** (Figure 1), a framework that jointly learns gaze prediction and attention-grounded caption generation from 90 annotated examples. To our knowledge, *this is the first work showing that attention-conditioned language generation can be learned effectively in a few-shot regime*. We also introduce a next-attention target prior derived from semantic cues via minimal-pair supervision. Unlike motion-based temporal forecasting, our method estimates future gaze positions from a single frame, conditioned on the current scene. We construct supervision by selecting frame pairs with clear attention shifts and annotating each transition with a structured caption describing scene context, current focus, anticipated next focus, and causal rationale. At inference, the model captures the forward-looking nature of human attention without requiring video input. Our key insight is to couple spatial attention and language understanding as complementary supervision signals through an explicit, structured attention-reasoning format. Building on this design, we propose a **dual-pathway architecture** that decouples gaze prediction from caption generation, enabling effective joint learning under extreme data scarcity. In summary, we make the following contributions:

- **First few-shot approach for attention-based generation.** We are **the first of its kind** to achieve joint spatial attention prediction and natural language explanation in a few-shot learning regime, training from 90 examples, two orders of magnitude more data efficient than existing joint modeling approaches [8, 62].
- **Structured attention reasoning formulation.** We reformulate driver attention modeling as structured anticipation rather than static saliency prediction. Specifically, we decompose attention into four explicit components, including scene context, current focus, anticipated next focus, and causal explanation, establishing a semantic representation that links spatial gaze patterns with forward-looking reasoning.

- **Dual-pathway architecture with training-only alignment.** We propose a decoupled gaze–language architecture that mitigates negative transfer under limited data. A training-only vision–language alignment mechanism injects semantic priors into spatial prediction without increasing inference complexity.
- **Data-efficient and transferable attention modeling.** Despite training in a few-shot regime, our method achieves competitive performance against fully supervised baselines trained on more data and demonstrates strong zero-shot generalization across multiple driver attention benchmarks.

2 Related Works

2.1 Driver Attention Modeling

Driver attention modeling has been studied as a visual saliency prediction task, estimating spatial gaze heatmaps from dashcam images or video sequences [62]. Early works applied CNNs to learn correlations between scene features and eye fixations, while subsequent models incorporated temporal dynamics and multimodal inputs using CNN-LSTMs and feature fusion from RGB, optical flow, and semantic segmentation [2, 58]. The SEEV model [52] outlines bottom- up (salience, effort) and top-down (expectancy, value) factors influencing attention allocation. Existing models typically emphasize either low-level visual features or task-level signals such as GPS and traffic semantics [17,25]. These approaches provide limited insights into the underlying causes of attention allocation beyond saliency.

2.2 Vision-Language Models for Explainable Attention

Recent work has applied VLMs to enhance attention interpretability through natural language. “Attention Neural Baby Talk" [39] generates captions highlighting hazardous elements by aligning attention masks with descriptions, while DRAMA [38] pairs videos with QA annotations explaining risk rationale. Large-scale VLMs like BLIP [31], Flamingo [1], and LLaVA [36] have enabled new approaches through vision-language pretraining and few-shot learning capabilities [54, 59, 61].

For driver attention specifically, LLada [62] proposes joint modeling of attention maps and textual descriptions through their W3DA dataset (∼70k samples), predicting both where drivers look and why attention is allocated. GazeXplain [8] generates natural language descriptions of gaze scanpaths for general visual attention.

2.3 Few-Shot Learning and Data-Efficient Adaptation

Few-shot learning addresses generalization from minimal examples. In semantic segmentation, PANet [56] achieves 48.1% mIoU with 5 images per class, while

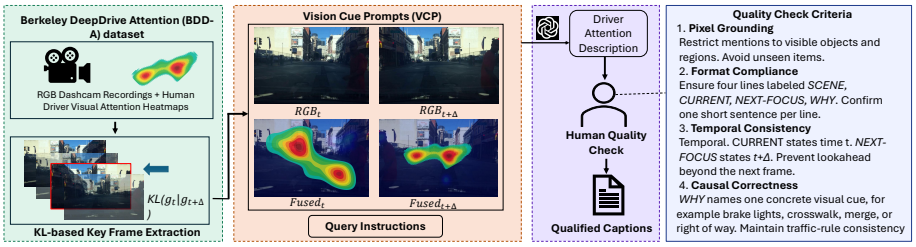

Fig. 2: Dataset curation pipeline. From BDD-A [58] videos to paired frames, GPT-4o captioning with fixed template, human verification, and final captions. Full details in supplementary material.

few-shot object detection methods report 15-30% AP with 10-30 examples [12, 23, 60]. In generative modeling, DreamBooth [48] personalizes Stable Diffusion using 3-5 images (CLIP-I 0.74), while UFC [24] achieves 87.3% accuracy with 30 examples versus 89.1% for fully-supervised baselines (10K+ examples).

Parameter-efficient fine-tuning enables adaptation with minimal updates. LoRA [21] matches full fine-tuning performance while updating 0.01% of parameters, reducing GPT-3's trainable parameters from 175B to 4.7M. For VLMs, LoRA adaptation maintains 95%+ performance with 10-20M trainable parameters [20]. Flamingo demonstrates in- context learning, improving VQAv2 from 49.2% (0-shot) to 63.1% (32-shot) [1], though this approach shows high variance and struggles with structured outputs like spatial maps [11].

