

Emotion-Aware Autonomous Driving: A Survey on Multimodal Emotion Recognition, LLM-Based Decision-Making, and Human-Machine Interaction

Xiye Tong *

Zhuhai College of Science and Technology, Zhuhai, Guangdong, 519040, China

* Corresponding author Email: xiyetong6@gmail.com

Abstract: Autonomous driving research has begun to move. Where the early emphasis fell on safety and efficiency, attention is now turning toward the people inside the car and how the ride actually feels. Most systems, though, still read the road without reading the passenger: real-time emotional states rarely enter the perception or decision pipeline, and that gap makes human-vehicle trust hard to build and the ride harder to improve. This review looks at three lines of work that bear on the problem—in-vehicle multimodal emotion recognition, driving decisions built on large language models (LLMs), and the mechanisms behind human-machine interaction and trust—and reads them against one another rather than in isolation. A consistent pattern emerges. Emotion recognition works well in controlled tests, yet its output seldom reaches the decision module. LLMs plan tasks and explain their choices, but no current system feeds passenger emotion into their reasoning. Interaction research cares about trust and explainability, only its content stays fixed and scripted, so it cannot bend to how a passenger feels in the moment. The three lines run in parallel and almost never meet. That missing coupling is the field’s central gap, and it is what keeps a vehicle from sensing, let alone answering, the subjective state of the person it carries. We then work through what integration would demand—latency, safety limits, and personalization—and set out five directions for future work: a unified perception-decision framework, dynamic explainable interaction driven by an LLM, lightweight in-vehicle deployment, personalized learning, and better datasets. Taken together, they sketch a reference for a closed loop that runs from emotion perception through decision generation to human-machine interaction.

Keywords: Autonomous Driving; Emotion Perception; Multimodal Emotion Recognition; Large Language Models; Human-machine Interaction.

1. Introduction

Autonomous driving has advanced quickly. Waymo and Baidu Apollo already run Level 4 vehicles commercially within mapped service areas [1, 2]. Yet adoption still stalls on a human problem: passengers stay anxious about, and distrustful of, how the car behaves, and affective computing has been named a prerequisite for making in-vehicle systems acceptable and workable in the first place [3]. And the discomfort is not vague—it shows up against concrete driving-style parameters like acceleration and following distance, which makes a poor emotional ride, not just crash statistics, a barrier to trust in its own right. Under the SAE scale, automation runs from L0 to L5 [1], and L4-and-above systems are now in commercial use within designated zones. Development, however, has long chased objective numbers—perception accuracy, path planning, collision avoidance—while the rider’s own experience and state of mind went largely unexamined [4, 5].

Driving leans heavily on cognition and emotion. Anxiety, tension, and anger all shift how occupants react and how the ride lands with them [4, 3]. Inside an autonomous vehicle, the triggers are specific: a sudden brake, a messy intersection, a decision the passenger cannot see the reason for—any of these can sour the mood and, with it, comfort and trust [3, 6]. Reading those states as they happen, then adjusting how the car drives and how it talks back, is therefore central to making human-machine collaboration work.

Several communities have taken the problem on. For sensing emotion in the cabin, researchers pair multimodal

data—face, voice, physiological signals—with deep learning to track how drivers and passengers feel in real time, and accuracy in controlled settings is high [4, 5]. On the decision side, frameworks like SayCan [7] and Inner Monologue [8] show that a language model can parse an instruction, reason through a situation, and hand back an executable plan, which opens a fresh route to driving task planning. Work on interaction and trust has taken yet another angle, using interface design, visual cues, and behavioral explanations to help passengers follow what the car is doing [9, 10]; one finding stands out—telling people “why” a decision was made builds more trust than telling them “how” [11].

What ties these efforts together is also what divides them: they have never been integrated, and three gaps follow. The first is at the perception end. Emotion recognition stops once it prints a label; it never builds the bridge from that label to an actual driving instruction [4]. The second sits in the decision layer. LLM-based systems take in objective environmental data and little else, and to our knowledge none has folded passenger emotion into strategy generation, so the strategies cannot bend to what any individual needs [1, 2]. The third is on the interaction side. Explanations are pre-scripted, generated from fixed templates rather than the live situation, and they ignore how a passenger’s mood changes whether the explanation even works [9, 10]. Put together, the picture is odd: a car that perceives its surroundings and plans a path with real skill, yet stays blind to the feelings of the person riding in it—and so cannot soften its driving or tailor a word of explanation.

