定义座舱情境：面向端到端智能座舱的感知框架

行业白皮书 · 2026 年 11 月 1 日

作者：Xu Dongchao（xudongchao@donfrontera.com）

摘要

智能座舱在语言上已经实现了端到端，在感知上却没有。语音识别、理解与生成如今已在量产车上由单一模型完成，但座舱对它所服务的人的了解——谁坐在哪里、每个人处于什么状态、彼此是什么关系、车辆和周围道路正在发生什么——仍然由彼此独立的功能各自拼凑，或者根本没有被整合起来。

本文认为，这一缺口解释了当今座舱的大部分痛点，而弥合它是实现真正端到端座舱的前提。本文聚焦感知一端，提出四点贡献。(1) **诊断**：我们把当前痛点追溯到缺失的情境，并借鉴自动驾驶从模块化架构、到端到端模型、再到视觉-语言-动作（VLA）模型的演进，指出自动驾驶与座舱正从相反的方向走向同一类架构，而座舱所缺的，恰恰是自动驾驶最先建立的统一感知。(2) **模型**：座舱情境模型围绕乘员，以三层十二维组织情境；在其中，角色由座位决定而非由人决定，乘员之间的关系决定谁可以听到什么，偏好按"人 × 角色 × 情境"分别记录。我们从情境理论和座舱中的实体两个方向论证，它覆盖了服务乘员所需的内容，并以用例覆盖检查加以说明。(3) **方法**：每个维度通过直接观测、即时推断或时间累积推断获得；每条情境信息都带有校准过的不确定性——分类状态用概率表示，连续量用区间表示——并附有证据和有效期。(4) **原则**：传感器按第一性原理选择，即每个量都以人感知它的方式去感知；我们据此给出一套参考传感器配置。最后，我们列出行业下一步需要开展的工作。

1. 引言

座舱 AI 在语言一侧进展迅速。基于端到端语音模型的车载助手已在 CES 2026 上实车展示（行业报道）。运行在新一代座舱芯片上的端侧全模态模型已宣布将于 2026 年车型上量产（行业报道）。参考设计描述了按意图和情境把请求分派给本地与云端智能体的车载智能体架构（技术博客）。

然而，驾驶员和乘客感受到的座舱，仍然要等人发出指令才行动，不知道谁在说话、为谁服务，还会在不该说的人面前说出私事。几乎每一次发布都会提到"情境感知"，但情境**究竟是什么**，始终没有被定义。

本文的核心论点是：**座舱在语言上已实现端到端，而它对所服务之人的感知仍然是模块化的、狭窄的。** 自动驾驶先建立了统一感知，直到最近才在视觉-语言-动作（VLA）模型中加入语言；座舱则从另一端起步：它已经有了语言，现在需要感知。两者正走向相似的架构，但座舱的感知问题性质不同：它必须理解多个人、他们的内在状态以及彼此之间的关系，而不仅仅是一个物理场景。

本文的四点贡献如下：

1. **诊断。** 把当前座舱的痛点追溯到缺失的情境，并以自动驾驶从模块化到端到端、再到 VLA 模型的演进作为架构类比，说明座舱缺少什么，以及座舱在感知与动作上与自动驾驶有何不同（第 2 至 4 节）。
2. **模型。** 定义围绕乘员组织的三层十二维座舱情境模型，并论证这十二个维度为何覆盖了座舱所需（第 5 节与附录 A）。
3. **方法。** 把每个维度的获取方式分为直接观测、即时推断和累积推断三类，并规定每条情境信息都带有校准过的不确定性：分类状态给出置信度百分比，连续量给出区间（第 6、7 节）。
4. **原则。** 提出按第一性原理选择传感器的方法，并据此给出参考传感器配置（第 6.1 节）。

本文是一个框架，而不是对每个维度的综述。对每种获取方式，我们只给出少量有证据支撑的示例；逐维度的感知与推断研究留待后续工作。

2. 当今座舱的痛点

下面七个痛点，失败的都是情境，而不是语言。座舱听懂了说的是什么，却不够了解是谁在说、说给谁听、在什么情况下说。

痛点	典型场景	缺失的情境
只会听，不会看	一位乘客发抖、搓手，直到有人说"把暖气调高"之前，系统毫无反应。	乘员生理状态
分不清是谁	后排乘客要求播放音乐，系统却播放了车主的歌单。	说话人的身份与座位
不顾情境	在空旷的高速上和拥挤的地下停车场里说"有点闷"，得到的是同样的回应。	场景、舱内环境、行驶阶段
不顾驾驶任务	驾驶员在拥堵中并线时，系统播报了一段很长的语音建议。	车辆与驾驶状态
不顾听众	上司坐在后排时，共享扬声器播报了用户求职面试的提醒。	乘员之间的关系
记错了对象	驾驶时设置的偏好，在同一个人坐副驾时被套用，或被套用到车上所有人。	按人、角色和情境记录的履历
主动服务失灵	几次时机不当的推荐之后，用户关闭了主动服务。	以上全部，以及置信度

最后一行是前面几行的结果。主动服务是 AI 座舱最明确的承诺，而它的好坏取决于它所依据的情境。

表格背后还有一个结构性痛点：**传感器数据各自为政**。驾驶员监测、乘员监测、舱内雷达和自动驾驶系统各自服务于自己的安全功能，它们的输出很少进入座舱 AI，尽管描述的恰恰是 AI 需要理解的情况。