In autonomous driving, few-shot learning remains underexplored. Most attention systems train on hundreds of thousands of frames [13, 42, 57].

3 Method

We present our few-shot dependent driver attention modeling framework that jointly models spatial driver attention prediction and structured natural language explanation. These two tasks depend on totally different kinds of spatial understanding, where captioning needs global semantic reasoning while gaze prediction requires localized spatial sensitivity. A shared cross-attention module tends to overfit one task and underfit another [16,29]. To address this imbalance, we introduce a dual-pathway architecture (Figure 3) in which gaze prediction and caption generation are handled by separate modules while still leveraging shared visual features. A vision-language alignment mechanism further provides semantic supervision to spatial prediction, ensuring predicted attention regions correspond to meaningful visual content. This design enables effective joint learning despite data scarcity.

3.1 Problem Formulation

Given a driving scene image $I \in \mathbb{R}^{H\times W\times 3}$, we jointly predict: (1) a spatial attention distribution $\widehat{G} \in \Delta^{S\times S}$ indicating where the driver looks, where $\Delta^{S\times S}$

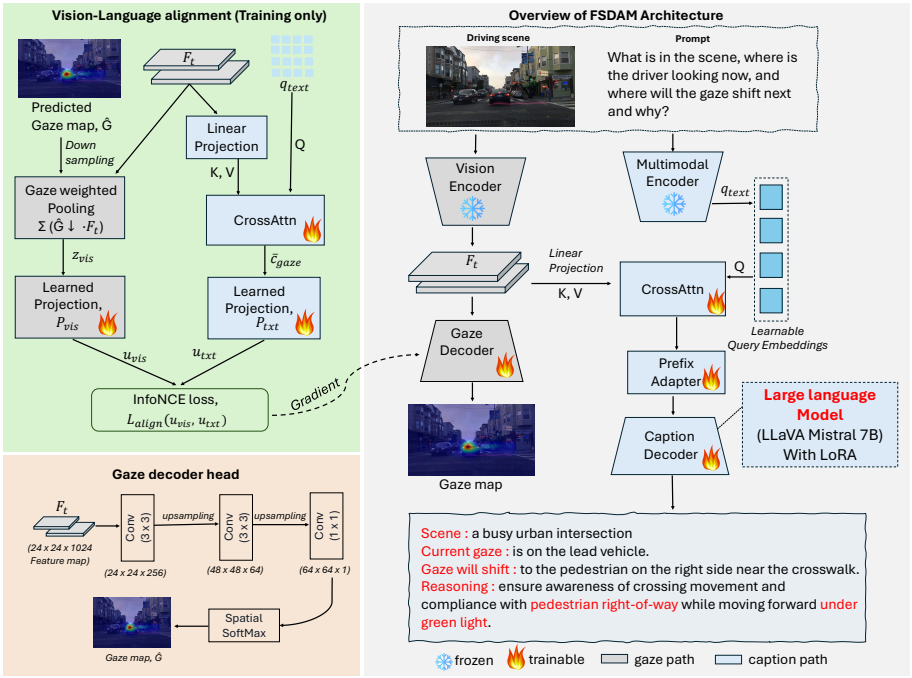

Fig. 3: Overview of the proposed FSDAM architecture (right). A frozen vision–language backbone extracts spatial features F_t and text embeddings q_{text}. The gaze pathway predicts the attention map $\widehat{G}$, while the explanation pathway performs cross-attention over F_t to generate structured reasoning. The training-only vision–language alignment module (top left) applies contrastive supervision between gaze-weighted pooled visual features and text features, and backpropagates gradients to the gaze decoder. The gaze decoder head (bottom left) upsamples F_t to produce $\widehat{G}$.

denotes the probability simplex over an $S \times S$ grid with S=64, and (2) a structured attention reasoning representation $\widehat{y}$ following the format:

$$\widehat{y} = (C_{scene}, C_{current}, C_{next}, C_{why}) \tag{1}$$

Here, each C_* is a natural language sentence describing one aspect of the attention state. Specifically, C_{scene} summarizes the global scene context, $C_{current}$ describes the driver's present focus of attention, C_{next} specifies the likely next focus of attention ("will check the crosswalk"), and C_{why} provides the causal rationale underlying this transition. Together, the four components form a structured attention reasoning representation that links spatial attention patterns with high- level semantic explanations, enabling the model to learn gaze behavior from single frames.

Our architecture builds upon LLaVA-Next-1.6 [35], which comprises a CLIP-ViT-L/14 vision encoder [46] extracting spatial features $F_t \in \mathbb{R}^{B\times 1024\times 24\times 24}$ for input frame t, and a Mistral-7B language model [22] producing image-conditioned

text embeddings $q_{\text{text}} \in \mathbb{R}^{B\times d_\ell}$, where d_ℓ is the language model's hidden dimension, through multimodal fusion [1,35]. These frozen features feed into specialized pathways for gaze prediction and caption generation, with task-specific adaptation enabled by LoRA [21]. We optimize 13.6M task-specific parameters (0.2% of the 7B backbone) while keeping pretrained components frozen to prevent overfitting.

