Defining Cabin Context: A Perception Framework for End-to-End Intelligent Cockpits

Industry White Paper · 1 November 2026

Author: Xu Dongchao (xudongchao@donfrontera.com)

Abstract

Intelligent cockpits have gone end-to-end on language but not on perception. Speech recognition, understanding and generation now run in a single model in production vehicles, yet what the cockpit knows about the people it serves — who sits where, what state each of them is in, how they relate to one another, what the vehicle and the road around it are doing — is still assembled from separate functions, or not assembled at all.

This paper argues that this gap explains most of today's cockpit pain points, and that closing it is the precondition for a genuinely end-to-end cockpit. We focus on the perception end and make four contributions. (1) **A diagnosis**: we trace current pain points to missing context and, drawing on autonomous driving's evolution from modular stacks through end-to-end models to vision-language-action (VLA) models, show that driving and the cockpit are converging on the same kind of architecture from opposite directions, and that the cockpit lacks exactly the unified perception that driving built first. (2) **A model**: the Cabin Context Model organizes context around the occupants in three layers and twelve dimensions, in which roles follow seats rather than people, relationships among occupants define who may hear what, and preferences are kept per person, role and situation; we argue from context theory and from the entities in a cabin that it covers what serving occupants requires, and illustrate this with a use-case coverage check. (3) **A method**: each dimension is obtained by direct observation, instant inference or inference accumulated over time, and every context item carries a calibrated uncertainty, a probability for categorical states and an interval for continuous values, with its evidence and validity. (4) **A principle**: sensors are chosen from first principles — each quantity is sensed the way a person would perceive it — and we give a reference sensor configuration on that basis. We close with the work the industry needs to do next.

1. Introduction

The cockpit's AI has advanced quickly on the language side. Assistants built on end-to-end speech models were shown in vehicles at CES 2026 (industry report). On-device omni-modal models running on the latest cockpit chips have been announced for 2026 vehicles (industry report). Reference designs describe in-vehicle agents that route requests to local and cloud agents by intent and context (technical blog).

Yet drivers and passengers still experience a cockpit that waits to be told what to do, does not know who is speaking or for whom, and says private things in front of the wrong people. "Context-aware" appears in almost every announcement; what the context *is* has not been defined.

Our thesis is that **the cockpit has gone end-to-end on language, while its perception of the people it serves is still modular and narrow**. Autonomous driving built unified perception first and has only recently added language, in vision-language-action (VLA) models. The cockpit started from the other end: it has

language and now needs perception. The two are converging on a similar kind of architecture, but the cockpit's perception problem is different in kind: it must understand several people, their inner states and their relationships, not only a physical scene.

This paper makes four contributions:

1. **Diagnosis.** It traces the pain points of current cockpits to missing context, and uses the evolution of autonomous driving, from modular stacks to end-to-end and VLA models, as an architectural analogy that explains what the cockpit lacks and where the cabin's perception and action problems differ from driving's (Sections 2 to 4).
2. **Model.** It defines the Cabin Context Model of three layers and twelve dimensions organized around the occupants, and argues why these twelve cover what a cabin needs (Section 5 and Appendix A).
3. **Method.** It classifies how each dimension is obtained — direct observation, instant inference or accumulated inference — and specifies that every context item carries a calibrated uncertainty: a confidence percentage for categorical states and an interval for continuous values (Sections 6 and 7).
4. **Principle.** It sets out a first-principles approach to choosing sensors and a reference sensor configuration built on it (Section 6.1).

The paper is a framework, not a survey of every dimension. For each way of obtaining context we give a few examples with evidence; dimension-by-dimension studies of sensing and inference are left to follow-up work.

2. Pain points of today's cockpits

Each of the seven pain points below is a failure of context, not of language. The cockpit understands what was said; it does not know enough about who said it, to whom, and in what situation.

Pain point	A typical scenario	Missing context
Listens but does not see	A passenger shivers and rubs her hands; nothing happens until someone says "turn up the heat".	Occupant physiological state
Does not know who	A rear passenger asks for music; the system plays the vehicle owner's playlist.	Identity and seat of the speaker
Ignores the situation	"It's stuffy" gets the same response on a quiet highway as in a crowded underground car park.	Scene, cabin environment, driving phase
Ignores the driving task	A long spoken suggestion arrives while the driver is merging in heavy traffic.	Vehicle and driving state
Ignores the audience	A reminder of a job interview is announced on the shared speakers with the user's manager in the back seat.	Relationships among occupants
Remembers the wrong thing	A preference set while driving is applied when the same person later rides as a passenger, or to everyone in the car.	History keyed to person, role and situation
Proactive features misfire	After a few badly timed suggestions, the user switches proactive services off.	All of the above, with confidence

The last row is the consequence of the others. Proactive service is the clearest promise of an AI cockpit, and it is only as good as the context it acts on.

A second, structural pain point sits behind the table: **sensor data stays in silos**. Driver monitoring, occupant monitoring, in-cabin radar and the automated driving system each serve their own safety function. Their outputs rarely reach the cockpit's AI, even though they describe exactly the situation the AI

needs to understand.

