Driver Attention as Competitive Allocation: A Scene–Task-Aware Dual-Branch Framework

Qianfang Wang*
School of Civil Engineering and Transportation
South China University of Technology
Cho Chun Shik Graduate School of Mobility
Korea Advanced Institute of Science and Technology
qianfangwang@163.com

Bin Rao*
SKL-IOTSC
University of Macau
binrao.1@connect.um.edu.mo

Pengpeng Xu†
School of Traffic and Transportation Engineering
Central South University
pengpengxu@yeah.net

Tiantian Chen
Cho Chun Shik Graduate School of Mobility
Korea Advanced Institute of Science and Technology
nicole.chen@kaist.ac.kr

Abstract

Driver attention prediction is crucial for interpretable driver monitoring and human-like autonomous driving. Existing methods typically formulate this task as a scene-to-heatmap regression problem, failing to reveal the underlying dynamic attention redistribution mechanism. To address this issue, we propose a scene-task-aware competitive allocation framework, which decomposes driver attention as a bottom-up visual saliency prior and a top-down scene-task utility representation. An adaptive product-of-experts allocation rule is designed to fuse these two heterogeneous branches in a competitive manner, where scene-task demand dynamically modulates the dependence on top-down task utility, and driving control urgency adjusts spatial attention concentration via a learnable attention budget temperature. To explicitly differentiate task-critical regions from visually prominent yet task-irrelevant distractors, we incorporate competition-aware contrastive learning to enhance feature discrimination in the latent space. Extensive experiments demonstrate that our method achieves strong and balanced performance, with a Kullback–Leibler divergence of 0.98 and a normalized scanpath saliency of 4.15 on the DR(eye)VE dataset, as well as a Pearson's correlation coefficient of 0.67 and a histogram similarity of 0.54 on the BDD-A dataset. Ablation studies and diagnostic analyses further validate that our framework yields interpretable attention allocation behaviors, including urgency-induced spatial entropy contraction and scene-task-demand-guided dynamic attention reallocation.

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

In practical driving scenarios, drivers are required to dynamically and continuously distribute their limited attention across various road elements, including potential traffic conflict areas, road geometric structures, and surrounding traffic participants. Such attention allocation supports the core driving capabilities of hazard perception, behavioral planning, and precise vehicle control [32]. Consequently, accurate and robust modeling of driver attention is of paramount significance for reliable driver state monitoring, proactive traffic risk warning, and human-like autonomous driving systems [21, 38].

Existing driver attention prediction methods fall into two mainstream paradigms. Bottom-up approaches infer driver gaze distributions and attention heatmaps directly from driving scene inputs by exploiting visual saliency, temporal dynamics, and semantic cues [2, 11, 21, 36, 39]. In contrast, top-down methods leverage prior domain knowledge and explicit behavioral models to achieve task-aware driver attention prediction [15, 16, 28, 34, 37]. Despite the progress achieved, most existing methods formulate driver attention prediction as a straightforward scene-to-heatmap regression task. Such a paradigm fails to explicitly characterize the dynamic attention redistribution mechanism when visually prominent yet task-irrelevant distractors compete with behavior-critical road regions, resulting in limited interpretability and practical applicability in real driving scenarios.

To address this critical limitation, we propose a novel scene-task-aware dual-branch framework. The bottom-up branch learns to generate visual saliency priors from raw driving scenes, while the top-down branch infers scene-task utility representations based on real-time vehicle states and contextual semantic information. To fully exploit complementary heterogeneous features, we adopt an adaptive Product-of-Experts (APoE) allocation rule to competitively fuse the outputs of the two branches on the image plane. Furthermore, to suppress the interference of visually salient distractors, we introduce a competition-aware contrastive learning strategy, which effectively distinguishes task-critical driving regions from visually prominent but task-irrelevant areas in the feature space. The overall design not only boosts the accuracy and robustness of attention prediction, but also supports explainable diagnostic analysis of dynamic attention redistribution under varying scene-task demands and driving control urgency. The main contributions of this work are summarized as follows:

- **Novel formulation.** We reformulate driver attention prediction as a scene-task-aware competitive allocation problem, where bottom-up visual saliency and top-down scene-task utility compete dynamically to yield the final attention distribution.
- **Effective framework design.** We present a dual-branch APoE framework with a disentangled demand-urgency modeling strategy. Enhanced by competition-aware contrastive learning, our method separates task-critical driving regions from visually salient yet task-irrelevant distractors in the feature space, effectively mitigating distracting interference.
- **Comprehensive validation.** We extensively evaluate the proposed framework on two benchmark datasets via quantitative comparison, systematic ablation studies, and in-depth diagnostic analyses. Experimental results demonstrate the superior predictive performance and favorable interpretable attention allocation behaviors of our approach.

2 Related Work

Scene-based driver attention prediction. Early studies on driver visual behavior mainly focused on in-cabin gaze estimation, including gaze direction, gaze zones, and fixation locations from facial appearance, eye regions or head pose [3, 17, 30, 35]. While effective for driver-state monitoring, these methods provide limited insight into how gaze is allocated over external driving scenes. This limitation has motivated scene-based driver attention prediction, where gaze-derived attention maps are estimated from front-view driving images or videos. Benchmarks such as DR(eye)VE [21], BDD-A [33], DADA [6], and DrFixD-rainy [29] have further supported this direction. Built upon these benchmarks, existing methods have improved prediction through saliency modeling [21, 36], temporal reasoning [11, 14], semantic enhancement [2, 20], unsupervised learning [39], and efficient architectures [36]. However, most methods still follow a scene-to-heatmap regression paradigm driven mainly by visual appearance, motion, and spatial saliency, offering limited explanation of attention redistribution when visually salient regions conflict with the current driving task.

Task-aware and explainable driver attention. To better capture the behavioral structure of driving, recent studies have enriched driver attention modeling with task-, context-, route-, semantic-, and behavior-related information, including analyses of task and contextual effects on attention prediction [16], route- and map-based task-driven gaze prediction [15], interpretable fixation modeling [28], explicit driving-task factors [34], and behavior-aware knowledge embedding [37]. Together, they show that top-down information is essential for identifying behaviorally relevant regions beyond appearance-driven saliency. More recently, explainable driver attention studies have leveraged the semantic reasoning capabilities of large language or vision-language models to relate predicted attention to scene entities, driving behaviors, task context, and why-level explanations [10, 38]. Nevertheless, these methods mainly offer post-hoc semantic explanations that relate scene-task context to attended regions through language or vision-language reasoning. In contrast, our work provides mechanistic interpretability by formulating attention prediction as limited-resource allocation.

Contrastive learning for dense prediction. Contrastive learning improves representations by pulling related samples together and pushing unrelated samples apart. Recent studies have extended it to dense prediction by contrasting pixel-, region-, context-, or encoder-decoder-level representations for better local discriminability and spatial consistency [22, 24, 25, 27, 31]. Prior work has shown that hard negatives can strengthen discriminative supervision by forcing models to distinguish semantically or visually similar samples [26]. Following this principle, dense prediction methods often construct hard negatives from boundary ambiguity, contextual discrepancy, prediction uncertainty, or difficult sample pairs [1, 24, 27]. In driver attention prediction, however, the most informative negatives are visually salient and semantically plausible regions that are irrelevant to the current driving task. Therefore, our competition-aware objective targets saliency–task competition, separating task-critical regions from task-irrelevant salient distractors.

3 Method

3.1 Problem Formulation and Overview

Problem formulation. Given a driving video clip $X_t = \{x_{t-L+1}, \dots, x_t\}$, vehicle-state sequence $E_t = \{e_{t-L+1}, \dots, e_t\}$, and semantic scene inputs $S_t = \{s_{t-L+1}, \dots, s_t\}$, our goal is to predict the driver's attention distribution on the image plane at time t:

$$\hat{A}_t = \mathcal{F}_\theta(X_t, E_t, S_t), \qquad \hat{A}_t \in \Delta^{N-1}, \quad N = H \times W. \tag{1}$$

where Δ^{N-1} denotes the spatial probability simplex over an $H \times W$ image. X_t, E_t, and S_t provide visual observations, ego-state dynamics, and semantic context, respectively.

Framework overview. As shown in Fig. 1, we formulate driver attention prediction as scene–task-aware competitive allocation. The framework constructs a bottom-up saliency prior p_t^{bu} for stimulus-driven visual attraction and a top-down scene-task utility distribution p_t^{td} for behaviorally relevant regions, which are integrated by an APoE allocation rule. Scene-task demand α_t controls top-down reliance, while control urgency ρ_t determines the budget temperature τ_t for allocation sharpness. During training, a competition-aware contrastive objective is applied to the dual-branch feature streams to align task-critical regions and suppress visually salient distractors.

3.2 Scene-Task-Aware Competitive Allocation

Definition 1: Bottom-up prior and top-down utility. At time t, the image plane is discretized into $N = H \times W$ locations, denoted by $\Omega = \{1, \dots, N\}$. We define the bottom-up saliency prior and top-down scene-task utility as strictly positive spatial distributions:

$$p_t^{bu}, p_t^{td} \in \Delta_{++}^{N-1}, \qquad \sum_{i=1}^{N} p_{t,i}^{bu} = \sum_{i=1}^{N} p_{t,i}^{td} = 1. \tag{2}$$

where p_t^{bu} captures visual saliency, while p_t^{td} represents task-dependent behavioral relevance.

