PDB-Eval：用于个性化驾驶行为描述和解释的大型多模态模型的评估

Junda Wu
Computer Science and Engineering
University of California San Diego
La Jolla, USA
juw069@ucsd.edu

Jessica Echterhoff
Computer Science and Engineering
University of California San Diego
La Jolla, USA
jechterh@ucsd.edu

Kyungtae Han
InfoTech Labs
Toyota Motor North America
Mountain View, USA
kt.han@toyota.com

Amr Abdelraouf
InfoTech Labs
Toyota Motor North America
Mountain View, USA
amr.abdelraouf@toyota.com

Rohit Gupta
InfoTech Labs
Toyota Motor North America
Mountain View, USA
rohit.gupta@toyota.com

Julian McAuley
Computer Science and Engineering
University of California San Diego
La Jolla, USA
jmcauley@eng.ucsd.edu

***Abstract*—理解驾驶员的行为和意图对于潜在风险评估和早期事故预防至关重要。可以根据个体驾驶员的行为来定制安全和驾驶员辅助系统，从而显著提高其效能。然而，现有的数据集在根据外部视觉证据描述和解释一般车辆运动方面存在局限性。本文介绍了一个用于详细理解个性化驾驶行为的基准，PDB-Eval，并将大型多模态模型（MLLMs）与驾驶理解和推理相结合。我们的基准由两个主要组件组成：PDB-X 和 PDB-QA。PDB-X 可以评估 MLLMs 对时间性驾驶场景的理解。我们的数据集旨在从外部视角寻找有效的视觉证据，以便从内部视角解释驾驶员行为。为了将 MLLMs 的推理能力与驾驶任务对齐，我们提出了 PDB-QA 作为针对 MLLM 指令微调的视觉解释问答任务。作为生成模型（如 MLLMs）的通用学习任务，PDB-QA 可以在不损害 MLLMs 泛化能力的情况下弥合领域差距。我们的评估表明，对 MLLMs 进行细粒度描述和解释微调可以有效弥合 MLLMs 与驾驶领域之间的差距，这能将问答任务的零样本性能提高至 73.2 %。我们进一步评估了在 PDB-X 上微调的 MLLMs 在 Brain4Cars 的意图预测和 AIDE 识别任务中的表现。我们观察到，在 Brain4Cars 的转向意图预测任务中性能提升高达 12.5 %，并且在 AIDE 的所有任务中性能提升一致最高达 11.0 %。**

***Index Terms*—MLLMs, Personalized Driving Behavior, Question-answering**

I. 引言

驾驶员的理解对于预测车辆运动 [?] 和评估道路上潜在风险至关重要。最近的研究显示了在交通事故、不确定性和车辆运动预测的识别任务方面的进展 [?] 。在这些任务中，关于车辆运动 [?] 和驾驶员行为 [?] 的文字解释在更可解释的理解方面变得愈加重要。由于大型多模态模型（MLLMs）可以基于视觉证据生成描述和解释 [?], [?], [?] ，MLLMs 被认为是驾驶任务的多模态推理器 [?], [?] 。然而，现有的 MLLMs 限于针对一般视觉理解和解释任务的微调和评估 [?], [?], [?] ，在适应驾驶任务时存在领域差距。

为了使 MLLMs（多模态语言模型）与驾驶员理解和推理任务保持一致，我们提出了一个评价数据集 PDB-Eval，用于个性化驾驶员行为理解。尽管许多现有工作专注于从车内视角 [?], [?] 或外部交通场景的视角 [?], [?] 分析驾驶员行为，但很少有尝试将这些视角理解为并行时间过程，其中一个视角的视频片段（即一个发生在时间边界内的事件）可以根据另一个视角的视觉证据进行解释。例如，在图 **??** 中，现有关于车辆运动的细粒度解释（例如，BDD-X [?] ）局限于单一视角理解。多视角驾驶理解数据集（例如，AIDE [?] ）能够理解多种类别的驾驶员行为和车辆运动，但在提供细粒度描述方面有所不足。为了统一来自内部和外部视觉证据的驾驶行为描述和解释，我们介绍了我们的第一个评估任务 PDB-X。例如，在图 **??** 中，我们展示了 PDB-X 基于多视角视觉证据提供的细粒度驾驶行为解释。我们进一步提出视觉解释问答任务 PDB-QA，以增强 MLLMs 在驾驶员行为背景下的解释和推理能力。