3.2 Spatial Attention Prediction

The gaze pathway predicts where drivers look. A convolutional decoder upsamples F_t from 24×24 to 64×64 resolution through successive convolution and bilinear upsampling layers [37,47], then applies spatial softmax to produce $\widehat{G} \in \Delta^{S\times S}$. We supervise with forward KL divergence to encourage covering all human fixation regions:

$$\mathcal{L}_{\text{KL}} = \text{KL}(G\|\widehat{G}) = \sum_{i,j} G(i,j) \log \frac{G(i,j)}{\widehat{G}(i,j)}. \tag{2}$$

Here, G is the ground-truth gaze distribution and i, j index spatial positions over the $S\times S$ grid. This alone produces overly diffuse predictions under limited data. We augment with a *blur-gap regularizer* that enforces spatial sharpness by penalizing predictions whose quality does not degrade under Gaussian smoothing (σ=1.0). Specifically, we smooth $\widehat{G}$ to obtain $\widetilde{G}$, then penalize cases where blurring does not degrade the prediction:

$$\mathcal{L}_{\text{gaze}} = \mathcal{L}_{\text{KL}} + \lambda \cdot \max\big(0, \text{KL}(G\|\widetilde{G}) - \mathcal{L}_{\text{KL}} + \epsilon\big), \tag{3}$$

with λ=0.3 and ϵ=0.05. This encourages confident spatial localization by exploiting the sharpness–blur relationship in saliency maps [28], where diffuse predictions degrade minimally under Gaussian smoothing while sharp, well-localized predictions do not. The gaze pathway receives gradients only from $\mathcal{L}_{\text{gaze}}$ during forward passes, with semantic supervision added through alignment (Section 3.4).

3.3 Attention-Grounded Caption Generation

Following prefix-tuning [32], we expand q_{text} into M queries via projection W_{cap}, then perform cross-attention [55] over F_t:

$$Q = W_{\text{cap}}(q_{\text{text}}) \in \mathbb{R}^{B\times M\times d}, \quad \text{CTX} = \text{CrossAttn}(Q, K, V). \tag{4}$$

Mean pooling aggregates M context vectors into $\bar{c}_{\text{cap}}$, which a Prefix Adapter Ψ projects to visual prefix tokens $P = \Psi(\bar{c}_{\text{cap}})$ conditioning the decoder. We supervise with autoregressive cross-entropy:

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, P, I), \tag{5}$$

where y contains scene description, current attention, next focus of attention, and causal reasoning. LoRA [21] enables task-specific tuning while keeping most parameters frozen.

3.4 Vision-Language Alignment

To ensure predicted attention regions align with semantic content, we introduce a training-only alignment mechanism. A dedicated cross-attention block processes text queries q_{text} to produce text-conditioned features $\bar{c}_{\text{gaze}}$. In parallel, we pool spatial features F_t using predicted gaze $\widehat{G}_{\downarrow}$ (downsampled to match F_t resolution) as soft weights:

$$z_{\text{vis}} = \sum_{i,j} \widehat{G}_{\downarrow}(i,j)\, F_t[:,:,i,j]. \tag{6}$$

Here, using the model's own predicted attention as pooling weights maintains consistency between training and inference [7, 19].

Both representations project into a shared 256-dimensional space via learned projections P_{vis} and P_{txt}, producing $u^{\text{vis}} = P_{\text{vis}}(z_{\text{vis}})$ and $u^{\text{txt}} = P_{\text{txt}}(\bar{c}_{\text{gaze}})$. We align these with InfoNCE contrastive loss [40]:

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(u_i^{\text{vis}}, u_i^{\text{txt}})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(u_i^{\text{vis}}, u_j^{\text{txt}})/\tau\right)}, \tag{7}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. Gradients pass through spatial pooling to the gaze decoder, pushing attention toward semantically relevant regions.

3.5 Training and Inference

We jointly optimize all three objectives through a weighted combination:

$$\mathcal{L} = w_g\, \mathcal{L}_{\text{gaze}} + w_c\, \mathcal{L}_{\text{cap}} + w_a\, \mathcal{L}_{\text{align}}, \tag{8}$$

where $(w_g, w_c, w_a) = (1.0, 1.0, 0.2)$ balance gaze prediction, caption generation, and vision-language alignment. The weights were chosen through a small-scale grid search on a validation split and were stable across experiments.

Training. We train using AdamW optimizer with separate learning rates for LoRA and task modules. Training uses mixed precision, gradient accumulation for an effective batch of 16, and input size 336×336. Only LoRA and lightweight adapters are updated, totaling 13.6M trainable parameters (∼0.2% of the backbone). Learning rate is adjusted via a ReduceLROnPlateau scheduler.

Inference. During inference, only gaze and caption branches are active. The gaze decoder predicts $\widehat{G}$ from visual features F_t, while the caption pathway generates structured explanations through cross-attention and prefix adaptation. The alignment module is used only during training and adds no cost at inference.

4 Experimental Setup

Dataset Preparation. We construct a gaze-language dataset from BDD-A dataset [58], which contains braking event videos selected from large-scale, crowdsourced driving video data combined with compiled human gaze data from 6.5

seconds prior and 3.5 seconds after each braking event. To identify frames capturing meaningful attention transitions, we employ a KL-divergence-based selection algorithm [50] that detects moments of maximum gaze distribution change between consecutive frames. Local peaks in the KL divergence curve mark anchor frames t where attention shifts abruptly. For each anchor, we select a target frame $t + \Delta$ within $[3, 18]$ frames that maximizes divergence from the anchor, capturing the strongest future attention transition. Following temporal sampling strategies from video action recognition [5], we filter clips shorter than 50 frames and retain the top-K = 2 anchor– target pairs per video for temporal diversity. Each resulting sample $(I_t, I_{t+\Delta}, G_t, G_{t+\Delta})$ consists of two frames and their corresponding gaze maps. We generate structured captions via GPT-4o [41] with human verification (Figure 2), yielding 90 training examples systematically sampled from the candidate key frame pairs to evenly cover eight driving scenario categories (see supplementary for the taxonomy and category-wise counts).

