arXiv:2602.01452v1 [cs.CV] 1 Feb 2026

面向智能车辆的基于注视的语义目标识别跨范式评估

Penghao Deng[1],**Jidong J. Yang**[1,*], **Jiachen Bian**[1]

[1]智能出行与基础设施实验室，工程学院，佐治亚大学，雅典，GA 30602，美国 {Penghao.Deng, Jidong.Yang, Jiachen.Bian}@uga.edu

*通讯作者

摘要

理解驾驶员在驾驶过程中将视觉注意导向何处（以注视行为为特征），对于开发下一代高级驾驶辅助系统并提升道路安全至关重要。本文将这一挑战视为一项语义识别任务，其对象是由车辆前视摄像头采集的道路场景。具体而言，本文采用三种不同的基于视觉的方法，研究注视点与目标语义的共位关系：直接目标检测（YOLOv13）、分割辅助分类（SAM2分别与EfficientNetV2和YOLOv13配对），以及基于查询的视觉语言模型（Qwen2.5-VL-7b与Qwen2.5-VL-32b的对比）。结果表明，直接目标检测（YOLOv13）和Qwen2.5-VL-32b显著优于其他方法，宏平均F1分数超过0.84。尤其是大型视觉语言模型（Qwen2.5-VL-32b）在识别交通信号灯等小型安全关键目标方面展现出更优的鲁棒性与性能，在不利夜间条件下尤为突出。相反，分割辅助范式受困于“部分与整体”语义鸿沟，导致召回率大幅下降。结果揭示了传统检测器的实时效率与大型视觉语言模型所提供的更丰富上下文理解及鲁棒性之间的根本权衡。这些发现为未来具有人因感知的智能驾驶员监测系统的设计提供了关键洞见与实践指导。

关键词 注视目标识别；高级驾驶辅助系统（ADAS）；视觉语言模型（VLM）；物体检测；驾驶员注意力

1 引言

尽管车辆工程和基础设施历经数十年的技术进步，确保道路安全仍是全球面临的首要挑战。每年，道路交通事故都会造成惊人的人员伤亡。美国国家公路交通安全管理局预计，2024年美国将有39,345人死于交通事故；这一数字虽略有下降，但与十年前相比仍高得令人担忧 [1]。从全球来看，道路交通伤害是5至29岁人群的首要死因，估计每年夺去119万人的生命 [2]。驾驶员注意力不集中是这一长期危机的重要诱因，这是一个多方面的问题，涵盖任何使驾驶员注意力从安全驾驶这一首要任务上转移的活动 [3, 4]。驾驶员分心是注意力不集中问题的一种表现，与相当大比例的警方报告交通事故有关；据估计，仅2023年，驾驶员分心就涉及8%的致命交通事故和13%的伤害性交通事故 [5]。

自然驾驶研究提供了有力的定量证据，表明注意力不集中与碰撞风险升高有关。研究显示，与专注驾驶相比，执行视觉或手动复杂的次要任务会使近碰撞或碰撞风险增加三倍 [6]。更重要的是，移开视线离开前方道路超过两秒，即被发现会使这一风险至少增加一倍 [7]。车载设备的激增

车载信息娱乐系统和个人移动设备的普及加剧了这一问题，使驾驶舱内出现了潜在的视觉、手动和认知分心来源 [8]。

为应对这一关键安全问题，汽车行业和监管机构积极推动驾驶员监测系统（DMS）的开发与部署。这些系统利用车内传感器和人工智能实时评估驾驶员的状态，检测困倦、分心和能力受损等情况 [9]。当前的 DMS 技术能够有效识别明显的注意力不集中迹象，例如长时间闭眼或使用手机；而车辆安全的下一个前沿方向，则是深入了解驾驶员的情境意识。这要求我们超越简单的分心检测，主动评估驾驶员对动态环境的感知。

这一工作的核心是分析驾驶员的注视。驾驶员注视点是其视觉注意最直接、最有力的指标，能够让我们了解其认知状态、意图以及对周边交通要素的认知情况 [10, 11]。通过准确识别驾驶员正在注视的语义对象，我们便能够构建新一代智能化、以人为中心的车辆系统。这项能力的应用范围广泛，且有望带来变革。

首先，在高级驾驶辅助系统（ADAS）领域，这使得创建情境感知预警系统成为可能。能够识别驾驶员未能通过视觉注意到正在接近的行人的高级驾驶辅助系统（ADAS），可以及时发出关键警报；而确认驾驶员已经在关注该危险的系统，则可以抑制多余且可能令人烦扰的警告，从而减少警报疲劳 [11]。这代表着一种从以车辆为中心的感知到以人为中心的感知的范式转变：前者中，系统只知道自己看到了什么；后者中，系统能够理解驾驶员看到了什么。

其次，对于以人为中心的自动驾驶车辆的开发，分析人的注视模式可为设计更直观、更值得信赖的人工智能驾驶策略提供极具价值的数据 [12]。在半自动驾驶车辆发生安全关键的接管事件期间，确认驾驶员的注意力是否恰当地指向相关危险，对于确保控制权安全交接至关重要 [13, 14]。此外，注视行为还可作为衡量驾驶员对自动化系统信任度的一种稳健且隐性的指标，而信任度是影响技术接受度和安全运行的关键因素 [15]。

第三，这项技术能够彻底革新车载人机界面（HMI）。人机界面（HMI）可设计为自适应地在驾驶员的视线中呈现信息，或利用细微提示将注意力引导至关键信息，从而在不增加认知负荷的情况下增强情境意识 [16]。

最后，对于驾驶员行为分析，例如事故后取证、驾驶员培训或保险远程信息处理，识别注视对象能够提供对驾驶员风险感知和决策过程的细粒度洞察，而这是速度和加速度数据等现有指标所无法获得的 [17]。

从形式上讲，本文所研究的问题是注视点目标识别：给定来自车辆前向摄像头的图像帧，以及驾驶员在图像平面上对应的时间同步的注视坐标，任务是确定该注视点处物体的语义类别，换言之，即驾驶员在特定时刻（帧）所关注的驾驶场景中的物体。尽管定义明确，该任务仍面临源于真实世界驾驶动态特性的重大技术挑战。

一项主要挑战是驾驶场景具有极大的多样性。任何基于视觉的系统都必须能够在各种环境条件下保持稳健。雨、雪和雾等恶劣天气会严重降低图像质量，带来噪声并遮挡物体特征 [18]。同样，光照条件变化极大，从直射阳光造成的强烈眩光，到夜间驾驶时的低光照和高对比度环境，每种情况都给感知算法带来独特的难题 [19]。行人、车辆和交通信号灯等目标物体本身，在外观、尺度和方向方面也千差万别。它们可能被其他道路使用者或基础设施部分遮挡，因此要完整、准确地识别它们并非易事 [20]。

一个更根本的挑战在于单个注视点本身存在歧义。从几何角度看，一个坐标可能指向多种对象：它可能落在远处一个小而独立的物体上，例如交通信号灯；也可能落在较大物体的某个特定部件上，例如卡车的车轮；还可能落在两辆相邻车辆之间的空隙中 [21]。这种歧义还受到语义维度的影响：驾驶员看向汽车的尾灯时，在语义上关注的是作为整体的“汽车”，而不只是尾灯。要正确推断驾驶员的注意焦点，有效的系统必须厘清这种部分与整体的关系。这一技术挑战与认知心理学中有充分记录的现象密切相关，例如“无意视盲”和“看而未见”（LBFTS）错误 [22]。当驾驶员的注视落在危险上，却未能有意识地察觉到危险时，就会发生看而未见事件；这通常是因为驾驶员缺乏预期或出现认知过载 [23, 24]。这表明，问题不仅在于几何定位，更在于推断潜在的人类认知状态，而其中的“真实标注”是驾驶员的语义意图。

最后，要使上述任何应用切实可行，尤其是涉及实时警告或控制权转换（例如在人工作驾驶与自动驾驶之间切换）的应用，整个识别过程都必须以极低的延迟完成。这对系统的计算预算提出了严格要求，因此更适合采用兼具准确性与效率的架构 [25]。如何在多种条件下兼顾高准确率与实时性能的要求，使注视点目标识别成为计算机视觉、人因工程和汽车工程交叉领域中一个独具挑战性的问题。

为解决注视点目标识别任务，可以借鉴计算机视觉和场景理解等更广泛领域中的多种不同方法范式。这些方法大致可分为基于目标检测、分割辅助以及基于视觉语言模型（VLM）的方法。

1.1 基于目标检测的方法

最直接的方法是利用成熟的现成目标检测器，例如 You Only Look Once（YOLO）系列或基于区域的卷积神经网络（R-CNN） [26, 27]。这些模型已广泛应用于交通场景理解和驾驶员监测，展现出单阶段检测器（如 YOLO）的高效率，以及两阶段检测器（如 Faster R-CNN [28]）的高准确率。其工作原理很简单：如果驾驶员注视坐标落在预测的边界框内，就将该框对应的类别标签作为注视目标。然而，在注视分析中，这种方法存在若干关键局限。首先是“注视落空”问题：如果注视点落在背景上或检测到的目标之间，该方法就会完全失效；即使驾驶员注视的是一个有效但未被检测到的实体，也无法提供任何信息 [29]。其次，标准检测器往往难以检测较小、较远或部分被遮挡的目标，而这些目标在驾驶场景中十分常见，且往往与安全密切相关 [30]。最后，边界框只能提供粗略定位，这意味着注视点可能落在边界框的边缘附近或角上，此时注视像素可能属于背景而非感兴趣对象，从而导致关联错误。

1.2 分割辅助分类方法

为了解决边界框带来的问题并实现更精确的定位，可以采用两阶段方法：首先使用图像分割模型分离注视坐标处的物体，然后使用分类模型识别生成的图像裁剪块。语义分割网络擅长对驾驶场景进行像素级分割，能够高保真地划分可行驶区域和常见物体类别 [31]。近年来，分割基础模型的发展，尤其是 Segment Anything Model（SAM），提供了一种强大的新工具 [32]。SAM 能够根据点提示或框提示为任意物体生成高质量的分割掩码，展现出卓越的零样本泛化能力 [33]。然而，这种范式也存在明显缺点。其主要局限在于类别无关分割：它能生成掩码，却不提供语义标签，因此还需要一个分类阶段 [34]。这种两阶段流水线会带来显著的计算开销，影响实时性能。更根本的是，这种方法容易受到前文所述“部分与整体”问题的影响。例如，如果驾驶员的注视落在汽车的车轮上，SAM 可能会准确地只分割出车轮，而不是整辆汽车。