3. The driving precedent

Autonomous driving moved from modular stacks to end-to-end models, and more recently to vision-language-action (VLA) models that add language reasoning. We use this path as an architectural analogy, not as a route the cockpit must follow step by step. Its lesson is that each stage was built on a unified perception of the scene, and that is the step the cockpit has not yet taken.

3.1 Three stages

Modular. Driving stacks separated perception (with sensor fusion), localization, prediction, planning and control, and planned with rules. The cockpit equivalent is the pipeline of speech recognition, language understanding, dialogue management, skills and speech synthesis, alongside driver and occupant monitoring functions that work on their own.

End-to-end. Driving first unified perception: BEVFormer lifted multi-camera features into one bird's-eye-view space with temporal fusion [2], and planning-oriented models such as UniAD trained perception, prediction and planning jointly on that shared representation [1]. Later models map sensor input to a trajectory in a single network. The cockpit equivalent is multi-sensor input feeding one shared **context**, on which decisions and actions are learned.

Vision-language-action. The current frontier adds language reasoning to perception and action. DriveVLM reasons through scene description, scene analysis and hierarchical planning with a vision-language model, and its Dual variant combines that reasoning with a conventional perception and planning pipeline [3]. EMMA maps camera input, with navigation and vehicle state expressed as text, to trajectories and other driving outputs in one multimodal language model [4]. The cockpit equivalent is a model that perceives the occupants, reasons in language, and acts across speech, devices and services.

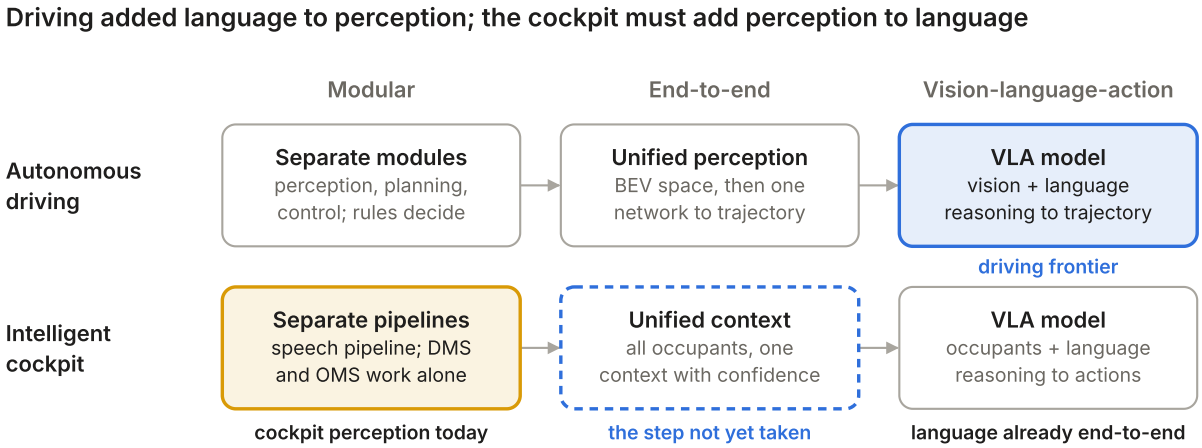


Figure 1. Driving and the cockpit converge on vision-language-action models from opposite directions

3.2 Converging from opposite directions

Driving and the cockpit are converging on the same kind of architecture from opposite ends. Driving began with perception and action and has now added language. The cockpit began with language, which is already end-to-end in production, and must now add perception and action. What the cockpit lacks is

exactly what driving built first: a unified perception of the scene, which in the cabin means a unified context of its occupants.

The VLA systems in driving also suggest what that context should look like. They keep explicit intermediate reasoning, such as a description of the scene, and some pair a slow reasoning model with a fast conventional pipeline [3]. A cockpit model can likewise keep an explicit, structured context with stated confidence as its intermediate representation, even when perception, reasoning and action are eventually trained together.

3.3 Where the cockpit differs

Aspect	Autonomous driving	Intelligent cockpit
Who is served	One driving task	Several occupants whose needs can conflict
What is perceived	The physical world outside: geometry, objects and the intentions of other road users, with largely objective ground truth	People's inner states, relationships and histories, often without objective ground truth
Time scale	Milliseconds to seconds for control, with maps and fleet data built up over longer periods	Seconds to months
Action space	A trajectory, executed as steering, throttle and brake	Speech (content, tone, which seat zone), cabin devices (zoned climate, seats, windows, lighting, displays), ecosystem content and services (music, video, navigation, payment, third-party apps), multi-step combinations, and choosing not to act
How success is judged	Safety-critical, with measurable safety and comfort metrics	Largely subjective

This is not a claim that the cockpit is harder overall: driving carries safety-critical stakes and a long tail of physical situations that the cabin does not. The cockpit's complexity is of a different kind: several occupants whose needs can conflict, subjective states without objective ground truth, relationships that build up over months, and a much broader action space. This paper addresses the perception end; defining the action space is left to future work.