Implementation Details. All experiments are implemented in PyTorch [44]. We build on the LLaVA-Next Mistral- 7B backbone, which integrates a CLIP-based vision tower pretrained on large-scale image–text pairs. The vision–language backbone remains frozen during training, and only lightweight task-specific components (gaze decoder, prefix adapter, alignment head, and LoRA adapters) are updated. Training is performed on a single NVIDIA GH200 GPU with input resolution 336×336, batch size 4, and gradient accumulation over 4 steps (effective batch 16). We use the *AdamW* optimizer with learning rates of 1×10^{-4} for LoRA and 2×10^{-4} for task heads, a *ReduceLROnPlateau* scheduler (factor 0.5, patience 2), and mixed precision of *bfloat16*.

Evaluation Metrics. We evaluate our model using standard saliency and captioning metrics. For gaze prediction, we adopt CC, KL, SIM, AUC-J, AUC-B, and NSS following the official MIT Saliency Benchmark [4]. For caption generation, we report BLEU, METEOR, ROUGE-L, CIDEr-R, and BERTScore using the COCO Caption Evaluation Toolkit [9]. All metrics are computed using their standardized implementations to ensure fair comparison with prior work.

5 Results & Analysis

5.1 Baselines

We evaluate gaze prediction on four datasets: BDD-A [58], DADA-2000 [15], DR(eye)VE [42], and W3D [62]. We compare against classical saliency/gaze baselines (U^2-Net [45], MINet [43], DBNet [53], DeepLabV3 [6]) and a joint gaze–language baseline (LLada [62]). Unless noted otherwise, reproduced baselines are trained on the full BDD-A training split using the official implementations and evaluated on each dataset's official test split. To provide a fair few-shot reference across both classical saliency and joint gaze–language baselines, we retrain U^2-Net and LLada on the same 90-sample BDD-A subset as FSDAM (Table 1, *Few-shot*), ensuring differences are not attributable solely to training-data scale.

For captioning, we compare on W3D against fully supervised baselines (GazeXplain [8], LLada [62]; ∼70k samples), zero-shot and in-context models (Qwen-

Table 1: In-domain driver attention prediction on BDD-A. **Bold** = best, underline = second best. * results from original paper (trained on BDD-A). † fine-tuned on 90 BDD-A samples.

Method	Training Data	CC↑	KL↓	SIM↑	AUC-J↑	AUC-B↑	NSS↑
Fully Supervised (Full BDD-A training set)							
U²-Net [45]	Full BDD-A	0.56	1.50	**0.47**	0.94	0.88	3.95
MINet [43]	Full BDD-A	0.46	10.30	0.16	0.89	0.82	3.67
DBNet [53]	Full BDD-A	0.57	1.30	0.41	0.95	**0.91**	**4.41**
DeepLabV3 [6]	Full BDD-A	0.47	9.62	0.21	**0.97**	0.75	2.56
LLada* [62]	Full BDD-A	**0.60**	1.16	**0.47**	–	–	–
Few-Shot (90 BDD-A samples)							
U²-Net† [45]	BDD-A-90	0.10	3.43	0.13	0.74	0.68	0.63
LLada† [62]	BDD-A-90	0.37	1.86	0.32	0.91	0.81	2.91
FSDAM (Ours)	BDD-A-90	**0.60**	**1.13**	0.43	0.96	**0.91**	4.10

VL [3], LLaVA [36]), and few-shot two-stage baselines pairing gaze models with LLaVA (DeepGazeIIE [34]+LLaVA, MLNet [10]+LLaVA), trained on our same 90 BDD-A samples. Additional analyses and experiments are provided in the supplementary material.

5.2 Quantitative Analysis

Spatial Attention Prediction Table 1 and Table 2 report quantitative results for driver attention prediction across four datasets. Table 1 evaluates in-domain performance on BDD-A under both fully supervised and few-shot regimes. Table 2 evaluates zero-shot transfer to DADA-2000, DR(eye)VE, and W3D, where all baselines are trained on full BDD-A and FSDAM uses only 90 samples. All reproduced models were trained on BDD-A for consistency. FSDAM provides a strong balance between sample efficiency and predictive accuracy. With only 90 training samples, it reaches performance that aligns with or exceeds models trained on much larger datasets, including W3D with ∼70K frames, DADA-2000 with 658K frames, and DR(eye)VE with 555K frames.

In-domain Table 1. On BDD-A [58], FSDAM obtains the lowest KL divergence at 1.13. This improves over LLada [62] by 2.6% and over DBNet [53] by 13%. It matches LLada [62] in CC at 0.60 and ranks second in NSS at 4.10, indicating better spatial attention prediction. Compared to the fully supervised MINet [43], FSDAM improves CC by 30% and reduces KL divergence by 89%, which highlights its effectiveness with limited data. In the few-shot setting, classical saliency models such as U²-Net† [45] degrade severely to CC=0.10, confirming that appearance-based features cannot generalize under extreme data scarcity. LLada† [62], despite sharing the same VLM backbone, also drops substantially to CC=0.37 and KL=1.86, demonstrating that vision-language alignment alone is insufficient without our dual-pathway few-shot design.