Definition 2: Scene-task demand and control urgency. We introduce two modulation variables:

$$\alpha_t \in [0, 1], \qquad \rho_t \in [0, 1]. \tag{3}$$

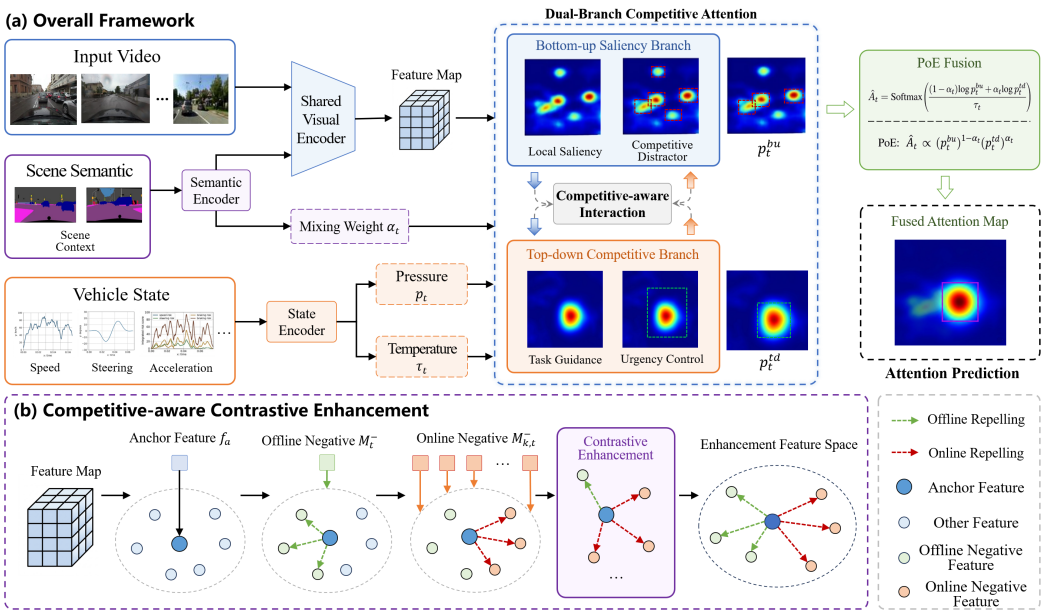


Figure 1: Overview of the proposed scene–task-aware competitive allocation framework.

Scene-task demand α_t controls the reliance on top-down utility, and control urgency ρ_t determines the budget temperature τ_t, where $\tau_{\min}$ and $\tau_{\max}$ define its lower and upper bounds:

$$\tau_t = \tau_{\max} - (\tau_{\max} - \tau_{\min})\rho_t, \qquad \tau_{\max} > \tau_{\min} > 0. \tag{4}$$

Thus, higher control urgency yields a lower temperature and a sharper allocation. The temperature bounds are empirically set to $\tau_{\min} = 0.8$ and $\tau_{\max} = 1.8$ for both datasets. τ_t is derived from the predicted ρ_t.

Definition 3: Budget-constrained attention allocation. Let $q_t = (q_{t,1}, \ldots, q_{t,N}) \in \Delta^{N-1}$ denote the attention allocation. For each location i, the combined allocation evidence is defined as

$$m_{t,i} = (1 - \alpha_t) \log p_{t,i}^{bu} + \alpha_t \log p_{t,i}^{td}. \tag{5}$$

The posterior attention allocation $q_t^\star$ is obtained by solving the entropy-regularized problem:

$$q_t^\star = \arg\max_{q \in \Delta^{N-1}} \left[\sum_{i=1}^{N} q_i m_{t,i} + \tau_t H(q) \right], \qquad H(q) = -\sum_{i=1}^{N} q_i \log q_i. \tag{6}$$

The evidence term assigns attention according to bottom-up and top-down cues, while the entropy term $H(q)$ controls allocation smoothness through τ_t.

Lemma 1: Closed-form APoE allocation. If $p_{t,i}^{bu} > 0$, $p_{t,i}^{td} > 0$, and $\tau_t > 0$ for all $i \in \Omega$, the optimization problem in Eq. (6) has a unique maximizer. For each location i,

$$q_{t,i}^\star = \frac{\exp(m_{t,i}/\tau_t)}{\sum_{j=1}^{N} \exp(m_{t,j}/\tau_t)} = \frac{\left(p_{t,i}^{bu}\right)^{(1-\alpha_t)/\tau_t} \left(p_{t,i}^{td}\right)^{\alpha_t/\tau_t}}{\sum_{j=1}^{N} \left(p_{t,j}^{bu}\right)^{(1-\alpha_t)/\tau_t} \left(p_{t,j}^{td}\right)^{\alpha_t/\tau_t}}. \tag{7}$$

The detailed proof is provided in Appendix A.1. This gives an APoE allocation rule, where α_t controls bottom-up/top-down dominance and τ_t controls posterior sharpness. In practice, the predicted attention map $\hat{A}_t$ implements this posterior allocation $q_t^\star$.

3.3 Neural Approximation

3.3.1 Dual-Branch Prediction

We instantiate the allocation formulation with a dual-branch dense predictor built on multi-scale visual features. Given video clip X_t, a visual backbone extracts multi-scale features B_t :

$$B_t = \{B_{t,\ell}\}_{\ell=1}^{M} = \Phi_v(X_t), \tag{8}$$

where $B_{t,\ell}$ is the feature map at level ℓ, and M is the number of feature levels.

The bottom-up stream directly preserves the visual feature stream:

$$B^{bu}_{t,\ell} = B_{t,\ell}, \qquad R^{bu}_t = D_{bu}(\{B^{bu}_{t,\ell}\}^M_{\ell=1}), \qquad p^{bu}_t = \phi_{\text{sp}}(R^{bu}_t). \tag{9}$$

where $B^{bu}_{t,\ell}$ is the bottom-up feature at level ℓ, R^{bu}_t is the bottom-up response map, D_{bu} is the bottom-up decoder, and $\phi_{\text{sp}}(\cdot)$ denotes spatial softmax over image locations.

For top-down stream, vehicle-state E_t and semantic scene S_t are encoded into context features $C_{t,\ell}$:

$$C_{t,\ell} = \Pi_\ell \left([\Phi_e(E_t)_\ell; \Phi_s(S_t)_\ell]\right), \tag{10}$$

where $\Phi_e(E_t)_\ell$ and $\Phi_s(S_t)_\ell$ are the vehicle-state and semantic features at level ℓ, Π_ℓ is a level-specific projection, $[\cdot;\cdot]$ denotes channel-wise concatenation, and $C_{t,\ell}$ represents the task-semantic context feature. These context features modulate visual features to form the top-down feature stream:

$$B^{td}_{t,\ell} = \phi_{\text{Attn}}\left(B_{t,\ell}, C_{t,\ell}\right). \tag{11}$$

where $B^{td}_{t,\ell}$ is the top-down feature at level ℓ, and $\phi_{\text{Attn}}(\cdot)$ denotes context-guided attention fusion.

The top-down decoder predicts the scene-task utility distribution:

$$U^{td}_t = D_{td}\left(\{B^{td}_{t,\ell}\}^M_{\ell=1}\right), \qquad p^{td}_t = \phi_{\text{sp}}(U^{td}_t). \tag{12}$$

where D_{td} is the top-down decoder and U^{td}_t is the scene-task utility response map.

Scene-task demand α_t is estimated from pooled context features, while control urgency ρ_t is estimated from control-related features P_t:

$$\alpha_t = \sigma\left(f_\alpha\left(\phi_{\text{pool}}(\{C_{t,\ell}\}^M_{\ell=1})\right)\right), \qquad \rho_t = \sigma\left(f_\rho(P_t)\right), \qquad \tau_t = \tau_{\max} - (\tau_{\max} - \tau_{\min})\rho_t. \tag{13}$$

where $\phi_{\text{pool}}(\cdot)$ denotes global pooling, f_α and f_ρ are MLP heads, and P_t contains ego-state and control-related cues such as speed, steering, acceleration, braking, or yaw-rate signals when available.

The final attention distribution is produced by the neural APoE rule, and details are provided in Appendix B.1.

$$\hat{A}_t = \phi_{\text{sp}}\left(\frac{1-\alpha_t}{\tau_t}\log p^{bu}_t + \frac{\alpha_t}{\tau_t}\log p^{td}_t\right). \tag{14}$$

3.3.2 Contrastive Enhancement

During training, we make saliency–task competition explicit in feature space by applying a contrastive objective to the dual-branch feature streams. At the contrastive level ℓ_c, the bottom-up saliency feature G^{bu}_t and task-conditioned top-down feature G^{td}_t are defined as:

$$G^{bu}_t = B^{bu}_{t,\ell_c}, \qquad G^{td}_t = B^{td}_{t,\ell_c}. \tag{15}$$

G^{bu}_t and G^{td}_t are the corresponding bottom-up and top-down feature maps.

Let M^+_t denote the attended-region mask derived from the ground-truth attention map. We use an auxiliary saliency response [5] to mine K salient distractor masks $\{M^-_{t,k}\}^K_{k=1}$ from residual saliency outside M^+_t. These negative masks correspond to regions that are visually salient but inconsistent with human attention annotations. Details of mask construction are provided in Appendix B.2.

We aggregate the dual-branch features over attended and distractor regions as:

$$z^{bu}_{t,+} = \mu_f(G^{bu}_t, M^+_t), \qquad z^{td}_{t,+} = \mu_f(G^{td}_t, M^+_t), \qquad z^-_{t,k} = \mu_f(G^{td}_t, M^-_{t,k}). \tag{16}$$

where $\mu_f(\cdot,\cdot)$ denotes mask-weighted feature averaging followed by feature normalization. The positive pair $(z^{bu}_{t,+}, z^{td}_{t,+})$ aligns bottom-up and top-down representations on task-critical regions, while $z^-_{t,k}$ denotes the top-down representation of the k-th task-irrelevant salient distractor.

Each distractor is weighted by a competition-aware hardness score:

$$h_{t,k} = \lambda_h h^{off}_{t,k} + (1-\lambda_h)h^{on}_{t,k}, \tag{17}$$

where $\lambda_h \in [0,1]$ balances offline and online hardness. The offline term $h^{off}_{t,k}$ is the normalized distractor score from residual-saliency mining, while the online term $h^{on}_{t,k}$ measures bottom-up/top-down disagreement:

$$h^{on}_{t,k} = \mu_m(p^{bu}_t, M^-_{t,k}) \left[1 - \mu_m(p^{td}_t, M^-_{t,k})\right], \tag{18}$$

where $\mu_m(\cdot,\cdot)$ denotes the mask-averaged response of a spatial distribution. Thus, a distractor receives a larger hardness score when it has high bottom-up saliency but low top-down utility.

The contrastive loss at time t is formualted as:

$$\mathcal{L}^{(t)}_{con} = -\log \frac{\exp\left(\kappa(z^{td}_{t,+}, z^{bu}_{t,+})/\tau_c\right)}{\exp\left(\kappa(z^{td}_{t,+}, z^{bu}_{t,+})/\tau_c\right) + \sum_{k=1}^{K} \exp\left(\kappa(z^{td}_{t,+}, z^{-}_{t,k})/\tau_c + \gamma\, \mathrm{sg}(h_{t,k})\right)}. \tag{19}$$

where $\kappa(\cdot,\cdot)$ denotes cosine similarity, τ_c is the contrastive temperature, γ controls hardness amplification, and $\mathrm{sg}(\cdot)$ denotes stop-gradient.

