Beyond Perception: A Survey and Future Directions of VLM-based Autonomous Driving Datasets

Chuheng Wei*, *Graduate Student Member, IEEE*, Hao Cheng, Ziyan Zhang,
Guoyuan Wu, *Senior Member, IEEE*, Ziran Wang, *Senior Member, IEEE*,
Matthew J. Barth, *Fellow, IEEE*

***Abstract*—Vision-Language Models (VLMs) have emerged as a transformative paradigm for autonomous driving, enabling semantic understanding, interpretable decision-making, and natural human-vehicle interaction beyond traditional perception-centric approaches. However, the scarcity of comprehensive datasets that integrate visual perception with linguistic annotations has become a critical bottleneck for advancing VLM-based autonomous driving research. This survey presents the first systematic review of VLM-based autonomous driving datasets, establishing a multi-dimensional taxonomy based on modality composition (image/video/multi-sensor-text), task coverage (perception/prediction/planning/reasoning), and annotation methodology (human/LLM/hybrid). We provide detailed analyses of 14 representative datasets, revealing their design choices, capabilities, and limitations across diverse applications, including visual question answering, instruction following, scene reasoning, and interactive dialogue. Furthermore, we discuss fundamental challenges facing this emerging field, including language ambiguity and visual grounding, cross-modal temporal alignment, annotation scalability, domain generalization, evaluation standardization, and ethical considerations. Finally, we outline four promising future directions: human-in-the-loop dataset construction, multi-sensor VLM integration, explainable decision reasoning, and unified evaluation frameworks. This survey aims to provide researchers with a comprehensive reference for understanding current progress and guide future developments toward semantically grounded, interpretable, and human-aligned autonomous driving systems.**

I. Introduction

Autonomous driving has evolved from modular perception pipelines based on specialized models for detection [1], segmentation [2], and tracking [3] to integrated systems capable of semantic understanding. While traditional perception approaches excel at geometric reasoning by outputting structured representations such as bounding boxes and trajectories, they fundamentally lack the ability to comprehend semantic contexts, interpret nuanced scenarios, and communicate understanding in natural language [4]–[6]. Vision-Language Models (VLMs) such as CLIP [7], BLIP [8], and LLaVA [9] address these limitations by bridging visual perception with linguistic reasoning through contrastive learning and cross-modal alignment. In autonomous driving, VLMs enable interpretable decision-making through natural language explanations [10], intuitive human-vehicle interaction via conversational interfaces [11], and complex reasoning that leverages commonsense knowledge [6], [12], [13], which are essential for handling long-tail situations beyond pure recognition.

However, the development of VLM-based autonomous driving systems faces a critical bottleneck: the scarcity of appropriate datasets. Existing large-scale AD datasets including KITTI [1], nuScenes [3], Argoverse [14], and BDD100K [15], which were primarily designed to support traditional perception tasks and provide rich sensor data (cameras, LiDAR, radar) with extensive geometric annotations (3D bounding boxes, semantic labels, HD maps). Yet they lack the comprehensive language annotations necessary for training and evaluating VLMs: semantic scene descriptions that capture contextual relationships, question-answer pairs that test multi-level understanding, natural language instructions for interactive control, reasoning explanations that articulate decision rationales, and conversational dialogues that enable human-vehicle communication. Recent efforts such as Talk2Car [11] and DriveLM [16] have begun to address this gap, but these datasets remain limited in scale, diversity of tasks, and annotation quality. The field lacks standardized benchmarks and evaluation protocols for assessing VLM performance in driving contexts.

This dataset infrastructure gap has profound implications for advancing VLM-based autonomous driving research. Without comprehensive, high-quality vision-language datasets, progress is constrained across multiple fronts: training end-to-end models that jointly perceive, understand, and reason about driving scenarios; developing interpretable systems that can explain their decisions in natural language; enabling sophisticated human-vehicle interaction through conversational AI; and establishing robust evaluation frameworks that measure both perceptual accuracy and language understanding quality. Furthermore, the diverse design choices in existing VLM datasets, including annotation methodologies (human-labeled vs. LLM-generated), task formulations (captioning vs. QA vs. reasoning), and data modalities (image-text vs. video-text vs. multi-sensor-text), necessitate systematic analysis to guide future dataset construction and model development. A comprehensive survey of VLM-based AD datasets is therefore essential to synthesize current efforts, identify best practices and limitations, establish taxonomies for dataset comparison, and chart directions for advancing this emerging field.

Contributions. This survey makes the following contributions:

- We provide the first systematic review of VLM-

Chuheng Wei, Hao Cheng, Ziyan Zhang, Guoyuan Wu, and Matthew J. Barth are with the College of Engineering, Center for Environmental Research and Technology, University of California at Riverside, Riverside, CA, 92507 USA.

Ziran Wang is with the College of Engineering, Purdue University, West Lafayette, IN, 47907 USA.