1.3 基于视觉语言模型（VLM）的方法

一种新兴但极具前景的范式是使用大型视觉语言模型。这些模型在海量图像与文本配对数据集上进行预训练，已展现出强大的视觉与语言联合推理能力 [35]。对于注视目标识别任务，可以向视觉语言模型输入完整图像、注视坐标以及自然语言查询，例如：“坐标 (x, y) 处的物体是什么？”这将问题重新表述为一种视觉问答（VQA）任务 [36]。现代开源视觉语言模型（如 Qwen-VL 系列）在空间理解、细粒度物体识别以及回答有关交通场景的复杂问题方面表现出色 [37]。它们具备视觉定位能力，能够将文本概念与图像中的特定区域关联起来，这一点尤为相关 [38]。这种端到端的查询式方法能够自然地推理上下文，并直接处理任务的语义本质。然而，文献中对视觉语言模型在这一特定细粒度问题上的应用仍鲜有探索。

1.4 研究空白与贡献

尽管这些不同范式各自具有优势与不足，但目前仍缺乏系统性的比较研究，以评估它们在识别驾驶员注视点目标这一特定任务上的表现。尚不存在统一的基准测试，能够在同等条件下评估这些根本不同的范式，尤其是在

真实世界驾驶中多样且具有挑战性的环境条件之下。此外，现代大型视觉语言模型作为一种直接的、基于查询的解决方案，相较于成熟的计算机视觉流程，在这一细粒度任务上的潜力仍是一个关键且尚未探索的问题。

本文旨在填补这一关键研究空白，对用于识别驾驶员注视点处语义对象的不同基于视觉的范式进行全面、系统的比较。本研究基于前文介绍的三种范式，评估了五种不同的方法，评估所用的是一个新开发的基准测试，旨在检验这些方法在真实且具有挑战性的驾驶场景中的性能。本文的主要贡献如下：

1. 设计并实现了一个综合框架，用于系统地评估和比较驾驶员注视目标识别的三种不同范式（基于目标检测、分割辅助和基于视觉语言模型）。2. 首次深入研究大型视觉语言模型（具体而言为Qwen-VL）在该任务中的应用，并将其性能与最先进的计算机视觉技术直接比较。3. 开发并推出一个基于 BDD100K 的全新人工标注基准数据集，涵盖多种具有挑战性的驾驶条件（日间/夜间、晴天/雨天），以支持稳健且贴近现实的评估。4. 详细分析各方法在不同环境场景和物体类别中的性能权衡、优势与不足，为未来驾驶员监测和场景感知系统的开发提供实用见解。

本文其余部分安排如下。第 2 节详细介绍我们的方法，包括基准数据集、用于比较的视觉范式以及评估指标。第 3 节对实验结果进行系统分析。第 4 节讨论主要发现，包括实际意义以及性能与效率之间的实际权衡。最后，第 5 节总结本文的研究发现，并展望未来研究方向。

2 基准数据集

为了严谨且有意义地评估不同的基于视觉的范式，必须采用多样且贴近现实的基准数据集。本研究基于BDD100K 数据集 [39] 创建了一个基准数据集，该数据集全面涵盖了各种驾驶场景，包括多种天气条件、一天中的不同时段以及不同类型的场景。这些特点为评估不同方法的性能和鲁棒性提供了贴近现实且具有挑战性的基础。

基准数据集通过仅从城市街道场景类别中选取图像创建而成。随后对这些图像进行筛选，构建出四种各具特色且富有挑战性的环境条件：晴朗白天、晴朗夜晚、雨天白天和雨夜。标注方案采用人工流程，在每张选定图像中感兴趣的物体上标记注视点。该流程旨在模拟合理的驾驶员注视点。最终数据集由图像—坐标—标签三元组组成，这些三元组作为后续所有实验的真实标注。

图1展示了四种标注场景各自的代表性示例，说明了标注方法以及任务在视觉上的复杂性。每种场景都给视觉算法带来了独特挑战。图1a所示的晴朗白天场景能见度良好，但可能出现强烈阴影和眩光。图1b所示的晴朗夜晚场景以环境光照较弱为特点，路灯和车辆前灯等人工光源则会形成高对比度。图1c所示的雨天白天场景中，能见度降低、湿润路面上的反光以及挡风玻璃上雨滴造成的遮挡都会降低图像质量。图1d所示的雨夜场景最具挑战性，因为它兼具夜间驾驶的困难和雨水导致的图像质量下降。每张图像中的红色标记都是用于模型评估的模拟注视点标注示例。

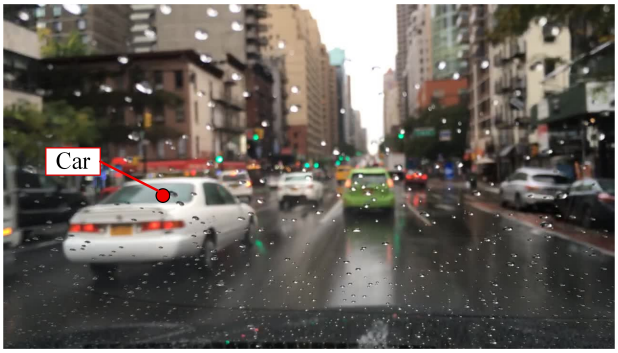
本研究选取了五种目标物体类别：行人、汽车、公交车、卡车和交通信号灯。这些类别在原始数据集中分别对应类别ID 0、2、5、7 和 9，选择依据是它们在典型城市驾驶环境中的常见程度及其安全重要性。类别ID与COCO数据集 [40], 保持一致，为评估提供了标准化基础。为确保所有方法的评估公平且一致，需要进行标签映射，以统一预训练模型采用的不同类别定义，具体包括YOLOv13目标检测器 [41] 基于COCO的分类体系，以及EfficientNetV2分类器 [42]基于ImageNet-1K的分类体系。



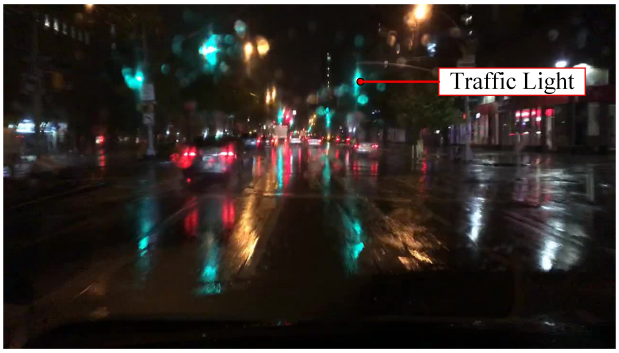
(a)晴朗白天



(b) 晴朗夜晚



(c) 雨天白天



(d) 雨夜

图1： 四种已标注场景及标注方案的代表性示例。

表1详细列出了用于将ImageNet-1K类别与五个较宽泛目标类别对应起来的映射关系。表中展示了如何将多个更细分的ImageNet-1K类别合并为一个目标类别。例如，旅行车、出租车和敞篷车等ImageNet-1K类别均映射至汽车类别。这一映射过程给基于分类器的方法带来了潜在挑战。例如，ImageNet-1K中行人类别的定义包含棒球运动员和新郎等类别，这些类别并不能代表驾驶场景中的典型行人，可能会限制分类器在该特定类别上的表现。

最终基准数据集由 400 张图像中的 1,355 个经人工标注的实例组成。为便于在不同条件下均衡比较模型性能，我们从前文所述的四种环境场景中各选取了 100 张图像。

图2展示了这五个目标类别中标注实例的分布情况。分布结果显示类别存在显著不平衡，其中汽车是最常见的类别（429个实例），公交车则最少（75个实例）。这种不平衡被认为能够代表真实世界的城市驾驶情况，因为某些车型比其他车型更常见。这一特点凸显了采用宏平均F1分数等能够适应类别分布不平衡的评估指标的重要性。

图3对四种场景中的实例分布和密度进行了更详细的分析。虽然每种场景的图像数量均衡，但其中标注物体的数量各不相同。图3a展示了每种场景的实例总数。图3b则通过展示每张图像中的平均实例数，提供了更细致的视角。分析表明，晴朗夜晚和雨天白天场景的目标密度最高，平均每张图像分别有3.57个和3.56个实例。这表明这些场景的视觉环境更为杂乱，可能会给受评估模型带来更大的感知挑战。

表1： 类别标签映射。

COCO类别ID	COCO类别名称	ImageNet-1K类别ID	ImageNet-1K类别名称
0	行人	981	棒球运动员
		982	新郎
		983	水肺潜水员
2	car	436	旅行车
		468	cab
		511	敞篷车
		627	豪华轿车
		661	T型车
		705	乘用车
		751	赛车
		817	跑车
5	bus	654	小型公共汽车
		779	校车
		874	无轨电车
7	卡车	555	消防车
		569	垃圾车
		675	搬家货车
		717	皮卡
		864	清障车
		867	半挂卡车
9	交通信号灯	920	交通信号灯

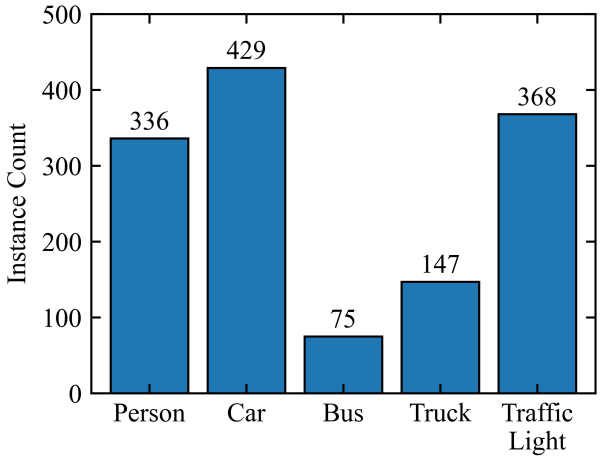
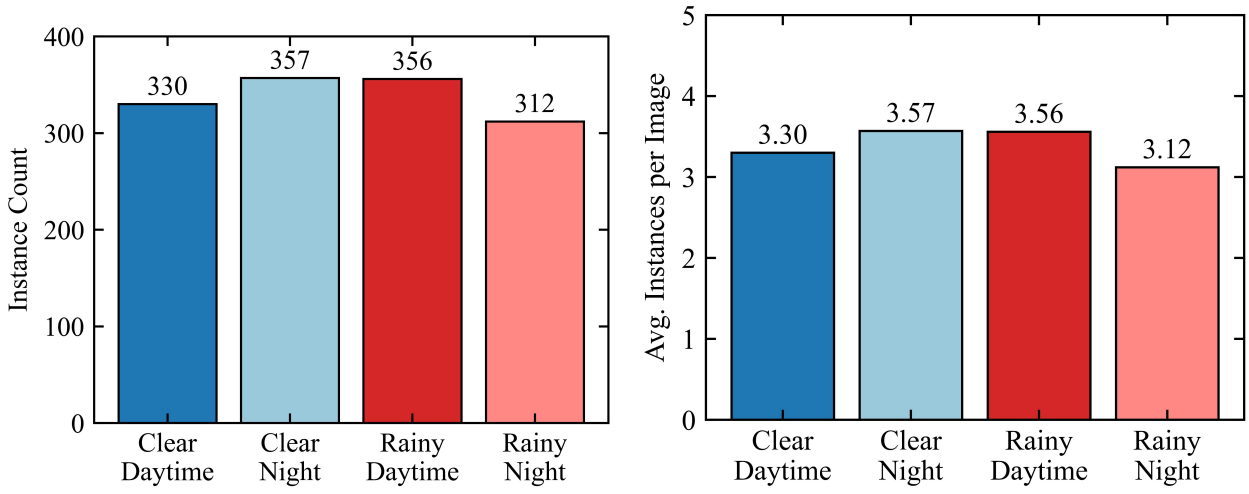