4. Where cockpits stand today

Current end-to-end cockpits perceive what occupants say and, increasingly, what the cabin camera sees; most other context is fetched on request or not used. We group 2025–2026 systems into three architecture classes rather than assess individual products, since public descriptions rarely state input coverage precisely.

The classes are not exclusive; one product can combine all three.

- **Class A — End-to-end speech.** One model replaces the pipeline of speech recognition, understanding and synthesis.
- **Class B — On-device omni-modal.** Audio, cabin vision and text in one model on the cockpit chip.
- **Class C — Agent orchestration.** A language model plans and calls vehicle, navigation and cloud agents, with memory retrieved per request.

Context dimension	Class A	Class B	Class C
Who sits where, who is speaking	Speaker zone	Partial (vision)	Via tool
Occupant state	Speech only	Partial (vision + speech)	Via tool (driver monitoring flags)
Relationships among occupants	Not confirmed	Not confirmed	Not confirmed
Vehicle and driving state	Not confirmed	Not confirmed	Via tool, on request
Scene (car park, off-road, tunnel)	Not confirmed	Not confirmed	Via tool (map)
External environment and trip	Not confirmed	Not confirmed	Via tool (navigation, weather, calendar)
History keyed to person, role, situation	Not confirmed	Limited	Account-level memory

Assessment based on public product descriptions. "Not confirmed" means the capability was not found in public material, not that it is known to be missing; coverage of individual programmes should be verified against their specifications.

Three observations follow, within the limits of public material. Context appears to be **pulled rather than perceived**: in descriptions of Class C systems, vehicle state reaches the model when the model calls for it, not as a continuous input. **No description we found models relationships among occupants**, which leaves the audience pain point in Section 2 without a visible answer. And **memory, where described, is keyed to an account**, not to a person in a role and situation.

5. The Cabin Context Model: three layers, twelve dimensions

We organize cabin context around the occupants in three layers and twelve dimensions. The **occupant layer** holds one context per person in the cabin, the **relationship layer** holds what lies between them, and the **shared layer** holds what is common to the whole cabin. Time is not a separate dimension but an axis that runs through all twelve. At time *t*:

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

Layer	#	Dimension	What it covers
Occupant	D1	Identity	Enrolled user, recognized regular, unknown guest, child; age group
Occupant	D2	Seat and role	Seat; role as seat × driving state; the resources that seat can use
Occupant	D3	Physiological state	Thermal comfort, fatigue, motion sickness, discomfort, heart and breathing rate
Occupant	D4	Cognitive state	Attention, workload, focus of attention
Occupant	D5	Affective state	Irritation, relaxation, tension and other emotions
Occupant	D6	Activity	Sleeping, on a call, working, eating, looking after someone
Occupant	D7	History and preferences	Preferences and habits keyed to person × role × situation
Relationship	D8	Relationships and interaction	Lasting relationships (family, colleague, manager, service) and current interaction (who is talking to whom)

Layer	#	Dimension	What it covers
Shared	D9	Vehicle and driving state	Speed, gear, doors, charge; automation level, takeover requests; driving load
Shared	D10	Scene and driving phase	Kind of place (highway, off-road, underground car park, tunnel); phase (boarding, cruising, parking, arrival)
Shared	D11	Physical environment	Cabin zone temperature, humidity, air quality, noise, light; outside weather, road surface, traffic
Shared	D12	Trip and task	Destination, purpose, calendar, tasks in progress; place semantics such as home and work