提取个性化的驾驶行为描述具有挑战性，因为人类标注者必须基于对不同类型司机的知识和观察来进行描述。我们建议使用多模态大模型（MLLMs）通过比较提示 [?] 来创建个性化的驾驶行为描述，这种方法强调具有相同意图的司机之间行为的差异，从而确保描述的具体性和相关性。尽管这种方法具有潜力，但关键是要解决 MLLM 生成内容中固有的幻觉问题，即模型可能会生成虚构或不相关的信息 [?], [?] 。这一问题在需要基于视觉证据进行详细和准确描述的任务中尤为重要 [?], [?] 。为降低此风险，我们对每种驾驶类型的提示答案进行分类，然后使用这些类别来引导更集中的 MLLM 提示，从而减少不相关信息的出现。

我们通过微调开源的多模态大模型（MLLMs），BLIP-2 和 VTimeLLM，来评估 PDB-Eval 中的两个任务，它们分别专注于图像理解和视频理解。我们还包括 GPT-4V 作为一个强大的零样本基线，以展示现有多模态大模型的性能。此外，我们进一步评估了 Brain4Cars 数据集中驾驶员意图预测任务 [?] （域内）和 AIDE 数据集中驾驶识别任务 [?] （跨域）。评估结果展示了多模态大模型通过在 PDB-X 上的训练所获得的描述和解释能力的有效性和泛化能力。

II. 相关工作

最近的驾驶理解研究涉及到包括事故时间、意图、事故预测、驾驶预期、分心和不确定性估计在内的任务 [?], [?] 。尽管取得了进展，许多系统缺乏细粒度的可解释性和推理能力，从而限制了对自动驾驶汽车等应用的信任。为了解决这个问题，最近的方法合并了文本解释、基于行车记录仪证据的问答 [?], [?] 、注意力图，甚至概念瓶颈框架 [?] 来表达司机的行为。与静态的、分类的描述相对比，我们的基准测试要求模型理解个性化的司机行为，并将其与动态的车辆和交通状况相关联。

大型多模态模型（MLLMs）在规划、导航、模拟和命令理解等驾驶相关任务中的应用日益增多 [?], [?] 。然而，尽管现有的基于视频的 MLLM 评估主要集中在单视角分析，一个关键挑战仍然存在：对两个因果关联的时间过程进行推理——驾驶员行为的内部动态和外部交通变化。因此，我们的综合评估任务要求 MLLMs 同步解释这两条流并提供基于视觉证据的解释。这种双视角方法不仅提高了驾驶理解系统的可解释性，还为更安全和更加透明的以人为中心的驾驶技术铺平了道路。

III. 方法

A. 流程概述

我们的数据集创建流程如图 1 所示，其中我们识别出五个数据处理步骤如下：

- 步骤 1 比较提示：为了理解个性化的司机行为，我们在 MLLMs 中提出了比较提示，用于描述执行相同动作时司机行为的差异。
- 步骤 2 驾驶员身份 & 意图一致性过滤：我们从生成的描述中提取意图和驾驶员身份，作为自动样本过滤的验证测量。
- 步骤 3 指南构建：我们不直接使用从比较提示中提取的描述，而是提取驾驶行为类型，并将描述分类为提示指南，这些指南由纯文本的 LLM 进行后处理。
- 第 4 步指南-指令提示：通过使用生成指南 [?] 提示 MLLMs，我们可以更有效地关注特定的驾驶行为特征。这种方法导致更细致和个性化的驾驶行为，从而降低幻想的可能性。
- 步骤 5 人工标注过滤：然而，MLLMs 在内部和外部行车记录仪的有限可视范围内仍可能臆测不存在的视觉环境。因此，人工标注涉及不可见和不相关行为描述检测的方面级和样本级过滤。