3.4 Allocation Properties

The closed-form APoE allocation yields several interpretable properties of bottom-up/top-down competition. We summarize the main theoretical properties below and provide all proofs in Appendix C.

Lemma 2: Log-odds Decomposition. For any two locations $i, j \in \Omega$, posterior allocation satisfies:

$$\log \frac{q^\star_{t,i}}{q^\star_{t,j}} = \frac{1-\alpha_t}{\tau_t} \log \frac{p^{bu}_{t,i}}{p^{bu}_{t,j}} + \frac{\alpha_t}{\tau_t} \log \frac{p^{td}_{t,i}}{p^{td}_{t,j}}. \tag{20}$$

Thus, relative attention preference decomposes into bottom-up and top-down evidence terms.

Theorem 1: Priority Inversion under Scene Task Demand. Consider a task-critical location i and a visually salient distractor j. Suppose the distractor has stronger bottom-up saliency, while the task-critical location has higher top-down utility:

$$\Delta s_{ij} = \log \frac{p^{bu}_{t,j}}{p^{bu}_{t,i}} > 0, \qquad \Delta u_{ij} = \log \frac{p^{td}_{t,i}}{p^{td}_{t,j}} > 0.$$

Then the pair-specific scene-task demand threshold is:

$$\alpha^\star_{t,ij} = \frac{\Delta s_{ij}}{\Delta s_{ij} + \Delta u_{ij}} \in (0,1), \tag{21}$$

and the task-critical region dominates the distractor if and only if:

$$q^\star_{t,i} > q^\star_{t,j} \quad \Longleftrightarrow \quad \alpha_t > \alpha^\star_{t,ij}. \tag{22}$$

This theorem characterizes when top-down scene-task demand overcomes the bottom-up saliency advantage of a distractor.

Theorem 2: Entropy Contraction under Control Urgency. Holding the allocation evidence m_t fixed, the entropy of $q^\star_t$ is non-decreasing with the budget temperature τ_t, as follows:

$$\tau_t = \tau_{\max} - (\tau_{\max} - \tau_{\min})\rho_t, \qquad \tau_{\max} > \tau_{\min} > 0, \tag{23}$$

A larger value of ρ_t induces a lower temperature and a more concentrated attention distribution:

$$\rho_t \uparrow \quad \Rightarrow \quad \tau_t \downarrow \quad \Rightarrow \quad H(q^\star_t) \downarrow. \tag{24}$$

3.5 Training Objective

The overall objective combines heatmap supervision, budget control, and contrastive enhancement:

$$\mathcal{L} = \mathcal{L}_{sal} + \lambda_{budget}\mathcal{L}_{budget} + \lambda_{con}\mathcal{L}_{con}. \tag{25}$$

Table 1: Quantitative comparison with state-of-the-art driver attention prediction methods on DR(eye)VE and BDD-A. Best results are shown in **bold**, and second-best results are underlined.

Method	DR(eye)VE				Method	BDD-A			
	KLD ↓	CC ↑	NSS ↑	SIM ↑		KLD ↓	CC ↑	NSS ↑	SIM ↑
DReyeVENet (2018) [21]	2.13	0.67	2.01	**0.56**	DReyeVENet (2018) [21]	1.84	0.62	1.61	0.46
BDD-ANet (2018) [33]	1.61	0.64	2.05	0.48	BDD-ANet (2018) [33]	2.07	0.48	3.45	0.35
CDNN (2019) [4]	2.00	0.67	2.03	0.54	CDNN (2019) [4]	1.84	0.51	1.30	0.39
UNISAL (2020) [5]	2.16	0.54	2.05	0.39	ViNet (2021) [13]	**1.04**	0.66	1.58	0.47
SCAFNet (2021) [6]	1.35	0.59	–	–	MVC-Net (2022) [12]	1.09	0.66	–	–
ViNet (2021) [13]	1.21	0.66	1.99	0.48	DrFixD-rainy (2022) [29]	1.09	0.64	–	0.47
MVC-Net (2022) [12]	1.64	**0.68**	–	–	ACT-Net (2022) [8]	1.11	0.63	–	0.45
ADA (2022) [7]	1.50	0.58	–	0.40	Unsupervised-v1 (2023) [39]	1.10	0.64	–	–
Unsupervised-v1 (2023) [39]	1.90	0.51	–	–	FBL-Net (2023) [2]	1.40	0.64	5.03	0.47
SCOUT (2024) [16]	1.08	0.63	3.89	0.52	SCOUT (2024) [16]	**1.04**	0.63	3.06	0.48
SalFoM-Student (2025) [19]	1.68	0.56	–	0.37	SalFoM-Student (2025) [19]	1.11	0.66	–	0.45
MTSF (2025) [14]	1.72	0.56	–	–	SalM2 (2025) [36]	1.08	0.64	–	0.47
CNN-ViT (2025) [34]	1.37	0.58	–	0.34	LLada (2025) [38]	1.23	0.58	**5.27**	0.42
Unsupervised-v2 (2025) [23]	1.90	0.49	–	–	Unsupervised-v2 (2025) [23]	1.10	0.64	–	–
LSTD (2026) [18]	1.67	0.57	–	0.39	FSDAM (2026) [9]	1.13	0.60	4.10	0.43
Ours	**0.98**	0.67	**4.15**	0.55	Ours	1.08	**0.67**	4.15	**0.54**

$$\mathcal{L}_{sal} = \frac{1}{T}\sum_{t=1}^{T} \ell_{sal}(\hat{A}_t, A_t^{gt}), \qquad \mathcal{L}_{budget} = \frac{1}{T}\sum_{t=1}^{T} \left\| \bar{H}(\hat{A}_t) - (1-\rho_t) \right\|_2^2, \qquad \mathcal{L}_{con} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{con}^{(t)}. \tag{26}$$

where $\ell_{sal}(\cdot,\cdot)$ is the prediction loss, and λ_{budget}, λ_{con} are loss weights.

The normalized entropy is calculated as:

$$\bar{H}(\hat{A}_t) = \frac{-\sum_{i=1}^{N} \hat{A}_{t,i} \log(\hat{A}_{t,i} + \epsilon)}{\log N}, \tag{27}$$

where ϵ is a small constant for numerical stability.

4 Experiments

4.1 Experimental Setup

We evaluate the proposed method on two real-world driver attention benchmarks: DR(eye)VE [21] and BDD-A [33]. DR(eye)VE contains large-scale naturalistic driving videos, while BDD-A covers diverse and safety-critical driving scenarios. For each sample, the model predicts the current driver attention map from a driving clip, vehicle-state sequence, and semantic scene representation. We report four standard metrics: Kullback–Leibler divergence (KLD), Pearson's correlation coefficient (CC), normalized scanpath saliency (NSS), and histogram similarity (SIM), where lower KLD and higher CC, NSS, and SIM indicate better performance. The model is implemented in PyTorch and trained with the objective in Sec. 3.5. Dataset protocols, preprocessing, metric definitions, network configurations, training schedules, and hyperparameters are provided in Appendix D.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares the proposed method with existing driver attention prediction methods on DR(eye)VE and BDD-A. On DR(eye)VE, our method achieves the best KLD and NSS, reducing KLD from 1.08 to 0.98 and improving NSS from 3.89 to 4.15 compared with SCOUT, while maintaining second-best CC and SIM scores. On BDD-A, our method obtains the best CC and SIM, with 0.67 CC and 0.54 SIM, and achieves competitive KLD and NSS. Although some methods perform better on individual metrics, their performance is less balanced across distributional similarity, spatial correlation, and fixation-level response. Overall, the proposed competitive allocation framework achieves strong performance across both benchmarks while maintaining a favorable balance among complementary attention metrics. Model-size and parameter comparisons are provided in Appendix D.4. Our model has 43.60M parameters, and the contrastive module is used only during training, introducing no additional inference-time cost.

Table 2: Ablation results on DR(eye)VE and BDD-A under the full test set (*All*), control-change scenarios (*Lat/Lon*), and stable lane-keeping scenarios (*Maintain*).

Dataset	Variant	All				Lat/Lon				Maintain			
		KLD ↓	CC ↑	SIM ↑	NSS ↑	KLD ↓	CC ↑	SIM ↑	NSS ↑	KLD ↓	CC ↑	SIM ↑	NSS ↑
DR(eye)VE	Top-down only	0.99	0.65	0.53	4.03	1.94	0.37	0.28	2.20	**0.75**	0.72	0.59	4.52
	Bottom-up only	1.07	0.63	0.52	3.89	2.09	0.35	0.29	2.08	0.81	0.70	0.58	4.38
	Dual-additive	1.03	0.65	0.52	4.00	1.97	0.35	0.27	2.10	0.80	0.72	0.58	4.47
	Dual-geometric ($\tau = 1$)	1.04	0.65	0.52	3.99	1.98	0.35	0.27	2.07	0.80	0.72	0.58	4.47
	APoE w/o contrastive	1.06	0.64	0.52	3.96	1.87	0.39	0.31	2.37	0.82	0.71	0.58	4.42
	APoE + vanilla contrastive	1.01	0.65	0.53	4.06	1.89	0.37	0.29	2.31	0.77	**0.73**	0.60	4.56
	APoE w/o offline	1.06	0.64	0.53	3.98	1.91	0.38	0.30	2.38	0.82	0.71	0.59	4.45
	APoE w/o online	1.01	0.65	0.53	4.05	1.86	0.39	0.31	2.41	0.78	0.72	0.59	4.51
	Full	**0.98**	**0.67**	**0.55**	**4.15**	**1.79**	**0.42**	**0.34**	**2.54**	0.76	**0.73**	**0.61**	**4.58**
BDD-A	Top-down only	1.11	0.62	0.49	3.74	1.19	0.55	0.42	2.92	1.04	0.64	0.51	3.95
	Bottom-up only	1.09	0.63	0.48	3.87	1.20	0.55	0.41	2.89	**1.03**	0.65	0.50	3.88
	Dual-additive	1.10	0.63	0.49	3.89	1.19	0.56	0.43	2.98	1.04	0.65	0.51	3.99
	Dual-geometric ($\tau = 1$)	1.11	0.63	0.48	3.89	1.20	0.56	0.42	2.95	1.05	0.65	0.50	4.08
	APoE w/o contrastive	1.22	0.63	0.49	3.67	1.30	0.56	0.44	2.96	1.15	0.65	0.52	3.98
	APoE + vanilla contrastive	1.10	0.64	0.50	3.84	1.18	0.57	0.44	3.02	**1.03**	0.66	0.52	4.05
	APoE w/o offline	1.11	0.64	0.48	3.81	1.22	0.55	0.42	2.93	1.05	0.66	0.54	4.13
	APoE w/o online	1.12	0.64	0.48	3.80	1.24	0.55	0.42	2.91	1.05	0.66	0.53	4.02
	Full	**1.08**	**0.67**	**0.54**	**4.15**	**1.15**	**0.58**	**0.46**	**3.11**	1.04	**0.69**	**0.59**	**4.35**

Table 3: Contextual modality ablation.