To close those gaps, this review reads the literature through

one lens: emotion perception, LLM decision-making, and interaction as a coupled whole. The sections follow that logic. Section 2 covers in-vehicle multimodal emotion recognition—data modalities, feature extraction, fusion—and marks where its output fails to reach the decision interface. Section 3 turns to LLMs in driving decisions, weighing what they manage and what they miss, and noting how passenger emotion drops out. Section 4 takes up interaction and trust:

how explainability, situational awareness, and emotion shape whether a passenger trusts the car. Section 5 asks what joining the three would take—latency, safety, personalization—and points to directions for unified framework design, lightweight deployment, and datasets. The target throughout is an emotion-aware closed loop.

The overall framework and cross-domain coupling structure of this review are illustrated in Figure 1.

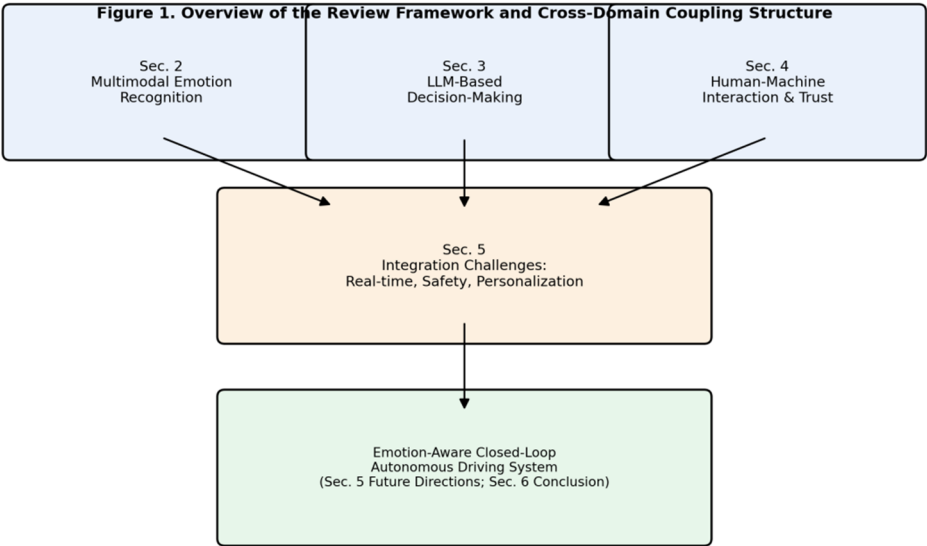


Figure 1. Overview of the review framework and cross-domain coupling structure

2. Advances in Multimodal Emotion Recognition Technology

At Level 4 and 5, the person in the seat is no longer driving. Operator has become passenger. Whether the car can read and answer their emotional state now decides how well the collaboration works, and whether they accept it at all [4, 3]. This section reviews in-vehicle multimodal emotion recognition along three lines: the data, how features are pulled

from it, and how modalities are fused.

2.1. Data Modalities and Acquisition Methods

Three modalities feed in-vehicle emotion recognition: facial expression, speech, and physiological signals. They differ in how they are captured, which datasets exist, and where each fit, as Table 1 lays out. None is complete on its own. Fusing two or more lets the modalities cover for each other, and the evidence is consistent—a fused system beats even the best single channel [4, 5].

Table 1. Comparison of Three Emotion Perception Modalities

Modality	Acquisition Method	Representative Datasets	Advantages	Major Limitations
Facial Expressions	RGB/Infrared Cameras	AffectNet, FER2013, CK+	Non-invasive, low cost	Susceptible to occlusion and lighting conditions [4]
Voice signals	In-vehicle microphones	Emo-DB, IEMOCAP	Natural acquisition	Performance degrades in high-noise environments [3]
Physiological signals	Wristband/chest strap/steering wheel electrodes	DEAP [12], DriveDB, WESAD	Difficult to fake; reflects true arousal	Requires skin contact; prone to motion artifacts [4]

2.2. Development of Feature Extraction Methods

Feature extraction has moved from hand-designed

descriptors toward deep learning, as Table 2 shows. CNNs, LSTMs, and Transformers [13] now dominate. The gain is generalization—they carry across scenarios far better than hand-engineered features. The cost is data and compute, both of which rise.