使用所提出的流程，我们从 Brain4Cars [?] 数据集中增强了双摄像头视频，并且加入了个性化的驾驶员行为描述和解释。原始的 Brain4Cars 数据集来自 10 位不同的驾驶员，分割成 700 个事件片段，其中包括 274 次换道、131 次转弯和 295 次直行驾驶实例 [?] 。

为了获得司机的个性化行为，我们提出在 MLLMs 中使用比较提示，以提取具有相同意图的两个司机之间的行为差异。对于每对具有相同意图的司机 u, v ，我们首先从车载摄像头视频帧 $\boldsymbol{I}_u^{in}, \boldsymbol{I}_v^{in} \in V_i$ 中提取 N 帧作为视觉证据。为了构建与 MLLM 提示兼容的视觉上下文，司机视频的时间帧被连接起来（例如，图 1 的步骤 1），

$$\boldsymbol{I}_{u,v}^{in} = \text{vconcat}\left[\text{hconcat}(\boldsymbol{I}_u^{in}), \text{hconcat}(\boldsymbol{I}_v^{in})\right],$$

其中，vconcat 和 hconcat 分别表示图像帧的纵向和横向连接。

由于 MLLMs 是从具有多模态对齐的 LLMs 发展而来的，MLLMs 可能拥有强烈的文本先验（即语言偏差）[?] ，这可能导致忽视视觉证据的虚假响应 [?], [?] 。具体来说，先前的研究指出，如果没有指导模型专注于视觉语义的细粒度方面，就可能导致幻觉 [?] 。因此，在这一步中，我们建议提示 MLLMs 仅生成司机之间的比较性描述，使 MLLMs 感知更多细粒度的视觉细节 [?] ，并可能防止幻觉问题 [?], [?] 。

我们的流程中的比较提示设计由 4 个主要指令组成。解释指令 $\boldsymbol{T}_x$ 是显式地对连接的视频帧序列进行作为视频的推理。然后，使用比较指令 $\boldsymbol{T}_c$ 提示 MLLM，以回答两个驾驶员行为之间的差异。为了得出驾驶员的身份和意图一致性指标，我们基于模型之前的比较描述，通过身份 $\boldsymbol{T}_{id}$ 和意图 $\boldsymbol{T}_{it}$ 提取指令来提示 MLLM。通过对驾驶员的身份和意图一致性评价，我们可以提高 MLLM 响应的准确性。结合视觉证据和文本指令，我们提示 MLLM 生成初步响应

$$\boldsymbol{R}_{u,v} = \text{MLLM}\left(\boldsymbol{I}_{u,v}^{in}, [\boldsymbol{T}_x; \boldsymbol{T}_c; \boldsymbol{T}_{id}; \boldsymbol{T}_{it}]\right),$$

，在该响应中我们确保驾驶员的意图相同。

我们引入了一个基于驾驶员身份和意图一致性的中间样本验证和过滤步骤。上下文指令 $\boldsymbol{T}_c'$ 是将比较描述重新格式化为驾驶员个性化行为类型及方面的字典，这利用了 LLM 的结构化文本生成能力。我们还提示文本 LLM 按照指令 $\boldsymbol{T}_{id}'$ 和 $\boldsymbol{T}_{it}'$ 分别提取身份和意图预测。通过使用指令和先前的响应 $\boldsymbol{R}_{u,v}$ 提示文本 LLM，我们可以分别提取个性化上下文 $\boldsymbol{C}_{u,v}$ 、驾驶员的身份指标 $\mathbf{1}_{u,v}^{ID}$ 和意图 $\boldsymbol{IT}_u, \boldsymbol{IT}_v$ 。然后，根据身份和意图的一致性，个性化上下文被过滤并聚合到 P 中，其中身份是检查生成的身份匹配结果 $\mathbf{1}_{u,v}^{ID}$ 是否与真实身份 $\text{ID}(u, v)$ 相同。由于比较提示比较的是具有相同意图的两个驾驶员，我们验证生成的意图 $\boldsymbol{IT}_u, \boldsymbol{IT}_v$ 是否也相同。

