Commun Transp Res

Research Article

ScenePilot-4K: A Large-Scale First-Person Dataset and Benchmark for Vision-Language Models in Autonomous Driving

Yujin Wang[1,†], Yutong Zheng[1,†], Wenxian Fan[1,†], Tianyi Wang[2], Hongqing Chu[1], Li Zhang[3], Bingzhao Gao[1,✉], Daxin Tian[4], Jianqiang Wang[5], Hong Chen[6]

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640069

ABSTRACT: In this paper, we introduce ScenePilot-4K, a large-scale first-person dataset for safety-aware vision-language learning and evaluation in autonomous driving. Built from public online driving videos, ScenePilot-4K contains 3,847 hours of video and 27.7M front-view frames spanning 63 countries/regions and 1,210 cities. It jointly provides scene-level natural-language descriptions, risk assessment labels, key-participant annotations, ego trajectories, and camera parameters through a unified multi-stage annotation pipeline. Building on this dataset, we establish ScenePilot-Bench, a standardized benchmark that evaluates vision-language models along four complementary axes: scene understanding, spatial perception, motion planning, and GPT-based semantic alignment. The benchmark includes fine-grained metrics and geographic generalization settings that expose model robustness under cross-region and cross-traffic domain shifts. Baseline results on representative open-source and proprietary vision-language models show that current models remain competitive in high-level scene semantics but still exhibit substantial limitations in geometry-aware perception and planning-oriented reasoning. Beyond the released dataset itself, the proposed annotation pipeline serves as a reusable and extensible recipe for scalable dataset construction from public Internet driving videos.

KEYWORDS: autonomous driving, vision-language models, dataset, benchmark

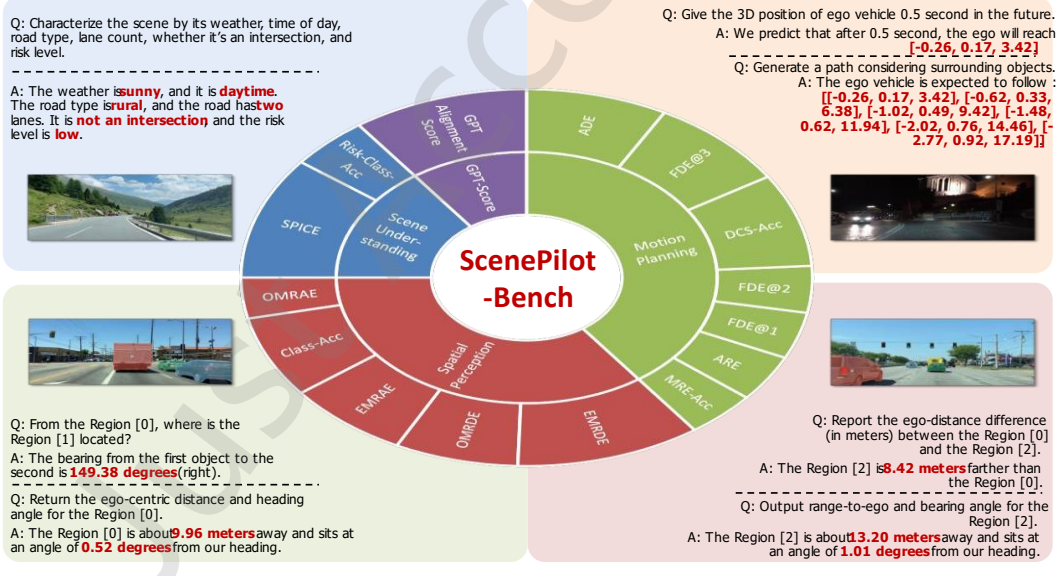


Fig. 1. The overall structure of ScenePilot-Bench, which is established based on the ScenePilot-4K dataset and emphasizes four complementary dimensions for VLM evaluation in autonomous driving: Scene Understanding, Spatial Perception, Motion Planning, and GPT-Score.

1 Introduction

Recent progress in multimodal understanding and reasoning has drawn growing interest in vision-language models (VLMs) for autonomous driving. Compared with conventional perception modules, VLMs offer a unified interface for describing scenes, reasoning about risks, identifying traffic participants, and providing decision-related explanations. The requirements of autonomous driving, however, differ substantially from those of general-domain

† Yujin Wang, Yutong Zheng, and Wenxian Fan contributed equally to this work.

[1] College of Automotive and Energy Engineering, Tongji University, Shanghai 201804, China. [2] Department of Civil, Architectural, and Environmental Engineering, The University of Texas at Austin, Austin 78712, USA.[3] School of Data Science, Fudan University, Shanghai 200438, China.[4] School of Transportation Science and Engineering, Beihang University, Beijing 100191, China. [5] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China. [6] College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China.

✉ Corresponding author. E-mail: gaobz@tongji.edu.cn

Received: June 1, 2026; Revised: August 12, 2026; Accepted: November 2, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

Just Accepted

multimodal evaluation. Describing visible content is only part of their role in a real driving scene. A useful model should also preserve spatial relations relative to the ego vehicle, reason about safety-critical participants, and provide planning-oriented interpretations across diverse road environments and traffic conventions.

Existing autonomous driving datasets and benchmarks differ in the tasks they support and the annotations they provide. nuScenes (Caesar et al., 2020) and nuPlan (Caesar et al., 2021) provide rich sensor data and annotations for perception, motion, and planning. BDD100K (Yu et al., 2020) focuses on large-scale visual perception. OpenDV-2K (Yang et al., 2024) provides first-person driving videos from diverse geographic regions, but does not align risk labels, metric object geometry, ego trajectories, and camera calibration at the clip level. DriveLM (Sima et al., 2024) and DriveBench (Xie et al., 2025) focus on language-grounded reasoning and reliability evaluation. None of these resources provides all of these annotations for the same first-person clips under a common evaluation protocol.

To address this problem, we introduce ScenePilot-4K, a large-scale first-person dataset for vision-language learning and evaluation in autonomous driving. We collected all source videos from YouTube and Bilibili, then trimmed and filtered them for quality. The final release contains 3,847 hours of video and 27.7M front-view frames from 63 countries/regions and 1,210 cities. Using a unified multi-stage annotation pipeline, we provide scene-level natural-language descriptions, risk assessment labels, key-participant annotations, ego trajectories, and camera parameters for the same clips. These annotations share a common clip-level data structure that links semantic, spatial, and planning-related supervision, allowing a broader evaluation of multimodal models for autonomous driving.

We build ScenePilot-Bench on ScenePilot-4K as a standardized benchmark for evaluating VLMs in first-person autonomous driving scenarios (Fig. 1). It covers four complementary axes: scene understanding, spatial perception, motion planning, and GPT-Score. The benchmark is designed to assess not only semantic plausibility but also driving-relevant abilities, including risk-sensitive reasoning, object-level spatial perception, and trajectory-oriented prediction. For models to be deployed in practice, they need to remain robust across regions with different traffic conventions and visual environments. We therefore include geographic generalization settings to test how well they meet this requirement.

We report baseline results for representative VLMs to support reproducible comparisons. The results show that these models generally perform well on high-level scene semantics but still have limitations in spatial perception, embodied reasoning, and cross-region transfer. These findings indicate that generic captioning or narrowly defined visual question answering (VQA) tasks alone are insufficient for large-scale multimodal evaluation in autonomous driving. Such evaluation should cover semantics, spatial perception, and motion-related reasoning. We therefore propose ScenePilot-4K and ScenePilot-Bench as an integrated dataset and benchmark for fine-tuning and evaluating VLMs in scene understanding, spatial perception, motion planning, and semantic alignment for autonomous driving.

Our contributions are as follows:

- We introduce ScenePilot-4K, a large-scale, geographically diverse first-person driving dataset with unified multi-task annotations, and provide a reproducible clip-level pipeline for annotation, quality control, and scalable dataset construction.
- We establish ScenePilot-Bench on ScenePilot-4K, with four evaluation axes and extensible protocols that support customized geographic generalization studies.
- We report baseline and geographic generalization results for representative VLMs, providing empirical evidence of their capability gaps and identifying open challenges in driving-oriented multimodal reasoning.

2 Related Work

Dedicated driving datasets are essential for improving scene understanding, spatial perception, and decision-related reasoning in autonomous driving. Early benchmarks, such as KITTI (Geiger et al., 2013) and Cityscapes (Cordts et al., 2016), focus on core perception and urban semantics, while larger-scale resources, including Waymo Open Dataset (Sun et al., 2020) and BDD100K (Yu et al., 2020), extend coverage to complex traffic scenes and multi-task video annotations. Language-grounded and risk-aware driving data have also emerged in grounding and command-following settings (Deruyttere et al., 2022; W. Li et al., 2021) and in risk-centric datasets such as DRAMA and DRAMA-ROLISP (Ding et al., 2025; Malla et al., 2023). For motion forecasting and planning, datasets such as nuScenes (Caesar et al., 2020), nuPlan (Caesar et al., 2021), and Argoverse 2 (Wilson et al., 2023) support trajectory prediction with multi-agent dynamics and map context. More recent efforts, such as OpenDV-2K (Yang et al., 2024), further highlight the importance of scalable and diverse first-person driving data.

Systematic benchmarks are equally important for evaluating how vision-language models can be used in autonomous driving. Existing studies mainly follow two paradigms: open-loop and closed-loop. Open-loop benchmarks assess perception and reasoning quality without directly controlling the vehicle, as represented by DriveLM (Sima et al., 2024), DriveBench (Xie et al., 2025), and object- or reasoning-centric benchmarks such as NuPrompt (Wu et al., 2025) and STRIDE-QA (Ishihara et al., 2026). Closed-loop evaluation instead measures the downstream control effect of model outputs, as exemplified by BENCH2ADVLM (Zhang et al., 2025). However, current VLMs more often serve as auxiliary modules for semantic reasoning, spatial grounding, and planning separately. Current driving datasets and benchmarks are typically organized around either sensor-based perception and motion or language-based reasoning. Their annotations and evaluation protocols are therefore not directly aligned at the clip level. ScenePilot-4K provides a shared first-person representation for scene semantics, risk assessment, key participants, metric spatial relations, ego trajectories, and camera parameters. ScenePilot-Bench evaluates the corresponding semantic, geometric, and planning capabilities under a common protocol.