Table 2. Comparison of Feature Extraction Methods

Method Type	Representative Methods	Applicable Modalities	Advantages	Limitations
Traditional Manual Features	LBP, HOG, Gabor, MFCC[14]	Face, Speech, Physiological	Low computational cost, high interpretability	Poor generalization
Deep Learning	ResNet, VGG[15], Deep-Emotion[16]	Facial	Automatic feature learning, strong generalization	Requires a large amount of labeled data
Sequence models	LSTM, Transformer[13]	Physiological, Speech	Modeling temporal dynamics	High computational resource requirements

2.3. Multimodal Fusion Strategies

Fusion strategies fall into three kinds—feature-level,

decision-level, and model-level. Each trades the depth of cross-modal interaction against stability differently, as Table 3 details.

Table 3. Comparison of Multimodal Fusion Strategies

Fusion Strategy	Fusion Stage	Representative Methods	Advantages	Limitations
Feature-Level Fusion	After Feature Extraction	Feature concatenation/weighting	Early detection of modal interactions	Requires strict alignment; engineering complexity [4, 5]
Decision-level fusion	After classification and decision-making	Voting/weighted average	Modular and solid	Ignores deep semantic relationships [5]
Model-level fusion	Within the network	Cross-modal attention [16, 15, 13]	Fully leverages modality complementarity	Requires large amounts of data and incurs high computational costs

2.4. Summary

This section has traced in-vehicle multimodal emotion recognition: how the three modalities—face, speech, and physiological signals—are captured and what marks them apart, how feature extraction moved from hand-crafted descriptors to deep learning, and how the three fusion strategies compare. The bottom line is mixed. Deep multimodal fusion reaches high accuracy in controlled tests, but the output still has nowhere to go—no working interface links it to the decision module, and emotional information never enters the driving control chain. That dead end raises a question worth chasing: can a language model, with its reasoning reach, become the bridge from emotion perception to driving decisions? Section 3 takes it up.

3. Current Status of Large Language Model Applications in Autonomous Driving Decision-Making

Section 1 showed that emotion recognition output never reaches the driving decision chain. This section turns to LLMs in driving decisions—what they manage in task planning, end-to-end control, and explanation, and where they fall short—and underlines how existing LLM-based systems leave passenger emotion out entirely.

A word on scope, since this review deliberately centers on LLM-based decision-making. Emotion-adaptive driving is not an LLM invention. Older and still-common approaches include rule-based adjustment, where an emotional reading past a set threshold trips a predefined change to speed or following distance; machine-learning classifiers that map affect labels onto fixed control parameters; and reinforcement-learning policies keyed to physiological or behavioral signals. These carry lower compute cost and are easier to verify, which suits safety-critical use. Their weakness is rigidity: they run a closed, predefined label-to-action map and cannot do the flexible contextual reasoning, the back-and-forth dialogue, or the plain-language explanation that a language model can. Those three abilities are exactly what the emotion–decision–interaction loop from Section 1 needs—turning a shaded emotional state into an adjusted strategy and then saying why. That is why the LLM route is the focus here, with the non-LLM methods kept in view as a baseline rather than the main object of study.

3.1. LLMs for Task Planning and Decision Generation

Built on the Transformer and pre-trained on huge text corpora, LLMs bring contextual understanding, multi-step reasoning, and task planning—a new way to generate high-level driving decisions. SayCan, from Ahn et al., was an early case of putting an LLM to robotic task planning: at each step it weighs the LLM’s judgment of what makes sense against a value function’s read on what is physically doable, then picks the best executable skill [7]. Inner Monologue went further, adding a language feedback loop so the model revises its plan as execution unfolds, which makes the planning sturdier [8]. Both target general robotics, not cars. Still, their core move—language reasoning paired with environmental feedback—carries over cleanly to driving task planning.