Variant	DR(eye)VE		BDD-A	
	KLD ↓	CC ↑	KLD ↓	CC ↑
w/o explicit semantics	1.02	0.64	1.16	0.63
w/o vehicle signals	1.06	0.63	1.15	0.64
Full	**0.98**	**0.67**	**1.08**	**0.67**

Table 4: Auxiliary saliency prior ablation.

Variant	DR(eye)VE		BDD-A	
	KLD ↓	CC ↑	KLD ↓	CC ↑
UNISAL (Full)	**0.98**	**0.67**	**1.08**	**0.67**
Spectral Residual	1.00	**0.67**	1.10	0.66
Random non-GT	1.02	0.66	1.12	0.65

4.3 Ablation Studies

Table 2 evaluates three ablation groups: single-branch variants, alternative fusion rules, and contrastive-learning variants, under the full test set (*All*), control-change scenes (*Lat/Lon*), and stable lane-keeping scenes (*Maintain*). The full model achieves the most consistent performance across datasets and metrics, with the clearest gains on control-change scenes. On DR(eye)VE *Lat/Lon*, it reduces KLD from 1.94/2.09 for the top-down-only and bottom-up-only variants to 1.79, and improves over additive fusion from 1.97 to 1.79 KLD and from 2.10 to 2.54 NSS. On BDD-A, removing contrastive learning increases KLD from 1.08 to 1.22, while removing either offline mining or online bottom-up/top-down disagreement also weakens performance. These results indicate that the two branches, adaptive APoE fusion, and competition-aware contrastive learning contribute complementary benefits.

To further assess the contributions of contextual inputs and auxiliary saliency priors, we conduct two additional ablation studies, as reported in Tables 3 and 4. For modality ablation, *w/o explicit semantics* removes semantic features while retaining vehicle-state inputs and dynamic urgency, whereas *w/o vehicle signals* removes vehicle-state inputs and fixes ρ_t to its training-set mean. For auxiliary saliency ablation, we replace UNISAL with Spectral Residual or randomly sampled non-GT regions while keeping the contrastive objective and negative-region sampling constraints unchanged.

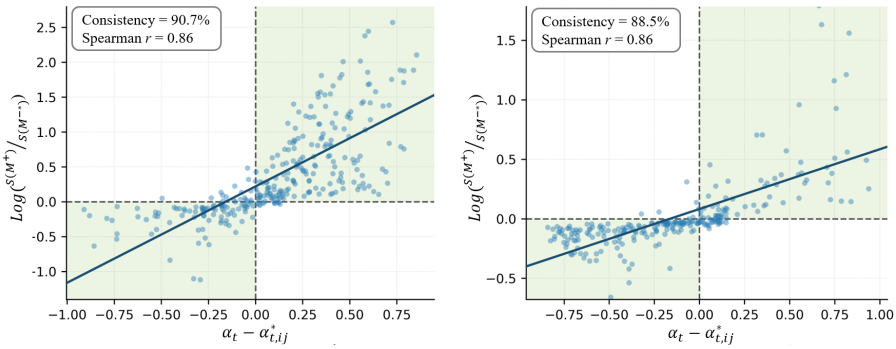
As shown in Table 3, removing either modality increases KLD and decreases CC on both datasets. On DR(eye)VE, removing vehicle signals causes a larger degradation than removing semantics, whereas their effects on BDD-A are comparable. Notably, removing vehicle signals also disables dynamic urgency modulation, so the observed degradation reflects their combined effect. Table 4 shows that replacing UNISAL with Spectral Residual increases KLD by only 0.02 on both datasets, indicating limited sensitivity to the choice of saliency prior. Random non-GT sampling also outperforms the variant without contrastive learning in Table 2, reducing KLD from 1.06 to 1.02 on DR(eye)VE and from 1.22 to 1.12 on BDD-A. These results indicate that contrastive learning remains effective without a learned auxiliary saliency prior, while UNISAL provides more informative hard negatives.

Table 5: Diagnostic check of control-urgency-induced entropy contraction.

Dataset	ρ-$\bar{H}$	τ-$\bar{H}$	$\bar{H}_{low}$	$\bar{H}_{mid}$	$\bar{H}_{high}$
DR(eye)VE	-0.59	0.59	0.87	0.82	0.79
BDD-A	-0.46	0.46	0.91	0.89	0.86

Table 6: Diagnostic of scene-task-demand-driven priority inversion under valid samples.

Dataset	N_{valid}	Acc. ↑	Δs	Δu	$\alpha^\star$
DR(eye)VE	300	90.67%	0.88	0.90	0.48
BDD-A	252	88.49%	0.23	0.26	0.48



(a) DR(eye)VE dataset (b) BDD-A dataset

Figure 2: Priority-inversion consistency on DR(eye)VE and BDD-A. Shaded quadrants indicate samples satisfying the theoretical condition in Theorem 1.

4.4 Interpreting Competitive Attention Allocation

We examine whether the learned model follows the theoretical properties in Sec. 3.4. Table 5 provides an observational diagnostic for the entropy contraction property in Theorem 2. Here, ρ-$\bar{H}$ and τ-$\bar{H}$ denote Spearman correlations with normalized attention entropy $\bar{H}$. The Low/Mid/High columns report mean $\bar{H}$ after grouping samples by ρ_t: bottom 15%, middle 70%, and top 15%, respectively. On DR(eye)VE and BDD-A, ρ_t shows moderate negative correlations with $\bar{H}$ $(-0.59/-0.46)$, while τ_t is positively correlated with entropy. As ρ_t increases from low to high, mean entropy decreases from 0.87 to 0.79 on DR(eye)VE and from 0.91 to 0.86 on BDD-A, supporting the interpretation that control urgency regulates allocation concentration.

Table 6 evaluates the priority inversion condition in Theorem 1 on valid conflict cases, where a salient distractor has a bottom-up advantage and the task-critical region has a top-down utility advantage, i.e., $\Delta s_{ij} > 0$ and $\Delta u_{ij} > 0$. Here, N_{valid} denotes the number of such cases, Acc. measures agreement between the learned allocation and the theoretical threshold condition. The learned allocation satisfies this condition in 90.67% and 88.49% of valid cases on DR(eye)VE and BDD-A, respectively, indicating that the predicted attention usually follows the demand threshold derived from the APoE formulation. To visualize this behavior at the sample level, we define the pooled attention response over a region mask M as $S(M) = \frac{\sum_i \hat{A}_{t,i} M_i}{\sum_i M_i + \epsilon}$. Figure 2 plots the demand margin $\alpha_t - \alpha^\star_{t,ij}$ against the log attention ratio $\log \frac{S(M^+)}{S(M^{-*})}$, where M^+ is the task-critical attended region and M^{-*} is the selected salient distractor. A positive demand margin means that scene-task demand exceeds the theoretical inversion threshold, while a positive log ratio means that the final prediction assigns more attention to the task-critical region than to the distractor. The positive trends on both datasets provide direct evidence that larger scene-task demand promotes attention reallocation from salient distractors to task-critical regions.

4.5 Sensitivity analysis

We analyze the sensitivity to three key hyperparameters: the contrastive weight λ_{con}, the budget-control weight λ_{budget}, and the offline/online hardness balance λ_h. As shown in Table 7, the model remains stable within a reasonable range of each parameter. Moderate λ_{con} improves distractor separation, while overly large values slightly degrade heatmap prediction. Similarly, a nonzero λ_{budget} improves urgency-aware concentration, and intermediate λ_h performs better than using only offline or online hardness, confirming their complementary roles.

Table 7: Sensitivity analysis of key hyperparameters.

Metric	λ_{con}					λ_{budget}					λ_h				
	0	0.05	0.10	0.15	0.20	0	0.05	0.10	0.15	0.20	0	0.25	0.50	0.75	1.00
KLD ↓	1.06	1.04	**0.98**	1.03	1.10	1.04	**0.98**	1.02	1.07	1.04	1.06	1.01	**0.98**	1.01	1.01
CC ↑	0.64	0.65	**0.67**	0.65	0.61	0.66	0.67	0.68	**0.69**	0.68	0.64	0.65	**0.67**	0.66	0.65

4.6 Qualitative Case Analysis

Figure 3 visualizes a 5-second nighttime turning sequence. As the vehicle approaches and executes the turn, the bottom-up branch remains sensitive to visually salient bright structures in the scene, whereas the top-down branch progressively emphasizes turning-related regions along the intended driving path. The fused prediction follows this task-guided shift and suppresses distractor-dominated responses, producing attention maps that are more consistent with the ground truth over time. These results provide qualitative evidence that the proposed dual-branch competitive allocation can adaptively resolve conflicts between visual saliency and driving relevance. Additional qualitative examples are provided in Appendix E.