3 ScenePilot-4K Dataset

3.1. Overview of the Dataset

We present ScenePilot-4K, a large-scale first-person dataset tailored for training and evaluating VLMs in autonomous driving. The dataset supports VLM training and establishes a unified benchmark for evaluating model performance across diverse driving conditions. To our knowledge, ScenePilot-4K is among the largest first-person, pure-vision driving video datasets with unified multi-task annotations in the field. It is constructed from high-quality first-person driving videos collected from Bilibili and YouTube. Building upon the publicly available OpenDV-2K dataset (Yang et al., 2024), which consists of over 1,700 hours of YouTube driving footage, we have augmented it with approximately 2,000 additional hours of driving videos. As a result, the dataset contains 3,847 hours of driving videos.

All videos are processed through a high-quality annotation pipeline. Specifically, raw videos are uniformly sliced at a fixed sampling rate, and each segment is annotated using a VLM for scene understanding and risk assessment. Key traffic participants are identified using YOLO11s (Jocher and Qiu, 2024). Furthermore, we integrate algorithms such as VGGT (J. Wang et al., 2025) to estimate both intrinsic and extrinsic camera parameters, as well as to generate accurate trajectory annotations.

Under Review

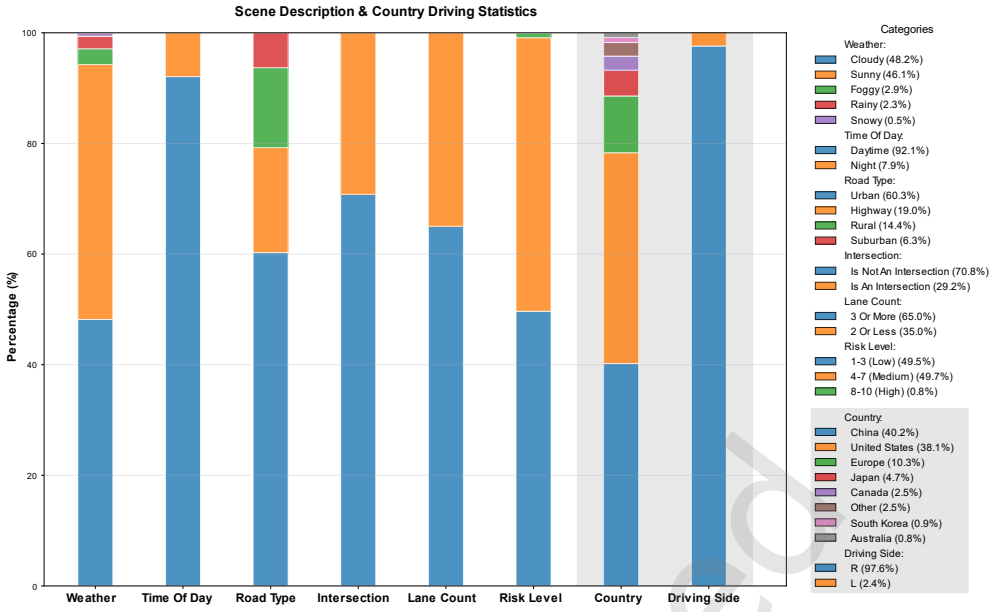


Fig. 2. ScenePilot-4K Dataset Statistics: Distribution of Scene Attributes and Driving Countries. This figure summarizes the overall distribution of scene attributes and geographic coverage in the ScenePilot-4K dataset.

Table 1. Comparison of Scale and Annotation Richness across Driving Datasets.

Dataset	Duration	Front-view Frames	Countries & Regions	Cities	NL Scene Desc.	Risk Assess.	Key Partic.	Ego Traj.	Cam. Params
KITTI (Geiger et al., 2013)	1.4h	15K	1	1	×	×	√	√	√
nuScenes (Caesar et al., 2020)	5.5h	214K	2	2	×	×	√	√	√
Waymo Perception (Sun et al., 2020)	11h	390K	1	3	×	×	√	√	√
Waymo Open Motion (Ettinger et al., 2021)	574h	20.7M	1	6	×	×	√	√	√
BDD100K (Yu et al., 2020)	1,000h	318K	1	2	×	×	√	√	×
Cityscapes (Cordts et al., 2016)	0.5h	25K	3	50	×	×	×	√	√
Argoverse 2 (Wilson et al., 2023)	4.2h	300K	1	6	×	×	√	√	√
ApolloScape (Huang et al., 2020)	103min	12K	1	1	×	×	√	√	√
nuPlan (Caesar et al., 2021)	1,282h	62.5M	2	4	×	×	√	√	√
ONCE (Mao et al., 2021)	approx. 144h	approx. 1M	1	-	×	×	√	√	√
Lyft Level5 (Houston et al., 2021)	approx. 1,000h	approx. 62.5M	1	1	×	×	√	√	√
Talk2Car (Deruyttere et al., 2022)	4.7h	9.2K	2	2	×	×	√	√	√
OpenDV-2K (Yang et al., 2024)	2,059h	65.1M	≥40	≥244	√	×	×	×	×
ScenePilot-4K (ours)	3,847h	27.7M*	63**	1,210**	√	√	√	√	√

Note: NL denotes natural-language. * Frame sampling rate is 2 Hz. ** Metadata-based and manually verified

3.1.1. Diversity and Advantages over Prior Relevant Datasets

As shown in Table 1, the ScenePilot-4K dataset offers substantial diversity and several advantages over prior relevant datasets, namely in the following aspects:

Scale. ScenePilot-4K aggregates 3,847 hours of driving videos with 27.7M front-view frames, making it among the largest video corpora for autonomous driving. Although some planning datasets report tens of millions of frames, none of them pair scale with the multi-task, language-grounded supervision offered here.

Geographic diversity. The dataset spans 63 countries/regions and 1,210 cities, far exceeding the predominantly single-country coverage of most benchmarks (often 1-6 cities). Even compared with geographically diverse prior work reporting coverage of ⩾40 countries and ⩾244 cities, ScenePilot-4K covers 63 countries/regions and 1,210 cities. This breadth naturally captures diverse road infrastructures, signage systems, traffic behaviors, and driving conventions (e.g., left- vs. right-hand traffic), supporting the evaluation of out-of-distribution robustness. The resulting variation in traffic rules, road design, and driving behavior also provides the basis for the cross-region and cross-traffic evaluations in Section 5.4.

Annotation richness. Among the datasets compared in Table 1, ScenePilot-4K jointly provides all of the following on the same clips: (i) scene-level natural-language descriptions, (ii) risk assessment labels, (iii) key-participant identification, (iv) ego-trajectory, and (v) camera parameters. In contrast, prior datasets typically offer only perception labels (boxes/segments) or metadata-derived ego trajectories, and when language or risk information is available it is usually limited ("only labels/metadata") and not aligned with trajectory or calibration information. This unified, multi-granularity supervision enables learning objectives that bridge vision-language understanding, risk assessment, agent saliency, and planning-aware prediction within a single benchmark.

3.1.2. Detailed Dataset Statistics

Fig. 2 shows the range of real-world conditions represented in

Just Accepted

ScenePilot-4K. The scenario statistics broadly reflect everyday driving. Clear weather is most common, while nighttime driving accounts for a small but non-negligible share. We deliberately retain a long-tailed risk distribution, including a small but important proportion of high-risk clips for stress-testing perception and reasoning. The dataset covers countries and regions with both right- and left-hand traffic, supporting out-of-distribution evaluation across regions.

3.2. The High-quality Annotation Pipeline

We use established pretrained models to annotate each first-person clip. Their outputs are aligned in time and organized in a common data structure. The procedure is described below:

1) Video Clipping and Frame Sampling: We trim the first and last 180 seconds from each raw first-person driving video, then uniformly sample the remaining footage at 2 Hz. The sampled frames are grouped into non-overlapping 5-second clips, each containing 10 frames. The implementation supports large-scale batch processing with a resume function and maintains a consistent data structure throughout.

2) Scene Description and Risk Assessment: We automatically generate a semantic scene description and a risk level score for each video clip using a cross-modal annotation module based on Qwen2-VL-72B-Instruct (P. Wang et al., 2024). The model receives only the 4th frame of each clip as the representative input, without additional temporal information, and assigns an ordinal risk score from 1 (safest) to 10 (most dangerous). The prompt asks the model to assess risk using the visual cues in that frame, including participant proximity and spatial configuration, apparent road occupancy, potential path conflicts inferred from the visible scene, road and intersection complexity, visibility, and environmental conditions. The score reflects the context visible in the representative frame rather than a single distance or speed threshold. Scores of 1 – 3 indicate low risk, 4 – 7 medium risk, and 8 – 10 high risk. The scene description follows this format: *The weather is sunny, and it is daytime. The road type is urban and the road has four lanes. It is an intersection, and the risk level score is 4.*

3) Key Participant Detection and Annotation: We use YOLO11s (Jocher and Qiu, 2024) to detect traffic participants. Because object scale and visual distinctiveness vary substantially across categories, each category $c \in C$ is assigned its own confidence threshold τ_c, where:

$$C = \{\text{vehicle, truck, bicycle, motorcycle, pedestrian}\} \quad (1)$$

The thresholds are set to $\tau_{vehicle} = 0.5, \tau_{truck} = 0.6, \tau_{bicycle} = 0.4, \tau_{motorcycle} = 0.55$, $\tau_{pedestrian} = 0.55$. To balance recall and false positives, we use lower thresholds for frequent, reliably detected classes and stricter filtering for rarer or more ambiguous ones.

For each detected object, we also store its ID, label, and normalized bounding box in a JSON record that can be combined directly with the camera intrinsics and extrinsics.

4) Camera Calibration and Ego-Trajectory Annotation: We use the pre-trained VGGT model (J. Wang et al., 2025) to estimate camera intrinsics, extrinsics, and the ego-trajectory at each frame. These estimates support frame-level geometric reasoning and future trajectory annotation without external sensors.

VGGT jointly infers the camera parameters from the 10 consecutive frames $\{I_t\}_{t=1}^{10}$ in each clip. It returns a pose encoding for each frame, which we convert into camera matrices using calibrated functions.