*Corresponding author. e-mail: chuheng.wei@email.ucr.edu

based autonomous driving datasets, establishing a multi-dimensional taxonomy based on modality composition, task coverage, and annotation methodology.
- We comprehensively review 14 datasets through a proposed taxonomy, identify key challenges in cross-modal alignment, annotation scalability, domain generalization, and evaluation standardization, and establish a roadmap for addressing these limitations.
- We outline promising future directions encompassing human-in-the-loop construction, multi-sensor integration, explainable reasoning, and unified evaluation frameworks.

II. Background and Foundations

A. Vision-Language Models (VLMs)

Vision-Language Models (VLMs) have emerged as a transformative paradigm that bridges visual perception and linguistic reasoning through large-scale multimodal pretraining. Early works such as CLIP [7] demonstrated the power of contrastive learning to align images and textual descriptions within a shared embedding space, enabling zero-shot recognition and open-vocabulary understanding. BLIP [8] extended this concept with multimodal pretraining objectives that unify image-text contrastive, matching, and captioning tasks, producing high-quality bidirectional alignment between visual and textual modalities. Subsequent models like LLaVA [9] integrated large language models (LLMs) such as Vicuna or LLaMA with visual encoders, allowing conversational understanding of images through visual instruction tuning. More recently, GPT-4V and other large-scale multimodal foundation models have pushed the boundary further by enabling reasoning over complex visual scenes and interacting with users in natural language dialogues.

Conceptually, VLMs operate on three key principles: 1) *Contrastive learning* aligns visual and textual representations by pulling semantically related pairs closer in embedding space while pushing unrelated ones apart, fostering robust semantic grounding; 2) *Cross-modal alignment* fuses heterogeneous modalities (visual tokens extracted from images and linguistic tokens from text) through transformer-based fusion or joint attention mechanisms, facilitating bidirectional understanding; 3) *Visual tokenization* converts spatial features into discrete embeddings, often through patch-level encoders (e.g., Vision Transformer [17]), enabling language models to process visual information analogously to text tokens.

In the context of autonomous driving, these capabilities open new research frontiers. VLMs can power *vision-language question answering* for interpreting dynamic traffic scenes, *semantic annotation* for automated labeling of complex interactions, *intention understanding* for driver behavior analysis, and *explainable reasoning* for natural language justification of vehicle actions. Moreover, they enable *human-vehicle interaction* via multimodal conversational interfaces, allowing users to query, instruct, or receive explanations in an intuitive and interpretable way, which represents an essential step toward socially aware and transparent autonomous driving systems.

B. Autonomous Driving Datasets and Multimodality

The development of autonomous driving systems has long relied on large-scale datasets that capture real-world diversity through multimodal sensor suites. Benchmark datasets such as KITTI [1], nuScenes [3], Argoverse [14], and BDD100K [15] provide synchronized data from multiple sensors, such as RGB cameras, LiDAR, radar, and GPS/IMU, alongside detailed annotations for detection, segmentation, tracking, and map-level localization. These resources have become foundational for perception and prediction research by supporting geometric reasoning and robust model evaluation.

However, despite their multimodality, these datasets remain primarily geometry-centric, focusing on visual and spatial annotations while lacking semantic and linguistic layers that describe scene intent or contextual relationships. They do not include high-level annotations such as textual scene descriptions, question-answer pairs, or decision rationales that could support VLM training and evaluation. Recent efforts have begun addressing this limitation: Talk2Car [11] introduced referring expression comprehension for localized natural language commands, and DriveLM [16] incorporated LLM-generated dialogues for scene understanding and reasoning. Nonetheless, these datasets remain limited in scale, scenario diversity, and annotation richness compared to traditional perception datasets.

This gap highlights the need for VLM-based autonomous driving datasets that integrate multimodal sensor data with linguistic annotations to enable joint perception, reasoning, and interaction, advancing autonomous driving toward semantically grounded and human-understandable autonomy.

III. Taxonomy of VLM-based Datasets for Autonomous Driving

VLM-based autonomous driving datasets can be systematically categorized along three major dimensions: **(1) modality composition**, which defines the sensory and linguistic sources of data; **(2) task coverage**, which determines the target application scope from perception to reasoning; and **(3) annotation methodology**, which reflects how multimodal labels are generated and aligned. Figure 2 illustrates the overall taxonomy framework, while Table I summarizes representative datasets according to these dimensions.

A. Modality Composition

The first dimension concerns the integration of visual, spatial, and linguistic modalities. Existing datasets can be categorized into three types based on their sensory inputs:

- **Image-text**: Early works such as Talk2Car [11] adopt this structure, where single RGB frames are paired with natural language commands referring to specific objects or actions.
- **Video-text**: More recent datasets, e.g., DriveLM [16], move toward this configuration, allowing temporal reasoning across sequential frames.
- **Multi-sensor-text**: Emerging research explores this fusion approach, incorporating LiDAR point clouds,