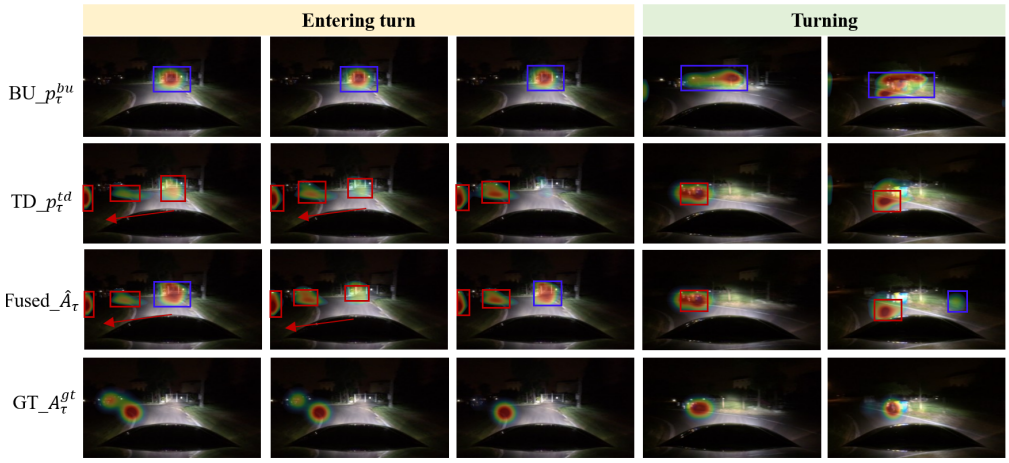


Figure 3: Temporal visualization of a valid priority-inversion sample in a nighttime turning scenario.

5 Conclusion

We presented a scene–task-aware competitive allocation framework for driver attention prediction, reformulating attention estimation as a resource-constrained competition between bottom-up saliency and top-down scene-task utility rather than direct scene-to-heatmap regression. By integrating dual-branch modeling, APoE allocation, demand–urgency modulation, and competition-aware contrastive learning, the proposed method improves both prediction accuracy and mechanism-level interpretability. Experiments on DR(eye)VE and BDD-A show strong performance, while ablation and interpretability analyses support the roles of entropy contraction, priority inversion, and distractor suppression. These results suggest that competitive allocation is an effective and interpretable mechanism for modeling driver attention, with potential value for transparent driver monitoring and human-centered intelligent driving. Limitations and broader impacts are discussed in Appendix F.

Acknowledgments

This research is supported by the National Key Research and Development Program of China (Project No. 2025YFB2505905)

References

[1] Maregu Assefa, Muzammal Naseer, Iyyakutti Iyappan Ganapathi, Syed Sadaf Ali, Mohamed L Seghier, and Naoufel Werghi. Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30850–30860, 2025.

[2] Yilong Chen, Zhixiong Nan, and Tao Xiang. Fblnet: Feedback loop network for driver attention prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13371–13380, 2023.

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1556–1565, 2024.

[4] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2146–2154, 2019.

[5] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified image and video saliency modeling. In *Proceedings of the European Conference on Computer Vision*, pages 419–435. Springer, 2020.

[6] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems*, 23(6):4959–4971, 2021.

[7] Shun Gan, Xizhe Pei, Yulong Ge, Qingfan Wang, Shi Shang, Shengbo Eben Li, and Bingbing Nie. Multisource adaption for driver attention prediction in arbitrary driving scenes. *IEEE transactions on intelligent transportation systems*, 23(11):20912–20925, 2022.

[8] Chao Gou, Yuchen Zhou, and Dan Li. Driver attention prediction based on convolution and transformers. *The Journal of Supercomputing*, 78(6):8268–8284, 2022.

[9] Kaiser Hamid, Can Cui, Khandakar Ashrafi Akbar, Ziran Wang, and Nade Liang. Fsdam: Few-shot driving attention modeling via vision-language coupling. *arXiv preprint arXiv:2511.12708*, 2025.

[10] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3347–3355, 2025.

[11] Daosong Hu, Mingyue Cui, and Kai Huang. Fifa: Fine-grained inter-frame attention for driver's video gaze estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18760–18769, 2025.

[12] Zhongxu Hu, Yiran Zhang, Qinghua Li, and Chen Lv. A novel heterogeneous network for modeling driver attention with multi-level visual content. *IEEE transactions on intelligent transportation systems*, 23(12):24343–24354, 2022.

[13] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3520–3527. IEEE, 2021.

[14] 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 Transactions on Intelligent Transportation Systems*, 26(2):1494–1509, 2024.

[15] Iuliia Kotseruba and John K Tsotsos. Scout+: Towards practical task-driven drivers' gaze prediction. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1927–1932. IEEE, 2024.

[16] Iuliia Kotseruba and John K Tsotsos. Understanding and modeling the effects of task and context on drivers' gaze allocation. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1337–1344. IEEE, 2024.

[17] Taiguo Li, Yingzhi Zhang, and Quanqin Li. Appearance-based driver 3-d gaze estimation using grm and mixed loss strategies. *IEEE Internet of Things Journal*, 11(23):38410–38424, 2024.

[18] Morteza Moradi, Mohammad Moradi, Ali Borji, Federica Proietto Salanitri, Giovanni Bellitto, Francesco Rundo, Simone Palazzo, and Concetto Spampinato. Learning long-and short-term dynamics for human attention prediction using large video models. *Computer Vision and Image Understanding*, page 104740, 2026.

[19] Morteza Moradi, Mohammad Moradi, Concetto Spampinato, Ali Borji, and Simone Palazzo. Distilling knowledge from large video models for driver visual attention prediction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[20] Anwesan Pal, Sayan Mondal, and Henrik I. Christensen. "Looking at the Right Stuff": Guided semantic-gaze for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11883–11892, 2020.

[21] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. 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.

[22] Zongshang Pang, Yuta Nakashima, Mayu Otani, and Hajime Nagahara. Revisiting pixel-level contrastive pre-training on scene images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1784–1793, 2024.

[23] Mengshi Qi, Xiaoyang Bi, Pengfei Zhu, and Huadong Ma. Towards robust unsupervised attention prediction in autonomous driving. *arXiv preprint arXiv:2501.15045*, 2025.

[24] Shoumeng Qiu, Jie Chen, Haiqiang Zhang, Ru Wan, Xiangyang Xue, and Jian Pu. Guided contrastive boundary learning for semantic segmentation. *Pattern Recognition*, 155:110723, 2024.

[25] Sébastien Quetin, Tapotosh Ghosh, and Farhad Maleki. Beyond the encoder: Joint encoder-decoder contrastive pre-training improves dense prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1000–1010, 2026.

[26] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

[27] Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, and Hyun Myung. Contextrast: Contextual contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3732–3742, 2024.

[28] Xing Tang, Jie Yu, and Yan Su. Modeling driver's visual fixation behavior using white-box representations. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15434–15449, 2022.

[29] 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.

[30] Sourabh Vora, Akshay Rangesh, and Mohan Manubhai Trivedi. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis. *IEEE Transactions on Intelligent Vehicles*, 3(3):254–265, 2018.

[31] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3024–3033, 2021.

[32] Christopher D Wickens, Jason S McCarley, and Robert S Gutzwiller. *Applied attention theory*. CRC press, 2022.

[33] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian Conference on Computer Vision*, pages 658–674. Springer, 2018.

[34] Chuan Xu, Bo Jiang, Yukun Wang, and Yan Su. A modeled study of driver visual attention driven by driving tasks. *Engineering Applications of Artificial Intelligence*, 157:111382, 2025.

[35] Yirong Yang, Chunsheng Liu, Faliang Chang, Yansha Lu, and Hui Liu. Driver gaze zone estimation via head pose fusion assisted supervision and eye region weighted encoding. *IEEE Transactions on Consumer Electronics*, 67(4):275–284, 2021.

[36] Chunyu Zhao, Wentao Mu, Xian Zhou, Wenbo Liu, Fei Yan, and Tao Deng. Salm^2: An extremely lightweight saliency mamba model for real-time cognitive awareness of driver attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1647–1655, 2025.

[37] Yuchen Zhou, Chao Gou, Zipeng Guo, Yihua Cheng, and Hyung Jin Chang. Behavior-aware knowledge-embedded model for driver attention prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[38] Yuchen Zhou, Jiayu Tang, Xiaoyan Xiao, Yueyao Lin, Linkai Liu, Zipeng Guo, Hao Fei, Xiaobo Xia, and Chao Gou. Where, what, why: Towards explainable driver attention prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2675–2685, 2025.

[39] Pengfei Zhu, Mengshi Qi, Xia Li, Weijian Li, and Huadong Ma. Unsupervised self-driving attention prediction via uncertainty mining and knowledge embedding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8558–8568, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly outline the contributions of our framework for driver attention in Section 1, and ensure all claims are supported by the empirical results presented in the subsequent sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our approach in the Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The full set of assumptions and proofs are provided in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We detail the network architectures, PyTorch implementation specifics, and hyperparameter configurations in the Appendix to ensure full reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to ongoing internal review and code clean-up, the source code is not included in the current submission. However, we commit to releasing the full codebase and configuration files publicly upon acceptance of the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Comprehensive training details, including dataset preprocessing along with optimizer settings, are specified in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the mean values of multiple experimental runs in the main text for clarity.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The hardware specifications and the total computational time required for training and inference are detailed in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This research complies fully with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts of our work are discussed in Appendix F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not involve the release of high-risk generative models or sensitive scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the original creators of the datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper leverages existing public datasets and does not introduce new data assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research utilizes anonymized, publicly available datasets and does not involve any human subject experiments or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects, hence IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large Language Models (LLMs) are not utilized as an original or non-standard component of the core trajectory prediction methodology in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix

A Allocation Proof

A.1 Proof of Lemma 1

We prove the closed-form solution of the entropy-regularized allocation problem in Definition 3. Recall that the feasible set is the probability simplex

$$\Delta_N = \left\{ q \in \mathbb{R}^N \mid q_i \geq 0,\ \sum_{i=1}^{N} q_i = 1 \right\}. \tag{28}$$

The objective is

$$J_t(q) = \sum_{i=1}^{N} q_i m_{t,i} + \tau_t H(q), \qquad H(q) = -\sum_{i=1}^{N} q_i \log q_i, \tag{29}$$

where

$$m_{t,i} = (1-\alpha_t) \log p_{t,i}^{bu} + \alpha_t \log p_{t,i}^{td}. \tag{30}$$

The first term in $J_t(q)$ is linear in q, and the entropy term is strictly concave in the interior of the probability simplex. Since $\tau_t > 0$, the objective is strictly concave and therefore has a unique maximizer over Δ_N.