VGGT returns a world-to-camera extrinsic matrix for each frame. From its rotation and translation components, we recover the camera center in the reconstructed world coordinate system. We then shift all positions so that the camera center in the fourth frame serves as the reference origin for the clip:

$$\hat{\boldsymbol{C}}_t = (-R_t^T t_t) - (-R_4^T t_4), \quad (2)$$

where R_4 and t_4 are the rotation and translation components for frame $t = 4$. We apply small positional offsets for stability. This gives a sequence of $T = 10$ positions in the re-centered world coordinate system:

$$\gamma_{ego} = \{\hat{\boldsymbol{C}}_t\}_{t=1}^{10} \in \mathbb{R}^{10\times 3}. \quad (3)$$

5) Further Processing: We apply lightweight geometric post-processing after VGGT estimates the camera intrinsics, extrinsics, and ego-trajectory for each frame. This stage lifts 2D evidence into 3D, recovers metric scale, and extracts the foregrounds of key participants. It also computes robust ego-centric distances, azimuths, and inter-agent proximities for use in VLM/VLA training and evaluation. The procedure is fully automatic and runs on each clip.

3D lifting from monocular geometry. For a pixel $u = (u, v)$ in frame t, we use its estimated depth $\hat{Z}_t(u)$ to back-project it into 3D:

$$p_t^c(u) = \hat{Z}_t(u) K_t^{-1} \tilde{u}, \quad \tilde{u} = [u, v, 1]^\top, \quad (4)$$

where $p_t^c(u) \in \mathbb{R}^3$ is the resulting point in camera coordinates, expressed in VGGT units. Here, $K_t \in \mathbb{R}^{3\times 3}$ is the camera intrinsic matrix, and $\tilde{\mathbf{u}}$ is the homogeneous pixel coordinate.

The point can optionally be transformed into the world coordinate system:

$$\tilde{p}_t^w(u) = T_t^{c\to w} [p_t^c(u)^\top, 1]^\top, \quad (5)$$

where $T_t^{c\to w} \in SE(3)$ is the camera-to-world transform, and $\tilde{\mathbf{p}}_t^w(\mathbf{u}) = [(\mathbf{p}_t^w)^\top, 1]^\top$ is the point in homogeneous world coordinates.

Metric scale recovery (frame-level with object-level fallback). We convert VGGT's arbitrary scale into metric distances by first estimating a ground-based scale for each frame, with class-specific object priors as an optional fallback. To estimate the ground scale, we sample points on a grid S in the lower half of the image and back-project each point into camera coordinates, $p_{t,k}^c = (x_k, y_k, z_k)$. We then select ground inliers using robust statistics on the absolute vertical coordinates $|y_k|$:

$$m = \text{median}(\{|y_k|\}), \quad (6)$$

$$\text{MAD} = \text{median}(||y_k| - m|), \quad (7)$$

$$\mathcal{G} = \{k: ||y_k| - m| \le \kappa \cdot \text{MAD}\}, \quad (8)$$

where $\kappa > 0$ is a fixed threshold (we use $\kappa \approx 2.5$).

For each target i in frame t, we select a single metric scale $s_{t,i}$ using the frame-level estimate when reliable and the object-level estimate otherwise:

$$s_{t,i} = \begin{cases} \dfrac{H_{cam}}{\text{median}(|y_k|: k \in \mathcal{G})}, & \text{if the frame-level estimate is reliable,} \\ \dfrac{H_{c(i)}}{(h_{px}/f_v)\tilde{z}}, & \text{otherwise,} \end{cases} \quad (9)$$

where $H_{cam} > 0$ is the nominal camera height, $H_{c(i)}$ is the canonical physical height for class c(i), h_{px} is the bounding-box height in pixels, f_y is the vertical focal length, and $\tilde{z}$ is the median foreground depth (in VGGT units) for object i. In practice, if $\mathcal{G}$ is empty or unstable, we approximate the denominator term in the first branch using m instead of $\text{median}(|y_k|: k \in \mathcal{G})$. The resulting scale $s_{t,i}$, expressed in meters per VGGT unit, is applied to all subsequent ego-centric and inter-object distance measurements.

Foreground extraction for key participants. We use each detected bounding box as a prompt for SAM (Kirillov et al., 2023) to obtain a binary mask $M_{t,i} \subset \Omega$. The mask is then refined using lightweight morphological operations:

$$M_{t,i} = \text{Close}\big(\text{Open}(\text{SAM}(\text{bbox}, I_t))\big), \quad (10)$$

where $I_t \in \mathbb{R}^{H\times W\times 3}$ is the RGB image, and Open/Close are standard 3×3 morphological operations. For vehicles, we exclude a thin bottom strip during seeding to reduce the inclusion of road pixels. If SAM inference fails or the bounding box is overly large, we use the rectangular bounding-box mask for $M_{t,i}$. The set of 3D foreground points is

$$\mathcal{P}_{t,i} = \{p_t^c(u): u \in M_{t,i}\}. \quad (11)$$

Robust ego-centric distance and azimuth. Define per-point radii $d_j = \|p_{t,i,j}^c\|_2$ for $p_{t,i,j}^c \in \mathcal{P}_{t,i}$. We report a robust distance using a small percentile:

$$\tilde{d}_{t,i} = \text{perc}_5(\{d_j\}), \quad D_{t,i} = \max\{0, s_{t,i}\tilde{d}_{t,i}\}, \quad (12)$$

where perc_5 is the 5th percentile operator, $\tilde{d}_{t,i}$ is in VGGT units, and $D_{t,i}$ is in meters. For directional context,

Just Accepted

$$\bar{p}^{c}_{t,i} = \frac{1}{|\mathcal{P}_{t,i}|}\sum_{p\in\mathcal{P}_{t,i}} p, \quad \theta_{t,i} = \operatorname{atan2}(\bar{x},\bar{z}), \tag{13}$$

where $\bar{p}^{c}_{t,i} = (\bar{x},\bar{y},\bar{z})$ is the foreground centroid in the camera frame and $\theta_{t,i} \in \left[-\frac{\pi}{2},\frac{\pi}{2}\right]$ is the left/right azimuth (radians) relative to the forward axis.

Inter-agent proximity. For targets i and j in the same frame,

$$\delta_{i\to j} = \operatorname{perc}_5\left(\left\{\min_{q\in\mathcal{P}_{t,i}} \|p-q\|_2 : p\in\mathcal{P}_{t,j}\right\}\right), \tag{14}$$

where $\delta_{i\to j}$ is a robust, directed separation in VGGT units. The symmetric metric proximity is

$$\Delta_t(i,j) = s^{\star}_t \cdot \min\{\delta_{i\to j},\delta_{j\to i}\}, \quad s^{\star}_t = \frac{1}{2}\left(s_{t,i}+s_{t,j}\right), \tag{15}$$

where $\Delta_t(i,j)$ is in meters and $s^{\star}_t$ averages the two targets' scales.

Output. For each detection i in frame t, we save $D_{t,i}, \theta_{t,i}$, the class label, and the bounding box. A scale_info record identifies whether s^{frame}_t or $s^{\text{obj}}_{t,i}$ was used and includes diagnostics such as inlier counts and MAD. For reproducibility, we also retain the masks and 3D points, storing the points in both VGGT units and metric coordinates obtained using $s_{t,i}$.

3.3. Error Control and Propagation Analysis

Two upstream factors dominate error in the geometric annotations: camera-pose estimation and metric-scale recovery. Errors in the predicted extrinsics E alter trajectory direction, curvature, and relative displacement, whereas monocular scale ambiguity affects metric distance. Calibrated intrinsics K and metric-depth alignment address the scale component but leave the predicted poses unchanged. The released ScenePilot-4K annotations use the frame-level ground-scale estimation with object-level fallback defined in Eq.(9).

As an additional diagnostic for metric-scale calibration, we evaluate Depth Anything 3 (DA3) (Lin et al., 2025) as an alternative metric-depth reference. Let D^{metric}_{DA3} and D_{VGGT} denote the aligned DA3 and VGGT depths at a valid road pixel p. We estimate the conversion factor from VGGT units to meters as

$$s = median_{p\in\Omega_{road}}\left(\frac{D^{metric}_{DA3}(p)}{D_{VGGT}(p)}\right). \tag{16}$$

where Ω_{road} denotes the set of valid road pixels.

We compare five scale-recovery configurations on the nuScenes (Caesar et al., 2020) validation set, for which calibrated camera intrinsics and reference ego poses are available. In the "GT intrinsics" setting, the VGGT-predicted K is replaced by the calibrated nuScenes intrinsics during geometric post-processing; "GT-history scale" instead derives the scale from the reference historical ego displacement. DA3, calibrated intrinsics, and GT-history scale are evaluated only as diagnostic alternatives and are not used to annotate the released ScenePilot-4K dataset.

Table 2. Diagnostic ablation of metric-scale recovery for VGGT-based ego-trajectory annotation.

Metric	Released pipeline	GT-history scale	GT intrinsics	DA3	DA3 + GT intrinsics
ADE	1.39	0.64	1.07	0.95	0.72
FDE@1	0.75	0.30	0.65	0.51	0.41
FDE@2	1.58	0.71	1.23	1.07	0.81
FDE@3	2.49	1.22	1.76	1.70	1.25

In this diagnostic experiment on nuScenes, replacing the released pipeline's scale recovery with DA3 reduces ADE from 1.39 m to 0.95 m. Adding calibrated intrinsics further reduces ADE to 0.72 m and FDE@3 to 1.25 m, improvements of 24.2% and 26.5% relative to DA3 alone. The GT-history setting gives the lowest ADE, but it relies on reference ego motion and is included only as an oracle for scale calibration.

The remaining gap between DA3 with calibrated intrinsics and GT-history scale shows that scale calibration alone does not account for all trajectory error. Because neither DA3 nor calibrated K changes the VGGT pose estimates, the residual error is consistent with extrinsic rotation and translation error, curvature drift, accumulated frame-to-frame pose error, and depth-domain mismatch.

In the released annotation pipeline, we apply frame-level ground-scale estimation with object-level fallback, together with class-specific detection thresholds, mask-quality checks, MAD-based ground selection, percentile-based distance estimates, and trajectory-consistency filters. Clips are rejected when reconstruction fails or when scale, pose changes, or trajectory kinematics are implausible; they are not manually corrected. Rejection reasons and fallback flags are retained for audit. These checks are quality-control filters rather than a vehicle-specific dynamic-feasibility model.