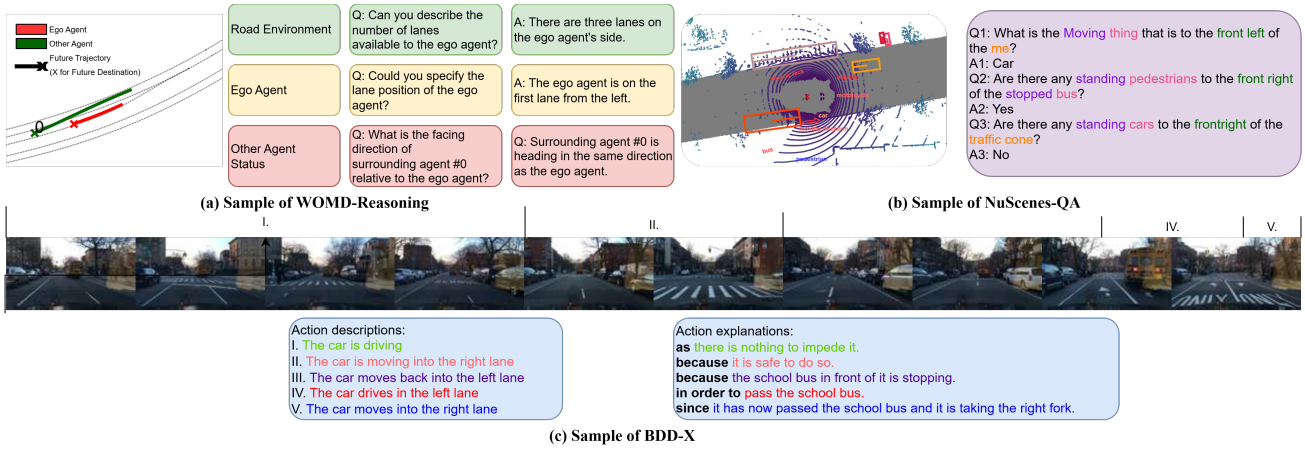

Fig. 1: Representative samples from three VLMbased autonomous driving datasets. **(a)** WOMD-Reasoning: multi-agent interaction reasoning with traffic rule-aware questions. **(b)** NuScenes-QA: template-based visual question answering for perception attributes. **(c)** BDD-X: action explanation linking driver decisions to visual observations.

BEV representations, and GPS/IMU signals aligned with linguistic descriptions to provide a comprehensive perception-language foundation.

Such multimodal integration enables VLMs to associate textual semantics not only with visual appearance but also with spatiotemporal and geometric context, representing a crucial step toward scene-level understanding.

B. Task Coverage

VLM-based datasets can also be categorized by the cognitive level of their target tasks:

- **Perception-level tasks**: image or scene captioning, referring expression comprehension, and object grounding that link natural language phrases to visual regions.
- **Prediction-level tasks**: describing dynamic behaviors, motion intentions, or environmental changes in text, bridging low-level perception and high-level semantics.
- **Planning and reasoning tasks**: answering "why" and "what-if" questions, generating natural language rationales for driving decisions, and supporting dialogue-based control or instruction following.

This hierarchical perspective highlights how datasets progressively evolve from passive visual description toward active reasoning and interaction. Future dataset designs are expected to integrate these levels into unified benchmarks that support end-to-end multimodal understanding.

C. Annotation Methodology

The third dimension involves how visual-linguistic pairs are created and verified. Existing datasets can be broadly divided into three categories:

- **Human-labeled**: manually annotated captions, QAs, or dialogues ensuring linguistic naturalness and contextual grounding (e.g., Talk2Car).
- **LLM-generated**: large language models used to automatically generate scene descriptions or QA pairs from metadata and sensor logs (e.g., DriveLM).
- **Hybrid**: combining automatic generation with human verification to balance scalability and accuracy.

Human annotation guarantees semantic precision but is costly and limited in scale, while LLM-based synthesis provides scalability but risks hallucination and weak grounding. Hybrid pipelines that integrate human-in-the-loop validation are becoming an emerging trend for constructing high-quality multimodal datasets.

IV. Vision-Language Driving Datasets

Diverse datasets have recently emerged to bridge vision and language in autonomous driving, pairing multimodal sensory inputs with linguistic annotations. Table I summarizes representative datasets by modality, task type, annotation methodology, and scale, forming the foundation for multimodal alignment and semantic reasoning in intelligent vehicles. Figure 1 illustrates representative examples demonstrating the diversity of annotation types across different cognitive levels.

a) BDD-X [18]: BDD-X pairs dashcam videos with human explanations of driver behavior. Each annotation contains a short description of the action and an explicit reason, for example "slowed down because the light is red" or "changed lane to avoid a parked car." The language is concise and casual, and it refers to events that the camera actually saw. Even though it is older than the recent large multimodal corpora, it is still the most convenient resource for teaching or testing models to verbalize why the ego moved the way it did, which is why later driving-language work keeps reusing it.

b) Talk2Car [11]: Built upon the nuScenes dataset, Talk2Car pairs real-world urban driving images with human-annotated language commands that refer to specific objects or regions in the scene. Each instruction corresponds to a concrete visual target, enabling models to learn fine-grained referring expression comprehension. This dataset provides a foundation for linking human verbal intent to visual perception in driving systems, promoting research into interpretable and interactive vehicle control.

