Thinking Ahead: Foresight Intelligence in MLLMs and World Model

Zhantao Gong[1], Liaoyuan Fan[2], Qing Guo[3,4], Xun Xu[5], Xulei Yang[✉5], and Shijie Li[✉5]

[1] College of Software, Nankai University, China
[2] The University of Hong Kong, China
[3] NKIARI, Shenzhen Futian, China
[4] AAIS & VCIP, Nankai University, China
[5] A*STAR Institute of Advanced Intelligence and Computing, Singapore

{li_shijie, yang_xulei}@a-star.edu.sg
https://huggingface.co/datasets/Gong-Grant/FSU-QA

Abstract. In this work, we introduce FSU-QA, a VQA dataset for autonomous driving scenarios designed to advance research on Foresight Intelligence—the ability to anticipate and reason about complex, long-horizon futures. Unlike existing benchmarks that mainly focus on immediate perception or reactive planning, FSU-QA evaluates future-oriented driving understanding through multi-agent-aware and rule-grounded counterfactual QA. Rather than holding surrounding agents fixed or targeting geometric path generation, our benchmark requires models to infer semantic future outcomes from front-view historical observations and past ego trajectories. A comprehensive evaluation on the accompanying FSU-Bench reveals that state-of-the-art VLMs still face significant challenges in anticipating future events. Furthermore, beyond model performance, we examine whether WM-generated predictions remain semantically consistent by using VLM-based proxy judges, and validate this evaluation protocol through shuffled control experiments. Fine-tuning models on FSU-QA leads to substantial improvements in foresight understanding, demonstrating the dataset's effectiveness and offering a principled foundation for future research.

Keywords: Visual Question-Answering · Vision-Language Models · World Models · Foresight Intelligence · Autonomous Driving

1 Introduction

In recent years, Vision-Large Language Models (VLMs) have achieved remarkable progress in visual understanding and cross-modal reasoning [22]. Typically, VLMs integrate a vision encoder with a Large Language Model (LLM) that has been pretrained on large-scale textual corpora rich in common knowledge. This design enables them to excel in both perception and understanding, leading to

* Equal contribution. ✉ Corresponding author.

their widespread application in domains such as robotic perception and scene understanding [42].

Most existing Vision–Language Model (VLM) studies focus on enabling VLMs to understand multimodal data such as 2D/3D scene understanding, spatial reasoning, and video comprehension [8,15,32]. However, these works mainly emphasize observed information while neglecting the ability to anticipate and reason about future, unobserved events. This is particularly challenging, as it requires reasoning over complex spatio-temporal information and dealing with multiple plausible futures, which introduces significant uncertainty. We term this concept Foresight Intelligence, which refers to the ability to anticipate and reason about complex future possibilities based on historical observations.

Foresight Intelligenceis conceptually distinct from the planning-oriented reasoning studied in prior benchmarks such as DriveLM [33] and OmniDrive [36]. Those benchmarks primarily ask models to issue immediate control decisions: given the current scene, what should the ego vehicle do next? Foresight Intelligenceposes a fundamentally different question: given a hypothetical action or event, what are its semantic consequences as the scene unfolds? This requires modeling multi-agent interactions under uncertainty, performing counterfactual simulation, and reasoning over extended temporal horizons—tasks that are simulation-level in nature rather than reactive.

In this work, we present `FSU-QA` (see Fig. 1), a Visual Question Answering (VQA) dataset, along with its accompanying benchmark, `FSU-Bench`, designed for training and evaluating Foresight Understanding. Together, they are designed to advance research on Foresight Intelligence. Considering that VLMs are increasingly being integrated into autonomous driving applications, where vehicles must operate in highly dynamic environments and are required to anticipate potential future events to ensure safety, we construct `FSU-QA` under the autonomous driving scenario. In `FSU-QA`, we meticulously design the tasks to both train models and comprehensively evaluate their Foresight Intelligence. Specifically, `FSU-Bench` requires the models to anticipate and interpret potential future scenarios across different temporal intervals based on the current situation, thereby enabling effective evaluation of long-horizon foresight capabilities. The designed tasks are organized according to their cognitive complexity, ranging from low-level perception tasks to mid-level imagination tasks and high-level reasoning tasks, which collectively provide a thorough assessment of foresight intelligence.