3.4. Cross-Dataset Validation and Responsible Data Curation

We assess annotation reliability through cross-dataset validation on nuScenes (Caesar et al., 2020) for trajectory annotation and STRIDE-QA (Ishihara et al., 2026) for spatial perception. Failed annotations are discarded without correction, prioritizing label precision and benchmark reliability over coverage. Most annotated trajectories follow the observed ego-vehicle motion, remain aligned with the road, and form smooth paths consistent with normal driving behavior. Fig. 3 illustrates a representative successful case, where the annotated trajectory closely follows the ego vehicle's forward motion along the lane. We remove trajectories with erratic points or clear path deviations using a rule-based filtering module that applies multiple criteria to improve statistical consistency and physical validity.



Fig. 3. Accurate annotation of trajectories in a driving scene of nuScenes.

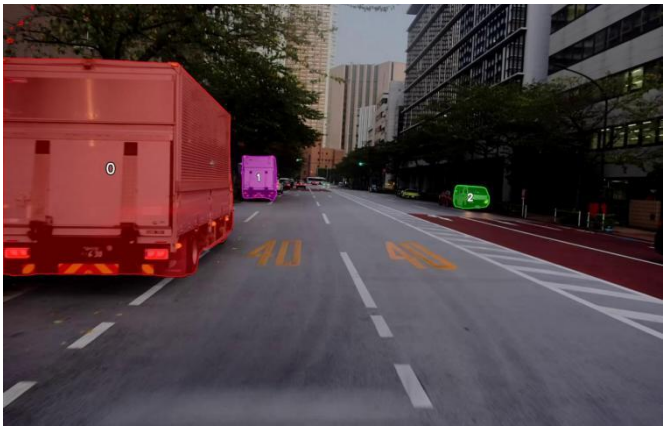


Fig. 4. Accurate annotation of spatial-perception labels.

We manually inspect 500 samples from STRIDE-QA to assess the spatial annotations. A sample fails if either its distance or azimuth error exceeds 8%. Under this criterion, the failure rate is approximately 3%.

Just Accepted

Fig. 4 illustrates a successful case with correctly localized key traffic participants and reliable spatial labels. The truck in region [2] has an annotated ego-distance of 45.61 meters, compared with a ground-truth distance of 43.47 meters, giving an error below 5%. Its annotated azimuth is −14.68 degrees, close to the ground-truth value of −15.00 degrees. These results indicate that the retained annotations are sufficiently reliable for training and evaluation.

For human verification, we selected a stratified sample of 500 clips covering adverse weather, nighttime driving, complex road layouts, all three risk levels, and multiple geographic regions. Three trained annotators reviewed each clip independently using the same rubric and then resolved disagreements through joint discussion. As shown in Table 3, the reviewers accepted scene descriptions and associated risk information for 485 clips (97.0%; two-sided 95% Wilson CI: 95.1 – 98.2%). Spatial annotations, including key-participant selection and object geometry, were accepted for 475 clips (95.0%; 92.7 – 96.6%), and ego trajectories for 465 clips (93.0%; 90.4 – 95.0%). These intervals express the uncertainty around the acceptance rates. We do not report an inter-annotator agreement coefficient because the individual judgments made before adjudication were not retained in a verifiable form.

Table 3. Human Verification Results.

Annotation type	n	Accepted	Acceptance Rate	95% Wilson CI
Scene description	500	485	97.0%	95.1-98.2%
Spatial annotation	500	475	95.0%	92.7-96.6%
Ego trajectory	500	465	93.0%	90.4-95.0%

We assign geographic labels in two stages. First, we infer preliminary region labels from video titles, descriptions, and available metadata. We then manually check these labels against visual cues, including road signs, street text, lane markings, traffic direction, landmarks, architectural style, and license plate patterns. Videos whose locations remain unclear are excluded from the geographic statistics. We also remove duplicate videos and clips with temporal discontinuities, severe occlusion from overlaid text, or poor visual quality. Visible faces and license plates are masked in the released visual examples. To support reproducibility within legal and privacy constraints, we release source-video links, clip construction logic, annotation code, prompt templates, question-answer templates, inference configurations, evaluation scripts, split files, and derived annotations. Raw videos and complete frame sequences remain accessible only through their original platforms.

3.5. Failure Case Analysis

Figs. 5-7 present representative errors from the semantic, ego-motion, and object-centric spatial branches. They are diagnostic examples of error propagation, not estimates of failure frequency.

In Fig. 5, the automatic annotation describes a three-lane suburban road with low risk, whereas manual review identifies a four-lane highway with medium risk. The error arises from an incorrect reading of both road structure and contextual risk.

Fig. 6 presents a nuScenes cross-dataset case for which reference ego poses are available. The reconstructed trajectory has an ADE of 2.77 m; FDE increases from 1.71 m at 1 second to 3.32 m at 2 seconds and 4.31 m at 3 seconds. The growing endpoint error is consistent with residual pose and scale drift over the reconstruction horizon.

Fig. 7 presents a STRIDE-QA failure for region [0]. The estimated distance is 5.13 m versus a reference of 3.99 m, a relative error of 28.5%. The estimated and reference azimuths are 18.09° and 15.58°, giving an absolute error of 2.51°.



Fig. 5. Failure case in scene-attribute annotation.

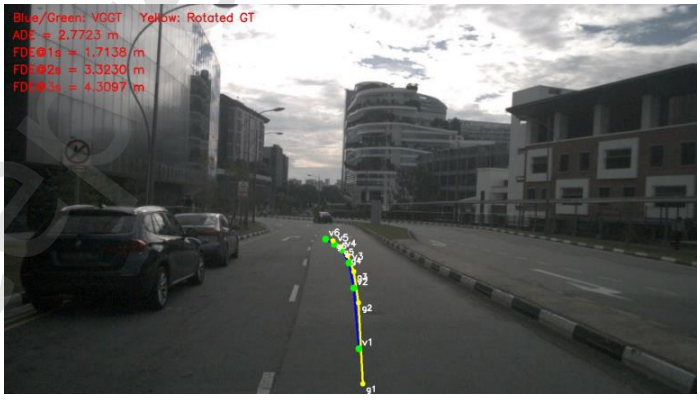


Fig. 6. Failure case in ego-trajectory annotation.

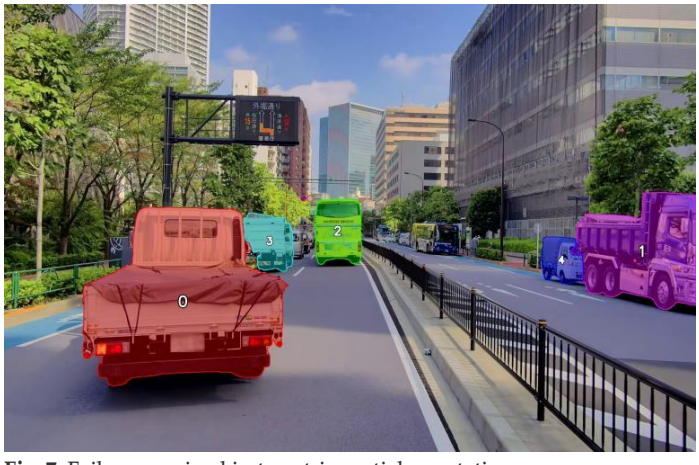


Fig. 7. Failure case in object-centric spatial annotation.

4 ScenePilot-Bench

To systematically evaluate multiple driving-related perception and reasoning capabilities of VLMs, we establish ScenePilot-Bench on top of ScenePilot-4K. Generally, we decompose VLMs' competence into four modules:

- Scene Understanding, which aims to assess VLMs' global semantic comprehension and risk reasoning ability;
- Spatial Perception, which is designed to measure the recognition and classification of key traffic participants, and spatial inference capability of VLMs;

Under Accepted

- Motion Planning, which evaluates VLMs' capability to predict future meta-actions and short-horizon ego trajectories;
- GPT-Score, which aims to evaluate the semantic alignment between the answer and the reference using AI tools.

Evaluation is organized at the clip level. Frame-level quantities, where applicable, are aggregated within each clip before computing clip-level scores, which are then averaged across clips. This clip-based protocol enhances temporal consistency and mitigates the influence of sporadic noisy frames, leading to a more stable and reliable assessment of model behaviors in real-world driving scenarios.

4.1. Scene Understanding

4.1.1. Refined SPICE

SPICE (Semantic Scene Consistency) evaluates the semantic propositional consistency between model-generated and reference scene graphs. We adopt a refined SPICE with a lightweight, Python-based parser tailored for autonomous driving. It uses a rule-based approach with domain-specific heuristics to extract key traffic participants and attribute tuples, avoiding dependency parsing. Relation tuples are deliberately omitted to reduce noise and improve robustness.

Formally, given a generated scene description c and a ground truth description S, they are both parsed into tuple sets $T(G(c))$ and $T(G(S))$. Each tuple includes a unary object tuple (o) and an attribute tuple (o, a). Precision (P), Recall (R), and the final SPICE score are defined as:

$$P(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}, \tag{17}$$

$$R(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}, \tag{18}$$

$$SPICE(c,S) = F_1(c,S) = \frac{2\,P(c,S)\cdot R(c,S)}{P(c,S)+R(c,S)}, \tag{19}$$

where $\otimes$ denotes tuple matching with stemming and synonym normalization. A higher SPICE indicates better global semantic consistency.

4.1.2. Risk-Class-Acc: Risk Reasoning Accuracy

We introduce Risk − Class − Acc to quantify VLMs' performance in risk classification tasks. Risk-Class-Acc uses the same low-, medium-, and high-risk intervals defined for annotation in Section 3.2:

$$\text{Risk}-\text{Class}-\text{Acc} = \frac{1}{N}\cdot\sum_{i=1}^{N} I\,(r_i^{pred} = r_i^{gt}), \tag{20}$$

where $I(\cdot)$ denotes the indicator function, while r_i^{pred} represents the predicted risk class and r_i^{gt} denotes the dataset-provided ground-truth risk class generated by the automatic annotation pipeline and used as the fixed evaluation target. Here, "ground truth" refers to the benchmark annotation used for evaluation rather than an independently authored human label. The reliability of these annotations is assessed through the stratified human verification described in Section 3.4. The risk classes are divided into three categories: low, medium, and high. This metric evaluates VLMs' risk classification performance against the dataset annotations.

4.2. Spatial Perception

This section assesses VLMs' spatial perception ability to detect and interpret key traffic participants, such as vehicles, trucks, pedestrians, bicycles and motorcycles, under complex traffic scenarios. It consists of two modules: Object Classification and Spatial Reasoning.