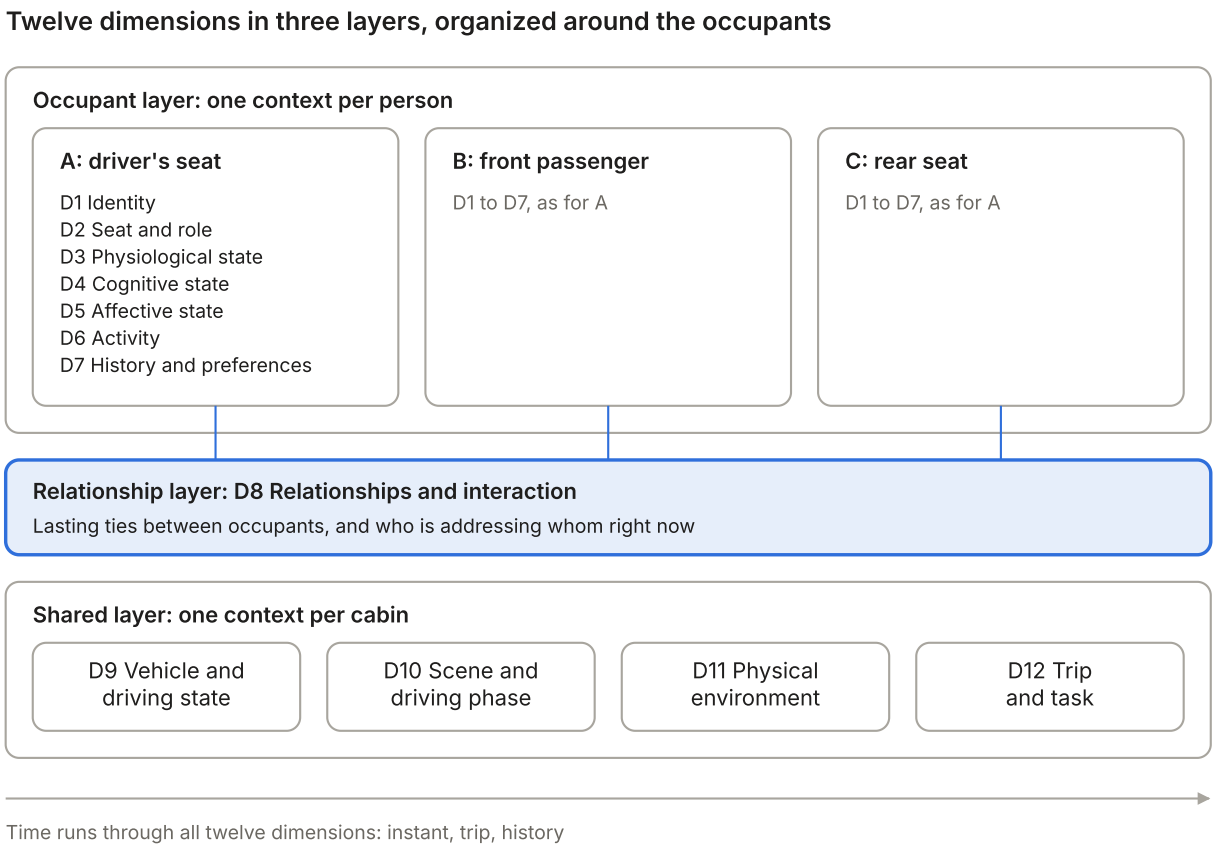


Figure 2. The Cabin Context Model: three layers, twelve dimensions

Each occupant carries the same seven dimensions; the relationship layer ties them together, and the shared layer frames them all.

5.1 Occupant layer

Identity is optional. The cockpit must serve people it cannot identify, so D1 is an attribute, not a precondition. Stable traits such as sensitivity to cold or to motion sickness attach to the person.

Roles follow seats, not people. Someone who drives today may sit in the back tomorrow. The role also depends on the vehicle's state, so we define **role = seat × driving state**.

Seat	Manual driving	Automated driving, human supervision or takeover required	Parked
Driver's seat	Driver	Supervisor or fallback driver, depending on automation level	Occupant

Seat	Manual driving	Automated driving, human supervision or takeover required	Parked
Front passenger seat	Passenger	Passenger	Occupant
Rear seats	Passenger	Passenger	Occupant
Child seat	Child passenger	Child passenger	Child left in vehicle if alone

In a vehicle that needs no human fallback, the driver's seat becomes a passenger seat and its occupant a passenger. The seat also defines which resources can serve that person: a climate zone, seat heating and ventilation, a display and a sound zone.

States are separated by kind. D3 to D6 distinguish physiological, cognitive, affective and activity states because they differ in how they are sensed and how reliably they can be inferred. Affective state receives heavy investment from industry and research, yet its reliability in real cabins remains the least established of the four, so it is kept as a dimension of its own to be tracked and evaluated separately.

History is keyed to person, role and situation. The same person may want a cool cabin to stay alert while driving and a warm one to sleep as a passenger; a preference recorded against the person alone would often be wrong.

5.2 Relationship layer

Relationships do not describe any one occupant, but they strongly constrain what may be said or shown in front of whom. Unlike a phone, whose output normally reaches one person, the cockpit's speakers and centre display reach everyone in the cabin. Every output therefore has an **audience**, and whether information may flow to that audience depends on the relationships within it. This is the cabin form of contextual integrity: whether an information flow is appropriate depends on its social context [5]. Schilit, Adams and Want already counted the social situation as context, such as whether one is with a manager or a co-worker [11].

Relationship	Examples	What it governs
Family	Partners, parent and child, elderly relatives	Whether memory and preferences may be shared
Guardianship	Parent and child, carer and elderly person	Whose instructions take precedence, content limits
Social	Friends, colleagues	Whether calendar or messages may appear on shared outputs
Professional	Manager, client	Formality; personal information withheld
Service	Ride-hailing driver and passenger, chauffeur and principal	Inverts roles: the main person served is not in the driver's seat

D8 also covers interaction on the instant time scale: who is currently addressing whom.

5.3 Shared layer

The shared layer separates **meaning** from **conditions**. D10 says what kind of place the vehicle is in and what phase the trip is in; D11 says what the physical conditions are. An underground car park (D10) and a cabin temperature of 17 °C (D11) are different kinds of facts, obtained in different ways. The driving phase matters for interpreting the occupant layer: closed eyes while parked and closed eyes at highway speed are different situations.