图2： 每个类别的标注实例分布。



(a)实例总数 **(b)** 平均实例密度

图3： 各场景中实例的分布与密度。

3 方法论

3.1 基于视觉的方法比较

本节介绍了本研究考察的三种范式，并对五种方法进行系统比较：范式1下的一种方法、范式2下的两种方法，以及范式3下的两种方法。

1. 范式 1（基于目标检测）：采用 YOLOv13 模型的直接识别方法。2. 范式 2（分割辅助分类）：通过两种特定方法进行评估的两阶段流水线：SAM2 与 EfficientNetV2 结合，以及 SAM2 与 YOLOv13 结合。3. 范式 3（基于视觉语言模型）：采用查询式方法，并使用两种模型规模（称为两种方法）进行评估：70亿参数版本（Qwen2.5-VL-7b）和 320亿参数版本（Qwen2.5-VL-32b）。

图4以高层示意图直观概述了三种范式各自的核心工作流程。图4a所示的基于目标检测范式是一个单步流程：目标检测模型处理输入图像和坐标，直接输出包含该坐标的物体类别。图4b所示的分割辅助范式采用两阶段流水线。第一阶段使用分割模型分割出注视坐标处的物体，第二阶段使用分类模型识别生成的裁剪图像。图4c所示的视觉语言模型范式是一个端到端流程：将输入图像、坐标和结构化文本提示输入视觉语言模型，由其直接输出语义类别。

3.1.1 范式1：基于物体检测

该范式代表了最直接且计算效率最高的解决方案。具体实现方法（方法1）采用了预训练 YOLOv13 模型。首先，在完整的输入图像上运行 YOLOv13 模型，生成一系列预测边界框，每个边界框都关联一个类别标签和一个置信度分数。随后，决策逻辑会检查给定的注视坐标是否落在某个已检测边界框的几何范围内。如果是，则将该边界框的类别标签指定为最终预测结果。在以下三种特定情况下，记录类别ID -1：（1）注视坐标不在任何已检测边界框内；（2）包含该坐标的边界框的置信度分数低于 0.4 的阈值；或（3）预测类别不属于预先定义的五个目标类别之一。该值表示无法识别指定位置处的有效物体。

3.1.2 范式2：分割辅助类别

该范式旨在通过像素级分割隔离感兴趣对象，再进行分类，从而提高精确率。该方法采用两种不同方法共用的两阶段流水线。第一阶段中，

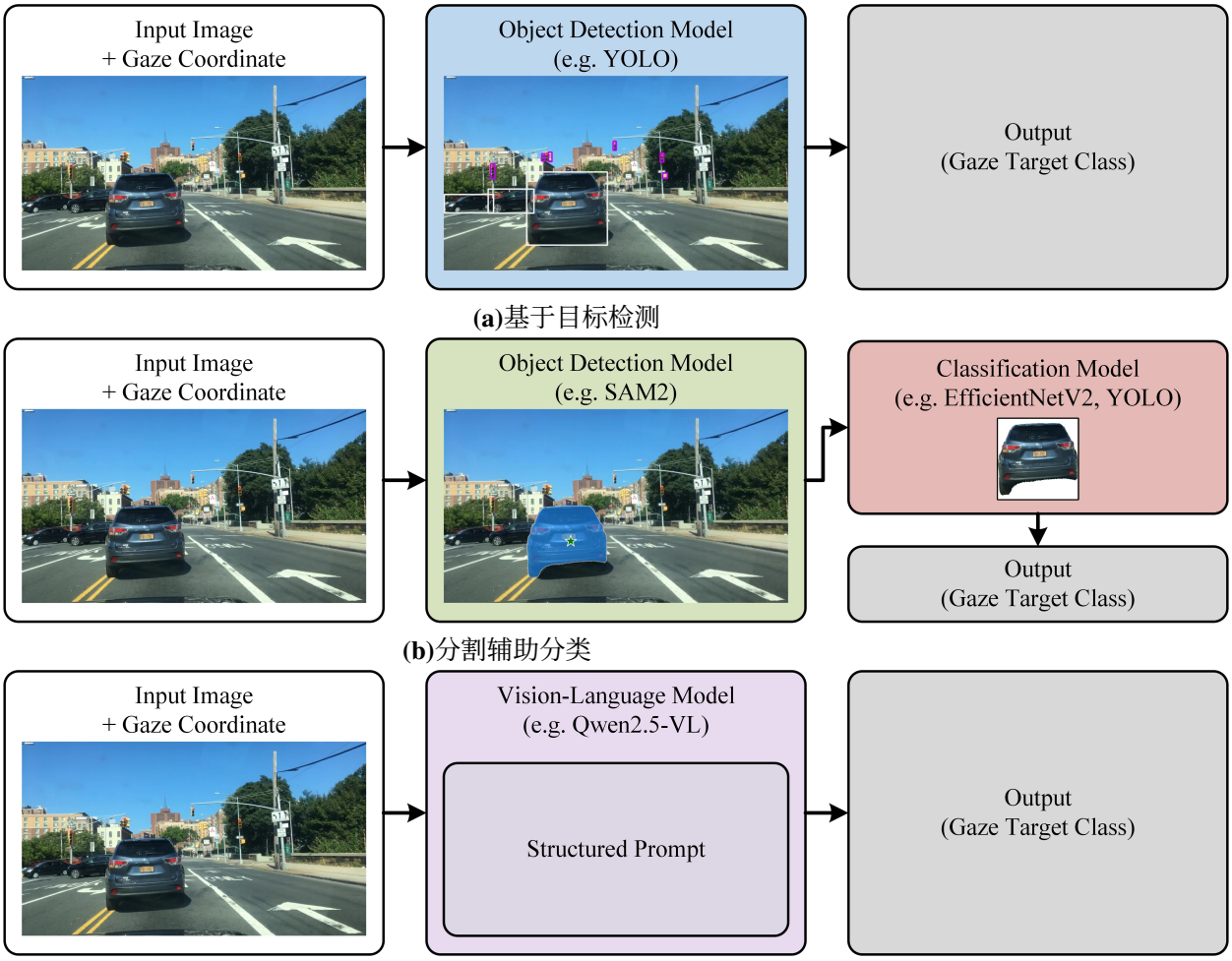


图4： 三种方法范式的可视化。

使用 SAM2 模型。将驾驶员注视坐标作为 正点提示提供给 SAM2，引导模型为该位置的物体生成精确的像素级分割掩码。完成分割后，执行标准化图像预处理步骤。根据生成的掩码，从原始图像中裁剪出物体。为了给后续分类阶段提供统一的输入，会构建一张正方形图像。裁剪出的物体置于新图像的中心，正方形内所有背景区域均填充为中性白色。

该范式中的两种方法仅在第二阶段有所不同：

1. 方法2-1（SAM2+EfficientNetV2）：将准备好的孤立物体方形图像输入预训练的 EfficientNetV2 模型，进行最终分类。如果根据表1中的映射，EfficientNetV2 预测的类别不属于五个目标类别之一，则将最终预测的类别ID设为 -1。2. 方法2-2（SAM2+YOLOv13）：此方法作为对照实验。其流程与方法2-1相同，但将分类模型替换为 YOLOv13 模型。如果 YOLOv13 模型预测的类别不属于五个目标类别之一，则将类别ID预测为 -1。此设计可用于分析该范式中的性能限制究竟源自初始分割阶段，还是后续分类模型。

3.1.3 范式3：基于视觉-语言模型

第三种范式探讨了大规模视觉语言模型通过上下文推理完成该任务的能力。这种方法将问题重新定义为视觉问答（VQA）任务。将完整的输入图像、（x，y）注视

坐标以及精心设计的文本提示作为一个组合输入提供给视觉语言模型。图5详细展示了该文本提示，其设计目的是明确要求模型识别给定坐标处的物体。至关重要的是，提示还将可能的答案限制为预定义的五种目标类别列表，并额外提供一个“不确定”选项。这个“不确定”类别实际上充当了 -1 类，为不确定的情况指定了输出，并确保模型的响应能够与其他方法直接比较。

> 你是一名精确的图像分析助手。你的任务是识别所提供图像中位于特定像素坐标处的物体。
>
> **请严格遵循以下规则：**
>
> 1. 你必须从以下列表中选择最匹配的类别：’[行人, 汽车, 公交车, 卡车, 交通信号灯, 不确定]’。2. 你的回答必须**严格**是列表中的标签之一（例如，“汽车”）。**不要**包含任何额外的解释、描述或无关字符。3. 优先选择这五个具体的物体类别。仅在你完全不确定时，才将“不确定”作为最后手段。
>
> 现在，请分析这张图像，并告诉我位于坐标’[{x}, {y}]’处的物体是什么。

图5：用于基于视觉语言模型的方法的结构化提示。

评估了两种方法，以研究模型规模对性能的影响：

1. 方法3-1（Qwen2.5-VL-7b）采用Qwen-VL模型的70亿参数版本。2. 方法3-2（Qwen2.5-VL-32b）采用该模型规模大得多的320亿参数版本。

对这两个模型进行比较，旨在深入了解模型规模如何影响这一特定且细粒度的视觉识别任务的性能。

3.2 评估指标

为定量评估所比较方法的性能，采用了一套全面的标准分类指标。用于评估的主要指标为准确率、精确率、召回率和F1分数。准确率定义为全部正确预测数与实例总数之比。为更详细地分析错误类型，精确率、召回率和F1分数分别按公式（1）至（3）的定义进行计算。

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

其中 TP、 FP和 FN分别表示给定类别的真正例、假正例和假负例计数。精确率衡量模型避免做出假正例预测的能力，而召回率衡量其找出所有真实实例的能力。F1分数通过对精确率和召回率取调和平均数，提供单一且均衡的性能度量。

对于处理与目标类别不对应的模型输出，我们进行了特殊考虑。类别ID为“-1”的预测表示模型未能将对象归入五个目标类别之一。这类失败可能由多种特定于方法的情况引起，包括置信度分数较低、注视点落在有效检测区域之外，或进行范围外分类。在评估中，任何产生“-1”预测的实例都被视为其真实类别的假阴性。由于真实标注数据集仅包含五个目标物体类别的标签，这种处理方式能够恰当地惩罚模型未能识别有效物体所导致的召回率损失。

此外，鉴于图2所示的数据集存在显著的类别不平衡，我们采用宏平均指标作为总体性能的主要衡量指标。宏平均精确率、宏平均召回率和宏平均F1分数的计算方式是：分别针对五个目标类别独立计算相应指标，再取非加权平均值。这种方法为每个类别赋予相同权重，不受其出现频率影响，从而能更公平、稳健地评估模型在所有物体类别上的表现。