4.2.1. Object Classification

In addition to recognition accuracy, VLMs are also required to possess precise category classification capability. We compute the overall accuracy directly on the entire test set as the core evaluation metric:

$$\text{Class-Acc} = \frac{1}{N}\sum_{i=1}^{N} I\,(c_i^{pred} = c_i^{gt}), \tag{21}$$

where **N** is the total number of samples across all traffic participant categories, namely vehicle, truck, pedestrian, bicycle and motorcycle. This metric comprehensively reflects the overall recognition ability of VLMs across all categories.

4.2.2. Spatial Reasoning

To evaluate spatial reasoning, we measure relative positional and angular errors between the ego vehicle and traffic participants, as well as among participants.

Ego-centric metrics, which include Mean Relative Distance Error to Ego (EMRDE) and Mean Relative Angle Error to Ego (EMRAE), evaluate the relative position of each traffic participant with respect to the ego vehicle:

$$\text{EMRDE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{d_i} - d_i|}{d_i}, \tag{22}$$

$$\text{EMRAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{\theta_i} - \theta_i|}{\max(|\theta_i|,\epsilon)}, \tag{23}$$

where d_i and θ_i are the ground truth distance and azimuth between ith participant and the ego vehicle, while $\widehat{d_i}$ and $\widehat{\theta_i}$ are the predicted values of VLMs, and ϵ is a small positive constant for numerical stability.

Object-centric metrics, which include Mean Relative Distance Error inter Object (OMRDE) and Mean Relative Angle Error inter Object (OMRAE), evaluate pairwise relationships among detected objects:

$$\text{OMRDE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{d_{ij}} - d_{ij}|}{d_{ij}}, \tag{24}$$

$$\text{OMRAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{\theta_{ij}} - \theta_{ij}|}{\max(|\theta_{ij}|,\epsilon)}, \tag{25}$$

where d_{ij} and θ_{ij} are the ground truth distance and azimuth between ith and jth participants, while $\widehat{d_{ij}}$ and $\widehat{\theta_{ij}}$ are the predicted values of VLMs, and ϵ is a small positive constant for numerical stability.

These indicators collectively assess spatial coherence and VLMs' ability to infer inter-object geometric dependencies.

4.3. Motion Planning

This module evaluates VLMs' dynamic reasoning ability, including both high-level meta-action prediction and low-level trajectory planning.

Table 4. Meta-action Classification Based on Acceleration and Change of Heading.

Behavior Type	Acceleration (a)	Heading Change ($\Delta\theta$)
Accelerating	$a \geq +0.15\,m/s^2$	-
Braking	$a \leq -0.15\,m/s^2$	-
Constant Speed	otherwise	-
Left Turn	-	$\Delta\theta \geq +8°$
Right Turn	-	$\Delta\theta \leq -8°$
Go Straight	-	otherwise

4.3.1. Meta-action Prediction

We define meta-actions based on acceleration and heading change trends derived from the annotated ground-truth 3-second future trajectory. Given a trajectory sequence (x_t, y_t, z_t), the velocity v_t and acceleration a_t are computed as:

$$v_t = \frac{\sqrt{(x_{t+1}-x_t)^2 + (y_{t+1}-y_t)^2 + (z_{t+1}-z_t)^2}}{t_{t+1} - t_t}, \tag{26}$$

$$a_t = \frac{v_{t+1} - v_t}{t_{t+1} - t_t}, \tag{27}$$

and the heading and its change are defined as:

$$\theta_t = \text{atan2}(z_{t+1} - z_t, x_{t+1} - x_t), \tag{28}$$

$$\Delta\theta_t = wrapToPi(\theta_{t+1} - \theta_t). \tag{29}$$

where $wrapToPi(\cdot)$ maps the heading difference to $[-\pi, \pi]$, avoiding discontinuities across the angle boundary.

We define six meta-actions describing longitudinal and lateral behaviors, as summarized in Table 4.

The Direction-Consistency Accuracy (DCS-Acc) is therefore

Just Accepted

introduced, which measures whether the VLM-predicted meta-action is consistent with the ground truth. DCS-Acc is formulated as follows:

$$\mathrm{DCS-Acc} = \frac{1}{N}\sum_{i=1}^{N} I\left(A_i^{pred} = A_i^{gt}\right), \tag{30}$$

where A_i^{pred} denotes meta-actions predicted by VLMs, and A_i^{gt} denotes the ground truth. A prediction is regarded as correct only when both its longitudinal action (accelerating, braking, or constant speed) and lateral action (left turn, right turn, or going straight) match the ground truth.

We further use relative and quantitative metrics, such as Mean Relative Acceleration Error (MRE-Acc) and Angular Relative Error (ARE) defined as follows:

$$\mathrm{MRE-Acc} = \frac{1}{N}\sum_{i=1}^{N} \frac{|a_i^{pred} - a_i^{gt}|}{max(|a_i^{gt}|, \epsilon)}, \tag{31}$$

$$\mathrm{ARE} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\Delta\theta_i^{pred} - \Delta\theta_i^{gt}|}{max(|\Delta\theta_i^{gt}|, \epsilon)}. \tag{32}$$

4.3.2. Trajectory Planning

The trajectory is expressed in the metric-scaled VGGT world frame (x right, y down, z forward) and translated so that the fourth-frame camera center is the origin, without additional rotation. At 2 Hz, the next six frames span a prediction horizon of 3 seconds. Trajectory quality control rejects non-finite outputs and clips with abrupt displacement or heading changes, abnormal acceleration, or curvature discontinuities; these checks are data-quality filters rather than a vehicle-specific dynamic-feasibility model. Two standard metrics are employed: Average Displacement Error (ADE) and Final Displacement Error (FDE@T):

$$\mathrm{ADE} = \frac{1}{N}\sum_{i=1}^{N} ||\hat{p}_i - p_i||_2, \tag{33}$$

$$\mathrm{FDE@T} = ||\hat{p}_T - p_T||_2. \tag{34}$$

ADE measures global spatial consistency, while FDE emphasizes endpoint accuracy. We report FDE@1s, FDE@2s, and FDE@3s to assess short-, mid-, and long-term prediction performance.

4.4. GPT-Score

We report GPT-Score, a semantic evaluation metric computed with GPT-4o (Hurst et al., 2024). For each instance, GPT-4o receives the prompt, ground truth answer, and VLM prediction, and outputs a scalar in [0,1] measuring their semantic alignment (higher is better). For full-scene description and risk-reasoning samples, we additionally provide SPICE, Risk-Class-Acc, and the ground-truth risk level as auxiliary inputs. For other question types, GPT-4o only sees the prompt, ground truth, and prediction.

4.5. Evaluation Weighting Strategy

4.5.1. Normalization

All metrics in the benchmark are normalized into a standardized score range of [0,100] for non-error-based metrics, and (0,100] for error-based metrics. We distinguish non-error-based and error-based metrics, and separate normalization strategies tailored to their mathematical characteristics.

Non-error-based Metrics: Non-error-based metrics include SPICE, Risk-Class-Acc, Class-Acc, DCS-Acc, and GPT-Score. These metrics naturally take values in [0,1], and exhibit a simple higher-is-better monotonic relationship.

For these metrics, we adopt a direct linear scaling to obtain a normalized score:

$$S_{non-error} = 100 \times M, \tag{35}$$

where M denotes the raw metric value.

Error-based Metrics: Error-based metrics include EMRDE, EMRAE, OMRDE, OMRAE, MRE-Acc, ARE, ADE, and FDE@T. These metrics quantify deviations between predictions and ground truth, and therefore follow a simple lower-is-better monotonic relationship. We therefore use a piecewise normalization, mapping raw error E to (0,100], which could be defined as:

$$S_{error} = \begin{cases} 100, & E < x_1, \\ 100 - 40 \cdot \dfrac{E - x_1}{x_2 - x_1}, & x_1 \le E < x_2, \\ 60 \cdot \exp(-k(E - x_2)), & E > x_2. \end{cases} \tag{36}$$

where x_1 is the high precision threshold, x_2 is the acceptable error threshold, and $k > 0$ is an exponential decay factor.

We first estimate the normalization parameters for each error-based metric using the empirical error distributions from a representative set of baseline models. These parameters are then fixed and applied consistently to all subsequent model evaluations. Specifically, we set the high-precision threshold x_1 to the 15th percentile, and the acceptable error threshold x_2 to the 75th percentile of the raw errors. The decay factor (k) is then determined by requiring the 90th-percentile error to yield a normalized score of 10. Thus, only the top 15% most accurate predictions receive a perfect score of 100, and the worst 25% predictions are mapped to the exponential penalty region. This setting preserves a broad linear range for distinguishing typical model errors while limiting the influence of extreme values. The percentile thresholds are empirical calibration choices rather than theoretically optimal constants.

This formulation provides the following desirable properties:

- **High-precision region** ($E < x_1$): Errors below a high-precision threshold are rewarded uniformly with a perfect score of 100, reflecting negligible deviation in practice.
- **Moderate-error region** ($x_1 \le E < x_2$): Scores decrease linearly from 100 to 60, enabling fine-grained comparison among VLMs with moderate errors.
- **High-error region** ($E \ge x_2$): Scores decay exponentially from 60 to 0, maintaining continuity while strongly penalizing large errors more strongly than small ones, in line with safety-critical evaluation in autonomous driving.

Table 5 lists the normalization parameters for all error-based metrics, showing how each maps into the unified scoring space.

Table 5. Normalization Parameters for Error-Based Metrics.

Metrics / Parameters	x_1	x_2	k
EMRDE	0.1135	0.3856	9.1380
EMRAE	0.1052	0.4055	2.2210
OMRDE	0.1244	0.4252	8.0125
OMRAE	0.1155	0.4152	2.0542
MRE-Acc	0.7250	1.3056	0.0216
ARE	0.7588	1.3319	0.0125
ADE	2.2850	5.2278	7.2514
FDE@1s	1.3595	3.7750	6.7157
FDE@2s	1.3595	3.7750	6.7157
FDE@3s	1.3595	3.7750	6.7157

4.5.2. The Overall Weighting Strategy

Table 6 lists the evaluation weights, which reflect the functional roles and safety relevance of the capabilities being assessed.

Within each module, we assign metric weights according to the role of each metric in driving evaluation. Scene Understanding assigns the greatest weight to SPICE and includes a separate risk term. Spatial Perception places more weight on ego-relative distance and angle than on inter-object relations. Motion Planning combines direction consistency, acceleration and angle errors, ADE, and FDE at 1, 2, and 3 seconds. The module weights express task priorities rather than a theoretically optimal choice. We therefore report all module-level metrics and use Overall as a compact summary. Section 4.5.3 examines how alternative weights affect model rankings.