3. 自动驾驶的先例

自动驾驶从模块化架构走向端到端模型，最近又走向加入语言推理的视觉-语言-动作（VLA）模型。我们把这条路径作为架构类比，而不是座舱必须逐步照走的路线。它的启示是：每个阶段都建立在对场景的统一感知之上，而这正是座舱尚未迈出的一步。

3.1 三个阶段

模块化。 自动驾驶系统把感知（含传感器融合）、定位、预测、规划和控制分开，并用规则做规划。座舱中对应的是语音识别、语言理解、对话管理、技能和语音合成组成的流水线，以及各自独立运行的驾驶员与乘员监测功能。

端到端。 自动驾驶首先统一了感知：BEVFormer 把多摄像头特征提升到同一个鸟瞰图空间并做时序融合 [2]，UniAD 等面向规划的模型在这一共享表征上联合训练感知、预测和规划 [1]。之后的模型用单一网络把传感器输入映射为轨迹。座舱中对应的是：多传感器输入汇入一个共享的**情境**，决策和动作在此之上学习。

视觉-语言-动作。 当前前沿是在感知和动作之上加入语言推理。DriveVLM 借助视觉-语言模型，经由场景描述、场景分析和分层规划进行推理，其 Dual 版本把这一推理与传统的感知和规划流水线结合起来 [3]。EMMA 把摄像头输入、以文本表达的导航和车辆状态，在一个多模态语言模型中映射为轨迹及其他驾驶输出 [4]。座舱中对应的是这样一个模型：感知乘员、用语言推理，并在语音、设备和服务之间执行动作。

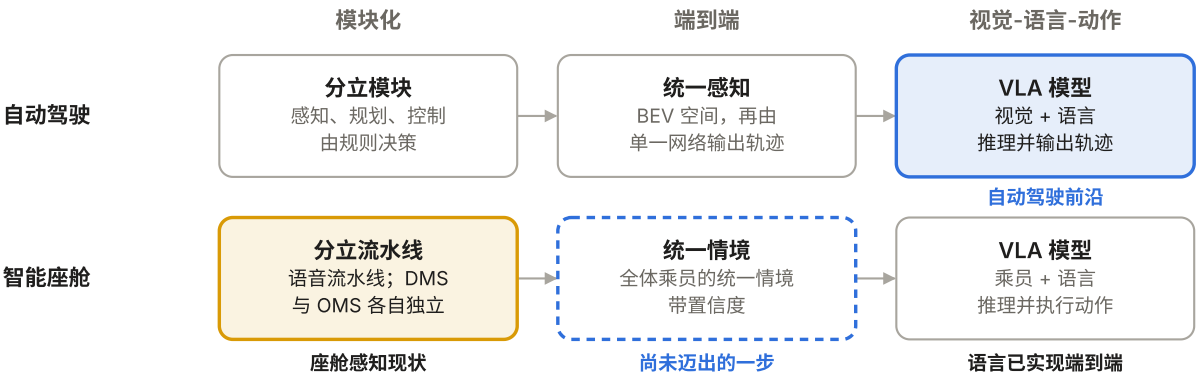


图 1. 自动驾驶与座舱从相反方向走向视觉-语言-动作模型

3.2 从相反的方向汇合

自动驾驶和座舱正从相反的两端走向同一类架构。自动驾驶从感知和动作起步，现在加入了语言；座舱从语言起步，语言在量产中已经端到端，现在需要补上感知和动作。座舱所缺的，正是自动驾驶最先建立的东西：对场景的统一感知，在座舱里就是对乘员的统一情境。

自动驾驶中的 VLA 系统也提示了这种情境应有的形态。它们保留了显式的中间推理，例如对场景的描述，有的还把慢速的推理模型与快速的传统流水线配合使用 [3]。座舱模型同样可以把显式、结构化、标明置信度的情境作为中间表示，即使感知、推理和动作最终会被联合训练。

3.3 座舱的不同之处

方面	自动驾驶	智能座舱
服务对象	一个驾驶任务	需求可能相互冲突的多个乘员
感知对象	车外物理世界：几何、物体以及其他道路使用者的意图，大体有客观真值	人的内在状态、关系和履历，往往没有客观真值
时间尺度	控制在毫秒到秒级，地图和车队数据在更长时间内积累	秒到数月
动作空间	一条轨迹，通过转向、油门和制动执行	语音（内容、语气、对哪个音区）、舱内设备（分区空调、座椅、车窗、灯光、屏幕）、生态内容与服务（音乐、视频、导航、支付、第三方应用）、多步组合，以及选择不行动
成败标准	安全攸关，有可量化的安全与舒适指标	很大程度上是主观的

这并不是说座舱整体上更难：自动驾驶承担着安全攸关的风险，面对座舱所没有的大量长尾物理场景。座舱的复杂性属于另一种类型：需求可能冲突的多个乘员、没有客观真值的主观状态、在数月间逐渐形成的关系，以及宽得多的动作空间。本文讨论感知一端，动作空间的定义留待后续工作。

4. 座舱的现状

当前的端到端座舱能感知乘员说了什么，并越来越多地感知舱内摄像头看到了什么；其他情境大多是按需获取，或者没有被使用。由于公开资料很少准确说明输入覆盖范围，我们不评估具体产品，而是把 2025–2026 年的系统归为三类架构。

这三类并不互斥，一款产品可以同时具备三者。

- **A 类——端到端语音。** 用一个模型取代由语音识别、理解和合成组成的流水线。
- **B 类——端侧全模态。** 音频、舱内视觉和文本在座舱芯片上由同一个模型处理。
- **C 类——智能体编排。** 由语言模型规划并调用车辆、导航和云端智能体，按请求检索记忆。