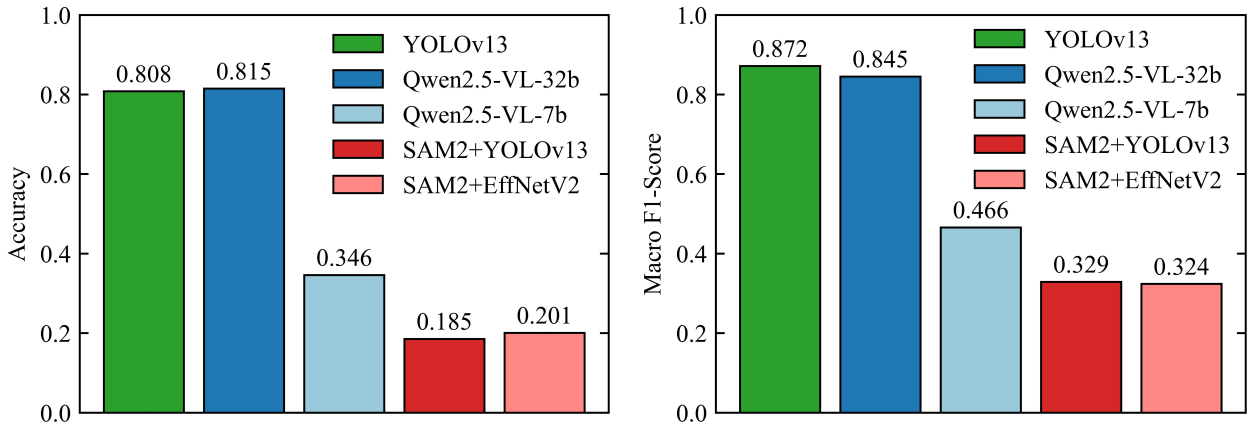
4 结果

本节对开发的基准数据集上五种对比方法进行了系统评估。分析首先概述总体性能，随后详细考察模型在不同场景下的鲁棒性，以及其在各种物体类别上的有效性。

4.1 总体性能比较

首先，在基准数据集上评估了五种基于视觉的方法的总体性能。为确保在存在固有类别不平衡的情况下进行全面、公平的评估，研究采用了四项评估指标：整体准确率、宏平均精确率、宏平均召回率和宏平均F1分数。

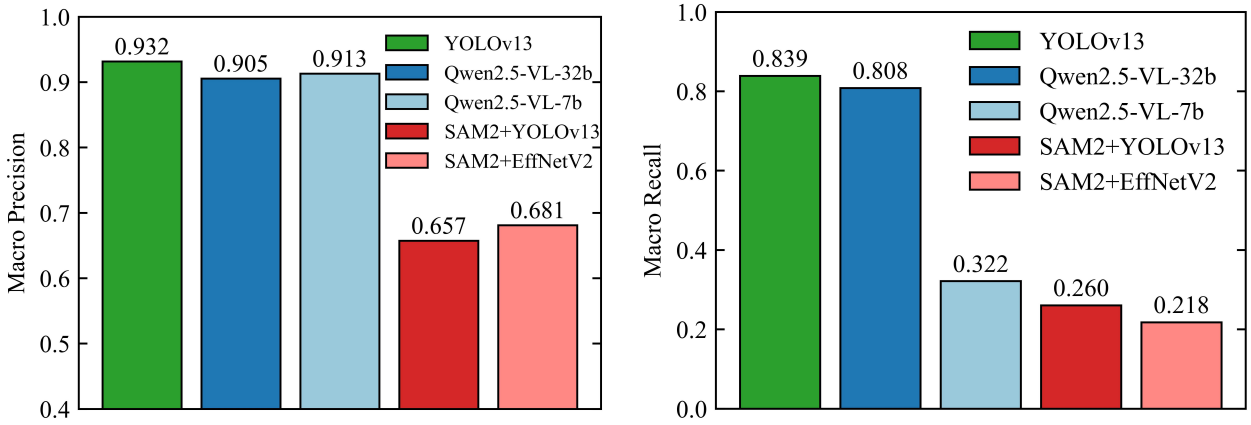
图6展示了本研究的主要宏观发现，对比了各评估方法的整体准确率和更具鲁棒性的宏平均F1分数。如图6a所示，整体准确率结果清楚地显示了表现最出色的两种方法。Qwen2.5-VL-32b模型的准确率最高，为0.815，紧随其后的是YOLOv13模型，准确率为0.808。相比之下，其他三种方法的表现明显较差，其中两种基于SAM2的方法得分低于0.201。尽管如此，在类别不平衡的情况下，整体准确率可能会产生误导。对于类别不平衡的数据集，更合适且更具鲁棒性的指标是宏平均F1分数，如图6b所示。根据这一指标，YOLOv13模型（0.872）的表现略优于Qwen2.5-VL-32b模型（0.845），而其余三种方法的表现仍存在较大差距。



(a) 准确率　　**(b)** 宏平均F1分数

图6： 整体性能比较。

为了更深入地理解这些模型的行为，我们进一步使用宏平均精确率和宏平均召回率评估了它们的性能，结果如图7所示。宏平均精确率量化了预测正例中正确的比例。图7a中的分析揭示了一个关键观察：排名前三的方法（YOLOv13为0.932，Qwen2.5-VL-32b为0.905，Qwen2.5-VL-7b为0.913）均取得了高于0.90的非常高的宏平均精确率。这表明，在各个类别中，当这些模型赋予物体标签时，该预测极有可能对应于真正例。相比之下，宏平均召回率衡量的是被成功识别的真实正例所占比例，反映了模型全面检测场景中存在物体的能力。如图7b所示，宏平均召回率存在显著差异。虽然YOLOv13（0.839）和Qwen2.5-VL-32b（0.808）保持了较高的宏平均召回率，但另外三种方法的宏平均召回率非常低，其中两条基于SAM2的流水线的宏平均召回率分别仅为0.260和0.218。这表明，尽管这些方法在发生检测时能产生相对准确的预测，但它们未能识别相当大一部分真实物体。



(a) 宏平均精确率　　**(b)** 宏平均召回率

图7： 性能比较。

4.2 跨场景性能分析

为了评估这些方法的鲁棒性，我们在四种环境场景下分析了它们的性能。总体准确率如图8所示。一个主要观察是，对于几乎所有方法而言，最高性能均在晴朗白天条件下取得，而雨夜条件被证明是最具挑战性的场景。在全部四种条件下，YOLOv13和Qwen2.5-VL-32b模型的表现均显著优于其他方法。这一准确率比较揭示了一个值得注意的规律。尽管YOLOv13在晴朗白天取得了略高的准确率（0.87对0.82），但Qwen2.5-VL-32b模型在其余三个更具挑战性的场景中表现最佳，尤其是在夜间环境中（晴朗夜晚：0.82对0.73；雨夜：0.78对0.75）。这一观察表明，视觉语言模型在低光照条件下表现出更强的韧性，因此有必要使用更稳健的评估指标进行更深入的考察。

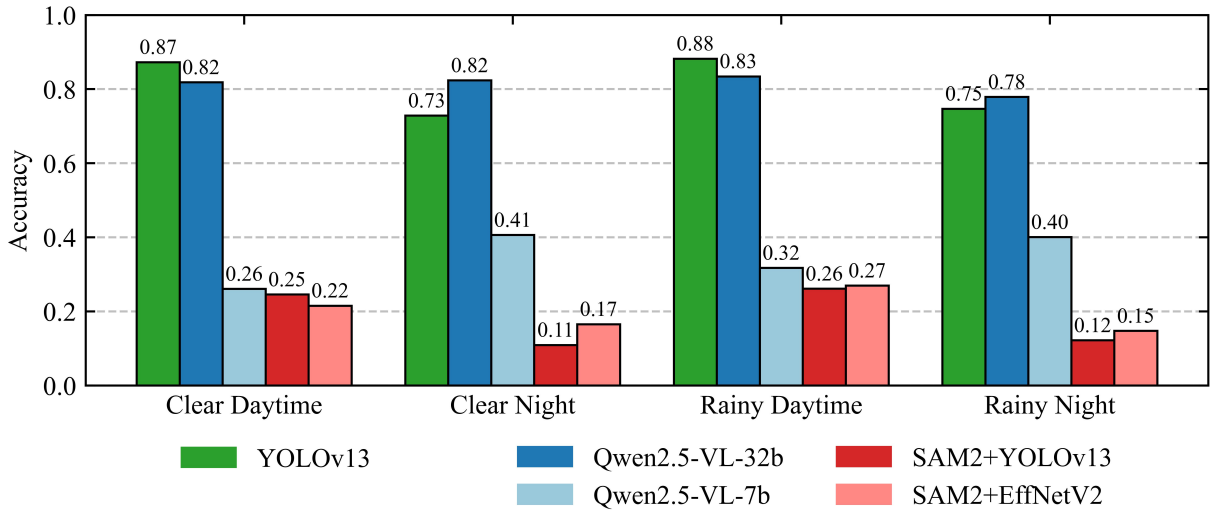


图8： 四种环境场景下的准确率比较。

我们使用宏平均F1分数对模型鲁棒性进行了更严格的分析，该指标考虑了类别不平衡。如图 9 所示，结果印证了准确率所反映的总体性能排名和下降趋势。不过，进一步考察发现，YOLOv13 在所有四种场景下始终取得最高的宏平均F1分数，凸显了其更优的总体性能。例如，在最具挑战性的雨夜场景中，YOLOv13 的宏平均F1分数达到 0.81，优于得分为 0.78 的 Qwen2.5-VL-32b。