Cross-dataset transfer Table 2. Our cross-dataset evaluation demonstrates the strong transferability of our approach. On DR(eye)VE [42], FSDAM reaches the lowest KL divergence at 0.77, improving over LLada [62] by 26% and

Table 2: Zero-shot cross-dataset transfer for driver attention prediction. All baselines are trained on the full BDD-A training set; FSDAM uses only 90 BDD-A samples. No target-domain fine-tuning is applied for any method. **Bold** = best, underline = second best.

Method	Training Data	DADA-2000						DR(eye)VE						W3D					
		CC↑	KL↓	SIM↑	AUC-J↑	AUC-B↑	NSS↑	CC↑	KL↓	SIM↑	AUC-J↑	AUC-B↑	NSS↑	CC↑	KL↓	SIM↑	AUC-J↑	AUC-B↑	NSS↑
U²-Net [45]	Full BDD-A	0.42	2.18	**0.37**	0.91	0.82	2.73	0.57	1.52	0.45	0.90	0.82	3.20	0.44	2.10	0.37	0.90	0.81	2.81
MINet [43]	Full BDD-A	0.32	10.45	0.14	0.82	0.73	2.02	0.44	8.87	0.36	0.84	0.80	2.65	0.35	10.51	0.14	0.83	0.77	2.54
DBNet [53]	Full BDD-A	0.41	1.89	0.29	0.93	0.85	**3.08**	0.48	1.79	0.29	**0.91**	**0.85**	3.71	0.47	1.77	0.33	0.93	0.86	3.50
DeepLabV3 [6]	Full BDD-A	0.42	10.26	0.19	**0.96**	0.73	2.23	**0.67**	8.78	0.47	**0.91**	0.84	2.74	**0.53**	9.70	0.32	**0.95**	0.81	2.35
FSDAM (Ours)	BDD-A-90 (few-shot)	**0.44**	**1.64**	0.33	0.92	**0.86**	2.95	0.61	**0.77**	**0.54**	**0.91**	0.84	**4.04**	0.53	**1.27**	**0.44**	0.91	0.87	3.50

DBNet [53] by 57%. It also achieves the highest SIM at 0.54, a 20% gain over U²-Net [45]. This shows that vision-language alignment helps capture domain stable attention cues that transfer to highway scenarios.

On DADA-2000 [15], FSDAM maintains competitive performance while using far fewer samples, achieving the best KL divergence at 1.64 with a 10% gain over LLada [62] and a 13% gain over DBNet [53]. On W3D, FSDAM matches DeepLabV3 [6] in CC at 0.53 and beats LLada in KL divergence at 1.27 by 13%. It also achieves the second best SIM at 0.44 and NSS at 3.50.

Together, these results reveal that strong driver attention modeling does not require exhaustive gaze supervision. FSDAM consistently matches or outperforms fully supervised baselines across diverse driving scenarios, indicating that our dual-pathway vision-language design captures transferrable, semantically grounded attention cues absent from data-intensive saliency models.

Caption Prediction Analysis Table 5 presents caption generation performance across three driving categories on the W3D dataset. Despite training under our few-shot regime, FSDAM achieves results approaching fully trained baselines trained on the full W3D dataset. In normal driving scenarios, FSDAM attains BLEU 0.42, METEOR 0.37, and CIDEr-R 0.83, closely approaching LLada (BLEU 0.44, METEOR 0.36, CIDEr-R 0.96) trained on ∼70k samples. Generalization holds across challenging scenarios, with competitive scores in safety-critical situations (BLEU: 0.35, METEOR: 0.33, CIDEr-R: 0.47) and traffic accidents (BLEU: 0.33, METEOR: 0.35, CIDEr-R: 0.84). Among few-shot baselines, two-stage approaches (DeepGazeIIE [34]+LLaVA, MLNet [10]+LLaVA) achieve BLEU scores no higher than 0.26 across all scenarios, while FSDAM consistently exceeds 0.33 across normal driving, safety-critical, and accident scenarios. Zero-shot and in-context models remain limited across all scenarios (BLEU < 0.21), confirming that task-specific grounding is necessary for reliable driver caption generation.

5.3 Qualitative Results

Figure 4 compares FSDAM against fully-supervised U2-Net [45] and DeepLabV3 [6]. In multi-agent scenarios (rows a, c, e), FSDAM produces broader attention cov-

Table 3: Comparison of captioning performance across driving scenarios on W3D dataset. Fully-trained baselines use the original W3D training data, Zero-shot and ICL models require no fine-tuning, and our FSDAM and few-shot baselines are trained on a 90-sample BDD-A subset. Higher is better.