TABLE I: Taxonomy of Representative VLM-based Autonomous Driving Datasets

Dataset	Year	Venue	Modality	Task Type	Annotation	Scale
BDD-X [18]	2018	ECCV	Video-Text	Planning & Reasoning	Manual	26K
Talk2Car [11]	2019	EMNLP-IJCNLP	Image-Text	Perception-level	Manual	11K
DRAMA [19]	2023	WACV	Video-Text	Prediction-level	Manual	17.8K
VLAAD [20]	2024	WACV-W	Video-Text	Planning & Reasoning	LLM	64K
DriveLM [16]	2024	ECCV	Video-Text	Prediction-level	LLM	100K
LingoQA [21]	2024	ECCV	Image-Text	Perception-level	Hybrid	60K
NuScenes-QA [22]	2024	AAAI	Image/LiDAR-Text	Perception-level	LLM	460K
NuScenes-MQA [23]	2024	WACV-W	Image-Text	Perception-level	Hybrid	1.46M
doScenes [24]	2024	arXiv	Video-Text	Planning & Reasoning	Manual	1K
OmniDrive [25]	2025	CVPR	Multi-view-Text	Planning & Reasoning	Hybrid	–
CoVLA [26]	2025	WACV	Video-Text	Planning & Reasoning	LLM	10K
Impromptu VLA [27]	2025	NeurIPS D&B	Video-Text	Planning & Reasoning	Hybrid	80K
WOMD-Reasoning [28]	2025	ICML	Motion/Map-Text	Prediction-level	LLM	2.94M
NuPlanQA [29]	2025	ICCV	Multi-view-Text	Planning & Reasoning	Hybrid	1M

c) DRAMA [19]: DRAMA focuses on the safety tail. It contains urban scenarios in which annotators point out the risky agent or region and write a sentence explaining the risk, such as a pedestrian emerging from occlusion or a vehicle cutting in from the side. All labels are human. Because these cases are rare in ordinary driving datasets, DRAMA is often used as a complementary benchmark to check whether a vision–language model actually attends to hazards and can articulate them, instead of only describing the easy, well-exposed parts of the scene.

d) VLAAD [20]: VLAAD offers language samples grounded in driving videos and scene metadata. Using detailed per-frame context as a prompt, the authors ask a strong language model to generate three types of outputs: short descriptions of the scene, user–assistant style instructions about what to do, and simple reasoning questions whose answers can be checked. Because the generation is always conditioned on the actual driving scene, the language stays on topic. Tasks are also labeled, so users can report separate numbers for description, instruction following, and reasoning. This makes VLAAD convenient when we want a single source to fine-tune or evaluate instruction-style driving VLAs.

e) DriveLM [16]: DriveLM extends multimodal reasoning in autonomous driving by framing perception, prediction, and planning tasks as graph-based visual question answering. Instead of isolated perception or control tasks, it formulates a dialogue-style reasoning process grounded in driving videos, where each question-answer pair reflects the decision-making flow of an autonomous driving pipeline. Annotated using large language models with human verification, DriveLM encourages the development of unified vision-language models capable of handling hierarchical reasoning and task decomposition in dynamic traffic environments.

f) LingoQA [21]: LingoQA provides a large-scale multimodal question answering benchmark tailored for autonomous driving videos. It encompasses a diverse collection of scene-level queries that cover perception, reasoning, and contextual understanding, thereby bridging low-level visual recognition and high-level semantic comprehension. Each question is manually annotated to ensure diversity and linguistic richness, promoting more generalizable multimodal understanding. LingoQA serves as an essential resource for evaluating vision-language models under realistic driving conditions where temporal reasoning and context awareness are critical.

g) NuScenes-QA [22]: NuScenes-QA starts from the observation that nuScenes already contains a complete 3D description of each frame, so language supervision can be generated directly from that structure. They write templates covering autonomous vehicle concerns such as object existence, quantity, relative position to ego, agent motion, weather, and time, then fill them using nuScenes annotations. Because everything is programmatically derived, researchers can easily map model errors to specific perception attributes, making it useful as an additional diagnostic layer on top of nuScenes perception.

h) NuScenes-MQA [23]: NuScenes-MQA keeps nuScenes as the visual base but changes the format so that one annotation can serve both generation and QA. Each item is a natural sentence where the answer is embedded and then marked with tags, so a model can be trained to output a full sentence and we can still check whether the answer span is correct. To make this work at scale, the authors combine rule-based construction from nuScenes metadata with LLM rewriting. Compared with plain template QA, the language here is longer and more varied, and compared with free-form captioning, it is still automatically gradable. That makes it a good training/evaluation source for driving VLMs that are expected to "speak" rather than only output a single token.