B. 指南指令构建和提示

MLLM 可以通过比较提示生成司机行为的详细对比特征。然而，在实际操作中，我们观察到视觉细节描述中存在幻觉。由于在视觉理解中的颗粒度不足，MLLM 可能会提供在实际视觉上下文中不存在的视觉细节 [?], [?], [?] 。为了使提示在描述某种类型的司机行为和突出的视觉特征时更具体，我们建议将收集的个性化上下文 D 作为提示的指导指令进行汇总。根据每种个性化行为的指导指令，生成的响应包含较少不相关的信息和幻觉 [?] 。

构建的指南指令是个性化驾驶员行为的类型，并配有这些行为类型的详细方面（例如，图 1 的步骤 4 中）。在指南指令提示期间，MLLM 将针对每种行为类型及其描述单独进行提示，以防止无关信息和幻觉。此外，我们还提示 MLLM 根据外部行车记录仪帧的视觉证据来解释驾驶员的行为描述（例如，图 1 的步骤 4 中）。为了使 MLLM 能够基于视频帧感知内部和外部的视觉证据，我们采用类似的方法来构建视觉证据。给定驾驶员的内部 $\boldsymbol{I}_u^{in} \in V_i$ 和外部 $\boldsymbol{I}_u^{ex} \in V_e$ 行车记录仪视频帧，

$$\boldsymbol{I}_u^{du} = \text{vconcat}\left[\text{hconcat}(\boldsymbol{I}_u^{in}), \text{hconcat}(\boldsymbol{I}_u^{ex})\right].$$

对于每一种行为类型 $k \in K$ 及其描述 $S(k)$ ，提示 $T_g(k, S(k))$ 是用于提示的指南指令。通过指南指令，我们可以提示 MLLM 获取描述 $\boldsymbol{D}_u$ 和解释 $\boldsymbol{E}_u$ ，

$$\boldsymbol{D}_u, \boldsymbol{E}_u = \text{MLLM}\left(\boldsymbol{I}_u^{du}, T_g(k, S(k))\right).$$

在图 2 中，我们展示了比较提示和指导指令提示下描述之间差异的一个示例。在比较提示中，我们可以观察到仅提到了一些粗略的方面（例如，座位位置、面部表情等），而没有进一步的论证。在这样的描述中，我们可以观察到可能由多模态大语言模型（MLLMs）生成的幻觉（例如，第一个司机比另一个司机更向前倾）。在指导指令提示的回应中，我们可以看到关于司机行为随时间变化的更细致的描述，而这些描述在某一特定方面上是具体的，没有不相关的信息。MLLM 还可以尝试根据来自外部行车记录仪帧的相应视觉证据解释生成的描述。

C. 人工标注者筛选

样本过滤任务之一是检测无关或不准确的行为类型，即方面级样本过滤。我们观察到一些提取的行为（例如，脚的位置、腿部运动）由于内部仪表盘摄像头的视野有限而无法被

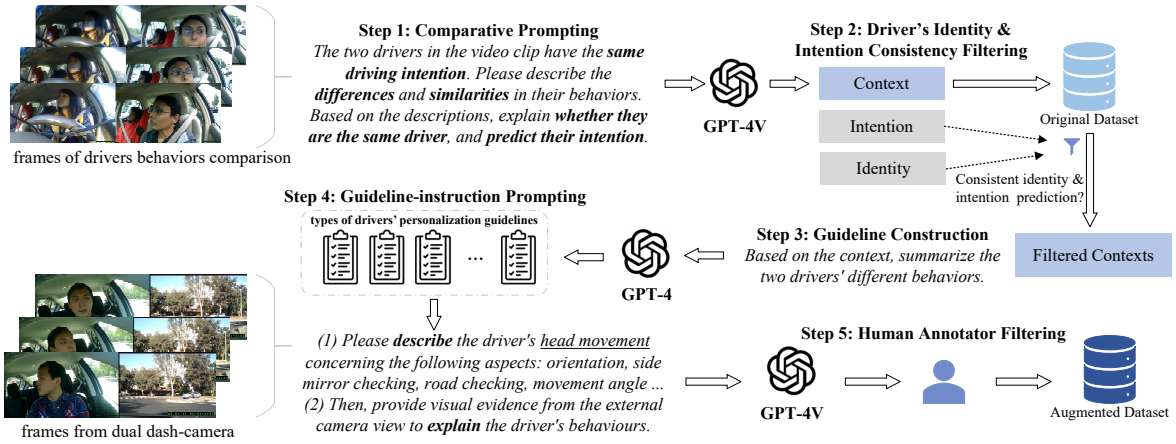

Fig. 1. 用于创建捕捉司机个性化行为特征的数据集的流程图示，包括五个关键步骤：比较提示、一致性样本过滤、总结生成指南、细粒度行为生成、以及人工注释过滤以确保质量