Method	Training Regime	Normal Driving				Safety-Critical Situation				Traffic Accident			
		BLEU	METEOR	ROUGE	CIDEr-R	BLEU	METEOR	ROUGE	CIDEr-R	BLEU	METEOR	ROUGE	CIDEr-R
Fully Trained on W3D													
GazeXplain* [8]	Full-data (W3D) [∼70k samples]	0.31	0.30	0.22	0.42	0.19	0.29	0.37	0.55	0.17	0.20	0.44	0.66
LLada* [62]	Full-data (W3D) [∼70k samples]	**0.44**	0.36	**0.58**	**0.96**	**0.44**	**0.38**	**0.59**	**1.23**	**0.38**	0.32	**0.52**	**1.00**
Zero-shot and In-Context Models													
Qwen-VL [3]	Zero-shot (no training)	0.10	0.19	0.28	0.34	0.19	0.21	0.29	0.13	0.08	0.21	0.29	0.12
LLaVA [36]	Zero-shot (no training)	0.12	0.14	0.23	0.35	0.13	0.19	0.11	0.10	0.17	0.26	0.19	0.13
Qwen-VL [3]	In-context learning (no fine-tuning)	0.13	0.18	0.22	0.36	0.21	0.17	0.30	0.23	0.12	0.24	0.33	0.15
Few-shot Learning (BDD-A 90 samples)													
DeepGazeI [27] + LLaVA	Few-shot	0.12	0.23	0.28	0.14	0.13	0.22	0.30	0.18	0.15	0.21	0.31	0.17
DeepGazeIIE [34] + LLaVA	Few-shot	0.11	0.18	0.26	0.17	0.11	0.20	0.32	0.13	0.11	0.19	0.34	0.10
MLNet [10] + LLaVA	Few-shot	0.13	0.19	0.27	0.31	0.26	0.20	0.32	0.12	0.13	0.18	0.33	0.28
FSDAM (Ours)	Few-shot	0.42	**0.37**	0.48	0.83	0.35	0.33	0.46	0.47	0.33	**0.35**	0.34	0.84

erage spanning peripheral pedestrians, intersection agents, and turning trajectories, whereas baselines concentrate on single targets or exhibit fragmented hot spots. In focused tasks (row d), all methods comparably localize attention to the lane-changing vehicle's brake lights. Row (b) reveals a shared limitation: all methods miss the distributed ground-truth pattern across both the lead vehicle and a right-side pedestrian. Overall, FSDAM achieves superior spatial coverage in complex scenes while maintaining competitive precision, validating that vision-language alignment enables effective few-shot attention learning.

5.4 Few-Shot Learning Analysis

Figure 5 presents the impact of varying support set sizes on gaze prediction performance across six metrics. FSDAM demonstrates efficient learning dynamics, with substantial improvements from 1-shot to 5-shot settings. Performance gains are most pronounced between 1-shot and 3-shot (e.g., NSS increases from 2.32 to 3.64, CC improves from 0.41 to 0.53), after which improvements plateau. Notably, our 5-shot model achieves performance (CC 0.58, SIM 0.43, NSS 4.09, KL 1.17) that closely approaches the full-data baseline (CC 0.60, SIM 0.43, NSS 4.10, KL 1.13), recovering 96.7% of full-data CC performance and 99.8% of NSS performance with only five support samples. KL divergence shows a consistent reduction from 1.74 to 1.17 as support size increases, indicating improved distributional alignment between predicted and ground-truth gaze. Importantly, the fact that SIM saturates early (SIM 0.43 at 5-shot matches the full-data SIM 0.43) suggests that the model rapidly learns the overall attention shape, while additional shots primarily refine correlation and peak alignment reflected by CC/NSS. This trend is consistent with our design goal: training-only vision–language alignment provides a strong semantic prior, and the gaze pathway then sharpens spatial localization as supervision increases. Overall, meaningful gaze predictions emerge from minimal supervision, with performance saturating after 3–5 examples.

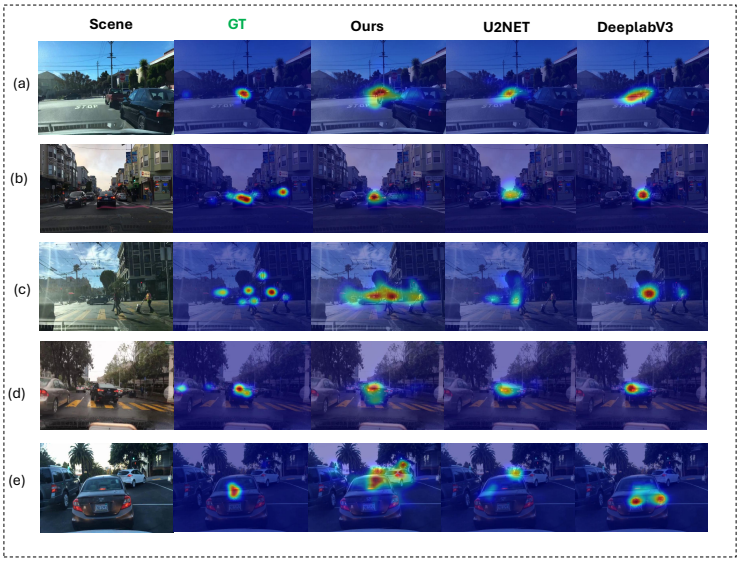

Fig. 4: Qualitative comparison of driver attention prediction on BDD-A test scenes showing input image (left) and attention heatmaps from ground truth, FSDAM (90 samples), U2-Net [45], and DeepLabV3 [6].

5.5 Ablations

We conduct systematic ablation experiments to assess each component's contribution in FSDAM. For gaze prediction, we evaluate on the official BDD-A test set. For caption generation, we evaluate on 49 curated samples with structured explanations, comprising all available test samples with complete four-component annotation. We compare four variants: **Gaze-Only** trains only gaze prediction; **Caption-Only** trains only caption generation; **Shared Cross-Attention** uses a single cross-attention module for both tasks; and **Full FSDAM** uses task-specific cross-attention modules with vision–language alignment.