情境维度	A 类	B 类	C 类
谁坐在哪里、谁在说话	说话人音区	部分（视觉）	通过工具调用
乘员状态	仅语音	部分（视觉 + 语音）	通过工具调用（驾驶员监测标志）
乘员之间的关系	未确认	未确认	未确认
车辆与驾驶状态	未确认	未确认	通过工具调用，按需获取
场景（停车场、越野、隧道）	未确认	未确认	通过工具调用（地图）
外部环境与行程	未确认	未确认	通过工具调用（导航、天气、日程）
按人、角色、情境记录的履历	未确认	有限	账号级记忆

评估基于公开的产品描述。"未确认"表示在公开资料中未找到该能力，并不表示已知其不存在；具体项目的覆盖情况应以其规格说明为准。

在公开资料的范围内，可以得出三点观察。情境看起来是**被调取的，而不是被感知的**：在 C 类系统的描述中，车辆状态是在模型调用时才到达模型，而不是持续输入。**我们找到的描述中，没有一个对乘员之间的关系建模**，因此第 2 节中"不顾听众"的痛点还看不到解决方案。而且，**在有所描述的系统中，记忆是挂在账号上的**，而不是挂在处于某个角色和情境中的人身上。

5. 座舱情境模型：三层十二维

我们围绕乘员，以三层十二维组织座舱情境。**乘员层**为舱内每个人各保存一份情境，**关系层**保存人与人之间的内容，**共享层**保存全舱共有的内容。时间不是单独的一个维度，而是贯穿全部十二个维度的一条轴。在时刻 **t**：

$$C_t = \left(\{c_{i,t}\}_{i \in \text{乘员}},\ R_t,\ S_t \right)$$

层	编号	维度	包含内容
乘员层	D1	身份	注册用户、已识别的常客、陌生访客、儿童；年龄段
乘员层	D2	座位与角色	座位；角色 = 座位 × 驾驶状态；该座位可调用的资源
乘员层	D3	生理状态	冷热舒适、疲劳、晕车、身体不适、心率与呼吸
乘员层	D4	认知状态	注意力、负荷、关注对象
乘员层	D5	情绪状态	烦躁、放松、紧张及其他情绪
乘员层	D6	活动	睡觉、通话、工作、进食、照看他人

层	编号	维度	包含内容
乘员层	D7	履历与偏好	按"人 × 角色 × 情境"记录的偏好和习惯
关系层	D8	成员关系与互动	长期关系（家人、同事、上级、服务关系）以及当下的互动（谁在对谁说话）
共享层	D9	车辆与驾驶状态	车速、档位、车门、电量；自动化等级、接管请求；驾驶负荷
共享层	D10	场景与行驶阶段	场所类型（高速、越野、地下停车场、隧道）；阶段（上车、巡航、泊车、到达）
共享层	D11	物理环境	舱内各温区的温度、湿度、空气质量、噪声、光照；车外天气、路面、交通
共享层	D12	行程与任务	目的地、出行目的、日程、进行中的任务；家、公司等地点语义