Comparative Prompting:
Both individuals are seated in the driver's seat of a vehicle and focused on an activity with their hands. The individual on the top row appears to be leaning forward slightly more than the individual on the bottom row. The individual on the top row has a more visible and expressive face, whereas the individual on the bottom row has a more neutral expression ...

Guideline Instruction Prompting (Head Movement):
Description *D*: Initially, the driver appears to be looking straight ahead. As the frames progress, his gaze shifts to the left, indicating the checking of the side mirror or a blind spot …
Explanation *E*: The traffic light is visible in the sequence, and the driver's initial straight-ahead gaze may correspond to watching the traffic light for changes. The sequence of images seems to capture the driver as he approaches and then waits at an intersection, possibly preparing for a left turn, as indicated by his head turning to check for oncoming traffic or pedestrians.

Fig. 2. 比较提示描述和指南指令提示描述的一个比较示例。上下文（带下划线且为红色）表明通过比较提示可能产生的幻觉。通过细粒度的指南指令，可以从描述 D 中提示 LMMs 提供更详细的视觉证据和时间信息（带下划线且为蓝色）。然后，LMMs 可以从外部视角捕获视觉证据，并在解释 E 中解释驾驶员的行为（带下划线且为绿色）。

准确描述。在这种情况下，MLLM 会产生不准确的视觉描述，并为这些描述提供错误的解释。此外，其他一些提取的行为（例如，驾驶员的外貌、服装等）与驾驶任务的表现无关。在这种情况下，解释不是基于视觉证据，而是基于 MLLM 的强文本先验（即语言偏见）[?] 。

经过方面级筛选，我们可以在表 ?? 中获得九种类型的行为。人工注释者的样本级筛选任务主要是从 MLLMs 中筛选出失败案例。由于这些生成的失败案例表达方式类似（例如，“抱歉，我无法提供……”），将这样的样本包含在模型微调中可能导致过拟合问题。因此，我们进一步对这些特定的失败案例进行样本级检查。

基于我们在第 III-A 节中的数据创建流程，我们从现有的 Brain4Cars 数据集 [?] 中开发了我们的评估数据集。在图 ?? 中，我们展示了在我们构建的数据集中针对这三项任务的一个示例。在比较提示中，我们从每个标注的驾驶意图子集中采样了 20 对驾驶员：右转、左转、右变道、左变道和直行。通过比较提示，收集的比较描述被总结为表 ?? 中的九种类型，每种类型平均有 19.11 条指导指令。我们在表 ?? 中总结了 PDB-Eval 的统计数据，其中报告了片段数量、描述-解释对（D/E）和问答对（QA）的数量。我们还报告了描述（Desc.）、解释（Expl.）、问题（Q）和答案（A）中平均的单词数量。

在 PDB-X 数据集中，我们观察到描述的平均长度比解释的平均长度要长，这是因为使用了指导说明。基于指导中的每个要求，MLLM 将为描述每条说明平均生成 6.54 个词。至于 PDB-QA 数据集，从 PDB-X 数据集中导出的平均问题数量是每个 D/E 对 7.05 个 QA 对，并且真实答案的平均长度大约是问题长度的 3 倍。基于 PDB-X 中的描述和解释，我们可以使用仅文本的 LLM 进行自动问答对生成 [?], [?] 。

我们评估了各种 MLLM：BLIP-2 [?] 、VTimeLLM [?] 和 GPT-4V，在 PDB-X 和 PDB-QA 上。通过在 PDB-X 上进行微调（BLIP-2 和 VTimeLLM），我们进一步在下游驾驶基准测试 Brain4Cars [?] （同领域）和 AIDE [?] （跨领域）上