Figure 6 shows a consistent performance hierarchy across both tasks. The shared cross-attention variant provides marginal gains over single-task baselines but degrades gaze prediction (4% higher KL divergence), indicating negative transfer. In contrast, full FSDAM achieves substantial improvements: **31.9% KL reduction**, **43.3% SIM improvement**, and **66.7% NSS improvement** for gaze prediction, alongside **32.4% ROUGE-L**, **7.1% CIDEr-R**, and **3.4% BERTScore improvements** for caption generation, all relative to single-task baselines. This hierarchy (single-task $<$ shared $<$ dual-pathway) validates three design principles: (1) multi-task training can improve overall performance by leveraging shared visual representations; (2) forcing both tasks to share a single cross-attention module induces negative transfer (e.g., higher KL divergence);

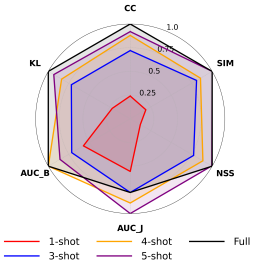

Fig. 5: Few-shot learning performance on BDD-A. Metrics CC, SIM, NSS, AUC-J, AUC-B, and KL (inverted) are min-max normalized to [0,1]; larger area = better.

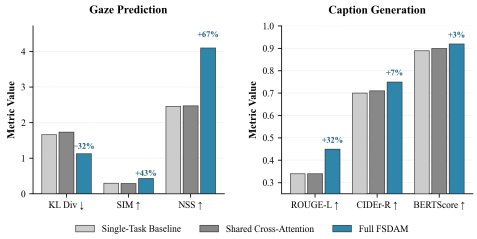

Fig. 6: Ablation study comparing four architectural variants. Dual-pathway FSDAM achieves the largest gains for both gaze prediction and caption generation.

and (3) decoupled pathways with training-only vision–language alignment yield the strongest and most consistent gains across both tasks.

6 Conclusions

We present FSDAM, a framework that achieves joint gaze prediction and natural language explanation from 90 training examples. Through a dual-pathway architecture with training-only vision-language alignment, FSDAM achieves competitive performance against fully-supervised baselines across four benchmarks (BDD-A, DR(eye)VE, DADA-2000, W3D). Ablation results confirm that task-specific cross-attention pathways prevent negative transfer, and that vision–language alignment provides meaningful semantic supervision to spatial prediction without increasing inference complexity. Few-shot learning analysis further shows that meaningful gaze predictions emerge from as few as 3–5 examples, with performance rapidly approaching the full-data regime. Our results demonstrate the effectiveness of this approach across multiple benchmarks. The model generates contextually grounded explanations across diverse driving scenarios, from normal conditions to safety-critical situations and accidents. Future directions include temporal modeling through video inputs and explicit handling of distributed attention patterns to further improve anticipation accuracy in dynamic scenarios. Generalization to extreme weather conditions such as heavy rain or fog remains an open direction for future validation.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022)
2. Alletto, S., Palazzi, A., Solera, F., Calderara, S., Cucchiara, R.: Dr(eye)ve: A dataset for attention-based tasks with applications to autonomous and assisted driving. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 54–60 (2016). https://doi.org/10.1109/CVPRW.2016.14
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? (2017), https://arxiv.org/abs/1604.03605
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset (2018), https://arxiv.org/abs/1705.07750
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017), https://arxiv.org/abs/1706.05587
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020), https://arxiv.org/abs/2002.05709
8. Chen, X., Jiang, M., Zhao, Q.: Gazexplain: Learning to predict natural language explanations of visual scanpaths (2024), https://arxiv.org/abs/2408.02788
9. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server (2015), https://arxiv.org/abs/1504.00325
10. Dodge, S.F., Karam, L.J.: Visual saliency prediction using a mixture of deep neural networks. IEEE Transactions on Image Processing **27**(8), 4080–4090 (2018). https://doi.org/10.1109/TIP.2018.2834826
11. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., Sui, Z.: A survey on in-context learning (2024), https://arxiv.org/abs/2301.00234
12. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector (2020), https://arxiv.org/abs/1908.01998
13. Fang, J., Yan, D., Qiao, J., Xue, J., Wang, H., Li, S.: Dada-2000: Can driving accident be predicted by driver attention? analyzed by a benchmark (2019), https://arxiv.org/abs/1904.12634
14. Fang, J., Yan, D., Qiao, J., Xue, J., Yu, H.: Dada: Driver attention prediction in driving accident scenarios. IEEE Transactions on Intelligent Transportation Systems **23**(6), 4959–4971 (2022). https://doi.org/10.1109/TITS.2020.3044678
15. Fang, J., Yan, D., Qiao, J., Xue, J., Yu, H.: Dada: Driver attention prediction in driving accident scenarios (2023), https://arxiv.org/abs/1912.12148