For the specifics of driving, researchers have extended these general frameworks. GPT-Driver treats motion planning as sequence generation over trajectory coordinates: the LLM predicts future path points autoregressively, and a low-level controller turns them into steering and acceleration commands [17]. It also writes out its reasoning in plain language as it goes, so a passenger can follow why the car did what it did—interpretability and control accuracy together. LanguageMPC takes another path, translating a spoken driving-style instruction like “drive smoothly” into weight changes in the MPC cost function. Driving style adapts online, and the LLM ends up acting as a bridge between a high-level semantic interface and a low-level controller [18].

3.2. Lack of Integration with Emotion Perception in Current Methods

These systems plan, control, and explain well enough, but one blind spot cuts across all of them: to our knowledge, none has ever fed the passenger’s live emotional state into the LLM’s reasoning. SayCan and Inner Monologue take task instructions and environmental states; GPT-Driver and LanguageMPC lean on road data and style instructions; DriveGPT4 works from raw visual frames [19]; other explainable end-to-end systems likewise rely on object-level or visual input with no affective channel at all [20]. Figure 2 lays this out. Nowhere does emotion enter the decision input, so it cannot touch the strategy that comes out. The result is a single strategy for everyone, unable to shift with the individual; and when driving style or road conditions leave a passenger uneasy, the car neither notices nor answers, which only deepens the one-sidedness of the collaboration.

System	Input Modality	Output	Passenger Emotion Input?
SayCan [10]	Task instruction + robot affordance value	Executable skill selection	No
Inner Monologue [11]	Task instruction + environment feedback	Closed-loop action plan	No
GPT-Driver [12]	Road/object features	Trajectory + NL justification	No
LanguageMPC [13]	NL driving-style instruction	MPC cost-function weights	No
DriveGPT4 [14]	Raw visual frames	Control signal + NL explanation	No

Figure 2. Input–output modalities of representative LLM-based driving decision systems; none incorporate passenger emotional state as an input.

3.3. Summary

This section has surveyed how LLMs contribute to driving decisions across three fronts—task planning, end-to-end control, and explanation. The evidence is clear that an LLM can turn a spoken instruction into an executable strategy and explain that strategy to the passenger at the same time. What no framework does, though, is represent passenger emotion, even though emotion is a chief driver of how the ride feels—a systemic hole in current LLM-based decision research. That points straight to the next question: at the level of interaction, has anyone weighed how emotion shapes trust and how well the exchange works? Section 4 takes it on.

4. Current State of Research on Human-Machine Interaction and Trust in Autonomous Driving

Section 3 showed that LLM-based decision systems overlook how passengers feel. Yet even setting emotion aside, interaction and trust are already what make or break a passenger’s acceptance of the technology. This section reviews the state of that research along four lines: how trust behaves over time, where current interaction methods fall short, how explanation is designed, and where emotion fits in.

4.1. Trust and Situational Awareness

Trust is the pivot in any human-machine team. Endsley’s six-month naturalistic study of Tesla Autopilot found that trust bends to mode confusion, shaky mental models, and behavior the driver did not see coming; the moment the car acts against expectation, trust falls fast, and drivers often cannot retake control when it counts [21]. For passengers, Reig et al. traced much of the distrust to a plain cause—people simply do not grasp what the vehicle intends to do [9]. Dillen et al. pinned it down in a real car: style parameters like longitudinal acceleration and following distance move in step with skin conductance, heart rate, and eye-tracking entropy, and those in turn color how comfortable and trusting the passenger reports feeling [6]. Read together, the studies say trust is not fixed. It shifts with what the car does, and a gap between driving style and expectation can drop it on the spot—which is the empirical footing for tying emotion perception to real-time strategy adjustment.

4.2. Limitations of Existing Interaction Methods

Today’s in-vehicle HMIs mostly speak through the eyes or the ears—a dashboard icon, an AR overlay, a beep, a clipped voice prompt [10]. Two limits are built in. One is bandwidth: a single icon or chime cannot carry a complex chain of reasoning. The other is direction: the channel runs one way, so a passenger cannot question the car or answer back, and

genuine dialogue never gets off the ground. An LLM changes the terms. It packs more information per turn and sustains a conversation, fielding questions and shaping an explanation to fit. The promise for passenger understanding and trust is real. The catch is maturity—putting LLMs to work on live in-vehicle explanation is early-stage, and the system design and evaluation still need working out.