Since $p_{t,i}^{bu} > 0$, $p_{t,i}^{td} > 0$, and $\tau_t > 0$, the optimum lies in the interior of the simplex. It is therefore sufficient to handle the normalization constraint explicitly. The Lagrangian is

$$\mathcal{L}(q, \lambda) = \sum_{i=1}^{N} q_i m_{t,i} - \tau_t \sum_{i=1}^{N} q_i \log q_i + \lambda \left(\sum_{i=1}^{N} q_i - 1 \right). \tag{31}$$

Taking the derivative with respect to q_i and setting it to zero gives

$$\frac{\partial \mathcal{L}}{\partial q_i} = m_{t,i} - \tau_t (\log q_i + 1) + \lambda = 0. \tag{32}$$

Rearranging and exponentiating both sides yields

$$q_i = \exp\left(\frac{\lambda - \tau_t}{\tau_t} \right) \exp\left(\frac{m_{t,i}}{\tau_t} \right). \tag{33}$$

The first exponential term is independent of i and serves as a normalization constant. Using $\sum_i q_i = 1$, we obtain

$$q_{t,i}^{\star} = \frac{\exp(m_{t,i}/\tau_t)}{\sum_{j=1}^{N} \exp(m_{t,j}/\tau_t)}. \tag{34}$$

Substituting the definition of $m_{t,i}$ gives

$$q_{t,i}^{\star} = \frac{\exp\left(\frac{(1-\alpha_t) \log p_{t,i}^{bu} + \alpha_t \log p_{t,i}^{td}}{\tau_t} \right)}{\sum_{j=1}^{N} \exp\left(\frac{(1-\alpha_t) \log p_{t,j}^{bu} + \alpha_t \log p_{t,j}^{td}}{\tau_t} \right)}. \tag{35}$$

This corresponds to the first equality in Eq. (7). Using $\exp(a \log x) = x^a$, we further obtain

$$q_{t,i}^{\star} = \frac{\left(p_{t,i}^{bu}\right)^{(1-\alpha_t)/\tau_t} \left(p_{t,i}^{td}\right)^{\alpha_t/\tau_t}}{\sum_{j=1}^{N} \left(p_{t,j}^{bu}\right)^{(1-\alpha_t)/\tau_t} \left(p_{t,j}^{td}\right)^{\alpha_t/\tau_t}}. \tag{36}$$

This gives the second equality in Eq. (7) and completes the proof.

B Neural Approximation Details

This section provides implementation details for the neural approximation in Sec. 3.3. We follow the notation in the main text: $B_t = \{B_{t,\ell}\}_{\ell=1}^{M}$ denotes multi-scale visual features, $B_{t,\ell}^{bu}$ and $B_{t,\ell}^{td}$ denote the bottom-up and top-down feature streams, and ϕ_{sp} denotes spatial softmax over image locations.

B.1 Dual-Branch Prediction Details

Visual backbone and bottom-up stream. Given an input video clip X_t, the visual backbone extracts multi-scale visual features:

$$B_t = \{B_{t,\ell}\}_{\ell=1}^{M} = \Phi_v(X_t), \tag{37}$$

where $B_{t,\ell}$ is the level-ℓ feature map and M is the number of feature levels. The bottom-up stream directly preserves the visual features:

$$B_{t,\ell}^{bu} = B_{t,\ell}. \tag{38}$$

The bottom-up decoder predicts a saliency response map and converts it into a spatial distribution:

$$R_t^{bu} = D_{bu}\left(\{B_{t,\ell}^{bu}\}_{\ell=1}^{M}\right), \qquad p_t^{bu} = \phi_{\mathrm{sp}}(R_t^{bu}). \tag{39}$$

Top-down context encoding. Vehicle-state inputs E_t and semantic scene inputs S_t are encoded into level-wise features $\Phi_e(E_t)_\ell$ and $\Phi_s(S_t)_\ell$. They are projected and concatenated to form task-semantic context features:

$$C_{t,\ell} = \Pi_\ell\left([\Phi_e(E_t)_\ell; \Phi_s(S_t)_\ell]\right), \tag{40}$$

where Π_ℓ is a level-specific projection and $[\cdot;\cdot]$ denotes channel-wise concatenation.

Context-guided attention fusion. The top-down feature stream is obtained by injecting context features into visual features:

$$B_{t,\ell}^{td} = \phi_{\mathrm{attn}}\left(B_{t,\ell}, C_{t,\ell}\right). \tag{41}$$

In implementation, ϕ_{attn} is instantiated as a residual attention fusion block. The visual and context features are projected to a shared dimension, attention-based fusion is applied, and the output is added back to the visual feature with normalization. This preserves the spatial structure of $B_{t,\ell}$ while injecting task-semantic context.

Top-down utility decoder. The top-down decoder predicts the scene-task utility response map:

$$U_t^{td} = D_{td}\left(\{B_{t,\ell}^{td}\}_{\ell=1}^{M}\right), \qquad p_t^{td} = \phi_{\mathrm{sp}}(U_t^{td}). \tag{42}$$

The spatial softmax operator maps a response map $Y \in \mathbb{R}^N$ to a spatial distribution:

$$\phi_{\mathrm{sp}}(Y)_i = \frac{\exp(Y_i)}{\sum_{j=1}^{N}\exp(Y_j)}, \qquad i \in \Omega. \tag{43}$$

where Y_i denotes the response value at the i-th image location after flattening the response map over $H \times W$ locations. This ensures that p_t^{bu} and p_t^{td} are strictly positive spatial distributions.

Demand–urgency heads. Scene-task demand is estimated from pooled context features, and control urgency is estimated from control-related features P_t:

$$\alpha_t = \sigma\left(f_\alpha\left(\phi_{\mathrm{pool}}(\{C_{t,\ell}\}_{\ell=1}^{M})\right)\right), \qquad \rho_t = \sigma\left(f_\rho(P_t)\right). \tag{44}$$

Here, $\phi_{\mathrm{pool}}(\cdot)$ denotes global pooling, f_α and f_ρ are multilayer perceptron (MLP) heads, and P_t contains ego-state and control-related cues such as speed, steering, acceleration, braking, or yaw-rate signals when available. Missing attributes are filled with zeros to keep a fixed input dimension. The budget temperature is then computed as

$$\tau_t = \tau_{\max} - (\tau_{\max} - \tau_{\min})\rho_t. \tag{45}$$

Neural APoE prediction. The final attention distribution is obtained by

$$\hat{A}_t = \phi_{\mathrm{sp}}\left(\frac{1-\alpha_t}{\tau_t}\log(p_t^{bu}+\epsilon) + \frac{\alpha_t}{\tau_t}\log(p_t^{td}+\epsilon)\right), \tag{46}$$

where ϵ is a small constant for numerical stability.

B.2 Competition-Aware Contrastive Details

This section details the mask construction and hardness estimation used by the competition-aware contrastive objective in Sec. 3.3.2. During training, the contrastive objective reuses the dual-branch feature streams at a selected contrastive level ℓ_c:

$$G_t^{bu} = B_{t,\ell_c}^{bu}, \qquad G_t^{td} = B_{t,\ell_c}^{td}. \tag{47}$$

Here, G_t^{bu} and G_t^{td} are the bottom-up and top-down feature maps used for contrastive learning.

Positive mask. Given the ground-truth attention map A_t^{gt}, we construct the attended-region mask:

$$M_t^+ = \mathcal{D}_{r_d}\left(\mathbb{I}(A_t^{gt} > \delta_{gt})\right), \tag{48}$$

where δ_{gt} is the attention threshold, $\mathbb{I}(\cdot)$ is the indicator function, and $\mathcal{D}_{r_d}(\cdot)$ denotes morphological dilation with radius r_d, which enlarges the attended region to tolerate fixation noise.

Residual saliency mining. An auxiliary saliency model, UNISAL [5], provides a saliency response R_t. We remove the attended region and obtain a residual saliency map:

$$R_t^{res} = R_t \odot (1 - M_t^+), \qquad \widetilde{R}_t = \eta_{01}\left(\mathcal{G}_\sigma(R_t^{res})\right), \tag{49}$$

where $\odot$ denotes element-wise multiplication, $\mathcal{G}_\sigma(\cdot)$ denotes Gaussian smoothing with standard deviation σ, and $\eta_{01}(\cdot)$ linearly normalizes a response map to $[0, 1]$. Local peaks in $\widetilde{R}_t$ are used as candidate salient distractor centers.

Negative masks and offline hardness. Let $B_{t,k} \in \{0, 1\}^N$ denote the k-th candidate distractor region generated from a local peak in $\widetilde{R}_t$. For each candidate region $B_{t,k}$, we discard invalid candidates that overlap with or lie too close to M_t^+:

$$\mathrm{Overlap}(B_{t,k}, M_t^+) \le \epsilon_{ov}, \qquad \mathrm{Clearance}(B_{t,k}, M_t^+) \ge \epsilon_{cl}. \tag{50}$$

The offline hardness is computed as

$$h_{t,k}^{off} = \lambda_r \,\mathrm{Resp}(B_{t,k}; \widetilde{R}_t) + (1 - \lambda_r)\,\mathrm{Prox}(B_{t,k}, M_t^+), \tag{51}$$

where $\mathrm{Resp}(\cdot)$ measures the average residual saliency inside $B_{t,k}$, $\mathrm{Prox}(\cdot)$ measures the normalized proximity to the attended region, and $\lambda_r \in [0, 1]$ balances the two terms. After diversity filtering, the top K candidates are converted into negative masks $\{M_{t,k}^-\}_{k=1}^K$, and $h_{t,k}^{off}$ is normalized across valid distractors. Algorithm 1 summarizes the offline competitive negative construction procedure.