Just Accepted

Table 6. Weighting structure of the evaluation framework of ScenePilot-Bench.

Module	Module Weight	Metric	Metric Weight
Scene Understanding	0.15	SPICE	0.70
		Risk-Class-Acc	0.30
Spatial Perception	0.35	Class-Acc	0.20
		EMRDE	0.30
		EMRAE	0.20
		OMRDE	0.20
		OMRAE	0.10
Motion Planning	0.40	DCS-Acc	0.20
		MRE-Acc	0.10
		ARE	0.10
		FDE@1	0.10
		FDE@2	0.10
		FDE@3	0.20
		ADE	0.20
GPT-Score	0.10	GPT Alignment Score	1.00

4.5.3. Sensitivity to Module Weights

To assess the sensitivity of the Overall ranking to the weighting scheme, we recompute scores for all 21 models under four alternative configurations. The Equal configuration assigns the same weight to each module. The other three give greater weight to spatial perception, motion planning, and semantic evaluation, respectively. We compare each ranking with the default ranking using Spearman's rank correlation and the maximum absolute rank shift.

Table 7. Sensitivity of Model Rankings to Module Weights.

Configuration	Scene	Spatial	Planning	GPT	Spearman ρ	Max shift
Default	0.15	0.35	0.40	0.10	1.000	0
Equal	0.25	0.25	0.25	0.25	0.968	4
Spatial-heavy	0.15	0.45	0.30	0.10	0.971	4
Planning-heavy	0.15	0.25	0.50	0.10	0.992	2
Semantic-heavy	0.20	0.30	0.35	0.15	0.987	3

Across the four alternative configurations, Spearman's ρ ranges from 0.968 to 0.992, and the maximum rank shift varies from 2 to 4 positions. The Spatial-heavy configuration changes which model ranks first, but the high correlations indicate that the overall ordering remains broadly stable within the tested range of weights.

4.6. Geographic Generalization

The robustness against domain shifts and geographic variations is crucial for real-world deployment. To evaluate VLMs' performance across geographic regions and traffic-direction conventions, we design a generalization assessment protocol comprising two cross-domain settings. No new metrics are introduced in this section.

China-to-Region Cross-Geographic Evaluation. We adopt a China-to-region setting to assess cross-geographic transfer. VLMs are trained solely on data from China and tested on four geographic regions: Europe (EU), Japan/Korea (JP/KR), the United States (US), and other countries and regions (OTH).

Right-to-Left Cross-Traffic Evaluation. Driving environments differ between left-hand- and right-hand-traffic systems. To evaluate VLMs' adaptability to this fundamental shift, we introduce a cross-traffic evaluation:

- **R → L**: Train exclusively on right-hand-traffic countries (e.g., China and the US) and test on left-hand-traffic countries (e.g., Japan and the UK).

We report the absolute performance under this cross-traffic setting because a corresponding right-hand-traffic in-domain baseline is not included.

4.7. Discussion

The proposed benchmark evaluates scene understanding, spatial perception, and motion planning in a single pipeline. It assesses both high-level reasoning and low-level geometric consistency. By considering objective spatial precision, subjective semantic plausibility, and geographic generalization together, the benchmark provides a comprehensive view of model performance in real-world driving contexts.

5 Experiments

5.1. Implementation Details of Experiments

For the comparative evaluation of VLMs on ScenePilot-Bench, we divide the dataset into training, validation, and test sets. We use the training set to fine-tune the VLMs, the validation set to tune hyperparameters and select models, and the test set to evaluate final performance. To balance representation across the three subsets, the split follows a principle of multidimensional conditional uniformity, jointly considering weather conditions, road types, lane configurations, intersection attributes, risk levels, traffic density, and geographic regions.

We sample the VQA training and test sets independently from the corresponding video splits. Each sample is linked to a unique video clip, and no clip appears in both sets. This prevents leakage between the two sets and ensures that the source clips used for testing are unseen during training. Unless otherwise specified, all comparative experiments use 200,000 VQA training samples and the same 100,000 test samples. Although the annotation pipeline produces nearly 400 million VQA instances, the data-scale ablation in Section 5.3 shows that 200,000 training samples achieve the highest overall score with the current model architecture and training setup. We therefore use 200k as a practical operating point and keep the test set fixed for consistent comparisons across experiments. The full annotated dataset remains available for future work on further training VLMs and improving their foundational capabilities.

We use the reference training framework for each model: LLaMA-Factory for LoRA fine-tuning of Qwen2.5-VL, Unsloth for ReasonDrive, and the respective InternVL-based implementations for ReCogDrive and DriveMonkey. Fine-tuning is performed on four NVIDIA A800 GPUs for all models except DriveMonkey, which uses four NVIDIA L40 GPUs. All experiments use bfloat16 precision and AdamW. For ReCogDrive and DriveMonkey, we apply rank-8 LoRA adapters with a scaling factor of 32 to the language model, keeping the vision backbone, multimodal projector, and base language-model weights frozen.

5.2. Baseline Results on ScenePilot-Bench

We evaluate representative commercial, open-source, and driving-specialized VLMs on ScenePilot-Bench, following the data splits and sampling settings in Section 5.1. Table 8 reports results for three baseline groups: VLMs without driving-specific fine-tuning, driving-specialized VLM variants, and Qwen backbones fine-tuned on ScenePilot-4K.

Just Accepted

Table 8. Baseline results on ScenePilot-Bench.

Model	Scene Understanding			Spatial Perception					
	SPICE	Risk-Class-Acc	total	Class-Acc	EMRDE	EMRAE	OMRDE	OMRAE	total
GPT-4o (Hurst et al., 2024)	92.93	74.82	87.50	91.93	38.51	22.49	26.17	20.63	41.73
GPT-5 (Bubeck et al., 2025)	92.18	71.64	86.01	91.83	46.32	25.82	19.37	22.31	43.53
Gemini-2.5-flash (Comanici et al., 2025)	91.70	69.44	85.02	**92.55**	43.94	22.63	21.46	22.74	42.78
Qwen3-VL-235B (Bai et al., 2025)	93.58	77.30	88.68	91.64	46.27	14.89	21.72	20.01	41.53
ReasonDrive-7B (Chahe and Zhou, 2025)	85.31	33.84	69.86	75.12	22.21	7.25	6.77	9.25	25.42
ReasonDrive-7B-trained	85.45	37.06	70.93	89.27	22.31	25.58	29.24	8.57	36.37
ReasonDrive-7B-CoT (Chahe and Zhou, 2025)	85.51	34.07	70.08	88.55	19.07	7.34	6.88	14.28	27.70
ReasonDrive-7B-CoT-trained	87.90	56.91	78.60	90.18	22.98	22.15	23.45	12.39	35.29
ReCogDrive-2B (Y. Li et al., 2026)	88.31	65.14	81.36	88.29	13.64	16.22	16.56	12.58	29.56
ReCogDrive-2B-trained	**95.15**	**85.86**	**92.36**	89.29	59.63	**75.53**	52.55	50.47	66.41
DriveMonkey-2B (Zhao et al., 2025)	62.47	57.68	61.03	84.12	14.30	3.94	15.24	5.89	25.54
DriveMonkey-2B-trained	89.12	68.48	82.93	78.95	22.47	30.37	23.28	9.09	34.17
ScenePilot-2-2B-200k	93.64	80.57	89.72	85.74	**69.43**	73.66	48.39	49.07	67.29
ScenePilot-2.5-3B-200k	93.65	78.67	89.15	86.19	68.24	73.68	**52.58**	**54.14**	**68.38**

Model	Motion Planning								GPT-Score	Overall
	DCS-Acc	MRE-Acc	ARE	ADE	FDE@1	FDE@2	FDE@3	total		
GPT-4o (Hurst et al., 2024)	15.79	82.16	**85.07**	58.68	**85.79**	42.71	24.15	49.30	22.56	49.71
GPT-5 (Bubeck et al., 2025)	22.39	81.32	79.59	57.53	85.25	44.88	28.20	50.73	36.51	52.08
Gemini-2.5-flash (Comanici et al., 2025)	19.59	76.73	79.89	43.77	76.40	34.10	19.63	43.31	33.64	48.41
Qwen3-VL-235B (Bai et al., 2025)	22.80	43.83	83.05	12.69	15.70	8.67	5.55	23.33	34.22	40.59
ReasonDrive-7B (Chahe and Zhou, 2025)	14.38	29.45	61.23	13.67	23.31	11.25	9.09	19.95	15.70	28.93
ReasonDrive-7B-trained	14.69	80.26	41.92	39.20	69.68	33.62	26.51	38.60	27.60	41.57
ReasonDrive-7B-CoT (Chahe and Zhou, 2025)	14.12	0.49	47.55	0.00	0.00	0.00	0.00	7.63	17.90	25.05
ReasonDrive-7B-CoT-trained	15.22	82.44	59.82	39.01	68.92	33.55	26.53	40.59	24.72	42.85
ReCogDrive-2B (Y. Li et al., 2026)	14.91	66.78	55.84	38.06	61.95	29.37	20.95	36.18	21.54	39.18
ReCogDrive-2B-trained	15.17	83.36	31.88	40.59	71.79	34.75	27.23	38.77	46.14	57.22
DriveMonkey-2B (Zhao et al., 2025)	15.60	79.57	77.03	36.16	59.83	27.42	21.04	38.94	12.16	34.89
DriveMonkey-2B-trained	16.74	78.13	80.91	41.19	62.85	30.71	23.90	41.63	18.87	42.94
ScenePilot-2-2B-200k	28.46	79.47	75.07	**59.31**	82.72	**47.78**	29.16	**51.89**	52.53	63.02
ScenePilot-2.5-3B-200k	**29.52**	**88.20**	70.09	53.07	79.94	43.64	**32.67**	51.24	**54.43**	**63.24**

Note: Scores are averaged over clips, then rounded to two decimal places.

Table 9. Ablation results for fine-tuning data scale and annotation components.