4.3. Explainability and Transparency

Explainable AI (XAI) is one of the main levers for building trust. Koo et al. ran the comparison directly: telling drivers “why” a decision happened lifted performance and satisfaction more than explaining “how” it worked, and giving both produced the safest result, though it taxed attention briefly [11]. Colley et al. pushed perception into view, showing passengers semantic-segmentation results through an AR interface; situational awareness rose without a real jump in cognitive load [22]. But all of this runs on canned content in a fixed format. It cannot flex to the live situation or the passenger’s state, and it leaves no room for questions or a second turn. Using an LLM’s language generation to make explanation dynamic and personal is one plausible way past that wall.

4.4. Emotional Factors in Human-Machine Collaboration

One assumption runs quietly through most explainability work: the user is emotionally neutral. How mood might change the outcome gets little attention. Reviewing affective-wellbeing technology in HCI, Sanches et al. note that most systems do the basics—collect data, report a result—without ever folding a user’s live emotional state into how feedback is shaped [23].

The link has rarely been studied head-on, yet three strands of nearby evidence converge on the same point: passenger emotion deserves to be a first-class variable in interaction design. Start with the tie between how the car drives and how the passenger feels. Dillen et al. showed in a real vehicle that acceleration and following-distance parameters move with skin conductance, heart rate, and gaze entropy, and shape the comfort and trust people report [6]. So arousal is not fixed background; it rises and falls with concrete actions of the car, and could in principle serve as a live input for timing and phrasing an explanation rather than being written off as noise. Dillen’s work stops at the physiological reading, though—it never routes the signal back into an explanation module, so the loop stays open. Next, consider how explanation should track a user’s cognitive-emotional load. Koo et al. found causal “why” explanations beat procedural “how” ones for trust and performance, with both together safest but heavier on attention [11]. Since negative affect tends to raise load, an anxious passenger may do better with a short, causal explanation while a calm one can take more detail—an

emotion-tuned policy no fixed interface can express. Finally, trust itself looks affect-dependent. Endsley showed trust collapses once the car breaks the user’s mental model [21], and Reig et al. traced distrust to not understanding the car’s intent [9]; both, though, treat trust mainly as a matter of reliability and transparency, and neither asks whether a repair—an apology, a correction, a visible change in driving—works better or worse depending on how the passenger feels when the breach lands. The threads point the same way. Emotion–interaction and emotion–trust are plausible and partly backed by adjacent findings, yet no one has built them into a single system. Emotion-knowledge-tuned models like DialogueLLM [24], which condition on inferred affect, and the wider case for in-vehicle affective computing from Eyben et al. [3], suggest the pieces now exist separately—which makes the missing integrated framework a matter of design, not of absent technology.

Nowhere does this bite harder than in a moving car. For an anxious passenger, a curt braking alert can make things worse; a word matched to how they feel can take the edge off. The lesson is that an explanation system has to read more than the road—it has to read the passenger, and shift the content, tone, and timing of what it says to match. Joining live emotion perception to LLM-generated explanation, however, is still an open problem and one the field urgently needs to work.

4.5. Summary

This section has reviewed the main advances in interaction and trust for autonomous driving. The work on dynamic trust, context-aware design, and explanation is substantial, yet three

limits keep recurring: the content is pre-scripted, the channel carries little, and passenger emotion is left out. Together they name one gap—how to build a system that senses how a passenger feels, generates a personal explanation on the fly, and holds a multi-turn exchange. Section 5 takes up the challenges and the road ahead for getting there, read through the lens of joining these strands.

5. Integration Challenges and Future Prospects

Sections 2 through 4 reviewed, in turn, multimodal emotion recognition, LLM-based driving decisions, and human-machine interaction. Each pointed to the same thing—the missing coupling among them, the field’s core gap. This section works through the technical hurdles to joining the three into a closed loop, then sets out where the research should go next.