评估 MLLM，以了解生成的驾驶员个性化行为描述和解释的有效性。

我们采用不同的视频预处理方法为三个基准提取输入流的兼容视觉表示。对于 BLIP-2，我们从双摄像头中以相等的时间间隔提取 10 帧图像并按等式 (III-B) 将它们连接起来，然后使用来自 BLIP-2 的 CLIP 编码器对图像进行编码 [**?**] 。对于 VTimeLLM，我们遵循 [**?**] 中的预处理方法，从视频的每一秒中提取 10 帧图像，并使用基于 ViT 的 CLIP 模型对视频进行编码。对于 GPT-4V，我们从视频中提取 10 帧连续图像，并将内部和外部视频的每帧图像连接在一起。

D. 描述和解释（PDB-X）

我们评估了在 PDB-X 中的描述和解释任务的 MLLM 基线，并在表 I 中报告了比较结果的 BLEU-4 指标。对于开源的 MLLM（BLIP-2 和 VTimeLLM），我们首先在 PDB-X 的训练集上微调这些模型，然后再进行评估。在推理过程中，我们采用与微调阶段相同的指令设计和视觉编码。比较微调后的 MLLM 和 GPT-4V 的表现，我们可以观察到在 PDB-X 上微调的一致优势。这种观察表明，使用简单指令提示 MLLM 无法直接实现精细化描述和解释，这展示了所提议的比较提示方法的有效性以及 PDB-X 任务所带来的挑战。

TABLE I
在所有类型的驾驶行为（在表格 **??** 中）及两种任务，即描述（Desc.）和解释（Expl.）上的 BLEU-4 指标性能。我们用粗体表示描述和解释任务的最佳性能。

	BLIP-2		VTimeLLM		GPT-4V	
Type	Desc.	Expl.	Desc.	Expl.	Desc.	Expl.
ACT	51.10	48.09	47.18	50.41	29.49	28.57
BOL	34.51	52.74	49.13	62.03	33.68	27.22
DRS	72.14	58.50	86.03	54.63	32.33	30.12
FAE	54.52	59.90	72.17	50.40	34.44	28.54
GEM	55.06	57.83	62.72	60.70	39.22	35.73
HAM	46.42	70.50	42.93	44.98	40.96	36.54
HEM	72.20	51.94	72.90	46.71	39.87	36.77
INT	56.52	60.97	53.80	49.22	25.69	26.67
IWP	76.39	66.53	73.80	65.95	37.23	32.04
AVE	57.65	58.55	62.30	53.89	34.77	31.35

通过比较 BLIP-2 和 VTimeLLM 的性能，我们可以观察到这两个多模态语言模型分别在解释和描述任务中表现优异。我们认为 BLIP-2 以增强的视觉知识推理能力 [**?**] 进行预训练，这有利于视觉解释，而模型输入的静态模态限制了其描述能力。另一方面，VTimeLLM 特别是在具有时间感知的描述能力上进行预训练，以便对细粒度的视频片段进行理解 [**?**] ，这解释了它在描述任务上的更好性能。然而，创造一个在这两种任务中都能实现稳健性能的多模态语言模型仍然是一个挑战。

E. 视觉解释问答 (PDB-QA)

我们进一步在 PDB-QA 上评估 MLLMs，以展示其在回答与个性化驾驶行为相关的复杂问题时的能力。在 PDB-QA 上相对较低的性能表明，MLLMs 在理解驾驶员的个性化信息方面存在缺陷。然而，通过模型微调，我们仍然可以观察到两个开源 MLLMs 的一致改进。

比较 BLIP-2 和 VTimeLLM 的性能，我们可以观察到与其零样本性能相比，VTimeLLM 在微调后的改进更大，这表明 VTimeLLM 的性能上限高于 BLIP-2。VTimeLLM 中的时间推理能力可以增强在特定时间边界 [**?**] 内的视频片段理解。然而，如何在不同时间边界内的视频片段中实现推理的交叉验证，以便更好地实现个性化理解仍然具有挑战性，这需要时间理解和因果推理能力。