Model	Scene Understanding			Spatial Perception					
	SPICE	Risk-Class-Acc	total	Class-Acc	EMRDE	EMRAE	OMRDE	OMRAE	total
ScenePilot-2.5-3B-50k	90.91	67.28	83.82	91.67	62.00	57.93	40.50	37.12	60.33
ScenePilot-2.5-3B-2M	**93.70**	**79.60**	**89.47**	90.16	69.86	**81.24**	56.53	57.41	**72.29**
ScenePilot-2.5-3B-20M	91.90	68.31	84.82	**92.23**	68.87	77.21	51.76	59.44	70.84
ScenePilot-2.5-3B-noscene-200k[1]	82.07	31.17	66.80	91.58	67.13	73.17	53.01	54.42	69.13
ScenePilot-2.5-3B-notraj-200k[2]	**93.70**	79.16	89.33	87.46	**71.11**	78.50	**55.05**	**61.60**	71.69
ScenePilot-2.5-3B-nospatial-200k[3]	93.23	77.77	88.59	88.78	0.24	0.00	0.10	0.01	17.85

Model	Motion Planning								GPT-Score	Overall
	DCS-Acc	MRE-Acc	ARE	ADE	FDE@1	FDE@2	FDE@3	total		
ScenePilot-2.5-3B-50k	24.27	86.44	61.43	53.74	80.93	45.46	32.53	49.53	42.75	57.78
ScenePilot-2.5-3B-2M	24.01	**88.63**	55.21	45.52	74.39	36.98	28.40	45.11	**54.69**	62.24
ScenePilot-2.5-3B-20M	29.93	85.75	64.32	53.68	80.75	44.15	32.00	50.62	52.17	**62.98**
ScenePilot-2.5-3B-noscene-200k[1]	**27.88**	87.06	69.23	56.63	**82.85**	**48.34**	**36.05**	**52.86**	40.96	59.46
ScenePilot-2.5-3B-notraj-200k[2]	12.50	66.52	54.09	39.38	69.95	33.73	26.53	38.11	41.04	57.84
ScenePilot-2.5-3B-nospatial-200k[3]	27.86	81.94	**69.49**	**59.94**	74.96	46.65	33.70	51.60	14.74	41.65

Notes: [1] *noscene* excludes scene-level annotations, including scene descriptions and risk labels, from fine-tuning. [2] *notraj* excludes ego-trajectory annotations. [3] *nospatial* excludes object-centric spatial annotations. All three variants use the same ScenePilot-2.5-3B backbone and a 200k-sample fine-tuning set, with the remaining training settings unchanged. Scores are averaged over clips, then rounded to two decimal

Just Accepted

places.

Table 10. Cross-region and cross-traffic evaluation of VLMs.

Model	Scene Understanding			Spatial Perception					
	SPICE	Risk-Class-Acc	total	Class-Acc	EMRDE	EMRAE	OMRDE	OMRAE	total
SP-2.5-3B-CN (All)	92.98	75.18	87.64	83.46	66.56	69.80	53.38	53.42	66.64
SP-2.5-3B-CN (EU)	92.57	75.26	87.37	83.44	70.99	67.33	54.37	52.87	67.61
SP-2.5-3B-CN (JP/KR)	92.48	71.01	86.04	83.84	68.61	68.50	53.91	54.31	67.27
SP-2.5-3B-CN (US)	92.74	74.75	87.34	83.18	69.53	69.04	56.72	54.81	68.13
SP-2.5-3B-CN (OTH)	91.99	70.14	85.44	83.09	**74.36**	71.16	**57.74**	**55.71**	**70.28**
SP-2.5-3B-R (L)	**93.72**	**78.79**	**89.24**	**90.51**	68.14	**74.17**	54.51	55.05	69.79

Model	Motion Planning								GPT-Score	Overall
	DCS-Acc	MRE-Acc	ARE	ADE	FDE@1	FDE@2	FDE@3	total		
SP-2.5-3B-CN (All)	30.14	85.77	74.78	56.94	76.36	43.63	31.89	51.85	46.36	61.85
SP-2.5-3B-CN (EU)	30.79	86.13	**78.70**	57.11	84.44	46.08	32.60	53.63	54.54	63.68
SP-2.5-3B-CN (JP/KR)	32.30	**87.33**	77.84	43.85	76.84	34.67	24.36	47.77	54.96	61.05
SP-2.5-3B-CN (US)	30.37	85.53	76.96	53.57	79.09	44.58	32.67	51.94	54.14	63.14
SP-2.5-3B-CN (OTH)	30.39	83.50	78.18	**63.46**	**90.24**	**52.16**	**37.94**	**56.76**	**56.22**	**65.74**
SP-2.5-3B-R (L)	**32.63**	86.35	76.68	56.94	86.00	46.17	33.25	54.08	54.61	64.91

Note: Scores are averaged over clips, then rounded to two decimal places.

The three groups show different strengths in semantic, spatial, and planning tasks. General-purpose VLMs obtain high SPICE scores and competitive results in risk understanding, but struggle with ego-centric spatial grounding and planning-oriented reasoning. This gap suggests that generic multimodal pretraining transfers well to coarse semantic understanding, but does not reliably induce the geometric calibration and reference-frame consistency needed for autonomous driving. Driving-specialized models, including ReasonDrive-7B (Chahe and Zhou, 2025), ReCogDrive (Y. Li et al., 2026), and DriveMonkey (Zhao et al., 2025), improve on several driving-oriented metrics after fine-tuning on ScenePilot-4K, although the gains vary across architectures and evaluation modules. The fine-tuned Qwen backbones achieve the most balanced performance across scene understanding, spatial perception, and motion planning. These results demonstrate that ScenePilot-4K provides effective supervision for jointly improving semantic reasoning, spatial grounding, and short-horizon planning, rather than indicating an advantage from benchmark-specific optimization. The comparison also demonstrates the diagnostic value of ScenePilot-Bench by distinguishing strengths in high-level semantics from weaknesses in embodied perception and planning.

5.3. Ablation Studies

We use the same Qwen2.5-VL-3B backbone and evaluation split to examine the effects of fine-tuning data scale and annotation components. Table 9 reports the module-level and overall scores for each configuration.

Effect of Fine-Tuning Data Scale. Performance improves as the fine-tuning set grows to 200k samples. Larger training sets improve some scene and spatial metrics, but planning performance fluctuates, and the overall score does not improve consistently. We use 200k samples in the comparative experiments because this setting provides the most balanced performance across the evaluated capabilities.

Effect of Annotation Components. Removing an annotation component mainly reduces performance in the corresponding evaluation module. Scene annotations primarily support scene understanding, trajectory annotations are essential for motion planning, and spatial annotations have the greatest effect on spatial perception. GPT-Score also drops substantially without spatial supervision, indicating that metric spatial grounding helps produce semantically consistent responses. These results confirm that the three annotation components provide distinct but complementary forms of supervision.

5.4. Geographic Generalization

We further evaluate cross-region robustness under geographic domain shift. In the China-to-region setting, we train the Qwen2.5-VL-3B baseline on 200,000 samples from China and test it on four geographic regions: Europe (EU), Japan/Korea (JP/KR), the United States (US), and other countries and regions (OTH), each with 100,000 samples. We also evaluate a right-to-left cross-traffic setting, where the model is trained only on right-hand-traffic countries and tested on left-hand-traffic regions. For the controlled global comparison, Global-to-All is the ScenePilot-2.5-3B-200k row in Table 8: the model is trained on 200,000 samples from the geographically diverse split and evaluated on the fixed global test set of 100,000 samples. China-to-All is the SP-2.5-3B-CN (All) row in Table 10. It uses the same backbone, training budget, and test set, but all training samples come from China. Table 10 summarizes the results.

The results show region-dependent variation across the evaluated modules. SPICE remains consistently high across regions, while risk understanding shows moderate variation. Spatial perception remains broadly functional, but the geometric metrics are more sensitive to regional differences in road layout, visual depth cues, and traffic organization. Motion-planning metrics also vary across the regional test sets, indicating sensitivity to geographic domain differences. The right-to-left evaluation further reports model performance under a change in traffic-direction conventions. Because a corresponding right-hand-traffic in-domain baseline is not included, we do not quantify the performance drop caused by this shift. The China-to-All comparison further shows that restricting the training data to a single country mainly affects semantic alignment, while motion-planning performance remains broadly comparable. This pattern suggests that geographically diverse training benefits cross-region performance, particularly for semantic interpretation across regional domains. These findings suggest that ScenePilot-Bench provides a practically relevant cross-region evaluation beyond standard in-domain testing. This also suggests that future driving-oriented VLMs should be evaluated not only by in-domain accuracy, but also by their ability to preserve spatial and planning consistency under realistic cross-region distribution shifts.

6 Conclusions

Just Accepted

We present ScenePilot-4K, a large-scale first-person driving dataset for vision-language learning and evaluation in autonomous driving. It contains 3,847 hours of driving videos and 27.7M front-view frames from 63 countries/regions and 1,210 cities. The dataset provides unified annotations for scene descriptions, risk assessment, key participants, ego trajectories, and camera parameters. With broad geographic coverage and detailed annotations, ScenePilot-4K supports the study of autonomous driving across diverse road environments.

Built on ScenePilot-4K, ScenePilot-Bench provides a standardized open-loop protocol for jointly evaluating scene understanding, spatial perception, motion planning, and semantic alignment, with GPT-based judgment used for the semantic assessment. Alongside in-domain evaluation, cross-region and cross-traffic settings assess model performance under geographic domain shifts. Together, ScenePilot-4K and ScenePilot-Bench provide a common framework for these analyses. Beyond the released benchmark, we expect the annotation pipeline and evaluation protocol to serve as extensible tools for building customized driving datasets and developing new robustness tests across diverse geographic settings.

Results from representative commercial, open-source, and driving-specialized vision-language models show that they perform relatively well on high-level scene semantics, but still have clear limitations in geometry-aware perception, planning-oriented reasoning, and generalization across geographic regions. These findings indicate that evaluating vision-language models for driving requires more than generic captioning or VQA.

The current release has four limitations. First, ScenePilot-Bench evaluates models in an open-loop setting and does not directly assess the effects of model outputs on downstream control in a closed-loop driving system. Second, some annotations are generated automatically using VLMs and geometric reconstruction. Although cross-dataset validation and conservative filtering improve annotation reliability, residual noise may remain in difficult scenes. Third, the geographic generalization settings cover representative cases but do not capture the full range of variations in camera setups, weather patterns, and regional distributions. Fourth, we have not conducted a strictly controlled ablation comparing single-frame and multi-frame inputs or repeated training with multiple random seeds for the large-scale configurations. The effects of temporal input and statistical variability remain to be quantified. We also plan to extend the benchmark through additional human verification, broader cross-domain evaluation, and closed-loop integration.