5.1. Technical Gaps in the Emotion-Decision-Interaction Closed-Loop System

The fragmentation traces back to the three disconnects named in Section 1. Their combined effect is a strange asymmetry: a car that decides safely through heavy traffic yet cannot sense how its passenger feels, and so cannot bend its driving or its explanations to match. Closing that means building the loop—emotion perception, then strategy generation, then a driving adjustment, then a plain-language account of it (Figure 3). That loop is the goal these gaps point toward [4, 1, 21].

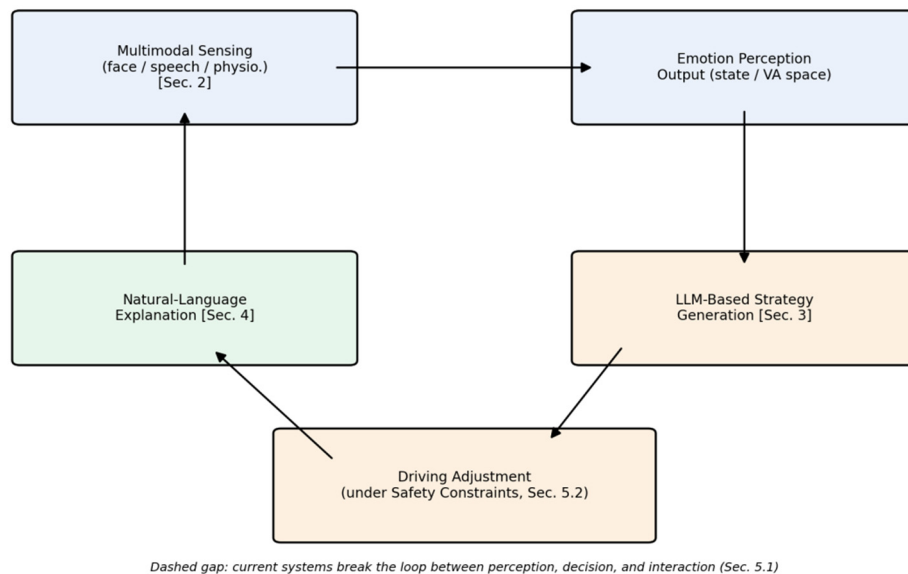


Figure 3. Emotion perception–decision–interaction closed-loop architecture

5.2. Key Challenges: Real-time Performance, Safety, and Personalization

Integrating these three elements to build an emotion-aware autonomous driving system faces the following key challenges.

5.2.1. Real-Time Performance

Both multimodal perception—video streams and physiological analysis especially—and LLM inference are heavy on compute. On a car, where onboard resources are tight and functional-safety demands are strict, running the

whole chain from emotion recognition through LLM inference to strategy generation inside a millisecond budget is a real engineering knot. Ways out exist: shrink the model through distillation, pruning, and quantization; split the load across edge and cloud; or build inference hardware made for affective-computing work.

5.2.2. Safety Constraints

Any emotion-driven adjustment must stay inside hard safety limits. Take a passenger who grows impatient under time pressure: mapping that straight onto more aggressive driving—higher acceleration, tighter following distance—would breach those limits, because it raises collision risk

outright. The signal has to be turned instead into a bounded response that keeps within functional-safety bounds. The design question, then, is how to build a verification layer that checks whether an LLM’s emotion-adaptive strategy is legitimate and holds its parameters in range, so nothing emotion-driven ever crosses the line set by vehicle dynamics or traffic law. Invariant-set methods and Robust MPC from control theory offer a starting toolkit.

5.2.3. Personalization and Privacy

Passengers differ, and not by a little—in where their emotional thresholds sit, how they show feeling, and what driving style they prefer [4, 6]. A pace one person finds easy can leave another tense. So the system has to model the individual, taking in feedback and updating its emotion-to-strategy map online to track each person. There is a second edge to this. Facial video, voice, and physiological traces are among the most sensitive biometric data there is. Getting consent, keeping storage local, and anonymizing under data-protection law is not a side issue—it is an ethical and legal condition the integrated system has to meet from the start.