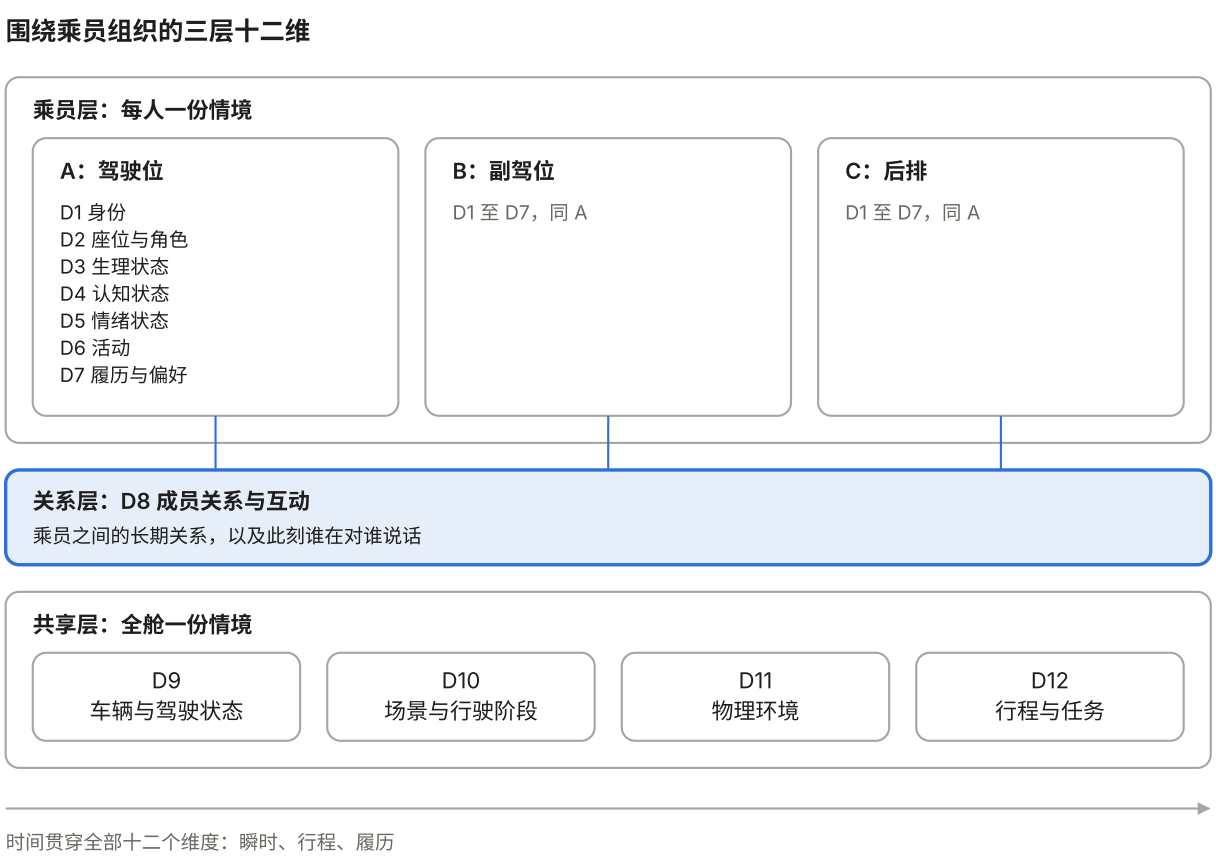


图 2. 座舱情境模型：三层十二维

每位乘员都有同样的七个维度；关系层把他们联系起来，共享层为他们提供共同的框架。

5.1 乘员层

身份是可选的。 座舱必须能服务它无法识别的人，因此 D1 是一个属性，而不是前提。怕冷、易晕车等稳定特征附着在人身上。

角色由座位决定，而不是由人决定。 今天开车的人，明天可能坐在后排。角色还取决于车辆状态，因此我们定义 **角色 = 座位 × 驾驶状态**。

座位	人工驾驶	自动驾驶，需要人类监督或接管	驻车
驾驶位	驾驶员	监督者或后备驾驶员，取决于自动化等级	乘员
副驾位	乘客	乘客	乘员

座位	人工驾驶	自动驾驶，需要人类监督或接管	驻车
后排座位	乘客	乘客	乘员
儿童座椅	儿童乘客	儿童乘客	若独自留在车内，即为遗留儿童

在不需要人类后备的车辆中，驾驶位就成了乘客座位，坐在上面的人就是乘客。座位还决定了能为这个人服务的资源：一个温区、座椅加热和通风、一块屏幕和一个音区。

状态按类型区分。 D3 至 D6 区分生理、认知、情绪和活动状态，因为它们的感知方式和推断的可靠程度都不同。情绪状态受到产业界和学术界的大量投入，但它在真实座舱中的可靠性仍是四者中最不成熟的，因此单列为一个维度，单独跟踪和评估。

履历按人、角色和情境记录。 同一个人开车时可能希望座舱凉一些以保持清醒，坐在副驾时又希望暖一些好睡觉；只按人记录偏好，常常会出错。

5.2 关系层

关系并不描述某一位乘员，但它强烈约束着在谁面前可以说什么、显示什么。手机的输出通常只到达一个人，而座舱的扬声器和中控屏会到达舱内所有人。因此，每一次输出都有**受众**，信息能否流向这些受众，取决于他们之间的关系。这就是情境完整性（contextual integrity）在座舱中的形式：一次信息流动是否恰当，取决于它所处的社会情境 [5]。Schilit、Adams 和 Want 早已把社会情境视为情境的一部分，例如一个人是和上级在一起，还是和同事在一起 [11]。

关系	例子	决定什么
家庭	伴侣、父母与子女、年长亲属	记忆和偏好能否共享
监护	父母与子女、照护者与老人	谁的指令优先、内容限制
社交	朋友、同事	日程或消息能否出现在共享输出上
职业	上级、客户	正式程度；不透露个人信息
服务	网约车司机与乘客、专职司机与雇主	角色倒置：主要服务对象不在驾驶位

D8 也包括瞬时时间尺度上的互动：此刻谁在对谁说话。

5.3 共享层

共享层区分**含义**与**条件**。D10 说明车辆处在什么样的地方、行程处于什么阶段；D11 说明物理条件如何。地下停车场（D10）和 17 °C 的舱内温度（D11）是不同类型的事实，获取方式也不同。行驶阶段对理解乘员层很重要：驻车时闭眼和高速行驶时闭眼，是完全不同的情况。

D9 包括驾驶负荷，它由车辆状态、场景和驾驶员的注意力推导而来。D12 包括家、公司等地点语义，它们是在多次行程中学到的，而不是在一次行程中感知到的。共享层中的条目仍然可以以某一位乘员为主体：目的地是共享的，但一条个人日程属于某一个人，并带有个人隐私等级。

5.4 时间是一条轴

每个维度都在三个时间尺度中的一个或多个上被观测：**瞬时**（秒级：一个手势、一句话、一次发抖）、**行程**（谁在车上、谁在二十分钟前睡着了）和**履历**（数周到数月）。D3 和 D6 主要在瞬时尺度上，D8 和 D12 在履历中积累，D7 则保存每位乘员积累下来的偏好和习惯。

5.5 为什么是这十二个维度

任何情境模型都无法被证明是完备的，因为情境本身是开放的。我们不主张完备性，也不主张严格互斥。我们从两个方向论证，这十二个维度覆盖了服务座舱乘员所需的内容；附录 A 用一组示例用例的覆盖检查加以说明。