16. Fifty, C., Amid, E., Zhao, Z., Yu, T., Anil, R., Finn, C.: Efficiently identifying task groupings for multi-task learning (2021), https://arxiv.org/abs/2109.04617
17. Fridman, L., Brown, D.E., Glazer, M., Angell, W., Dodd, S., Jenik, B., Terwilliger, J., Patsekin, A., Kindelsberger, J., Ding, L., Seaman, S., Mehler, A., Sipperley, A., Pettinato, A., Seppelt, B.D., Angell, L., Mehler, B., Reimer, B.: Mit advanced vehicle technology study: Large-scale naturalistic driving study of driver behavior and interaction with automation. IEEE Access **7**, 102021–102038 (2019). https://doi.org/10.1109/ACCESS.2019.2926040
18. Ghosh, A., Zheng, S., Tamburo, R., Vuong, K., Alvarez-Padilla, J., Zhu, H., Cardei, M., Dunn, N., Mertz, C., Narasimhan, S.G.: Roadwork: A dataset and benchmark for learning to recognize, observe, analyze and drive through work zones. In: ICCV (2025)
19. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning (2020)
20. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp (2019), https://arxiv.org/abs/1902.00751
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
22. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
23. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting (2019), https://arxiv.org/abs/1812.01866
24. Kim, D., Kim, J., Cho, S., Luo, C., Hong, S.: Universal few-shot learning of dense prediction tasks with visual token matching (2023), https://arxiv.org/abs/2303.14969
25. Kotseruba, I., Tsotsos, J.K.: Scout+: Towards practical task-driven drivers' gaze prediction. In: 2024 IEEE Intelligent Vehicles Symposium (IV). pp. 1927–1932 (2024). https://doi.org/10.1109/IV55156.2024.10588743
26. Kröger, J.L., Lutz, O.H.M., Müller, F.: What does your gaze reveal about you? on the privacy implications of eye tracking. In: IFIP International Summer School on Privacy and Identity Management. pp. 226–241. Springer (2020)
27. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet (2015), https://arxiv.org/abs/1411.1045
28. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Deepgaze ii: Reading fixations from deep features trained on object recognition (2016), https://arxiv.org/abs/1610.01563
29. Li, D., Sharma, A., Zhang, H.R.: Scalable multitask learning using gradient-based estimation of task affinity. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 1542–1553. KDD '24, ACM (Aug 2024). https://doi.org/10.1145/3637528.3671835, http://dx.doi.org/10.1145/3637528.3671835

30. Li, G., Wang, Y., Zhu, F., Sui, X., Wang, N., Qu, X., Green, P.: Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in china. Journal of safety research **71**, 219–229 (2019)
31. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
32. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation (2021)
33. Li, Y., Xiong, K., Guo, X., Li, F., Yan, S., Xu, G., Zhou, L., Chen, L., Sun, H., Wang, B., et al.: Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. arXiv preprint arXiv:2506.08052 (2025)
34. Linardos, A., Kümmerer, M., Press, O., Bethge, M.: Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling (2021), https://arxiv.org/abs/2105.12441
35. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
36. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
37. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2015), https://arxiv.org/abs/1411.4038
38. Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J.: Drama: Joint risk localization and captioning in driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1043–1052 (2023)
39. Mori, Y., Fukui, H., Hirakawa, T., Nishiyama, J., Yamashita, T., Fujiyoshi, H.: Attention neural baby talk: Captioning of risk factors while driving. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 4317–4322 (2019). https://doi.org/10.1109/ITSC.2019.8917187
40. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2018)
41. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S.,

Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774

42. Palazzi, A., Abati, D., Solera, F., Cucchiara, R.: Predicting the driver's focus of attention: the dr (eye) ve project. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1720–1733 (2018)
43. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection (2020), https://arxiv.org/abs/2007.09062
44. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019), https://arxiv.org/abs/1912.01703
45. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition **106**, 107404 (2020)
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
47. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597
48. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2023), https://arxiv.org/abs/2208.12242
49. Sharma, P.K., Chakraborty, P.: A review of driver gaze estimation and application in gaze behavior understanding. Engineering Applications of Artificial Intelligence **133**, 108117 (2024)
50. Shlens, J.: Notes on kullback-leibler divergence and likelihood (2014), https://arxiv.org/abs/1404.2000

51. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering (2023)
52. Steelman, K.S., McCarley, J.S., Wickens, C.D.: Theory-based models of attention in visual workspaces. International Journal of Human–Computer Interaction **33**(1), 35–43 (2017)
53. Tian, H., Deng, T., Yan, H.: 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). https://doi.org/10.1109/JAS.2022.105716
54. Tian, X., Gu, J., Li, B., Liu, Y., Zhao, Z., Wang, Y., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
56. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment (2020), https://arxiv.org/abs/1908.06391
57. Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., Whitney, D.: Predicting driver attention in critical situations (2018), https://arxiv.org/abs/1711.06406
58. Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., Whitney, D.: Predicting driver attention in critical situations. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds.) Computer Vision – ACCV 2018. pp. 658–674. Springer International Publishing, Cham (2019)
59. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model (2024), https://arxiv.org/abs/2310.01412
60. Yan, Z., Fang, Q., Lv, W., Su, Q.: Anomalysd: Few-shot multi-class anomaly detection with stable diffusion model (2024), https://arxiv.org/abs/2408.01960
61. Zhou, X., Liu, M., Yurtsever, E., Zagar, B.L., Zimmer, W., Cao, H., Knoll, A.C.: Vision language models in autonomous driving: A survey and outlook (2024), https://arxiv.org/abs/2310.14414
62. Zhou, Y., Tang, J., Xiao, X., Lin, Y., Liu, L., Guo, Z., Fei, H., Xia, X., Gou, C.: Where, what, why: Towards explainable driver attention prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025), https://arxiv.org/abs/2506.23088