5.3. Future Research Directions

5.3.1. Unified Multimodal Emotion-Decision Framework

One direction is to encode what the emotion module outputs—say, a continuous point in valence-arousal space [5]—as structured input the LLM can reason over, so passenger emotion enters strategy generation directly. The open questions cluster there: how to map continuous emotional values into language the model reads, how sensitive an LLM is to numeric emotional input, and how to jointly encode emotion with environmental perception. From there, an end-to-end framework joining perception and decision could carry the whole chain, from recognizing emotion to generating a driving strategy.

5.3.2. LLM-Based Dynamic Explainable Interaction

An LLM’s language generation and multi-turn dialogue [1, 2] open the way to an explanation module the passenger can actually engage. Rather than only receiving a canned account of what the car did, they could ask a follow-up about a specific event and get an answer shaped by the conversation so far and by how they seem to feel. The hard parts are known: holding context across turns, balancing detail against brevity, and giving the system a way to voice uncertainty when a situation is ambiguous, so an overconfident explanation does not quietly erode trust.

5.3.3. Lightweight Design and In-Vehicle Deployment

Small LLMs like LLaMA-7B and Qwen-1.8B, paired with on-device inference frameworks, have brought lightweight models within reach of in-vehicle hardware. That opens a concrete task. The next step is to design small models built for one joint job—sensing emotion and deciding how to drive. Task-specific fine-tuning, architecture search, and compression can cut latency while the core reasoning survives, which is what putting a closed-loop emotion system on a real car actually requires.

5.3.4. Dataset Construction in Natural Driving Environments

Most in-vehicle emotion datasets come from labs or simulators. Few annotate emotion, driving behavior, and system decisions together for real-world driving. Field collection could change that—recording a passenger’s multimodal physiological and behavioral data, the vehicle’s motion, and the passenger’s own read on specific events at

once. The payoff is an open benchmark to train and test against.

5.3.5. Phased Research Roadmap for Closed-Loop System Development

From all of this, we suggest three phases. The first sets the emotion-decision interface: a standard format—structured prompt templates, for instance—by which the recognition module hands emotional information to the LLM, tested in small controlled scenarios to confirm the model responds to emotional input at all. The second builds an emotion-adaptive prototype: with an LLM at the core, an in-vehicle dialogue system that reshapes its explanation’s content and tone to the passenger’s live state, checked against real users for whether it actually builds trust. The third moves to the field and to open data—gathering joint emotion, decision, and interaction data on real roads to form a shared benchmark for training and evaluation. The phases feed one another, and the endpoint is a deployable, emotion-aware closed loop.

5.4. Summary

This section worked through the main hurdles to joining emotion perception, LLM decision-making, and interaction—latency, safety, personalization—and set out four directions: unified framework design, dynamic explainable interaction, lightweight deployment, and dataset construction. None of it is settled. Closing the gap between sensing emotion and deciding how to drive, and building a car that puts the passenger first, is one of the field’s defining problems and a central route to better human-machine collaboration.

6. Conclusion

Emotion-aware autonomous driving is a young, cross-disciplinary field. What this review contributes is a first attempt to hold multimodal emotion recognition, LLM-based decisions, and interaction-and-trust in one frame and read the break between them plainly—emotion perception halts at a label, LLM decisions carry no representation of how the passenger feels, and interaction stays fixed and mood-blind. We treat that break not as a local flaw in any one line of work but as the fundamental bottleneck on human-machine collaboration in today’s systems. That coupled, cross-domain reading is what single-track reviews have not offered.

References

- [1] Cui, C., Ma, Y., Cao, X., Ye, W., & Wang, Z. (2024). A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (pp. 1–10). Waikoloa, HI, USA. <https://doi.org/10.1109/WACVW60836.2024.00002>.
- [2] Yang, Z., Jia, X., Li, H., & et al. (2024). LLM4Drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*. <https://doi.org/10.48550/arXiv.2311.01043>.
- [3] Eyben, F., Wollmer, M., Poitschke, T., & et al. (2010). Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in Human-Computer Interaction*, 2010, 263473. <https://doi.org/10.1155/2010/263473>.
- [4] Zepf, S., Hernandez, J., Schmitt, A., Minker, W., & Picard, R. W. (2020). Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys*, 53(3), 1–30. <https://doi.org/10.1145/3381008>.