Algorithm 1 Offline Competitive Negative Construction

Require: Ground-truth attention map A_t^{gt}, auxiliary saliency response R_t, number of negatives K
Ensure: Negative masks $\{M_{t,k}^-\}_{k=1}^K$ and offline hardness scores $\{h_{t,k}^{off}\}_{k=1}^K$

```
Construct the attended-region mask M_t^+ = Dilate(1(A_t^gt > δ_gt))
Compute residual saliency R_t^res = R_t ⊙ (1 − M_t^+)
Smooth and normalize R̃_t = η_01(G_σ(R_t^res))
Extract local peak set P_t from R̃_t
for each peak (y_k, x_k) ∈ P_t do
    Estimate a local scale and construct a candidate region B_{t,k}
    if Overlap(B_{t,k}, M_t^+) > ε_ov or Clearance(B_{t,k}, M_t^+) < ε_cl then
        Discard B_{t,k}
    else
        Compute h_{t,k}^off from residual response and proximity
    end if
end for
Apply diversity filtering to remove highly overlapping candidates
Select the top K valid candidates according to h_{t,k}^off
Convert selected candidates into masks {M_{t,k}^-}_{k=1}^K
return {M_{t,k}^-}_{k=1}^K, {h_{t,k}^off}_{k=1}^K
```

Masked feature and response aggregation. For a feature map G and a mask M, we define mask-weighted feature aggregation as

$$\mu_f(G, M) = \mathrm{Norm}\left(\frac{\sum_i \widetilde{M}_i G_i}{\sum_i \widetilde{M}_i + \epsilon}\right), \tag{52}$$

where $\widetilde{M}$ is the mask resized to the feature resolution, $\mathrm{Norm}(\cdot)$ denotes feature normalization, and i indexes spatial locations. For a spatial distribution p, we define the mask-averaged response as

$$\mu_m(p, M) = \frac{\sum_i \widetilde{M}_i p_i}{\sum_i \widetilde{M}_i + \epsilon}. \tag{53}$$

The positive and negative representations are then

$$z^{bu}_{t,+} = \mu_f(G^{bu}_t, M^+_t), \qquad z^{td}_{t,+} = \mu_f(G^{td}_t, M^+_t), \qquad z^-_{t,k} = \mu_f(G^{td}_t, M^-_{t,k}). \tag{54}$$

Online hardness. The online hardness measures whether a distractor is emphasized by bottom-up saliency but suppressed by top-down utility:

$$h^{on}_{t,k} = \mu_m(p^{bu}_t, M^-_{t,k}) \left[1 - \mu_m(p^{td}_t, M^-_{t,k})\right]. \tag{55}$$

The final competition-aware hardness is

$$h_{t,k} = \lambda_h h^{off}_{t,k} + (1 - \lambda_h) h^{on}_{t,k}, \tag{56}$$

where $\lambda_h \in [0, 1]$ balances offline residual mining and online bottom-up/top-down disagreement.

Algorithm 2 Competition-Aware Contrastive Learning

Require: Feature maps G^{bu}_t, G^{td}_t, spatial distributions p^{bu}_t, p^{td}_t, attended mask M^+_t, negative masks $\{M^-_{t,k}\}_{k=1}^K$, offline hardness $\{h^{off}_{t,k}\}_{k=1}^K$

Ensure: Sample-wise contrastive loss $\mathcal{L}^{(t)}_{con}$

Aggregate task-critical representations $z^{bu}_{t,+} = \mu_f(G^{bu}_t, M^+_t)$ and $z^{td}_{t,+} = \mu_f(G^{td}_t, M^+_t)$
Compute positive similarity $s^+_t = \kappa(z^{td}_{t,+}, z^{bu}_{t,+})$
for each negative mask $M^-_{t,k}$ **do**
    Aggregate negative representation $z^-_{t,k} = \mu_f(G^{td}_t, M^-_{t,k})$
    Compute online hardness $h^{on}_{t,k} = \mu_m(p^{bu}_t, M^-_{t,k})[1 - \mu_m(p^{td}_t, M^-_{t,k})]$
    Combine hardness $h_{t,k} = \lambda_h h^{off}_{t,k} + (1 - \lambda_h) h^{on}_{t,k}$
    Compute negative similarity $s^-_{t,k} = \kappa(z^{td}_{t,+}, z^-_{t,k})$
    Construct negative logit $\ell^-_{t,k} = s^-_{t,k}/\tau_c + \gamma\,\mathrm{sg}(h_{t,k})$
end for
Compute $\mathcal{L}^{(t)}_{con}$ using the positive logit and all hardness-biased negative logits
return $\mathcal{L}^{(t)}_{con}$

Contrastive loss. Let $\kappa(a, b)$ denote cosine similarity:

$$\kappa(a, b) = \frac{a^\top b}{\|a\|_2 \|b\|_2 + \epsilon}. \tag{57}$$

The positive and negative similarities are

$$s^+_t = \kappa(z^{td}_{t,+}, z^{bu}_{t,+}), \qquad s^-_{t,k} = \kappa(z^{td}_{t,+}, z^-_{t,k}). \tag{58}$$

The hardness-biased negative logit is

$$\ell^-_{t,k} = \frac{s^-_{t,k}}{\tau_c} + \gamma\,\mathrm{sg}(h_{t,k}), \tag{59}$$

where τ_c is the contrastive temperature, γ controls hardness amplification, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. The sample-wise contrastive loss is

$$\begin{aligned}\mathcal{L}_{con}^{(t)} &= -\log \frac{\exp(s_t^+/\tau_c)}{\exp(s_t^+/\tau_c) + \sum_{k=1}^{K} \exp(\ell_{t,k}^-)} \\ &= -\log \frac{\exp\left(\kappa(z_{t,+}^{td}, z_{t,+}^{bu})/\tau_c\right)}{\exp\left(\kappa(z_{t,+}^{td}, z_{t,+}^{bu})/\tau_c\right) + \sum_{k=1}^{K} \exp\left(\kappa(z_{t,+}^{td}, z_{t,k}^{-})/\tau_c + \gamma\,\mathrm{sg}(h_{t,k})\right)}.\end{aligned} \tag{60}$$

The final contrastive loss is averaged over valid samples with at least one mined distractor. Algorithm 2 summarizes the full competition-aware contrastive learning procedure.

C Proofs of Competitive Allocation Properties

C.1 Proof of Lemma 2

We prove the log-odds decomposition between two locations. Recall that the optimal posterior allocation is

$$q_{t,i}^\star = \frac{\exp(m_{t,i}/\tau_t)}{\sum_{k=1}^{N} \exp(m_{t,k}/\tau_t)}, \tag{61}$$

where

$$m_{t,i} = (1-\alpha_t)\log p_{t,i}^{bu} + \alpha_t \log p_{t,i}^{td}. \tag{62}$$

For any two locations $i, j \in \Omega$, the normalization term cancels in the ratio:

$$\frac{q_{t,i}^\star}{q_{t,j}^\star} = \exp\left(\frac{m_{t,i} - m_{t,j}}{\tau_t}\right). \tag{63}$$

Taking the logarithm and substituting $m_{t,i}$ gives

$$\log \frac{q_{t,i}^\star}{q_{t,j}^\star} = \frac{1-\alpha_t}{\tau_t} \log \frac{p_{t,i}^{bu}}{p_{t,j}^{bu}} + \frac{\alpha_t}{\tau_t} \log \frac{p_{t,i}^{td}}{p_{t,j}^{td}}. \tag{64}$$

This proves Lemma 2.

C.2 Proof of Theorem 1

Let i denote a task-critical location and j denote a visually salient distractor. By Lemma 2,

$$q_{t,i}^\star > q_{t,j}^\star \quad \Longleftrightarrow \quad \log \frac{q_{t,i}^\star}{q_{t,j}^\star} > 0. \tag{65}$$

Using Eq. (20), this condition becomes

$$\frac{1-\alpha_t}{\tau_t} \log \frac{p_{t,i}^{bu}}{p_{t,j}^{bu}} + \frac{\alpha_t}{\tau_t} \log \frac{p_{t,i}^{td}}{p_{t,j}^{td}} > 0. \tag{66}$$

Define

$$\Delta s_{ij} = \log \frac{p_{t,j}^{bu}}{p_{t,i}^{bu}} > 0, \qquad \Delta u_{ij} = \log \frac{p_{t,i}^{td}}{p_{t,j}^{td}} > 0. \tag{67}$$

Then

$$\log \frac{p_{t,i}^{bu}}{p_{t,j}^{bu}} = -\Delta s_{ij}, \qquad \log \frac{p_{t,i}^{td}}{p_{t,j}^{td}} = \Delta u_{ij}. \tag{68}$$

Substituting these terms and using $\tau_t > 0$, we obtain

$$-(1-\alpha_t)\Delta s_{ij} + \alpha_t \Delta u_{ij} > 0. \tag{69}$$

Rearranging gives

$$\alpha_t(\Delta s_{ij} + \Delta u_{ij}) > \Delta s_{ij}. \tag{70}$$

Since $\Delta s_{ij} > 0$ and $\Delta u_{ij} > 0$, this is equivalent to

$$\alpha_t > \frac{\Delta s_{ij}}{\Delta s_{ij} + \Delta u_{ij}} = \alpha_{t,ij}^\star. \tag{71}$$

Because the denominator is strictly larger than the numerator, $\alpha_{t,ij}^\star \in (0,1)$. This proves Theorem 1.

C.3 Proof of Theorem 2

We first establish the entropy–temperature monotonicity used by Theorem 2. Holding the evidence scores $\{m_{t,i}\}_{i=1}^{N}$ fixed, define the inverse temperature

$$\beta_t = \frac{1}{\tau_t}. \tag{72}$$

The posterior allocation can be written as the Gibbs distribution

$$q^\star_{t,i} = \frac{\exp(\beta_t m_{t,i})}{Z(\beta_t)}, \qquad Z(\beta_t) = \sum_{j=1}^{N} \exp(\beta_t m_{t,j}). \tag{73}$$

The entropy of $q_t^\star$ is

$$H(q_t^\star) = -\sum_{i=1}^{N} q^\star_{t,i} \log q^\star_{t,i} = \log Z(\beta_t) - \beta_t \mathbb{E}_{q_t^\star}[m_t]. \tag{74}$$

Differentiating with respect to β_t, and using the standard identities

$$\frac{d \log Z(\beta_t)}{d\beta_t} = \mathbb{E}_{q_t^\star}[m_t], \qquad \frac{d\mathbb{E}_{q_t^\star}[m_t]}{d\beta_t} = \mathrm{Var}_{q_t^\star}(m_t), \tag{75}$$

we obtain

$$\frac{dH}{d\beta_t} = -\beta_t \,\mathrm{Var}_{q_t^\star}(m_t) \le 0. \tag{76}$$

Since $\beta_t = 1/\tau_t$,

$$\frac{d\beta_t}{d\tau_t} = -\frac{1}{\tau_t^2}. \tag{77}$$

By the chain rule,

$$\frac{dH}{d\tau_t} = \frac{dH}{d\beta_t}\frac{d\beta_t}{d\tau_t} = \frac{\mathrm{Var}_{q_t^\star}(m_t)}{\tau_t^3} \ge 0. \tag{78}$$

Thus, holding m_t fixed, posterior entropy is non-decreasing with respect to τ_t.