i) doScenes [24]: doScenes looks at short urban driving clips and asks annotators to write what the car should do right now, in natural language, while staying grounded in what is visible. Instead of just saying "a truck is on the right," labels say "slow down before the truck on the right" or "wait until the pedestrian finishes crossing," sometimes with an explicit reference to an agent. The scale is modest compared with synthetic QA, but every instruction is human-written and directly actionable. This makes the dataset fill a very specific hole: most nuScenes language works describe or query, but do not tell the ego how to act; doScenes gives exactly that supervision and is therefore handy for language-to-driving and

for interactive assistants in cars.

j) OmniDrive [25]: OmniDrive is designed so that a single driving scene supports multiple language objectives. For each scene, the authors extract structured descriptions of agents, lanes, and relations, then generate multiple textual views: neutral scene descriptions, perception questions, and "what if" questions built from alternative trajectories. Generation uses rules for scene fidelity and an LLM for language richness, with human filtering for quality. The result is a corpus enabling models to perform ordinary VQA, higher-level reasoning, and counterfactual thinking within one dataset, particularly useful for pretraining driving VLAs before task-specific fine-tuning.

k) CoVLA [26]: CoVLA provides real driving clips with paired supervision: dense scene descriptions and corresponding ego actions or trajectories. The labels are produced by an automatic pipeline that uses sensor data and a vision–language model to generate scene text, enabling the dataset to scale beyond fully manual annotation while remaining tied to actual video. A key difference from caption-style datasets is that CoVLA explicitly links observation to behavior; descriptions explain what is happening, and actions specify how to respond. This makes it well-suited for training and benchmarking vision–language–action models for driving.

l) Impromptu VLA [27]: Impromptu VLA targets the messy side of human instructions in cars. The authors collect clips from several open driving sources, focus on situations that are more difficult or atypical, and then use a VLM plus human filtering to attach language that asks for a driving decision. Unlike in template-based sets, prompts here are often underspecified ("can I turn?", "is it clear on the right?"), so the model has to connect the utterance to the visual evidence. The answers are short and action-oriented. This makes the dataset valuable for testing robustness: models that only saw clean, fully specified questions usually stumble on this kind of real phrasing, while a model that handles Impromptu VLA well is more likely to work in an actual in-car dialogue.

m) WOMD-Reasoning [28]: WOMD-Reasoning builds directly on the motion histories and HD maps from the Waymo Open Motion Dataset and asks questions that arise in multi-agent traffic: who has priority at this junction, which car will yield, whether a collision is plausible under current intent, and why. The pipeline first turns the structured scene into symbolic facts, applies traffic and interaction rules, and then prompts a large model to phrase the questions and answers. The QAs are grouped into factual, interaction, and intention levels so that we can evaluate progressively. Because supervision is derived from actual trajectories rather than from single images, the dataset covers behaviors that ordinary driving VQA almost never touches, making it a strong candidate for training prediction- and rule-aware vision–language models.

n) NuPlanQA [29]: NuPlanQA reframes nuPlan's multi-view episodes as language queries about scene dynamics and ego behavior. Episodes are reorganized into conditionable formats (images, maps, trajectories, ego state), then an LLM generates questions covering scene description, nearby agents, traffic rules, safety, and planner intent, with low-quality samples filtered by annotators. Originating from a planning benchmark, the questions explicitly probe decision rationale and traffic compliance, distinguishing it from generic VQA for testing reasoning beyond object

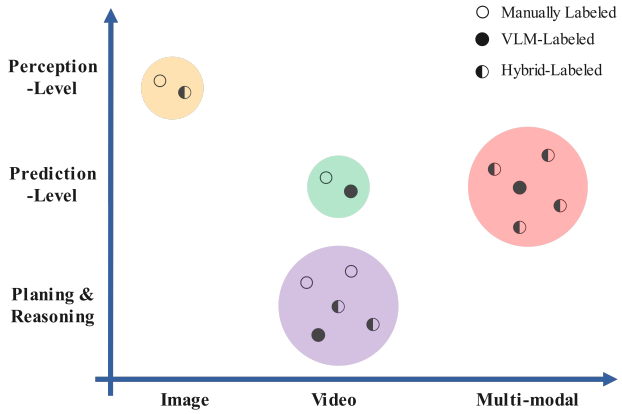

Fig. 2: Taxonomy of VLM-based Autonomous Driving Datasets: Distribution across modality, task level, and annotation methodology.

V. Discussion and Remaining Challenges

Building on the datasets summarized above, we now reflect on the current state of VLM-based autonomous driving. This discussion evaluates how existing datasets and models address different levels of multimodal perception, reasoning, and interaction, and where their limitations still lie. The section concludes by outlining both technical and non-technical challenges that define the next phase of research and development.

A. Discussion

VLM-based research in autonomous driving has achieved substantial progress in bridging vision and language through multimodal learning. As illustrated in Table I and Figure 2, the landscape of VLM-based datasets has evolved significantly over the past seven years. The chronological progression reveals a clear shift from small-scale, human-annotated datasets focused on perception tasks (Talk2Car, BDD-X with 11K-26K samples) toward large-scale, LLM-assisted corpora supporting diverse cognitive levels (NuScenes-MQA, WOMD-Reasoning with up to 2.94M samples). This scaling trend has been accompanied by a methodological transition: among the 14 datasets reviewed, only 5 (36%) remain purely human-labeled, while 6 (43%) adopt hybrid strategies and 3 (21%) rely on fully automated generation, reflecting both the economic necessity of scaling annotation and the maturation of LLM capabilities.