D9 includes driving load, derived from the vehicle's state, the scene and the driver's attention. D12 includes place semantics such as home and work, which are learned over many trips rather than sensed on one. Items in the shared layer can still name a single occupant as their subject: the destination is shared, but a personal calendar entry belongs to one person and carries a personal privacy class.

5.4 Time as an axis

Every dimension is observed on one or more of three time scales: **instant** (seconds: a gesture, a sentence, a shiver), **trip** (who is aboard, who fell asleep twenty minutes ago) and **history** (weeks to months). D3 and D6 live mostly on the instant scale, D8 and D12 accumulate over history, and D7 holds the preferences and habits that accumulate for each occupant.

5.5 Why these twelve dimensions

No model of context can be proven complete, because context is open-ended. We do not claim completeness or strict mutual exclusivity. We argue instead, from two directions, that the twelve cover what serving the people in a cabin requires; Appendix A illustrates this with a coverage check over example use cases.

They cover the classic elements of context. Dey and Abowd named location, identity, activity and time as the primary types of context [12]. Schilit, Adams and Want framed context-aware computing around where you are, who you are with and what resources are nearby [11]. Each maps onto the model: identity to D1; location to D2 (seat) and D10 (place); activity to D6; time to the axis that runs through all dimensions; who you are with to D8; nearby resources to the seat resources in D2 and the conditions in D11. What the model adds — inner states (D3 to D5), history (D7) and vehicle state (D9) — is what is specific to a cabin.

They follow from the entities in a cabin. The entities are each occupant, the pairs of occupants, the vehicle, the world outside and the trip. Asking what can be known about each yields the model:

- **An occupant:** who they are (D1); where they sit and in what role (D2); their body (D3), mind (D4) and feelings (D5); what they are doing (D6); what they prefer and habitually do (D7). The split of D3 to D6 follows the familiar human-factors distinction between physiological, cognitive, affective and behavioural states.
- **Occupants together:** how they are related and how they are interacting (D8).
- **The vehicle:** what it is doing (D9).
- **The world outside:** what kind of place it is (D10) and its physical conditions (D11).
- **The trip:** where it is going, why, and what is to be done (D12).

Three boundary rules keep the dimensions as separate as possible. D10 holds meaning (what kind of place, which phase) and D11 holds physical conditions. D7 holds only what a person prefers or habitually does; facts learned over time about other dimensions, such as a relationship (D8) or a home location (D12), stay in those dimensions with history as their time scale. And every item names its subject, so a personal item in a shared dimension still belongs to one occupant. Some overlap remains at the edges, and the boundaries should be refined as the dimensions are studied. The twelve form a stable top level; each can be refined into sub-dimensions without changing the structure.

6. How context is obtained

Every context dimension is obtained in one of three ways, and the way determines where its confidence comes from and how that confidence changes over time. We assume the reference sensor configuration of Section 6.1, so that the question is what can be known, not what a given vehicle has fitted.

Mode	What it is	Confidence comes from	Over time
Direct observation	A visible attribute or a system state is read from a single source	Signal validity and, for perceived attributes, the recognition model's calibrated per-item probability	Becomes stale and unknown once not refreshed
Instant inference	A hidden state is inferred by fusing several current signals	How many independent signals agree	Rises as signals corroborate; drops when they conflict
Accumulated inference	A fact is inferred from patterns over many trips	How much consistent evidence has accumulated	Grows with evidence; decays with counter-evidence or silence

The line between the first two modes is not whether a model is involved: recognizing an adult in a camera image uses a model too. Direct observation covers attributes that are visible on the surface or reported by a system, such as who sits where or the vehicle's speed, and it still carries recognition error. Instant inference covers states that cannot be seen directly and must be reasoned from several signals, such as feeling cold.

The examples below illustrate each mode. They are not a survey of every dimension, and the evidence cited shows that each kind of inference is feasible, not how well it performs in a production cabin. Most figures come from controlled studies with their own class definitions and settings, and patents show proposed approaches rather than validated results.

6.1 Reference sensor configuration

Sensors are chosen from first principles: each quantity is sensed the way a person would perceive it. We see who sits where and whether they are a child; we see someone shiver and feel how warm a face is; we hear who is speaking. A further sensor is justified when a quantity is physically unavailable to the primary one — a heartbeat is not visible, a child under a blanket is hidden. It is also justified when a safety function needs an independent second source, or when a less intrusive sensor can do the job, as radar can detect presence and breathing without capturing images. What does not justify a further sensor is that today's algorithms happen to score higher on it: algorithms improve, and a sensor architecture fixed to their current weaknesses does not. This is a design principle for the reference configuration below; cost, packaging and supplier constraints, which differ by programme, are outside its scope.