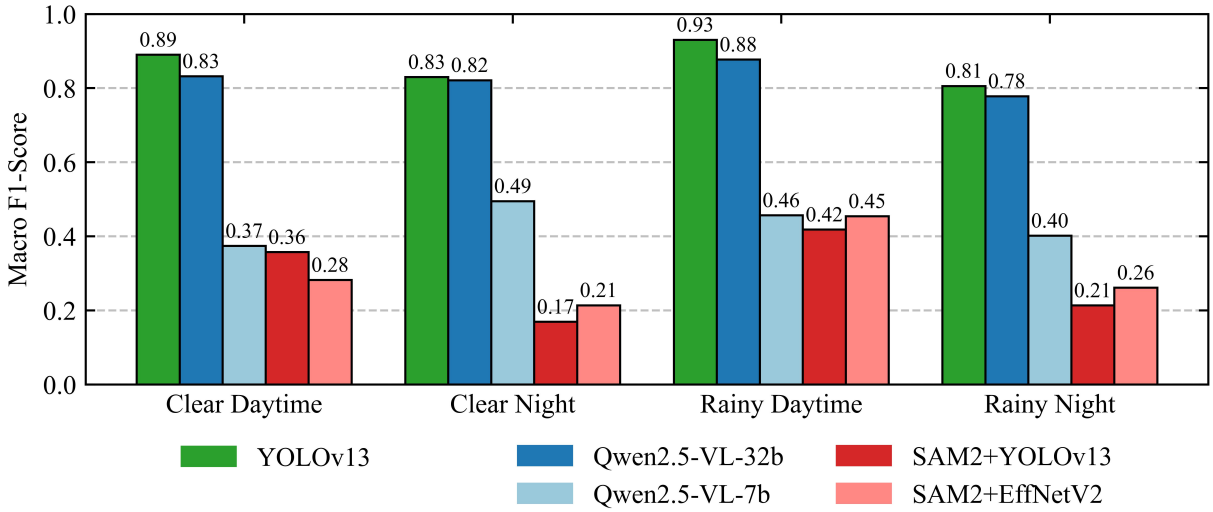


图 9： 四种环境场景下宏平均 F1 分数的比较。

除了宏平均F1分数外，图 10 还细致比较了宏平均精确率和宏平均召回率。图 10a 显示，表现最好的模型（YOLOv13 和 Qwen2.5-VL-32B）在所有场景下的宏平均精确率始终较高。相比之下，随着环境条件恶化，宏平均召回率呈现明显且持续的下降趋势（图 10b）。例如，YOLOv13 的召回率从晴朗白天的 0.88 降至雨夜的 0.76。这一比较表明，恶劣条件下的性能下降主要由漏检（假负例）导致，说明在具有挑战性的视觉条件下，模型越来越难以检测到物体。

4.3 不同物体类别的性能分析

在分析不同环境场景下的性能后，本节考察各方法在五种物体类别上的表现。目标是揭示不同范式针对不同目标类型的优势与局限，例如大型车辆与小型信号灯，以及常见类别与稀有类别之间的差异。同样，分析首先从整体准确率入手，随后通过 F1 分数、精确率和召回率进行更严格的评估，以提供均衡而全面的评价。

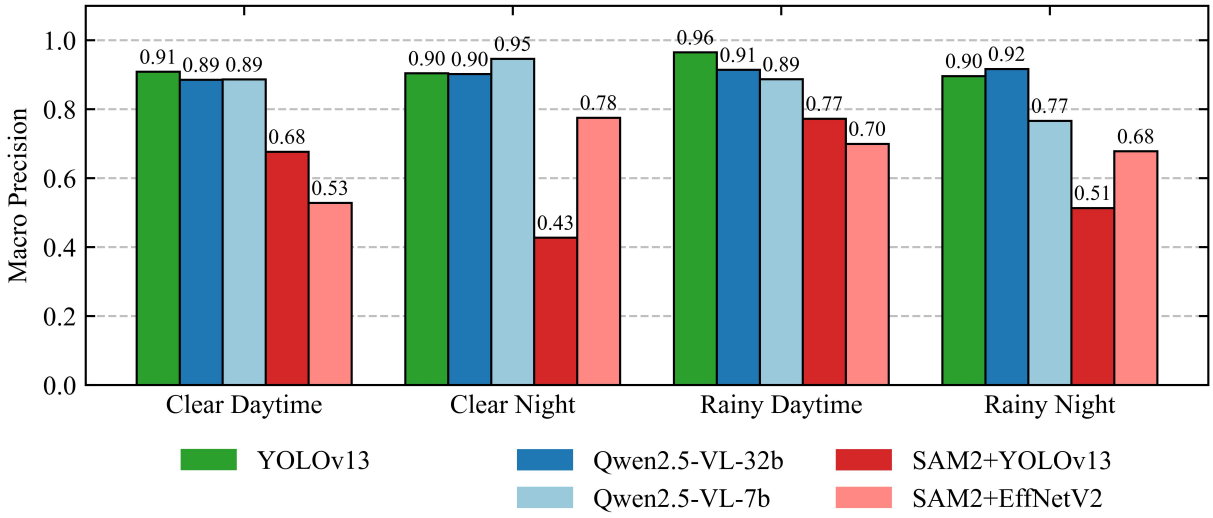
图 11 展示了五种物体类别的整体准确率。再次，YOLOv13 和 Qwen2.5-VL-32b 在大多数类别中表现最佳，同时也展现出互补优势。YOLOv13 在汽车（0.93）和公交车（0.96）等车辆类别上的表现出色，而 Qwen2.5-VL-32b 在交通信号灯检测方面明显更好，准确率达到 0.82，远高于 YOLOv13 的 0.57。

图12展示了五类物体的 F1分数，证实 YOLOv13 和 Qwen2.5-VL-32b 整体表现占优。YOLOv13 在常见的大型物体上取得最高的 F1分数：行人（0.91）、汽车（0.94）、公交车（0.91）和卡车（0.87），凸显了其识别定义清晰类别的优势。值得注意的是，在交通信号灯类别上，性能出现了明显反转：Qwen2.5-VL-32B 的 F1分数（0.90）显著高于 YOLOv13（0.73），体现了大型视觉语言模型范式在识别尺寸较小、视觉上更具挑战性的物体时具有明显优势。

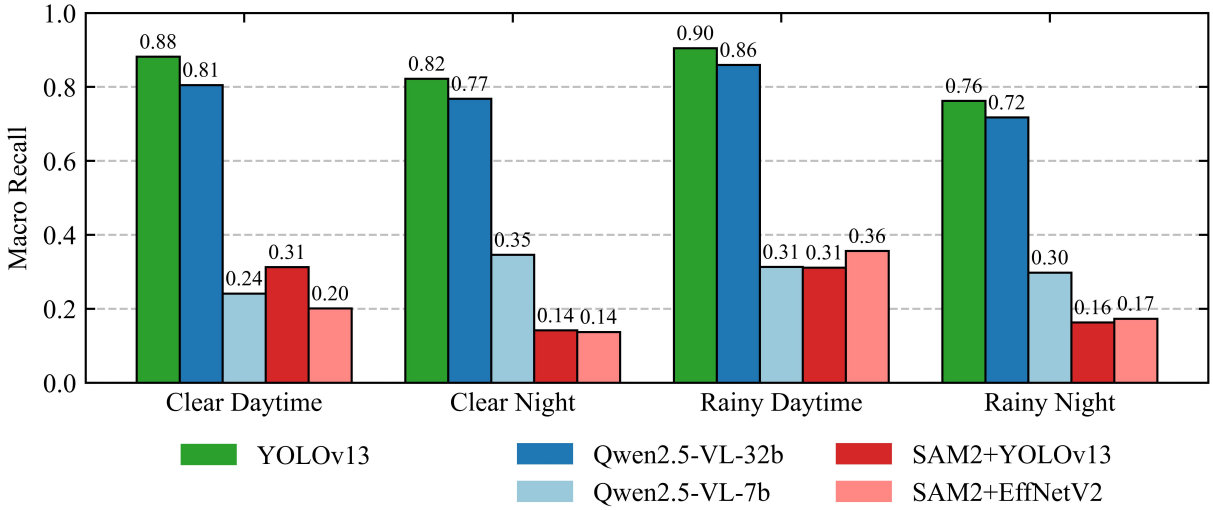
图13展示了类别级别的精确率和召回率。YOLOv13 和 Qwen2.5-VL-32B 在交通信号灯类别上的精确率均达到满分（1.00），表明没有假正例。两者的性能差距主要源于召回率：Qwen2.5-VL-32B 的召回率为 0.82，显著高于 YOLOv13 的 0.57。这解释了 YOLOv13 的 F1分数较低，其原因是漏检，而非误分类。视觉语言模型更高的召回率可能反映了其利用全局场景上下文和隐含世界知识的能力，从而能够更可靠地检测尺寸较小、视觉上更具挑战性的物体。

4.4 模型行为的深入分析

为补充定量结果，本节对所评估模型的内部行为与定性误差模式进行更深入的分析。该分析分为两部分：首先，使用混淆矩阵来



(a) 宏平均精确率



(b)四种环境场景下的宏平均召回率

图 10: 性能比较。

可视化类别特定错误模式；其次，考察关键模型组件的置信度分数分布，以了解模型如何做出决策，并诊断失败的根本原因。

图14展示了所有五种评估方法的混淆矩阵，可直接比较它们的错误模式。表现最佳的模型 YOLOv13（图14a）和 Qwen2.5-VL-32b（图14e）的矩阵呈现出明显的对角占优，直观地证实了它们的正确分类率较高。对于这些模型，少数非对角线错误属于合乎逻辑的误分类，例如将卡车（类别ID 7）误认为汽车（类别ID 2）；鉴于两者在视觉上的相似性，这种情况并不意外。然而，最关键的发现来自每个矩阵的第一列，该列表示预测类别ID -1。两种分割辅助方法 SAM2+EfficientNetV2（图14b）和 SAM2+YOLOv13（图14c）的矩阵在这一列中都显示出极高的数值。例如，在图14b中，336 个行人实例中有 335 个被错误拒绝。同样，与规模更大的对应模型相比，较小的视觉语言模型 Qwen2.5-VL-7b（图14d）的矩阵也显示出这一列中的拒绝数量非常高。这直接以可视化方式印证了此前得出的结论：分割辅助方法的灾难性失效，以及 7b 视觉语言模型相对较弱的表现，主要是由于无法做出正类分类（即召回率问题），而非误分类率过高。