它们覆盖了情境的经典要素。 Dey 和 Abowd 把位置、身份、活动和时间列为情境的基本类型 [12]。Schilit、Adams 和 Want 围绕"你在哪里、你和谁在一起、附近有什么资源"来界定情境感知计算 [11]。这些要素都能映射到本模型：身份对应 D1；位置对应 D2（座位）和 D10（场所）；活动对应 D6；时间对应贯穿所有维度的时间轴；和谁在一起对应 D8；附近的资源对应 D2 中的座位资源和 D11 中的环境条件。本模型多出的部分——内在状态（D3 至 D5）、履历（D7）和车辆状态（D9）——正是座舱特有的内容。

它们可以从座舱中的实体推导出来。 座舱中的实体是每位乘员、乘员之间、车辆、车外世界和这次行程。追问每个实体能被知道什么，就得到了这个模型：

- **一位乘员：** 是谁（D1）；坐在哪里、担任什么角色（D2）；身体（D3）、心智（D4）和感受（D5）如何；正在做什么（D6）；偏好什么、习惯做什么（D7）。D3 至 D6 的划分沿用了人因工程中常见的生理、认知、情绪和行为状态的区分。
- **乘员之间：** 是什么关系、此刻如何互动（D8）。
- **车辆：** 正在做什么（D9）。
- **车外世界：** 是什么样的地方（D10），物理条件如何（D11）。
- **行程：** 去哪里、为什么去、要做什么（D12）。

三条边界规则让各维度尽量分开。D10 保存含义（是什么地方、处于哪个阶段），D11 保存物理条件。D7 只保存一个人偏好什么、习惯做什么；关于其他维度、随时间学到的事实，例如一段关系（D8）或家的位置（D12），留在各自的维度中，以履历作为其时间尺度。此外，每个条目都指明其主体，因此共享维度中的个人条目仍然属于某一位乘员。维度的边缘仍有一些重叠，这些边界应随各维度的深入研究而细化。十二个维度构成稳定的顶层结构；每个维度都可以细分出子维度，而不改变整体结构。

---

6. 情境如何获取

每个情境维度都通过三种方式之一获得，获取方式决定了它的置信度从何而来、如何随时间变化。我们假设采用第 6.1 节的参考传感器配置，因此讨论的问题是"能够知道什么"，而不是"某辆车装了什么"。

方式	是什么	置信度来自	随时间
直接观测	从单一来源读取一个外显属性或系统状态	信号有效性；对于感知得到的属性，还有识别模型经校准的单条概率	未刷新时变为过期、取值未知
即时推断	融合多个当前信号，推断一个不可直接看到的状态	有多少独立信号相互印证	信号相互印证时上升，相互矛盾时下降
累积推断	从多次行程的规律中推断一个事实	积累了多少一致的证据	随证据增长；遇到反例或长期无新证据时衰减

前两种方式的分界并不在于是否用到模型：在摄像头图像中识别出成人，同样要用到模型。直接观测涵盖外显的、或由系统报告的属性，例如谁坐在哪里、车速多少，它同样带有识别误差。即时推断涵盖无法直接看到、必须由多个信号推理得出的状态，例如"感到冷"。

下面的示例分别说明每种方式。它们不是对每个维度的综述；所引用的证据说明各类推断是可行的，而不说明它们在量产座舱中的表现如何。大多数数据来自受控研究，有各自的类别定义和实验条件；专利展示的是被提出的方案，而不是经过验证的结果。

6.1 参考传感器配置

传感器按第一性原理选择：每个量都以人感知它的方式去感知。我们用眼睛看谁坐在哪里、是不是孩子；看到有人发抖，感受到脸上的温度；用耳朵听是谁在说话。当某个量在物理上无法被主传感器获取时，增加传感器是合理的——心跳看不见，盖着毯子的孩子被遮挡。当安全功能需要独立的第二信息源，或者有一种侵入性更低的传感器可以胜任时，增加传感器同样合理，例如雷达可以在不拍摄图像的情况下检测存在和呼吸。不能作为增加传感器理由的，是当前算法恰好在它上面得分更高：算法会进步，而为迁就算法当前弱点而固定下来的传感器架构不会。这是一条针对下面参考配置的设计原则；成本、布置和供应商约束因项目而异，不在其讨论范围内。

传感器	安装位置	感知内容	服务的维度
驾驶员摄像头（近红外）	转向管柱或 A 柱	驾驶员的面部、眼睛、头部姿态、视线	D1、D3、D4、D5、D6
舱内摄像头（RGB-IR，广角）	前排顶置控制台；每排后座上方车顶	每个座位：占用、体型、姿态、面部、手势、衣着、发抖、表情、手持物品、谁面向谁	D1、D2、D3、D4、D5、D6、D8
热成像摄像头（长波红外）	顶置，覆盖所有座位	面部和手部皮肤温度、出汗	D3
毫米波雷达（60 GHz）	车顶中央	呼吸和心率；视线被遮挡的乘员是否存在	D3、D2
麦克风阵列	每个座位一个音区	语音、说话人音区、语调、对话轮次	D1、D2、D5、D6、D8
舱内环境传感器	每个温区	温度、湿度、CO_2、颗粒物、光照	D11
安全带与座椅传感器	每个座位	安全带是否扣上、座椅调节	D2
人机交互事件	屏幕、旋钮、方向盘按键	手动调节及其撤销	D6、D7
配对的可穿戴设备（可选）	乘员身上，经本人同意	皮肤电导、心率变异性	D3、D4
车辆总线	—	车速、档位、车门、车窗、空调、电量、车身运动	D9
自动驾驶系统	其车外摄像头、雷达、激光雷达及融合单元	目标物、道路与车道、场景及运行设计域判断、自动化等级、接管请求	D9、D10、D11
定位、地图与云服务	GNSS 与 IMU；地图；天气与日程服务	位置、场所类型、天气、路线、日程	D10、D11、D12