TABLE II
BLEU-4 指标在 PDB-QA 数据集上的表现，包括预训练（PT）、微调（FT）和零样本（ZS）。

	BLIP-2		VTimeLLM		GPT-4V
Metric	PT ↑	FT ↑	PT ↑	FT ↑	ZS ↑
BLEU-4	30.29	33.64	28.07	48.61	16.08

在本节中，我们评估视觉描述和解释在几项驾驶任务中的有效性，以微调的 MLLMs 输出作为文本证据，相应的视频特征作为视觉证据。为了评估的一致性，我们对每个 MLLM 使用第 **??** 节中描述的相同视频特征，以及 MLLMs 在 PDB-X 上微调的文本证据。

我们首先在 Brain4Cars [**?**] 数据集上评估意图预测任务，作为域内评估。为了展示 MLLMs 的泛化能力，我们不结合 Brain4Cars 使用的低级信号（例如，车辆速度和 GPS 信息）。

在表格 III 中，我们观察到司机的个性化行为可以成为预测司机转向动作的良好互补文本证据，因为这些动作通常显示出司机更明显的行为特征。另一方面，Brain4Cars 中的原始方法通过融合各种感官信息来源，在车道变更预测中仍优于仅使用视觉语言信息的 MLLMs。我们认为，在需要额外感官信息的任务中，我们的方法仍然可以将这些信息纳入到一般推理过程中。

TABLE III
在意图预测任务 [**?**] 上的精准度和召回率指标表现。BLIP-2 和 VTimeLLM 列是由在 PDB-X 上微调的 LMMs 的文本证据输出获得的。S-RNN 列是从原始论文中的方法获得的 [**?**] 。

	BLIP-2		VTimeLLM		S-RNN	
	Prec.	Recall	Prec.	Recall	Prec.	Recall
Turn	84.57	76.43	83.85	77.18	75.20	75.30
Change	76.47	72.73	76.04	72.11	85.40	86.00

*1) 在 AIDE [**?**] 上的评估结果。:* 我们进一步评估了最初在 PDB-X 上微调而没有在 AIDE 上进行额外微调的 MLLMs 作为跨域评估。AIDE 提出的任务包括驾驶员行为识别（DBR）、驾驶员情绪识别（DER）、交通环境识别（TCR）和车辆状况识别（VCR）。除我们方法中使用的双摄像头视频外，AIDE [**?**] 还额外包含来自左右摄像头的视频，以及驾驶员面部、身体、手势和姿势的多模态注释。

TABLE IV
在 AIDE 数据集 [**?**] 上 Acc 和加权 F1 指标的表现。BLIP-2 和 VTimeLLM 列由最初在 PDB-X 上微调然后用于 AIDE 而未进行额外微调的 LMMs 获得。

	BLIP-2		VTimeLLM		AIDE	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
DER	72.74	72.73	74.06	73.41	71.26	68.71
DBR	64.86	64.41	65.52	65.34	65.35	63.29
TCR	90.15	89.86	90.15	90.19	83.74	81.28
VCR	76.35	76.00	78.82	77.67	77.12	75.23

在表 IV 中，我们可以观察到经过微调的多模态大模型（MLLMs）在跨领域驾驶任务中具有稳健的泛化能力，其中微调后的 VTimeLLM 在所有任务中均实现了比 AIDE 更好的准确性和加权 F1。在 AIDE 的四个驾驶任务中，我们观察到驾驶员情绪识别（DER）和交通状况识别（TCR）有更好的提升，分别得益于驾驶员的个性化行为描述和外部视觉解释。

我们引入了驾驶员个性化行为评估数据集（PDB-Eval），该数据集利用车内和外部视角进行个性化驾驶行为分析。我们的基准由两个组成部分组成——PDB-X 和 PDB-QA——通过在多模态大语言模型（MLLMs）中利用视觉比较提示来捕捉行为差异并提供描述性解释。对各种 MLLMs 的微调使这些任务的性能相较于 GPT-4V 的零样本结果提升了最高达 73.2%。然而，相对较低的 BLEU-4 分数表明在细粒度、时间感知的多模态推理方面仍存在挑战。未来的工作应侧重于使 MLLMs 具备识别具有时间感知的局部视觉证据的能力。