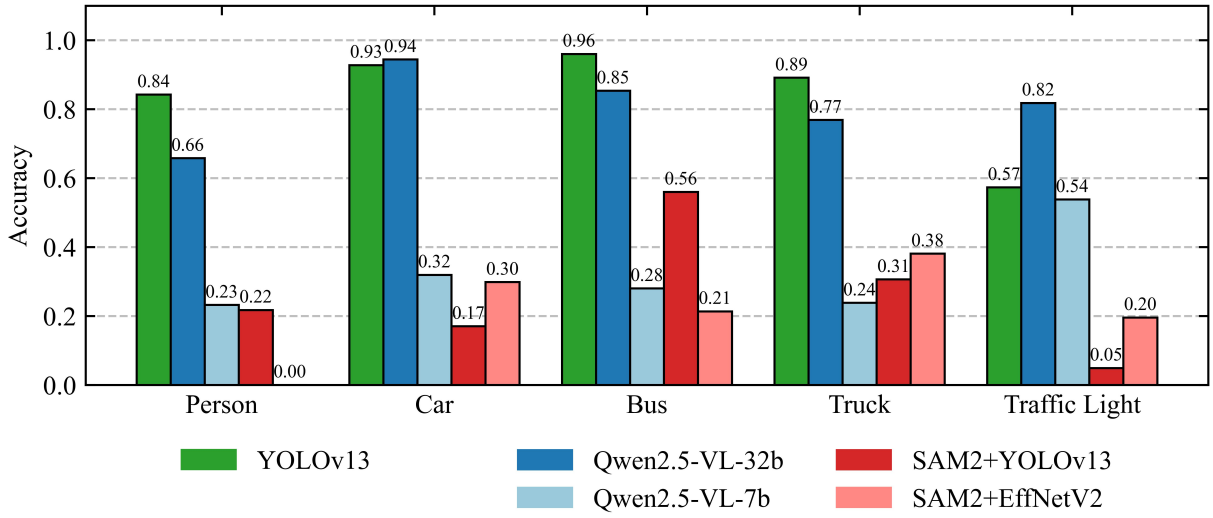


图 11： 五种物体类别上的方法性能（准确率）。

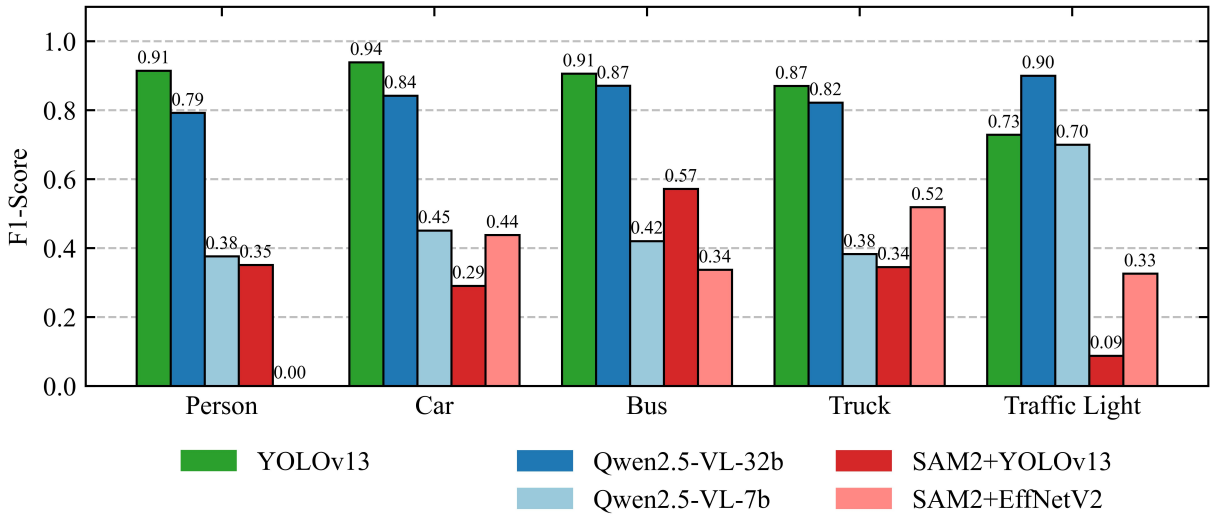
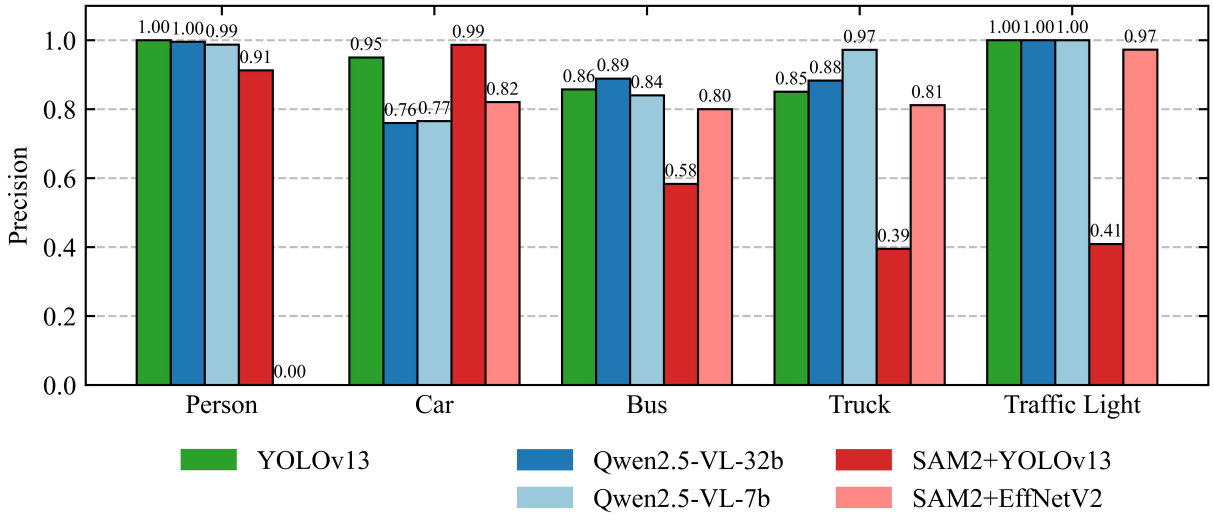


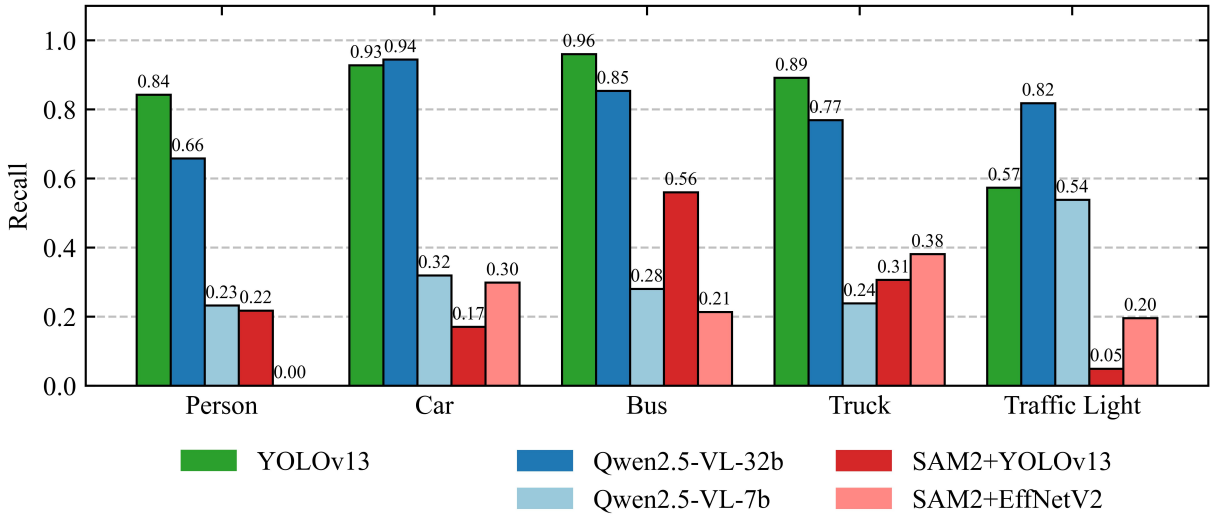
图12： 五种物体类别上的方法性能（F1分数）。

为进一步诊断分割辅助范式失效的根本原因，研究人员分析了其各组成部件的置信度分数，如图15所示。图15a中的独立 YOLOv13 模型直方图显示，该模型通常对其预测非常有信心，分布明显偏向 0.8 至 1.0 之间的高分。关键证据见图15b：SAM2 分割阶段同样信心十足，其分数高度集中在 0.8 以上。然而，在 SAM2 生成的裁剪图像上运行分类器时，它们的置信度显著下降。如图15c和图15d所示，经过 SAM2 处理后的分类器，其置信度分布更加分散、不确定；其中 EfficientNetV2 模型的置信度分数集中在 0.1 至 0.3 的较低范围。

这些观察结果清晰地揭示了“部分与整体”问题。基于SAM2的流程失效，并非源于分割效果不佳或置信度低；事实上，SAM2通常对其输出具有很高的置信度。问题在于，SAM2会以很高的置信度分割出错误的语义实体。例如，当驾驶员的注视落在汽车尾灯上时，SAM2可能会准确地将尾灯分割为一个物体部件。然而，下游分类器此时接收到的是一个孤立的部件，却必须从汽车等整体物体类别中进行选择，从而导致判断不确定或误分类。这种语义不匹配解释了置信度分数偏低以及频繁预测为背景（类别ID：-1）的现象，也表明部分与整体问题是分割辅助范式应用于此任务时的一项根本局限。



(a) 精确率



(b)各物体类别的召回率

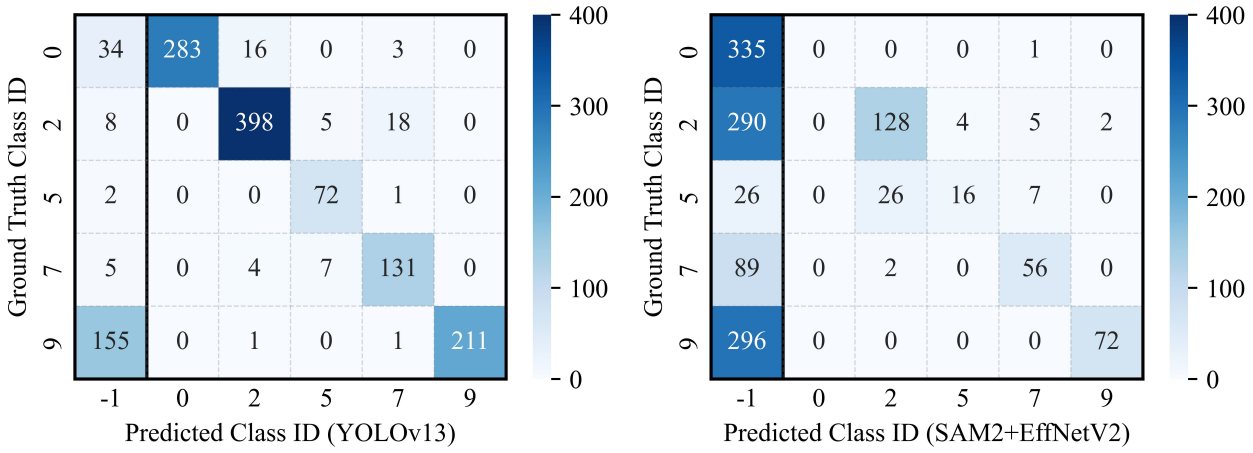
图13：性能比较。

4.5 计算成本分析

除模型性能外，计算效率也是车辆实际部署的关键因素。为评估这一点，所有实验均在单块NVIDIA A6000 GPU上进行。推理速度（帧每秒）通过对100张测试图像的处理时间取平均值计算得出，而GPU峰值内存占用则反映了处理过程中观察到的最大内存消耗。