- [5] Wang, Y., Song, W., Tao, W., & et al. (2021). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83–84, 1–32. <https://doi.org/10.1016/j.inffus.2022.02.011>.
- [6] Dillen, N., Ilievski, M., Law, E., Nacke, L. E., Czarnecki, K., & Schneider, O. (2020). Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–13). Honolulu, HI, USA. <https://doi.org/10.1145/3313831.3376247>.
- [7] Ahn, M., Brohan, A., Brown, N., & et al. (2022). Do as I can, not as I say: Grounding language in robotic affordances. In *Proceedings of the Conference on Robot Learning (CoRL)* (pp. 287–318). <https://doi.org/10.48550/arXiv.2204.01691>.
- [8] Huang, W., Xia, F., Xiao, T., & et al. (2022). Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of the Conference on Robot Learning (CoRL)* (pp. 1–15). <https://doi.org/10.48550/arXiv.2207.05608>.
- [9] Reig, S., Norman, S., Morales, C. G., Das, S., Steinfeld, A., & Forlizzi, J. L. (2018). A field study of pedestrians and autonomous vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)* (pp. 1–10). Toronto, ON, Canada. <https://doi.org/10.1145/3229889.3234043>.
- [10] Ayoub, J., Zhou, F., Bao, S., & et al. (2019). From manual driving to automated driving: A review of 10 years of AutoUI. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '19)* (pp. 1–22). Utrecht, Netherlands. <https://doi.org/10.1145/3342197.3351888>.
- [11] Koo, J., Kwac, J., Ju, W., & et al. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing*, 9(4), 269–275. <https://doi.org/10.1007/s12008-014-0227-2>.
- [12] Koelstra, S., Muhl, C., Soleymani, M., & et al. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 18–31. <https://doi.org/10.1109/T-AFFC.2011.15>.
- [13] Vaswani, A., Shazeer, N., Parmar, N., Polosukhin, I., & Jones, L. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 5998–6008). Long Beach, CA, USA. <https://doi.org/10.48550/arXiv.1706.03762>.
- [14] Li, S., & Deng, W. (2018). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3), 1465–1491. <https://doi.org/10.1109/T-AFFC.2018.2871424>.
- [15] Wang, K., Peng, X., Yang, J., Lu, S., & Qiao, Y. (2020). Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6896–6905). Seattle, WA, USA. <https://doi.org/10.1109/CVPR42600.2020.00693>.
- [16] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9), 3046. <https://doi.org/10.3390/s21093046>.
- [17] Mao, J., Qian, Y., Zhao, H., & Wang, Y. (2023). GPT-driver: Learning to drive with GPT. *arXiv preprint arXiv:2310.01415*. <https://doi.org/10.48550/arXiv.2310.01415>.
- [18] Sha, H., Mu, Y., Jiang, Y., & et al. (2023). LanguageMPC: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*. <https://doi.org/10.48550/arXiv.2310.03026>.
- [19] Xu, Z., Zhang, Y., Xie, E., & et al. (2023). DriveGPT4: Interpretable end-to-end autonomous driving via large language models. *arXiv preprint arXiv:2310.05730*. <https://doi.org/10.48550/arXiv.2310.05730>.
- [20] Chen, L., Sinavski, O., Hünermann, J., & et al. (2023). Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1–19). London, United Kingdom. <https://doi.org/10.1109/ICRA48891.2023.10161128>.
- [21] Endsley, M. R. (2017). Automatic driving systems: A preliminary naturalistic study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making*, 11(4), 339–354. <https://doi.org/10.1177/1555343417711490>.
- [22] Colley, M., Eder, B., Rixen, J. O., & Rukzio, E. (2021). Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–11). Yokohama, Japan. <https://doi.org/10.1145/3411764.3445315>.
- [23] Sanches, P., Janson, A., Karpashevich, P., & et al. (2019). HCI and affective health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–17). Glasgow, UK. <https://doi.org/10.1145/3290605.3300470>.
- [24] Zhang, Y., Wang, M., Wu, Y., & et al. (2024). DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*. <https://doi.org/10.48550/arXiv.2310.11374>.