Figure 2 spatially organizes datasets along modality composition (x-axis) and task cognitive level (y-axis), revealing distinct clustering patterns. Image-text datasets concentrate on perception-level tasks, reflecting their origins in object grounding and referring expression comprehension.

Video-text datasets exhibit the broadest distribution, spanning from planning and reasoning (BDD-X [18], doScenes [24], CoVLA [26]) to prediction-level tasks (DriveLM [16], DRAMA [19]), demonstrating the versatility of temporal visual information. Multi-modal datasets cluster predominantly in the planning and reasoning space [30], suggesting that integration of heterogeneous sensors (LiDAR, HD maps, motion trajectories) naturally supports higher-level decision-making. However, this spatial distribution also exposes significant gaps: prediction-level tasks receive comparatively less attention, with only three datasets explicitly targeting motion intention understanding and multi-agent interaction reasoning. Moreover, no existing dataset comprehensively integrates all three cognitive levels of perception, prediction, and planning within a unified benchmark.

Nevertheless, several technical and non-technical obstacles continue to impede VLM-based autonomous driving from reaching full maturity. From a technical perspective, current datasets exhibit limitations in multimodal diversity, temporal grounding accuracy, and scalability, whereas models frequently encounter issues pertaining to alignment, data efficiency, and generalization. From a broader perspective, the field must address ethical, privacy, and trust concerns to ensure responsible deployment. The following subsections delineate these remaining challenges in detail.

B. Technical Challenges

a) Ambiguity and Contextual Grounding. Natural language employed in driving scenarios frequently exhibits vagueness and strong context dependence [31]. Expressions such as "the car in front" or "the pedestrian on the right" may reference distinct entities contingent upon temporal context, ego-vehicle position, or map topology. Although current datasets capture spatial annotations, few explicitly encode pragmatic cues or temporal dependencies. Future research should incorporate contextual grounding signals, including reference trajectories or ego-motion information, to facilitate model disambiguation of language expressions within dynamic scenes.

b) Multimodal and Temporal Alignment. Synchronizing visual, spatial, and linguistic modalities constitutes a persistent challenge [4]. Camera, LiDAR, and radar sensors operate at disparate frequencies, resulting in temporal misalignment between linguistic tokens and visual frames. This temporal discrepancy compromises the coherence of multimodal reasoning, particularly for dialogue-based or causal explanation tasks. Temporal transformers and adaptive cross-modal attention mechanisms represent promising approaches for enabling continuous alignment across heterogeneous data streams.

c) Data Efficiency and Annotation Scalability. Training large-scale VLMs necessitates substantial quantities of paired data, yet acquiring human-verified text annotations for driving scenes remains costly and time-intensive. Recent approaches exploit large language models for automatic caption generation [32] or employ parameter-efficient fine-tuning techniques [33]. However, these methods may introduce hallucinated descriptions or inconsistencies. A hybrid pipeline combining automatic generation with targeted human verification could balance scalability and quality, while active learning strategies may further reduce annotation redundancy.

d) Realism and Generalization. Synthetic or simulation-based datasets often fail to capture the linguistic variability and environmental complexity characteristic of real-world traffic scenarios [34]. Models trained on curated or LLM-generated language may overfit to idealized phrasing and exhibit poor performance with spontaneous human instructions. Addressing this realism gap requires incorporating in-the-wild driving videos, implementing cross-domain adaptation techniques, and developing language grounding approaches robust to noisy, imperfect sensor conditions.

e) Evaluation Frameworks and Standardization. Current evaluation practices remain fragmented and lack consistency. Whereas perception tasks employ established metrics such as mAP and IoU, no consensus exists regarding the assessment of multimodal reasoning, dialogue quality, or interpretability [35]. Developing standardized benchmarks that jointly evaluate visual perception accuracy, linguistic coherence, and reasoning validity represents a critical research priority. Such benchmarks would facilitate fair model comparisons, promote open competitions, and guide dataset design toward unified multimodal evaluation protocols.

f) Personalized and Human-aligned Modeling. Existing VLM systems treat all drivers and queries uniformly, neglecting individual behavioral and linguistic variations [36]. Driving decisions are inherently influenced by personal preferences, risk tolerance, and situational interpretation. Incorporating personalized representations that link visual perception with driver intent, affective states, or linguistic patterns could enable adaptive, human-aligned reasoning [37]. This research direction bridges VLM development with driver behavior modeling and offers potential for enhancing comfort, safety, and interpretability.