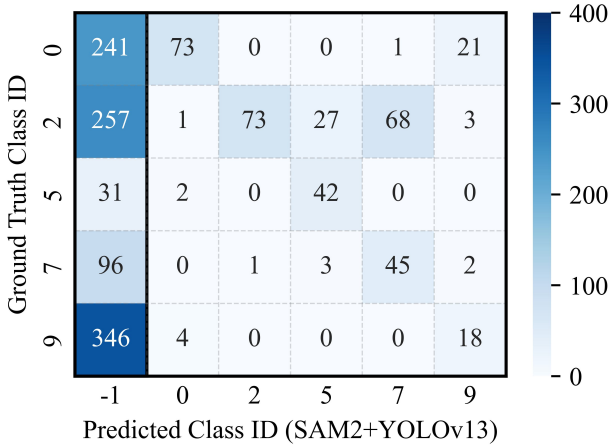
表2总结了五种方法的计算成本和资源需求，显示出速度与内存方面的显著差异。YOLOv13是唯一实现实时性能（36.1帧每秒）且内存占用适中（27.6M参数、1.53吉字节显存）的模型，因而适合车载部署。基于SAM2的流水线运行速度约为10帧每秒（10.3-10.4帧每秒），对近实时任务或许可以接受，但对延迟关键型应用则不足。视觉语言模型的吞吐量极低，其中Qwen2.5-VL-7b为2.1帧每秒，Qwen2.5-VL-32b为0.07帧每秒（约每帧14秒），同时显存需求巨大（17.6-45吉字节显存），因此需要顶级或定制GPU。

将这些结果与先前的性能分析相结合，可以清楚地看出性能-效率权衡。YOLOv13在准确率与速度之间取得了良好平衡，是实时部署中最实用的选择。

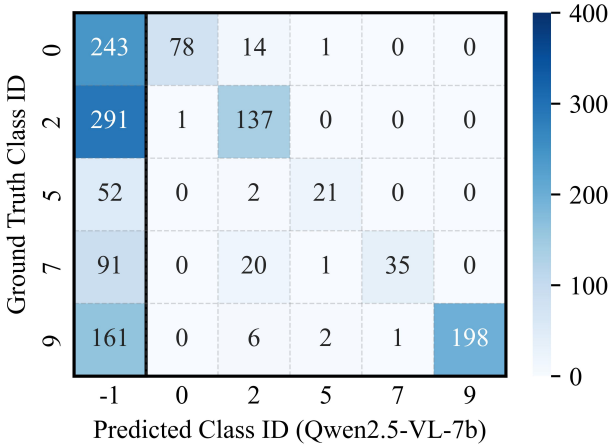


(a) YOLOv13

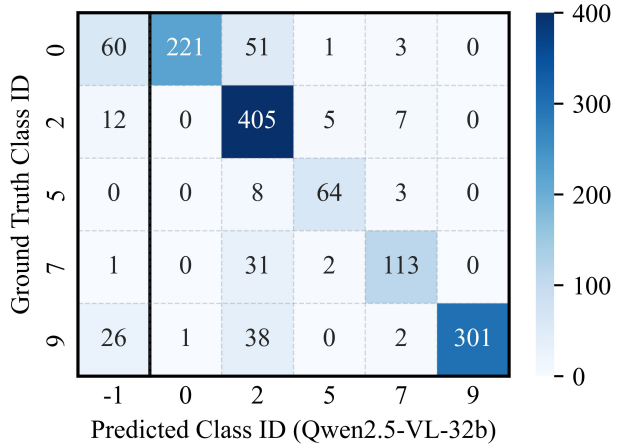
(b) SAM2+EfficientNetV2



(c) SAM2+YOLOv13

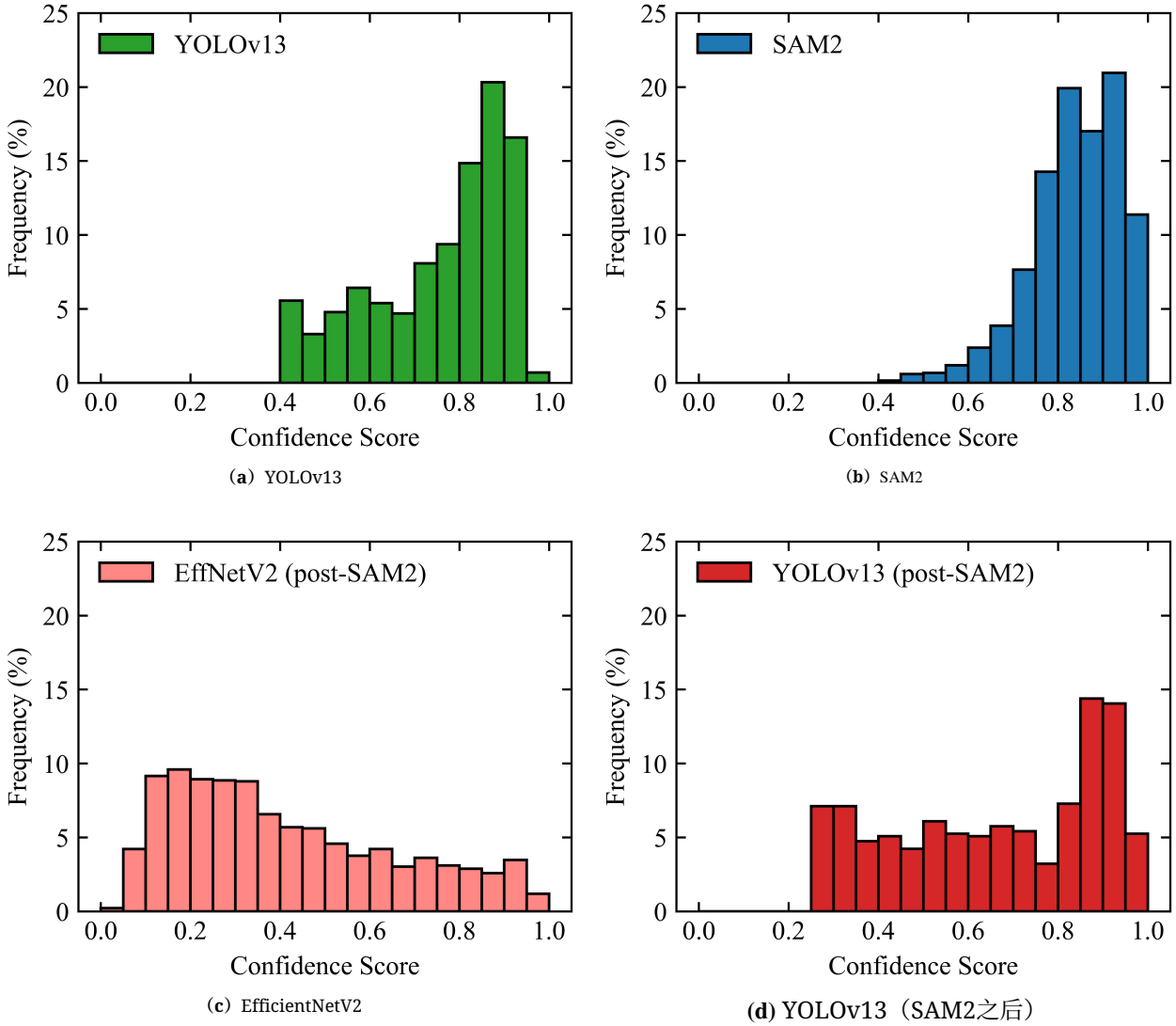


(d) Qwen2.5-VL-7b



(e) Qwen2.5-VL-32b

图14： 所有五种方法的混淆矩阵。



(a) YOLOv13 (b) SAM2

(c) EfficientNetV2 (d) YOLOv13（SAM2之后）

图15：模型组件的置信度分数分布。

Qwen2.5-VL-32B虽然在某些条件下稳健且准确，但会产生过高的计算成本，从而限制其实时适用性。基于SAM2的方法在效率和准确率方面表现均较差，因此是最不可行的选择。

表2：所有方法的计算成本比较。

范式	方法	模型大小	参数 (M)	推理速度（帧每秒）	GPU峰值内存使用量（吉字节）
物体基于检测	YOLOv13	107 MB	27.6	36.1	1.53 吉字节
分割-辅助分类	SAM2 +EfficientNetV2	963 MB	343.0	10.4	11.83 吉字节
	SAM2 +YOLOv13	1.3 吉字节	252.1	10.3	9.20 吉字节
视觉-语言基于模型	Qwen2.5-VL-7b	15.5 吉字节	8,292.2	2.1	17.63 吉字节
	Qwen2.5-VL-32b	63.6 吉字节	33,452.7	0.07	44.97 吉字节

5 讨论

本研究的主要发现是，对于基于点的目标识别，直接目标检测和大规模视觉语言模型范式优于分割辅助分类方法。这表明，利用整体场景上下文的方法比依赖分析式分阶段策略的方法更有效。尽管 YOLOv13 和 Qwen2.5- VL-32B 的架构截然不同，但二者都会分析整幅图像来做出预测，因此能够比两阶段流程作出更可靠的推断；后者会先隔离出一个缺乏上下文的小区域，再进行分类。

分割辅助方法因存在根本性的“部件与整体”语义鸿沟而失效。置信度分析（图15）显示，SAM2生成的分割结果置信度很高，而下游分类器的置信度却较低且不稳定，这表明错误源于模型虽然确信地分割了物体部件，却没有分割完整物体。对于车辆，注视点通常会得到车牌等部件的精确分割结果，但这些部件无法被归类为汽车。对于小目标（行人、交通信号灯），分割区域往往过小且缺乏上下文信息，无法可靠识别。ImageNet-1K中的类别体系不匹配，也进一步限制了行人类别的性能。相比之下，对于较大的物体（公交车、卡车），当注视点落在物体的大面积区域上时，模型能够分割完整物体并进行更可靠的分类，因此取得了一定成效。

相比之下，大型视觉语言模型表现出色，尤其是在具有挑战性的情形下，这凸显了其上下文推理和常识推理能力的优势。Qwen2.5-VL-32b的优势在小目标（如交通信号灯，图12）上的稳健表现，以及在恶劣条件下的相对稳定性（图 9）中体现得尤为明显。交通信号灯类别的精确率-召回率分析（图13）尤其具有启发性：尽管视觉语言模型的精确率达到完美水平，其召回率却显著更高，这表明它能够运用隐含的世界知识，有效推断交叉路口处的小型彩色灯光是交通信号灯。这一优势在夜间最为突出，此时Qwen2.5-VL-32b的表现显著优于传统检测器。整体场景推理使视觉语言模型即使在局部视觉线索因弱光或恶劣天气而退化时，仍能保持有效。

这些发现为系统工程师带来了明确的实际启示，总结见表3。结果凸显了一种根本性的性能-效率权衡。对于当前需要实时反馈（>10 帧每秒）的车载应用，像YOLO这样高效的目标检测器是最务实且可部署的选择，能以极低的计算成本在常见场景中提供较高的准确率。然而，对于未来的安全关键型应用，例如 3+ 级高级驾驶辅助系统或事故后取证分析，在不利和罕见条件下的鲁棒性至关重要，大型视觉语言模型正成为技术上的继任者。尽管其巨大的计算需求目前限制了其实时适用性，但其卓越的上下文推理能力以及在挑战性环境中的韧性，指向了以人为中心的车辆感知的未来。硬件与模型优化的持续进步有望缩小这一差距。