Sensor	Placement	What it senses	Dimensions served
Driver camera (near-infrared)	Steering column or A-pillar	Driver's face, eyes, head pose, gaze	D1, D3, D4, D5, D6
Cabin cameras (RGB-IR, wide angle)	Overhead console for the front row; roof for each rear row	Every seat: occupancy, body size, posture, face, gestures, clothing, shivering, expressions, objects in hand, who faces whom	D1, D2, D3, D4, D5, D6, D8
Thermal camera (long-wave infrared)	Overhead, covering all seats	Facial and hand skin temperature, sweating	D3

Sensor	Placement	What it senses	Dimensions served
Millimetre-wave radar (60 GHz)	Roof centre	Breathing and heart rate; presence of occupants hidden from view	D3, D2
Microphone array	One zone per seat	Speech, speaker zone, voice prosody, conversation turns	D1, D2, D5, D6, D8
Cabin environment sensors	Per climate zone	Temperature, humidity, CO_2, particulates, light	D11
Seat-belt and seat sensors	Every seat	Belt latched, seat adjustments	D2
HMI events	Displays, knobs, wheel buttons	Manual adjustments and their reversals	D6, D7
Paired wearables (optional)	On the occupant, with consent	Skin conductance, heart-rate variability	D3, D4
Vehicle bus	—	Speed, gear, doors, windows, climate, charge, motion	D9
Automated driving system	Its exterior cameras, radar, lidar and fusion	Objects, road and lane, scene and operating-domain judgment, automation level, takeover requests	D9, D10, D11
Positioning, maps and cloud	GNSS and IMU; map; weather and calendar services	Location, place type, weather, route, schedule	D10, D11, D12

6.2 Direct observation

Who sits where, and adult or child. A person tells an adult from a child by looking, and so does the cabin camera: body size and proportions, face, and whether a child seat is present. Vision-based occupant classification has a long record. A single greyscale camera separated rear-facing infant seats, children, adults and empty seats with about 95% accuracy as early as 2003 (CVPR 2003), and a later system distinguished adults from forward- and rear-facing child seats with 98.9% accuracy on over 84,000 frames (BMVC 2004). Non-visual sensors struggle with exactly this distinction: a child on a booster seat can be mistaken for an adult by overhead ultrasound, and inductive sensors cannot tell occupant types apart (Cost-based Feature Transfer). Radar complements the camera where vision cannot see, such as a child covered by a blanket, and where a second, non-imaging source is wanted for safety or privacy.

Which seat a speaker is in. Multi-zone microphone systems identify the zone of the active speaker by fusing timing and signal-power cues across distributed microphones (US 10332545).

Vehicle and automation state. Speed, gear, doors, climate and charge come from the vehicle bus; automation level and takeover requests come from the automated driving system. Its perception is already described at object level with standardized semantics between sensors and the fusion unit (ISO 23150-1:2026), so the cockpit can inherit both the objects and their existence probabilities.

6.3 Instant inference

Thermal comfort: why one signal is not enough. Facial thermography can estimate comfort: a building study with low-cost thermal cameras reported about 85% accuracy for personal comfort models, with ears, nose and cheeks most indicative (NSF-funded study). But facial temperature alone can mislead: in both hot and cold conditions comfort can vary widely while facial skin temperature hardly changes, and in

the cold the body regulates by shivering (US 11458803). On the warm side, sweating can be detected with thermal imaging (Scitepress 2020). A credible estimate of "this passenger is cold" therefore combines facial temperature, visible shivering or hand-rubbing, clothing, zone temperature and outside temperature.

Underground car park: weak signals, strong conclusion. GNSS loss alone could mean a tunnel. Add a ceiling, pillars and parking-bay lines seen by the automated driving system's cameras, low speed with large steering angles, parking assist active, and a sharp drop in ambient light, and the conclusion becomes near-certain. Off-road driving is inferred the same way: no lane markings, unpaved surface, an off-road drive mode or locked differential, large body pitch and roll, and wheel slip. Automated driving already maintains this kind of scene judgment to check its operational design domain; today it is rarely passed to the cockpit.

Motion sickness in the rear seat. About a third of passengers experience dizziness or motion sickness on long journeys, with small children particularly susceptible (industry study). An industrial research programme found body temperature and skin conductance among the markers most correlated with subjective sickness (eeNews Europe). In the reference configuration, temperature comes from the thermal camera; skin conductance needs contact sensing, such as a paired wearable. Proposed detectors combine physiological signals with vehicle motion data and passengers' gaze (US 11820402), adding seat position as a prior: rear passengers looking at screens are most at risk.

Even the most mature inference is probabilistic. For drowsiness, the guidance values pre-filled in Euro NCAP's driver-monitoring dossier template for 2026 are 0.5 false alerts per hour and a 40% true-positive rate (Euro NCAP). A context item that hides this uncertainty misleads whatever uses it.

6.4 Accumulated inference

Home and work. Repeated arrivals at a place in the morning peak suggest a workplace; evening arrivals followed by an overnight stop suggest home. On smartphone location data from 5,099 people in 68 countries, the HoWDe algorithm detects home and work with accuracy up to 97% and 88% respectively (de Sojo et al.). Vehicle trips are a different and sparser source; on vehicle trip logs from 467 drivers, a simpler method failed to detect home in 2.5% of cases and work in 6.2% (US 11248924). The same patterns reveal a home garage: the same car park, entered in the evening, left the next morning.