6.2 直接观测

谁坐在哪里，是成人还是儿童。 人靠看来分辨成人和儿童，舱内摄像头也一样：看体型和身体比例、面部，以及有没有儿童座椅。基于视觉的乘员分类由来已久。早在 2003 年，单个灰度摄像头就能以约 95% 的准确率区分后向婴儿座椅、儿童、成人和空座位（CVPR 2003）；之后的一个系统在超过 84,000 帧图像上，以 98.9% 的准确率区分成人与前向、后向儿童座椅（BMVC 2004）。非视觉传感器恰恰难以做出这种区分：坐在增高座椅上的儿童，可能被顶置超声波传感器误判为成人，而电感式传感器无法区分乘员类型（Cost-based Feature Transfer）。雷达在视觉看不到的地方补充摄像头，例如盖着毯子的孩子，也在需要为安全或隐私提供第二个非成像信息源时发挥作用。

说话人在哪个座位。 多音区麦克风系统通过融合分布式麦克风之间的时间差和信号能量线索，识别当前说话人所在的音区（US 10332545）。

车辆与自动化状态。 车速、档位、车门、空调和电量来自车辆总线；自动化等级和接管请求来自自动驾驶系统。自动驾驶的感知结果在传感器与融合单元之间已有标准化的目标级语义描述（ISO 23150-1:2026），因此座舱可以同时继承目标物及其存在概率。

6.3 即时推断

冷热舒适：为什么一个信号不够。 面部热成像可以估计舒适度：一项在建筑环境中使用低成本热像仪的研究显示，个人舒适度模型的准确率约为 85%，耳朵、鼻子和脸颊的温度最具指示性（NSF 资助研究）。但只看面部温度可能会误判：无论热还是冷，舒适度都可能变化很大，而面部皮肤温度几乎不变；在寒冷中，人体通过发抖来调节体温（US 11458803）。在热的一侧，出汗可以通过热成像检测（Scitepress 2020）。因此，要可信地判断"这位乘客觉得冷"，需要结合面部温度、可见的发抖或搓手、衣着、温区温度和车外温度。

地下停车场：弱信号，强结论。 仅凭 GNSS 信号丢失，也可能是在隧道里。再加上自动驾驶摄像头看到的顶棚、立柱和车位线，低速大转角，泊车辅助已激活，以及环境光骤降，结论就几乎可以确定。越野行驶也以同样的方式推断：没有车道线、非铺装路面、越野驾驶模式或差速锁已锁止、车身俯仰和侧倾大、车轮打滑。自动驾驶本来就维护着这类场景判断，用来确认自己是否处于运行设计域内；只是如今它很少被传给座舱。

后排晕车。 在长途行程中，约三分之一的乘客会出现头晕或晕车，幼儿尤其容易出现（行业研究）。一项产业研究项目发现，体温和皮肤电导是与主观晕车感受相关性最高的指标之一（eeNews Europe）。在参考配置中，温度来自热成像摄像头；皮肤电导需要接触式感知，例如配对的可穿戴设备。已有方案把生理信号与车辆运动数据和乘客视线结合起来（US 11820402），并以座位位置作为先验：看屏幕的后排乘客风险最高。

即便是最成熟的推断也是概率性的。对于疲劳检测，Euro NCAP 2026 年驾驶员监测文档模板中预填的指导值是每小时 0.5 次误报、40% 的真阳性率（Euro NCAP）。隐藏这种不确定性的情境条目，会误导使用它的任何系统。

6.4 累积推断

家与公司。 反复在早高峰到达某地，说明那里是工作地点；傍晚到达并停留过夜，说明那里是家。在来自 68 个国家、5,099 人的智能手机定位数据上，HoWDe 算法识别家和公司的准确率最高分别可达 97% 和 88%（de Sojo 等）。车辆行程是一种不同且更稀疏的数据源；在 467 名驾驶员的车辆行程日志上，一种更简单的方法识别家的失败率为 2.5%，识别公司的失败率为 6.2%（US 11248924）。同样的规律还能识别出家里的车库：同一个停车场，傍晚驶入，次日早晨驶出。

关系。 没有哪一次行程能表明某位乘客是驾驶员的上级，但规律可以：两个人多久同乘一次、在什么时段、坐在哪些座位、彼此如何称呼、谈论什么。来自两人电话通话的证据表明，根据对话内容区分商务与私人通话的准确率约为 88%，区分家人与非家人约为 82%（Stark 等）。我们没有找到关于车内多名乘员之间关系推断的研究。

情境中的偏好。 偏好按人、角色和情境积累：一个人冬天早上开车上班时的设置，是关于那种情境的证据，而不是关于每一次出行的。

累积推断遵循两条原则。第一，推断可以自由地让座舱变得**更谨慎**，但只有在用户确认之后，才能让座舱变得**更开放**：误以为某位乘客是同事，代价很小；误以为某人是配偶，则可能泄露隐私。第二，累积推断的结果应当对用户可见，并可删除。

7. 带置信度的情境

座舱感知的输出是一组情境条目，每个条目都带有校准过的不确定性：分类状态给出置信度百分比，连续量给出区间。这样，无论使用这些情境的是今天的决策模型，还是将来的单一端到端模型，都可以权衡一个不确定的估计，而不是把它当作事实。