C. Non-technical Challenges

a) Ethical and Bias Concerns. Language models inevitably inherit biases from pretraining corpora [38], [39]. When deployed in autonomous driving systems, these biases may manifest through descriptive preferences or decision rationales that inadvertently reflect cultural or demographic stereotypes. Bias-aware dataset auditing, balanced annotation strategies, and explainable reasoning pipelines are essential for ensuring fairness and accountability across diverse driving environments.

b) Privacy and Data Governance. Compliance with privacy regulations such as General Data Protection Regulation (GDPR) is imperative [40]. Notably, GDPR requires prior consent for recording individuals, rendering post-hoc anonymization legally insufficient. Dataset construction pipelines must therefore integrate consent-based data collection alongside technical safeguards including real-time visual anonymization (face and license plate blurring), spatial obfuscation, and differential privacy to protect both participants and bystanders.

c) Human Factors and Trust. The interpretability afforded by VLMs introduces novel interaction dynamics between humans and artificial intelligence systems [41]. Although natural language explanations enhance transparency, users may exhibit over-reliance or misinterpret model outputs when they appear human-like yet contain factual errors. Establishing calibrated trust models, confidence-aware explanation systems, and adaptive user interfaces will be essential for safe integration of these technologies into real-world driving environments.

d) Standardization and Collaboration. The current ecosystem of VLM-based driving datasets exhibits fragmentation, with substantial variation in annotation methodologies, data formats, and evaluation criteria. This lack of interoperability impedes progress and comparability across research efforts [42]. Developing open standards analogous to those established for perception benchmarks such as nuScenes or Waymo Open Dataset would encourage reproducibility and cross-institutional collaboration. International consortia and shared infrastructure can accelerate dataset construction while ensuring consistent quality and broader accessibility.

Collectively, these technical and non-technical challenges underscore the complexity of extending VLMs to safety-critical autonomous driving applications. Addressing these challenges necessitates a multidisciplinary effort integrating advances in multimodal learning, ethics, data governance, and human-centered design to achieve interpretable, adaptive, and trustworthy autonomy.

VI. Future Directions

Building upon current progress and identified challenges, we outline four key research directions for advancing VLM-based autonomous driving datasets and systems.

*a) **Human-in-the-Loop Dataset Construction**:* While LLM-generated annotations provide scalability, they often suffer from hallucination and weak visual grounding. Conversely, purely human-annotated datasets remain costly and limited in scale. A hybrid approach that combines automated generation with targeted human verification can balance quality and efficiency. Active learning frameworks can identify uncertain samples for human review, while confidence-calibrated LLMs handle straightforward cases. Interactive annotation interfaces that allow annotators to refine LLM outputs rather than creating labels from scratch can significantly reduce cognitive load and construction costs.

*b) **Multi-sensor VLM Integration**:* Current VLM datasets predominantly focus on camera-text pairs, neglecting information from LiDAR, radar, and bird's-eye-view (BEV) representations. Future research should explore multi-sensor VLMs that align textual descriptions with heterogeneous sensory inputs. LiDAR-text alignment can enable language-grounded 3D object reasoning, while BEV-text fusion can support scene-level spatial relationship understanding. Integrating HD map priors with language can further facilitate geographic and rule-aware reasoning, bridging perception-centric systems toward semantically aware, explainable autonomy.

*c) **VLM-based Explainable Decision Reasoning**:* Beyond perception and prediction, VLMs hold promise for interpretable planning and decision-making. Future datasets should incorporate counterfactual reasoning annotations that explain not only what the ego vehicle did, but also what it could have done under alternative conditions and why certain actions were chosen or rejected. Training VLMs on such rationale-rich data can enable vehicles to generate human-understandable justifications for decisions, thereby enhancing transparency, trust, and debuggability in safety-critical applications.

*d) **Open-ended Evaluation Frameworks**:* The lack of standardized evaluation metrics hinders progress and comparability. Future work should develop unified evaluation frameworks that jointly measure visual grounding accuracy, linguistic fluency, reasoning validity, and task-specific performance. Benchmark suites tailored to driving contexts should encompass diverse tasks including VQA, instruction following, counterfactual reasoning, and interactive dialogue. Incorporating human-centric dimensions such as interpretability and trust calibration will ensure VLMs align with real-world deployment requirements.

VII. Conclusion

This paper presents the first systematic survey of VLM-based autonomous driving datasets. We established a comprehensive taxonomy categorizing datasets by modality composition, task coverage, and annotation methodology, and provided detailed reviews of 14 representative datasets spanning image-text, video-text, and multi-sensor-text configurations. Our analysis reveals significant progress in bridging vision and language for driving applications, from early referring expression comprehension to recent large-scale question answering and reasoning datasets.

We critically discussed fundamental challenges facing this emerging field, including language ambiguity and visual grounding, cross-modal temporal alignment, annotation scalability and quality trade-offs, domain gaps between synthetic and real data, lack of standardized evaluation benchmarks, and ethical concerns regarding bias and privacy. We further outlined four promising future directions: human-in-the-loop dataset construction, multi-sensor VLM integration, explainable decision reasoning, and open-ended evaluation frameworks. By fostering standardization and collaboration, the community can advance VLM-based autonomous driving toward safe, transparent, and human-aligned autonomy.

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[4] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao *et al.*, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 958–979.