Relationships. No single trip reveals that a passenger is the driver's manager, but patterns do: how often two people ride together, at what times, in which seats, how they address each other and what they talk about. Evidence from two-party telephone calls shows that conversation content separates business from personal calls with about 88% accuracy and family from non-family with about 82% (Stark et al.). We found no study of relationship inference among several occupants inside a vehicle.

Preferences in context. Preferences accumulate per person, role and situation: what a person sets when driving to work on a winter morning is evidence about that situation, not about every trip.

Two principles govern accumulated inference. First, an inference may freely make the cockpit **more cautious**, but should make it **more open** only after the user confirms it: wrongly assuming a passenger is a colleague costs little, wrongly assuming a spouse may expose private information. Second, accumulated inferences should be visible to the user and deletable.

7. Context with confidence

The output of cabin perception is a set of context items, and every item carries a calibrated uncertainty: a confidence percentage for categorical states, an interval for continuous values. Whatever consumes the context — a decision model today, a single end-to-end model later — can then weigh an uncertain estimate rather than treat it as fact.

7.1 The context item

Field	Meaning	Example
Subject	The occupant, seat or whole cabin the item is about	Rear-left occupant
Dimension	Which context dimension	Thermal comfort
Value	The estimated value	Cold
Confidence	For a categorical value, the calibrated probability that it is correct, as a percentage; for a continuous value, an uncertainty interval	78%
Mode	Direct observation, instant inference or accumulated inference	Instant inference
Evidence	The signals or past events that support it	Shivering detected; zone temperature 17 °C; outside 2 °C; no coat
Observed at	When the item was produced	08:12:05
Valid until	When it expires unless refreshed	08:14:05
Privacy class	Non-personal, personal or sensitive	Sensitive

7.2 What the percentage means

A confidence percentage is useful only if it is **calibrated**: across all items labelled 80%, about 80% should be correct. Raw model scores are usually not calibrated, so each dimension needs a calibration step against labelled data. The percentage applies to categorical states such as "cold" or "asleep". A continuous quantity such as cabin temperature or heart rate is better reported as an estimate with an uncertainty interval, for example 72 ± 4 beats per minute.

Two kinds of confidence should not be confused. **Method maturity** describes how reliably a dimension can be estimated in general; accuracies reported in the literature, mostly from laboratories and small samples, are evidence of this and are not per-item confidence. **Item confidence** describes how far this estimate, for this person, now, can be trusted. Only item confidence goes into the context item.

7.3 How confidence behaves by mode

- **Direct observation** — set by signal validity and, for perceived attributes such as adult or child, by the recognition model's calibrated per-item probability. When the signal is not refreshed within its validity window, the item becomes stale and its value unknown; it is not judged false.
- **Instant inference** — rises as independent signals agree and falls when they conflict. Conflicts are recorded in the evidence field rather than averaged away.
- **Accumulated inference** — grows with each consistent observation and decays with counter-evidence or with time without new evidence. After a house move, the old "home" should lose confidence on its own.

7.4 Worked examples

Subject	Dimension	Value	Confidence	Mode	Evidence
Rear-right seat	Occupancy	Adult present	99%	Direct observation	Cabin camera: adult body size, face and posture in the seat; belt latched
Front passenger	Speaking	Active speaker	95%	Direct observation	Microphone zone timing and power
Cabin	Scene	Underground car park	96%	Instant inference	GNSS lost; ceiling and bay lines in camera view; 8 km/h; parking assist on; ambient light dropped
Rear-left occupant	Thermal comfort	Cold	78%	Instant inference	Shivering detected; zone 17 °C; outside 2 °C; facial temperature unchanged
Rear-left occupant	Motion sickness	Onset likely	55%	Instant inference	Facial skin temperature dropping; winding road; gaze on tablet for 20 min
Front passenger to driver	Relationship	Colleague	70%	Accumulated inference	34 shared weekday commutes; formal address; work topics
Driver	Place semantics	Destination is workplace	93%	Accumulated inference	41 weekday arrivals between 08:00 and 09:30 over 60 days

Values are illustrative. They show the intended spread: observations near certainty, instant inferences varying with how much evidence agrees, and accumulated inferences that start low and strengthen trip by trip.

8. Privacy considerations

A model that perceives every occupant, their state and their relationships is only acceptable if privacy is built into the context itself. The context item format helps: it lets raw sensitive data stay where it was sensed while only minimal, labelled conclusions move on.