7.1 情境条目

字段	含义	示例
主体	条目所描述的乘员、座位或全舱	左后乘员
维度	属于哪个情境维度	冷热舒适
取值	估计的取值	偏冷
置信度	对分类取值，是其正确的校准概率，以百分比表示；对连续值，是不确定性区间	78%
获取方式	直接观测、即时推断或累积推断	即时推断
证据	支持该条目的信号或历史事件	检测到发抖；温区 17 °C；车外 2 °C；未穿外套
观测时间	条目产生的时间	08:12:05
有效期至	未刷新时失效的时间	08:14:05
隐私等级	非个人、个人或敏感	敏感

7.2 百分比的含义

置信度百分比只有在**经过校准**时才有用：在所有标为 80% 的条目中，应当大约有 80% 是正确的。原始模型分数通常未经校准，因此每个维度都需要用标注数据做一次校准。百分比适用于"偏冷""睡着"这类分类状态；舱内温度或心率这类连续量，更适合表示为带不确定性区间的估计，例如每分钟 72 ± 4 次。

有两种置信度不应混淆。**方法成熟度**描述一个维度在总体上能被多可靠地估计；文献中报告的准确率大多来自实验室和小样本，是方法成熟度的证据，而不是单条结果的置信度。**单条置信度**描述的是：这一次、对这个人做出的这个估计，在多大程度上可信。只有单条置信度会写进情境条目。

7.3 不同获取方式下置信度的变化

- **直接观测**——由信号有效性决定；对于成人或儿童这类感知得到的属性，还由识别模型经校准的单条概率决定。信号在有效期内未刷新时，条目变为过期、取值未知，而不是被判定为假。
- **即时推断**——独立信号相互印证时上升，相互矛盾时下降。矛盾会记录在证据字段中，而不是被平均掉。
- **累积推断**——每出现一次一致的观测就增长，遇到反例或长期没有新证据时衰减。搬家之后，原来的"家"应当自动降低置信度。

7.4 示例

主体	维度	取值	置信度	获取方式	证据
右后座位	占用	有成人	99%	直接观测	舱内摄像头：座位上有成人体型、面部和坐姿；安全带已扣
副驾乘员	说话	正在说话	95%	直接观测	麦克风音区的时间差与能量
全舱	场景	地下停车场	96%	即时推断	GNSS 丢失；摄像头视野中有顶棚和车位线；8 km/h；泊车辅助已开启；环境光骤降
左后乘员	冷热舒适	偏冷	78%	即时推断	检测到发抖；温区 17 °C；车外 2 °C；面部温度未变
左后乘员	晕车	可能开始	55%	即时推断	面部皮肤温度下降；弯道路段；注视平板 20 分钟
副驾乘员与驾驶员	关系	同事	70%	累积推断	34 次工作日同乘通勤；称呼正式；谈论工作话题
驾驶员	地点语义	目的地是工作地点	93%	累积推断	60 天内 41 次在工作日 08:00 至 09:30 之间到达

数值仅为示意，用来展示预期的分布：直接观测接近确定，即时推断随相互印证的证据多少而变化，累积推断起初较低，并随行程逐次增强。

8. 隐私考量

一个感知每位乘员、他们的状态和关系的模型，只有把隐私内建于情境本身，才能被接受。情境条目的格式对此有帮助：原始敏感数据留在被感知的地方，只有最少量的、带标签的结论向外传递。

- **原始数据在端侧处理。** 视频帧、语音音频和生理信号留在车内；离开感知层的只有情境条目。这符合欧洲数据保护委员会对联网汽车优先在车内处理数据的立场 [6]，符合中国汽车数据相关规定中把车内处理和默认不收集列为指导原则的要求 [7]，也符合日本《个人信息保护法》对使用目的的限制 [8]。
- **保留结论，而非对话。** 关于关系的累积推断依据的是乘员说的话，但应当只保留推断出的关系、它的置信度和证据计数，绝不保留对话文本。
- **尊重没有同意的人。** 乘客并没有同意被分析。一位未被识别的乘员，仍然可以被感知为"在场"和"某个音区的说话人"，而不必被识别身份；乘客也应当可以选择不被记住。
- **让累积推断可见。** 推断出的地点和关系，应当能被其所涉及的用户查看和删除。
- **对情绪格外谨慎。** 根据欧盟《人工智能法》，情绪识别负有透明度义务 [9]，而且它在座舱中的准确性是所有维度中最不成熟的。

9. 相关工作

本文建立在四条研究脉络之上，与它们的不同之处在于：本文同时围绕多名乘员来组织。

情境的一般定义。 Dey 把情境定义为可用于刻画与交互相关的实体所处情况的任何信息 [10]。这一定义刻意保持一般性；我们把它具体化到一个共享的座舱中，其中的实体是坐在各个座位上的多个人以及车辆本身。

舱内监测与 AI。 关于座舱 AI 的综述按用例来组织已有工作（Artificial Intelligence Methods in In-Cabin Use Cases），Drive&Act 等数据集聚焦于对驾驶员活动的细粒度识别（Drive&Act）。驾驶员面部表情识别已有大量综述（Saadi 等）。这些工作大多以驾驶员为中心，并分别处理每一种状态；我们则把这些状态放入一个涵盖乘客、角色和关系的统一模型中。