The budget temperature is defined as

$$\tau_t = \tau_{\max} - (\tau_{\max} - \tau_{\min})\rho_t, \qquad \tau_{\max} > \tau_{\min} > 0. \tag{79}$$

Therefore,

$$\frac{d\tau_t}{d\rho_t} = -(\tau_{\max} - \tau_{\min}) < 0. \tag{80}$$

Combining the two derivatives gives

$$\frac{dH(q_t^\star)}{d\rho_t} = \frac{dH(q_t^\star)}{d\tau_t}\frac{d\tau_t}{d\rho_t} = -(\tau_{\max} - \tau_{\min})\frac{\mathrm{Var}_{q_t^\star}(m_t)}{\tau_t^3} \le 0. \tag{81}$$

Thus, the entropy of the posterior allocation is monotonically non-increasing with respect to control urgency. This proves Theorem 2.

D Experimental Details

D.1 Benchmarks

We evaluate the proposed model on two real-world driver attention benchmarks, DR(eye)VE [21] and BDD-A [33]. Both datasets provide front-view driving videos with gaze-derived driver attention annotations. For each sample, the model takes a driving clip, vehicle-state inputs, and semantic scene inputs as input, and predicts the attention map at the target frame. All input images and semantic maps are resized to 224×224.

DR(eye)VE. DR(eye)VE is a large-scale driver attention benchmark collected in naturalistic driving environments. We follow the official video-level split: videos 1–35 are used for training, videos 36–38 for validation, and videos 39–74 for testing. This split ensures that training, validation, and test samples come from disjoint driving videos. The front-view images, gaze-derived attention maps, and semantic scene inputs are temporally aligned, and the model predicts the attention distribution of the current frame from a short input clip.

BDD-A. BDD-A contains diverse dashcam driving videos with driver attention annotations, covering more varied and safety-critical traffic scenarios. We follow the official train/validation/test split for model training, best-checkpoint selection, and final evaluation. The same preprocessing and input-output protocol as DR(eye)VE are adopted: input images and semantic maps are resized to 224×224, and the model predicts the target-frame driver attention map from the corresponding driving clip, vehicle-state inputs, and semantic scene representation.

D.2 Metrics

Kullback–Leibler divergence. KLD measures the distributional discrepancy between the ground-truth attention distribution and the predicted distribution:

$$\mathrm{KLD}(Q, P) = \sum_{i=1}^{N} Q_i \log \frac{Q_i + \epsilon}{P_i + \epsilon}. \tag{82}$$

Lower KLD indicates better distribution-level alignment.

Pearson's correlation coefficient. CC measures the linear correlation between the predicted and ground-truth attention maps:

$$\mathrm{CC}(P, Q) = \frac{\sum_{i=1}^{N} (P_i - \mu_P)(Q_i - \mu_Q)}{\sqrt{\sum_{i=1}^{N} (P_i - \mu_P)^2} \sqrt{\sum_{i=1}^{N} (Q_i - \mu_Q)^2 + \epsilon}}, \tag{83}$$

where μ_P and μ_Q are the mean values of P and Q. Higher CC indicates stronger spatial correlation.

Normalized scanpath saliency. NSS evaluates the normalized predicted saliency at human fixation locations. Let $F_i \in \{0, 1\}$ denote the binary fixation map and

$$\bar{P}_i = \frac{P_i - \mu_P}{\sigma_P + \epsilon}$$

denote the zero-mean, unit-variance prediction. NSS is defined as

$$\mathrm{NSS}(P, F) = \frac{1}{\sum_i F_i + \epsilon} \sum_{i=1}^{N} \bar{P}_i F_i. \tag{84}$$

Higher NSS means the model assigns higher normalized responses to actual fixation locations.

Histogram similarity. SIM measures the overlap between two normalized attention distributions:

$$\mathrm{SIM}(P, Q) = \sum_{i=1}^{N} \min(P_i, Q_i). \tag{85}$$

Higher SIM indicates greater distributional overlap between prediction and ground truth.

D.3 Implementation Details

The model is implemented in PyTorch and trained on NVIDIA RTX 6000 Pro GPUs. We use AdamW as the optimizer and train the model with a batch size of 32 for at most 30 epochs. Early stopping is applied when the validation loss does not improve for 5 consecutive epochs, and the checkpoint with the best validation performance is used for testing. The input clip length is set to 16 frames with a frame stride of 8. Semantic labels are converted to one-hot maps when category-level semantic inputs are used.

Table A.1: Model size and parameter comparison with representative driver attention prediction models.

Model	Size (MB)	Params (M)
UNISAL [5]	15.50	4.07
Xu [34]	61.34	16.08
ViNet [13]	124.00	32.54
MTSF [14]	193.50	50.79
SCAFNet [6]	207.94	54.51
FBLNet [2]	333.71	87.48
Student [19]	622.00	163.10
LSTD [18]	1310.00	343.83
Ours	166.32	43.60

For competition-aware contrastive learning, we use $K = 6$ mined distractor regions for each sample. The contrastive temperature is set to $\tau_c = 0.07$, the contrastive loss weight is set to $\lambda_{con} = 0.10$, and the offline/online hardness balance is set to $\lambda_h = 0.50$. The hardness amplification coefficient is denoted by γ. The budget-control loss weight is set to $\lambda_{budget} = 0.05$.

D.4 Model Complexity

Table A.1 reports the model size and parameter count of representative driver attention prediction models. Our model contains 43.60M parameters with a model size of 166.32MB, which is smaller than several recent large-scale models such as MTSF, SCAFNet, FBLNet, Student, and LSTD. In addition, the competition-aware contrastive learning module is used only during training and does not introduce extra inference-time computation.

E Additional Qualitative Visualizations

We provide additional qualitative visualizations to further examine the behavior of the proposed competitive allocation mechanism under different driving scenarios. Figures A.1–A.3 show three representative temporal sequences, including an intersection turning-before-entry scene, a curved-road multi-vehicle interaction scene, and a traffic-light multi-vehicle interaction scene. In each example, columns correspond to consecutive frames, and rows show the bottom-up prior p_τ^{bu}, top-down utility p_τ^{td}, fused prediction $\hat{A}_\tau$, and ground-truth attention map A_τ^{gt}. All heatmaps are visualized with the same colormap and overlaid on the corresponding front-view image.

Across scenarios, the bottom-up branch tends to respond to visually salient structures such as bright lights, lane markings, roadside objects, or high-contrast vehicles. In contrast, the top-down branch emphasizes regions that are more relevant to the current driving context, such as the intended path, turning direction, leading vehicle, or potential conflict area. The fused prediction generally follows the top-down branch when bottom-up saliency is inconsistent with driving relevance, while remaining close to the bottom-up prior when both branches are aligned. These examples further show that the proposed competitive allocation mechanism can adaptively resolve conflicts between visual saliency and driving relevance.

F Limitations and Broader Impacts

Limitations. Although the proposed framework provides an interpretable formulation of driver attention prediction, several limitations remain. First, the learned scene-task demand and control urgency variables depend on the quality of visual, semantic, and vehicle-state inputs. Their estimates may become less stable when semantic parsing is noisy, vehicle-state signals are missing, or driving intent is ambiguous. Second, the competition-aware contrastive objective relies on mined salient distractors. While residual saliency mining and online bottom-up/top-down disagreement provide informative negatives, the quality of negative mining may still affect the strength of contrastive supervision. Third, the diagnostic analyses focus on theoretically valid conflict cases and specific

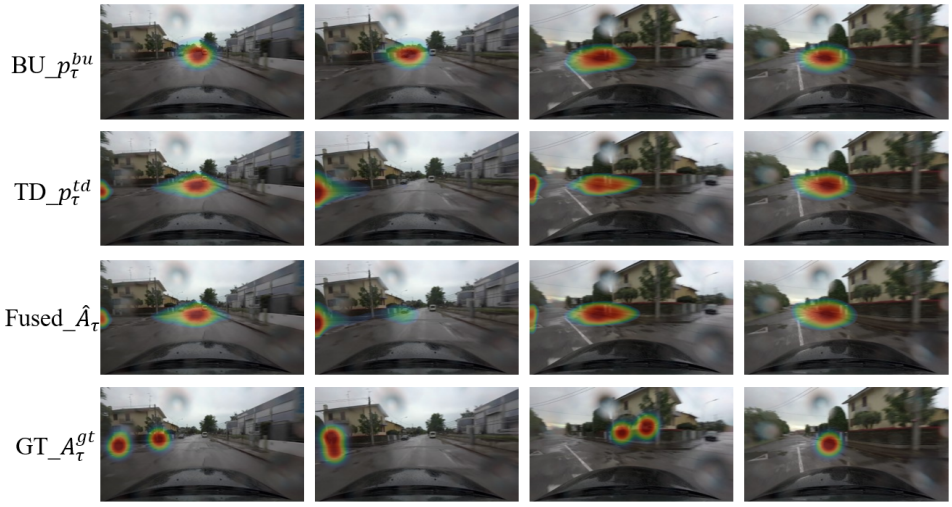


Figure A.1: Intersection turning-before-entry scene.



Figure A.2: Curved-road multi-vehicle interaction scene.



Figure A.3: Ttraffic-light multi-vehicle interaction scene.

allocation properties, which support the proposed mechanism but do not cover all possible attention behaviors in open-world driving. Finally, the current evaluation is conducted on existing driver attention benchmarks, and broader validation across different weather, lighting, geographic regions, and driver populations would further strengthen real-world applicability.

Broader impacts. This work contributes to safer and more interpretable intelligent driving systems by moving driver attention prediction beyond black-box heatmap regression. By modeling attention as scene–task-aware competitive allocation, the proposed framework not only predicts where drivers attend, but also provides a principled explanation of how visual saliency, driving relevance, scene-task demand, and control urgency jointly shape attention allocation. This mechanism-level interpretability can help diagnose distracting interference, analyze attention shifts in complex traffic scenes, and improve the transparency of driver monitoring and autonomous driving models. The framework may also benefit human-centered driving assistance by supporting more understandable warning systems and more human-like perception modules that prioritize behaviorally relevant regions rather than merely visually salient objects. More broadly, the idea of adaptive competition between bottom-up saliency and top-down task utility may inspire interpretable attention modeling in other dynamic embodied-AI settings. At the same time, predicted attention maps should be treated as auxiliary decision-support signals rather than direct measurements of driver intent or cognition. Real-world deployment should require rigorous validation, uncertainty awareness, and privacy-preserving handling of driver-related data.