[5] C. Wei, G. Wu, and M. J. Barth, "Cooperative perception for automated driving: A survey of algorithms, applications, and future directions," *Proceedings of the IEEE*, 2025.

[6] X. Zhou, M. Liu, E. Yurtsever, B. L. Zagar, W. Zimmer, H. Cao, and A. C. Knoll, "Vision language models in autonomous driving: A survey and outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[8] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[10] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024.

[11] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M. F. Moens, "Talk2car: Taking control of your self-driving car," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2088–2098.

[12] C. Pan, B. Yaman, T. Nesti, A. Mallik, A. G. Allievi, S. Velipasalar, and L. Ren, "Vlp: Vision language planning for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 760–14 769.

[13] Z. Qin, X. Yao, C. Wei, A. Ji, G. Wu, and Z. Sun, "Contextual-personalized adaptive cruise control via fine-tuned large language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 1754–1762.

[14] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.

[15] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[16] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *European conference on computer vision*. Springer, 2024, pp. 256–274.

[17] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[19] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 1043–1052.

[20] S. Park, M. Lee, J. Kang, H. Choi, Y. Park, J. Cho, A. Lee, and D. Kim, "Vlaad: Vision and language assistant for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 980–987.

[21] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton *et al.*, "Lingoqa: Visual question answering for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 252–269.

[22] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4542–4550.

[23] Y. Inoue, Y. Yada, K. Tanahashi, and Y. Yamaguchi, "Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 930–938.

[24] P. Roy, S. Perisetla, S. Shriram, H. Krishnaswamy, A. Keskar, and R. Greer, "doscenes: An autonomous driving dataset with natural language instruction for human interaction and vision-language navigation," *arXiv preprint arXiv:2412.05893*, 2024.

[25] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *CoRR*, 2024.

[26] H. Arai, K. Miwa, K. Sasaki, K. Watanabe, Y. Yamaguchi, S. Aoki, and I. Yamamoto, "Covla: Comprehensive vision-language-action dataset for autonomous driving," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 1933–1943.

[27] H. Chi, H.-a. Gao, Z. Liu, J. Liu, C. Liu, J. Li, K. Yang, Y. Yu, Z. Wang, W. Li *et al.*, "Impromptu vla: Open weights and open data for driving vision-language-action models," *arXiv preprint arXiv:2505.23757*, 2025.

[28] Y. Li, C. Fan, S. Z. Zhao, C. Li, C. Xu, H. Yao, M. Tomizuka, B. Zhou, C. Tang, M. Ding *et al.*, "Womd-reasoning: A large-scale dataset for interaction reasoning in driving," in *Forty-second International Conference on Machine Learning*, 2024.

[29] S.-Y. Park, C. Cui, Y. Ma, A. Moradipari, R. Gupta, K. Han, and Z. Wang, "Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models," *arXiv preprint arXiv:2503.12772*, 2025.

[30] C. Wei, Z. Qin, Z. Zhang, G. Wu, and M. J. Barth, "Integrating multi-modal sensors: A review of fusion techniques for intelligent vehicles," in *2025 IEEE Intelligent Vehicles Symposium (IV)*, 2025, pp. 1817–1824.

[31] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, "The rise and potential of large language model based agents: A survey," *Science China Information Sciences*, vol. 68, no. 2, p. 121101, 2025.

[32] N. Petrovic, K. Lebioda, V. Zolfaghari, A. Schamschurko, S. Kirchner, N. Purschke, F. Pan, and A. Knoll, "Llm-driven testing for autonomous driving scenarios," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, 2024, pp. 173–178.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[34] Z. Qin, S. Li, C. Wei, G. Wu, M. J. Barth, A. Abdelraouf, R. Gupta, and K. Han, "Investigating personalized driving behaviors in dilemma zones: Analysis and prediction of stop-or-go decisions," *IEEE Robotics and Automation Letters*, 2025.

[35] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente *et al.*, "Holistic evaluation of text-to-image models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 69 981–70 011, 2023.

[36] C. Wei, Z. Qin, S. Li, Z. Zhang, X. Zhao, A. Abdelraouf, R. Gupta, K. Han, M. J. Barth, and G. Wu, "Pdb: Not all drivers are the same–a personalized dataset for understanding driving behavior," *arXiv preprint arXiv:2503.06477*, 2025.

[37] C. Wei, Z. Qin, G. Wu, M. J. Barth, A. Abdelraouf, R. Gupta, and K. Han, "Dilemma zone: A comprehensive study of influential factors and behavior analysis," in *2024 Forum for Innovative Sustainable Transportation Systems (FISTS)*. IEEE, 2024, pp. 1–8.

[38] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[39] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100211, 2024.

[40] F. D. Protection, "General data protection regulation (gdpr)," *Intersoft Consulting, Accessed in October*, vol. 24, no. 1, 2018.

[41] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2425–2452, 2022.

[42] J. Janai, F. Güney, A. Behl, A. Geiger *et al.*, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and trends® in computer graphics and vision*, vol. 12, no. 1–3, pp. 1–308, 2020.