表3： 三种范式及其实际权衡总结。

范式	总体性能（宏平均F1）	鲁棒性(夜间/雨天)(宏平均F1)	小目标性能(交通信号灯召回率)	速度（帧每秒）	VRAM(GB)	主要缺点
基于检测	Good（0.87）	High（≥ 0.81）	弱（0.57）	实时（36.1）	Low（1.53）	未检测到小目标
分割-辅助	Poor（< 0.35）	Low（≤ 0.45）	差（≤ 0.20）	Near实时（10.3）	High（9.20）	“部分与整体”语义鸿沟
基于视觉语言模型	尺度依赖因变量（0.47–0.85）	尺度依赖依赖(0.40–0.88)	尺度依赖(0.54–0.82)	Slow(0.07)	非常高(44.97)	High计算cost

6 结论

本文对识别驾驶员注视点处语义对象的方法进行了系统的跨范式评估。主要发现总结如下：

1. 范式有效性：研究发现，直接目标检测（YOLOv13）和大型视觉语言模型（Qwen2.5-VL-32b）范式是完成该任务最有效的方法。相比之下，分割辅助范式存在根本性的“部分与整体”语义鸿沟。

2. 性能-效率权衡：两种领先范式之间存在明确的权衡。YOLOv13 具备出色的实时效率，是当前车载部署最务实的选择。Qwen2.5-VL-32b 虽然计算开销较大，但鲁棒性和上下文理解能力更强，尤其是在处理小目标（例如交通信号灯）以及应对不利条件（例如夜间）时表现突出，因此有望成为未来安全关键型应用的有力候选方案。

3. 主要失效模式：在所有模型中，具有挑战性的条件下性能下降主要是由漏检（召回率问题）导致的。大型视觉语言模型（Qwen2.5-VL-32b）在这些条件下始终表现出更高的召回率，凸显了其整体且具备上下文感知的感知能力所带来的优势。

总体而言，这些发现为新一代以人为本的智能车辆系统的开发提供了路线图。尽管如此，若干局限性也指明了未来的研究方向。该基准数据集虽然涵盖多种情况，但规模有限，且依赖模拟注视点，无法反映真实眼动追踪硬件的噪声特性。此外，目前的分析基于单帧，未利用时间上下文。因此，未来研究应使用真实世界注视数据验证这些发现，探索兼顾效率与鲁棒性的混合架构，并将基于视觉语言模型的方法扩展到视频输入，以利用时间信息。

数据可用性

该数据集可通过 http://bdd-data.berkeley.edu 公开访问

作者声明不存在任何利益冲突。

参考文献

[1] 国家统计与分析中心。2024年机动车交通事故死亡人数初步估计。[2] 世界卫生组织。《2023年道路安全全球现状报告》。[3] Christer Ahlström、Katja Kircher、Marcus Nyström和Benjamin Wolfe。驾驶员注意力研究中的眼动追踪——注视数据的解读如何影响我们的认知。2:778043。

[4] Susana García-Herrero、Juan Diego Febres、Wafa Boulagouas、José Manuel Gutiérrez和Miguel Ángel Mariscal Saldaña。技术导致的分心驾驶对驾驶员违规行为的影响及其对交通事故严重程度的后续影响评估。18(13):7155。[5] 国家统计与分析中心。2023年分心驾驶。[6] Sheila G. Klauer、Johnathon P. Ehsani、Daniel V. McGehee和Michael Manser。次要任务参与对青少年驾驶表现和碰撞风险的影响。57:S36–S43。[7] Bruce G. Simons-Morton、Feng Guo、Sheila G. Klauer、Johnathon P. Ehsani和Anuj K. Pradhan。请注视道路：年轻驾驶员的碰撞风险随分心时长增加。54:S61–S67。[8] Feng Guo、Sheila G Klauer、Youjia Fang、Jonathan M Hankey、Jonathan F Antin、Miguel A Perez、Suzanne E Lee和Thomas A Dingus。年龄对与驾驶员分心相关的碰撞风险的影响。46(1):258–265。[9] Yanchao Dong、Zhencheng Hu、Keiichi Uchimura和Nobuki Murayama。面向智能车辆的驾驶员注意力不集中监测系统：综述。12(2):596–614。[10] Muhammad Qasim Khan和Sukhan Lee。注视与眼动追踪：高级驾驶辅助系统中的技术与应用。19(24):5540。[11] Agapito Ledezma、Víctor Zamora、Óscar Sipele、M. Paz Sesmero和Araceli Sanchis。实施注视追踪算法以改进高级驾驶辅助系统。10(12):1480。[12] Yiheng Feng、Yunfeng Chen、Jiansong Zhang、Chi Tian、Ran Ren、Tianfang Han和Robert W. Proctor。结合网联自动驾驶车辆的下一代交通基础设施的以人为中心的设计：系统之系统视角。25(3):287–315。[13] Shabnam Haghzare、Jennifer L. Campos和Alex Mihailidis。对年长驾驶员在自动驾驶与非自动驾驶期间注视行为的分类：迈向检测模式混淆的初步步骤。40(2):241–254。[14] Jong-Bae Kim。用于提升自动驾驶汽车安全的可持续实时驾驶员注视监测。17(9):4114。[15] F. Walker、J. Wang、M. H. Martens和W. B. Verwey。注视行为与皮肤电活动：自动化车辆中驾驶员信任的客观测量。64:401–412。[16] Yancong Zhu、Yunke Geng、Ruonan Huang、Xiaonan Zhang、Lu Wang和Wei Liu。驶向未来：探索面向自动驾驶车辆的玻璃投影显示系统的以人为中心的设计与实验。40(15):4087–4102。[17] Manda Winlaw、Stefan H. Steiner、R. Jock MacKay和Allaa R. Hilal。利用远程信息处理数据发现高风险驾驶行为。131:131–136。[18] Debasis Kumar和Naveed Muhammad。通过数据融合与YOLOv8实现自动驾驶在恶劣天气下的目标检测。23(20):8471。[19] Genchao Liu、Kun Wu、Wei Lan和Yunjie Wu。YOLO-FDCL：用于复杂光照条件下驾驶员疲劳检测的改进型YOLOv8。25(15):4832。[20] Tuisku Tammi、Jami Pekkanen、Benjamin Ultan Cowley和Otto Lappi。以锚定于任务表现的综合注视指标量化遮挡期间的追踪质量。15:31858。[21] Diako Mardanbegi、Tobias Langlotz和Hans Gellersen。通过前庭眼反射深度估计解决三维注视交互中的目标歧义。载于*2019*年计算机系统人因国际会议论文集，CHI ’19，第1–12页。计算机协会。[22] Jeremy M Wolfe、Anna Kosovicheva和Benjamin Wolfe。正常盲视：当我们看了却未能看见。26(9):809–819。[23] Mai-Britt Herslund和Niels O Jørgensen。交通中的“看了却未能看见”错误。35(6):885–891。[24] Kellie D. Kennedy和James P. Bliss。模拟驾驶任务中的无意视盲。57(1):1899–1903。[25] João Eduardo Hoffmann、Hilkija Gaïus Tosso、Max Mauro Dias Santos、João Francisco Justo、Asad Waqar Malik和Anis Ur Rahman。面向自动驾驶车辆的实时自适应目标检测与追踪。6(3):450–459。[26] Zhongzhou Wang、Keming Yao和Fuao Guo。基于改进YOLOv5的驾驶员注意力检测。13(11):6645。[27] Njayou Youssouf。使用卷积神经网络进行交通标志分类，并使用更快区域卷积神经网络和YOLOv4进行检测。8(12):e11792。

[28] Madhusri Maity、Sriparna Banerjee 和 Sheli Sinha Chaudhuri。基于 Faster R-CNN 和 YOLO 的车辆检测：综述。载于*2021*年第五届计算方法与通信国际会议（*ICCMC*），第 1442–1447 页。[29] Francesco Tonini、Nicola Dall’Asen、Cigdem Beyan 和 Elisa Ricci。目标感知注视目标检测。载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第 21803–21812 页。[30] Behzad Mirzaei、Hossein Nezamabadi-pour、Amir Raoof 和 Reza Derakhshani。小目标检测与跟踪：全面综述。23(15):6887。[31] Huaqi Zhao、Su Wang、Xiang Peng、Jeng-Shyang Pan、Rui Wang 和 Xiaomin Liu。面向自动驾驶的路面语义分割。10:e2250。[32] Firas Hazzaa、Innocent Udoidiong、Akram Qashou 和 Sufian Yousef。Segment Anything：综述。2024年:150–161。[33] Zhongkai Yuan。Segment Anything模型的原理、应用与进展。53:73–78。[34] Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman Rädle、Chloe Rolland、Laura Gustafson、Eric Mintun、Junting Pan、Kalyan Vasudev Alwala、Nicolas Carion、Chao-Yuan Wu、Ross Girshick、Piotr Dollár 和 Christoph Feichtenhofer。SAM 2：在图像和视频中实现 Segment Anything。[35] Mohammed Elhenawy、Huthaifa I. Ashqar、Andry Rakotonirainy、Taqwa I. Alhadidi、Ahmed Jaber 和 Moham-mad Abu Tami。用于自动驾驶的视觉语言模型：基于 CLIP 的动态场景理解。14(7):1282。[36] Aryan Keskar、Srinivasa Perisetla 和 Ross Greer。评估多模态视觉语言模型提示策略在道路场景理解中用于视觉问答的效果。载于*2025 IEEE/CVF*冬季计算机视觉应用会议研讨会（*WACVW*），第 937–946 页。[37] Jinze Bai、Shuai Bai、Shusheng Yang、Shijie Wang、Sinan Tan、Peng Wang、Junyang Lin、Chang Zhou 和 Jingren Zhou。Qwen-VL：一种用于理解、定位、文本阅读等任务的多功能视觉语言模型。[38] Shijie Wang、Dahun Kim、Ali Taalimi、Chen Sun 和 Weicheng Kuo。从生成式视觉语言模型中学习视觉定位。载于 *2025 IEEE/CVF* 冬季计算机视觉应用会议（*WACV*），第 8057–8067 页。[39] Fisher Yu、Haofeng Chen、Xin Wang、Wenqi Xian、Yingying Chen、Fangchen Liu、Vashisht Madhavan 和 Trevor Darrell。BDD100K：用于异构多任务学习的多样化驾驶数据集。[40] Tsung-Yi Lin、Michael Maire、Serge Belongie、Lubomir Bourdev、Ross Girshick、James Hays、Pietro Perona、Deva Ramanan、C. Lawrence Zitnick 和 Piotr Dollár。Microsoft COCO：上下文中的常见物体。[41] Mengqi Lei、Siqi Li、Yihong Wu、Han Hu、You Zhou、Xinhu Zheng、Guiguang Ding、Shaoyi Du、Zongze Wu 和 Yue Gao。YOLOv13：采用超图增强自适应视觉感知的实时目标检测。[42] Mingxing Tan 和 Quoc Le。EfficientNetV2：更小的模型与更快的训练。载于第*38*届国际机器学习会议论文集，第 10096–10106 页。PMLR。