We hope that ScenePilot-4K and ScenePilot-Bench will provide useful open resources for research on multimodal driving understanding, spatial grounding, and safety-aware evaluation. To support this work, we release annotations, metadata, split files, source-video indices, and reproduction code, making it easier to inspect and reproduce the experiments and examine performance differences across geographic regions.

Author contribution

Yujin Wang: Writing – original draft, Conceptualization, Methodology, Software, Investigation, Validation, Data curation. Yutong Zheng: Writing – original draft, Methodology, Software, Investigation, Validation, Data curation. Wenxian Fan: Writing – original draft, Investigation, Validation, Data curation, Visualization. Tianyi Wang: Writing – original draft, Validation, Writing – review & editing. Hongqing Chu: Methodology, Formal analysis, Writing – review & editing. Li Zhang: Methodology, Writing – review & editing. Bingzhao Gao: Supervision, Methodology, Formal analysis, Funding acquisition, Writing – review & editing. Daxin Tian: Formal analysis, Writing – review & editing. Jianqiang Wang: Formal analysis, Funding acquisition, Writing – review & editing. Hong Chen: Supervision, Formal analysis, Writing – review & editing.

Replication and data sharing

The source codes and replication package are available on ETS-Data at https://doi.org/10.26599/ETSD.2026.9190094.

Acknowledgments

This research is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM123), National Key R&D Program of China (2023YFB2504400), the National Natural Science Foundation of China (No.62373289), Shanghai Municipal Science and Shanghai Automotive Industry Science and Technology Development Foundation (No. 2407), and the Fundamental Research Funds for the Central Universities.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. The author Jianqiang Wang is the Editorial Board member of this journal, but he is not involved in the peer review or decision of this article.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve the language and readability of the manuscript. After utilizing this tool, the authors thoroughly reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., et al., 2025. Qwen3-VL Technical Report. https://doi.org/10.48550/arXiv.2511.21631

Bubeck, S., Coester, C., Eldan, R., Gowers, T., Lee, Y.T., Lupsasca, A., et al., 2025. Early science acceleration experiments with GPT-5. https://doi.org/10.48550/arXiv.2511.16072

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., et al., 2020. nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11621-11631.

Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., et al., 2021. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. https://doi.org/10.48550/arXiv.2106.11810

Chahe, A., Zhou, L., 2025. ReasonDrive: Efficient visual question answering for autonomous vehicles with reasoning-enhanced small vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 3870-3879.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., et al., 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. https://doi.org/10.48550/arXiv.2507.06261

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al., 2016. The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3213-3223.

Deruyttere, T., Grujicic, D., Blaschko, M.B., Moens, M.F., 2022. Talk2Car: Predicting physical trajectories for natural language commands. IEEE Access 10, 123809-123834.

Ding, X., Han, J., Xu, H., Zhang, W., Li, X., 2025. HiLM-D: Enhancing MLLMs with multi-scale high-resolution details for autonomous driving. Int J Comput Vis 133, 5379-5395.

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., et al., 2021. Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9710-9719.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. Int J Robot Res 32, 1231-1237.

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., et al., 2021. One thousand and one hours: Self-driving motion prediction dataset. In:

Just Accepted

Proceedings of the Conference on Robot Learning, 409-418.

Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R., 2020. The ApolloScape open dataset for autonomous driving and its application. IEEE Trans Pattern Anal Mach Intell 42, 2702-2719.

Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., et al., 2024. GPT-4o system card. https://doi.org/10.48550/arXiv.2410.21276

Ishihara, K., Sasaki, K., Takahashi, T., Shiono, D., Yamaguchi, Y., 2026. STRIDE-QA: Visual question answering dataset for spatiotemporal reasoning in urban driving scenes. Proceedings of the AAAI Conference on Artificial Intelligence 40, 5257-5266.

Jocher, G., Qiu, J., 2024. Ultralytics YOLO11. https://github.com/ultralytics/ultralytics

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015-4026.

Li, W., Qu, Z., Song, H., Wang, P., Xue, B., 2021. The traffic scene understanding and prediction based on image captioning. IEEE Access 9, 1420-1427.

Li, Y., Xiong, K., Guo, X., Li, F., Yan, S., Xu, G., et al., 2026. ReCogDrive: A reinforced cognitive framework for end-to-end autonomous driving. In: International Conference on Learning Representations.

Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., et al., 2025. Depth Anything 3: Recovering the visual space from any views. https://doi.org/10.48550/arXiv.2511.10647

Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J., 2023. DRAMA: Joint risk localization and captioning in driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1043-1052.

Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., et al., 2021. One million scenes for autonomous driving: ONCE dataset. https://doi.org/10.48550/arXiv.2106.11037

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., et al., 2024. DriveLM: Driving with graph visual question answering. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland, 256-274.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., et al., 2020. Scalability in perception for autonomous driving: Waymo Open Dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2446-2454.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025. VGGT: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 5294-5306.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., et al., 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. https://doi.org/10.48550/arXiv.2409.12191

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., et al., 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. https://doi.org/10.48550/arXiv.2301.00493

Wu, D., Han, W., Liu, Y., Wang, T., Xu, C.Z., Zhang, X., Shen, J., 2025. Language prompt for autonomous driving. Proceedings of the AAAI Conference on Artificial Intelligence 39, 8359-8367.

Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., et al., 2025. Are VLMs ready for autonomous driving? An empirical study from the reliability, data and metric perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 6585-6597.

Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., et al., 2024. Generalized predictive model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14662-14672.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., et al., 2020. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2636-2645.

Zhang, T., Jin, T., Wang, L., Liu, J., Liang, S., Zhang, M., et al., 2025. Bench2ADVLM: A closed-loop benchmark for vision-language models in autonomous driving. https://doi.org/10.48550/arXiv.2508.02028

Zhao, Z., Fu, H., Liang, D., Zhou, X., Zhang, D., Xie, H., et al., 2025. Extending large vision-language model for diverse interactive tasks in autonomous driving. https://doi.org/10.48550/arXiv.2505.08725

Author biography



Yujin Wang is currently a Ph.D. candidate at the College of Automotive and Energy Engineering, Tongji University, China. He received the B.E. degree in vehicle engineering from Tongji University, China, in 2023. His research interests include vision-language models, deep learning, and foundation intelligence in autonomous driving.



Yutong Zheng is currently a master's student at the College of Automotive and Energy Engineering, Tongji University, Shanghai 201804, China. He received his B.S. degree in vehicle engineering from Tongji University, Shanghai, China, in 2025. His research interests include autonomous driving, vision-language models and deep learning.



Wenxian Fan is currently a Ph.D. student at the College of Automotive and Energy Engineering, Tongji University, China. She received her bachelor's degree from Tongji University, China, in 2026 and began her doctoral studies in September 2026. Her research interests include vision-language models, deep learning, and foundation intelligence in autonomous driving.



Tianyi Wang is currently a Ph.D. student at the Department of Civil, Architectural, and Environmental Engineering, The University of Texas at Austin, USA. He received the M.S. degree in mechanical engineering and materials science from Yale University, USA, in 2025. He earned the B.E. degree in vehicle engineering from Tongji University, China, in 2024. His research interests include intelligent transportation systems and connected and automated vehicles, particularly in decision-making, trajectory planning and cooperative control.



Hongqing Chu received the B.S. degree in Vehicle Engineering from Shandong University of Technology, China, in 2012 and his Ph.D. degree in Control Theory and Control Engineering from Jilin University, China, in 2017. He is currently an Associate Professor with the College of Automotive and Energy Engineering, Tongji University. His current research interests include autonomous driving, vehicle motion control, and advanced driver-assistance systems.

Just Accepted



Li Zhang received the PhD degree in computer science from the Queen Mary University of London. He is now a full professor with Fudan University. Previously, he was a postdoctoral research fellow with the University of Oxford. His research interests include computer vision and embodied AI.



Bingzhao Gao received the B.E. and M.E. degrees in vehicle engineering from Jilin University, China, in 1998 and 2002, respectively, and the Ph.D. degree in mechanical engineering from Yokohama National University, Japan, and in control engineering from Jilin University, in 2009. He is currently a Professor with the College of Automotive and Energy Engineering, Tongji University, China. His research interests include vehicle powertrain control and vehicle stability control.



Daxin Tian is currently a Professor with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His research interests include mobile computing, intelligent transportation systems, vehicular ad hoc networks, and swarm intelligent systems. He is also an IEEE Intelligent Transportation Systems Society Member and IEEE Vehicular Technology Society Member. He is a fellow of IEEE.



Jianqiang Wang received the B.E. and M.E. degrees from Jilin University of Technology, China, in 1994 and 1997, respectively, and the Ph.D. degree from Jilin University, in 2002. He is currently a Professor with the School of Vehicle and Mobility, Tsinghua University, China. He has authored more than 150 papers and is a coinventor of more than 140 patent applications. He was involved in more than ten sponsored projects. His research interests include intelligent vehicles, driving assistance systems, and driver behavior. He was a recipient of the Best Paper Award in the 2014 IEEE Intelligent Vehicle Symposium, the Best Paper Award in the 14th ITS Asia-Pacific Forum, the Best Paper Award in the 2017 IEEE Intelligent Vehicle Symposium, the Changjiang Scholar Program Professor in 2017, the Distinguished Young Scientists of NSF China in 2016, and the New Century Excellent Talents in 2008.



Hong Chen received the B.E. and M.E. degrees in process control from Zhejiang University, China, in 1983 and 1986, respectively, and the Ph.D. degree in system dynamics and control engineering from the University of Stuttgart, Germany, in 1997. In 1986, she joined the Jilin University of Technology, China. From 1993 to 1997, she was a Wissenschaftlicher Mitarbeiter with the Institut fuer Systemdynamik und Regelungstechnik, University of Stuttgart. Since 1999, she has been a Professor with Jilin University, and hereafter a Tang Aoqing Professor. From 2015 to 2019, she served as the Director of the State Key Laboratory of Automotive Simulation and Control, Jilin University. She is currently a Distinguished Professor at Tongji University, China. Her current research interests include model predictive control, nonlinear control, and applications in mechatronic systems focusing on automotive systems. She is a Fellow of IEEE.

Just Accepted