自动驾驶的场景描述。 自动驾驶通过运行设计域描述其运行条件，并以目标级语义对感知输出进行标准化（ISO 23150）。我们把这些用作共享情境的来源，而不是重新定义它们。

情境中的隐私。 Nissenbaum 的情境完整性理论认为，信息流动是否恰当取决于它所处的社会情境 [5]。我们通过"输出的受众"这一概念，把它应用到座舱中。

---

10. 下一步与结论

定义座舱情境只是第一步，要把它付诸实践，需要整个行业的工作。

1. **统一的情境接口。** 自动驾驶已经标准化了传感器与融合单元之间的接口；座舱还没有对应的标准，用来把带有置信度和隐私等级的情境条目传给使用它们的系统。
2. **向座舱开放自动驾驶系统的场景层。** "地下停车场""越野"这类场景判断已经存在于自动驾驶系统内部，但从其融合输出到座舱的通道还没有标准化。
3. **关于乘客的研究。** 法规和消费者评测都聚焦于驾驶员；乘客的疲劳、晕车和舒适度还没有可比的基准。
4. **关于"谁在对谁说话"的研究。** 系统能够判断说话人在哪个座位。判断一句话是否是对系统说的，已是一个成熟的研究方向 [13]；但判断一位乘客是在对哪位乘员说话，以及推断多名乘员之间的关系，在很大程度上仍是空白。
5. **逐维度研究。** 针对每个维度，研究哪些传感器和方法能在真实车辆中达到什么样的置信度。我们计划在本文之后开展此类研究。
6. **动作空间。** 端到端座舱的另一端——语音、设备和生态服务——需要像本文对情境那样进行同样细致的定义。

自动驾驶只有在学会把道路看成一幅完整的图景之后，才实现了端到端。座舱已经学会了说话，却还没有学会以同样的方式看见它的乘员。一个以乘员为中心的情境模型——角色由座位决定，关系决定受众，每个条目都说明它是如何获得的、在多大程度上可信——正是座舱所需要的那幅图景。

附录 A. 用例覆盖

下列用例所需的每一项情境，都可以映射到所列维度之一，而且每个维度至少被一个用例用到。这是对示例的一致性检查，而不是实证验证；更完整的检查应把同样的方法应用到具体项目自己的用例清单上。

用例	所需情境	维度
一位乘客觉得冷	发抖与皮肤温度；哪个座位；温区与车外温度	D3、D2、D11
后排乘客要求播放音乐	谁在说话、在哪个座位；其偏好	D1、D2、D7
一项私人日程即将到时	车上有谁；他们与用户的关系；这项日程	D1、D8、D12
长途驾驶时建议休息	驾驶员疲劳；行程进度与下一个停靠点；车速	D3、D12、D9
孩子在后排睡着了	有儿童在车上；正在睡觉；哪个座位	D1、D6、D2
艰难并线时收到消息	驾驶负荷；场景与阶段	D9、D10

用例	所需情境	维度
指着说"把那扇窗关上"	手势；视线目标；说话人的座位	D6、D4、D2
后排乘客晕车	生理指标；在看平板；座位；弯道路段	D3、D6、D2、D10
回到家里的车库	地下停车场；地点语义；泊车	D10、D12、D9
驾驶员在拥堵中变得烦躁	情绪状态；拥堵阶段；车辆状态	D5、D10、D9
网约车乘客想要暖和一点	角色与座位；服务关系；温区温度	D2、D8、D11
儿童被留在驻车的车内	有儿童在车上；座位；驻车阶段；舱内温度	D1、D2、D10、D11
驾驶员看书时收到接管请求	接管请求；驾驶员角色；注意力；活动	D9、D2、D4、D6
为早晨通勤预先调节座舱	通勤习惯；目的地与出发时间；车外温度	D7、D12、D11

参考文献

1. Y. Hu et al., "Planning-oriented Autonomous Driving" (UniAD), CVPR 2023. arXiv:2212.10156
2. Z. Li et al., "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers", ECCV 2022. arXiv:2203.17270
3. X. Tian et al., "DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models", 2024. arXiv:2402.12289
4. J.-J. Hwang et al., "EMMA: End-to-End Multimodal Model for Autonomous Driving", 2024. arXiv:2410.23262
5. H. Nissenbaum, "Privacy as Contextual Integrity", **Washington Law Review** 79(1), 2004.
6. European Data Protection Board, Guidelines 01/2020 on processing personal data in the context of connected vehicles and mobility related applications, version 2.0, 2021.
7. 国家互联网信息办公室等，《汽车数据安全管理若干规定（试行）》，2021。
8. 日本《个人信息保护法》（Act on the Protection of Personal Information），2020 年修订，2022 年施行。
9. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 50.
10. A. K. Dey, "Understanding and Using Context", **Personal and Ubiquitous Computing** 5(1), 2001.
11. B. Schilit, N. Adams, R. Want, "Context-Aware Computing Applications", IEEE Workshop on Mobile Computing Systems and Applications, 1994.
12. A. K. Dey, G. D. Abowd, "Towards a Better Understanding of Context and Context-Awareness", GVU Technical Report GIT-GVU-99-22, Georgia Institute of Technology, 1999.
13. I. Siegert, N. Weißkirchen, A. Wendemuth, "Acoustic-Based Automatic Addressee Detection for Technical Systems: A Review", **Frontiers in Computer Science**, 2022. doi:10.3389/fcomp.2022.831784

正文中作为证据引用的行业报道、标准、专利和研究，已在引用处附上链接。