- **Process raw data on the device.** Video frames, voice audio and physiological signals stay in the vehicle; only context items leave the perception layer. This matches the European Data Protection Board's preference for in-vehicle processing in connected vehicles [6], China's automotive data provisions, which make in-vehicle processing and no collection by default guiding principles [7], and purpose limits under Japan's APPI [8].
- **Keep conclusions, not conversations.** Accumulated inference about relationships draws on what occupants say. Only the inferred relationship, its confidence and its evidence counts should be kept, never transcripts.
- **Respect people who have not agreed.** Passengers have not consented to being analysed. An unidentified occupant can still be perceived as present and as a speaker in a seat zone without being identified, and passengers should be able to opt out of being remembered.
- **Make accumulated inferences visible.** Inferred places and relationships should be viewable and deletable by the user they concern.
- **Treat affect with extra care.** Emotion recognition carries transparency obligations under the EU AI Act [9], and its in-cabin accuracy is the least established of all dimensions.

9. Related work

This paper builds on four lines of work and differs from each in being organized around several occupants at once.

General definitions of context. Dey defined context as any information that characterizes the situation of an entity relevant to an interaction [10]. That definition is deliberately general; we specialize it to a shared cabin, where the entities are several people in seats and the vehicle itself.

In-cabin monitoring and AI. Surveys of AI in the cabin organize work by use case (Artificial Intelligence Methods in In-Cabin Use Cases), and datasets such as Drive&Act focus on fine-grained recognition of the driver's activities (Drive&Act). Driver facial expression recognition has been surveyed extensively (Saadi et al.). This work is largely driver-centred and treats each state separately; we place these states in one model that covers passengers, roles and relationships.

Scene description for automated driving. Automated driving describes its operating conditions through operational design domains and standardizes perception outputs at object level (ISO 23150). We reuse these as sources of shared context rather than define them anew.

Privacy in context. Nissenbaum's contextual integrity holds that the appropriateness of an information flow depends on its social context [5]. We apply it to the cabin through the notion of an output's audience.

10. Next steps and conclusion

Defining cabin context is a first step; turning it into practice needs work across the industry.

1. **A common context interface.** Automated driving standardized the interface between sensors and the fusion unit; the cockpit has no equivalent for passing context items, with their confidence and privacy class, to whatever consumes them.
2. **Opening the automated driving system's scene layer to the cockpit.** Scene judgments such as "underground car park" or "off-road" already exist inside the automated driving stack, but the path from its fusion output to the cockpit is not standardized.
3. **Research on passengers.** Regulation and consumer testing focus on the driver; passenger fatigue, motion sickness and comfort have no comparable benchmarks.
4. **Research on who speaks to whom.** Systems can tell which seat a speaker is in. Detecting whether speech is directed at a system is an established research topic [13], but working out which occupant a passenger is addressing, and inferring relationships among several occupants, remains largely open.
5. **Dimension-by-dimension studies.** For each dimension, which sensors and methods reach which confidence in real vehicles. We intend to follow this paper with such studies.
6. **The action space.** The other end of the end-to-end cockpit — speech, devices and ecosystem services — needs a definition as careful as the one proposed here for context.

Autonomous driving became end-to-end only after it learned to see the road as one picture. The cockpit has learned to talk; it has yet to see its occupants that way. An occupant-centric model of context, in which roles follow seats, relationships define the audience, and every item states how it was obtained and how far it can be trusted, is the picture it needs.

Appendix A. Use-case coverage

Every item of context needed by the use cases below can be mapped to one of the listed dimensions, and every dimension is needed by at least one use case. This is a consistency check on examples, not an empirical validation; a fuller check would apply the same test to a programme's own use-case catalogue.

Use case	Context needed	Dimensions
A passenger feels cold	Shivering and skin temperature; which seat; zone and outside temperature	D3, D2, D11
A rear passenger asks for music	Who spoke and from which seat; their preferences	D1, D2, D7
A private appointment is due	Who is aboard; how they relate to the user; the appointment	D1, D8, D12
A long drive suggests a break	Driver fatigue; trip progress and next stop; speed	D3, D12, D9
A child falls asleep in the back	Child present; sleeping; which seat	D1, D6, D2
A message arrives during a difficult merge	Driving load; scene and phase	D9, D10
"Close that window", with a point	Gesture; gaze target; speaker's seat	D6, D4, D2
A rear passenger feels sick	Physiological markers; reading a tablet; seat; winding road	D3, D6, D2, D10
Arriving at the home garage	Underground car park; place semantics; parking	D10, D12, D9
The driver grows irritated in congestion	Affective state; congestion phase; vehicle state	D5, D10, D9
A ride-hailing passenger wants it warmer	Role and seat; service relationship; zone temperature	D2, D8, D11
A child is left in a parked car	Child present; seat; parked phase; cabin temperature	D1, D2, D10, D11
A takeover request while the driver reads	Takeover request; driver's role; attention; activity	D9, D2, D4, D6
Pre-conditioning for the morning commute	Commute habits; destination and departure; outside temperature	D7, D12, D11

References

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. Several Provisions on Automotive Data Security Management (Trial)（汽车数据安全管理若干规定（试行）), Cyberspace Administration of China et al., 2021.
8. Act on the Protection of Personal Information (Japan), as amended 2020, in force 2022.
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

Industry announcements, standards, patents and studies used as evidence are linked where they are cited in the text.