`FSU-Bench` is also designed to evaluate World Models (WMs)' ability to produce semantically coherent data. While VLMs excel at interpreting observed scenes, WMs incorporate physical realism into generative modeling to synthesize plausible future scenarios. Although such data may appear visually realistic, it remains an open question whether they are semantically coherent and can contribute to Foresight Intelligence. To address this, `FSU-Bench` adopts a proxy-evaluation paradigm: VLMs serve as discriminative probes to measure whether WM-generated futures encode semantically meaningful predictive content, quantified by downstream QA performance gains. This indirect evaluation is more

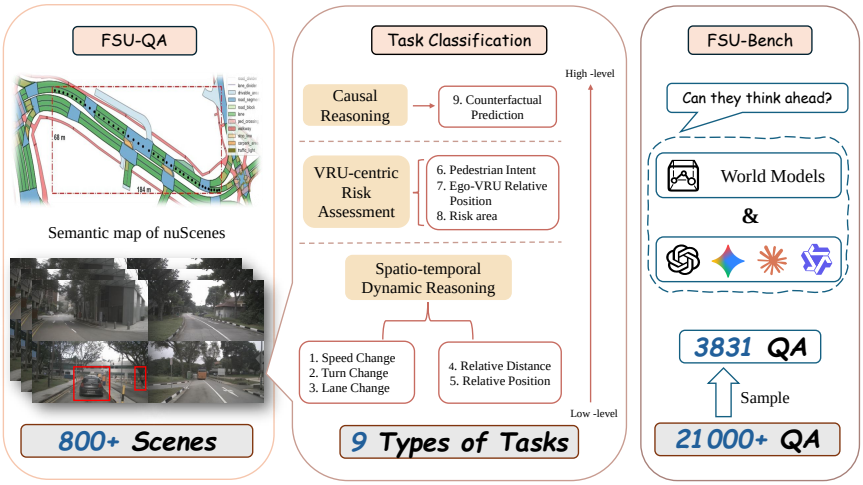


Fig. 1: Overview of `FSU-QA` and `FSU-Bench`.

diagnostic of real-world utility than standard pixel-level generation metrics (e.g., FVD, FID). With such a joint evaluation of VLMs and WMs, a comprehensive assessment of Foresight Intelligence is conducted, aiming to answer two fundamental questions: (1) What is the capability of VLMs in Foresight Intelligence? (2) Can the data generated by WMs contribute to Foresight Intelligence?

Our contribution can be summarized as:

- We propose `FSU-QA`, a dataset designed to train and evaluate the Foresight Intelligence of VLMs through tasks requiring counterfactual simulation and multi-agent causal reasoning—capabilities that go beyond the reactive decision-making scope of existing VQA benchmarks. Our experiments further demonstrate that `FSU-QA` can effectively enhance VLMs' foresight reasoning.
- We introduce a proxy WM-evaluation protocol that measures whether predicted future videos and trajectories provide useful semantic cues for downstream foresight QA, while acknowledging that this protocol is probe-model dependent.
- A comprehensive evaluation across multiple VLMs and WMs on `FSU-Bench` reveals that current models consistently struggle with high-level counterfactual reasoning, while WM-generated futures provide measurable but architecture-dependent semantic gains.

2 Related Work

2.1 Spatio-Temporal Reasoning in VLMs

Modern VLMs have progressed significantly beyond single-frame understanding [38, 46]. Advanced models [24] now demonstrate a strong capacity to cap-

ture both temporal information and static spatial relations in videos, achieving robust performance in open dialogue and multimodal reasoning. For instance, Qwen2.5-VL [3] leverage timestamp encoding for long video understanding, while InternVL-3.5 [37] employs a visual resolution router to efficiently process fine-grained video details. Notably, recent work such as GPT-Driver [25], LanguageMPC [26], EMMA [49], and SimLingo [29] has demonstrated that VLMs can perform chain-of-thought reasoning for autonomous driving, yet these models are primarily evaluated on present-state understanding rather than anticipating unobserved future events. Beyond VLM-based driving reasoning, trajectory forecasting methods model future motion using interaction-aware and spatio-temporal consistency cues [18,19]. However, these methods primarily output geometric trajectories or short-term motion states, whereas Foresight Intelligencerequires language-grounded semantic reasoning about long-horizon, multi-agent, and counterfactual future consequences.

2.2 Benchmarks

To evaluate VLMs' capabilities, a series of benchmarks have been proposed, but these largely focus on understanding the present state.

- **Static Spatial Reasoning:** Existing benchmarks range from 2D relational reasoning datasets such as CLEVR [16] and GQA [13] to static 3D spatial reasoning benchmarks such as ScanQA [2], VSI-Bench [44], and MMSI-Bench [45]. Recent monocular and open-vocabulary works [9, 17, 35] further advance spatial reasoning and grounding from visual observations, but primarily target current-scene understanding rather than long-horizon foresight.
- **Dynamic Spatial Reasoning:** Acknowledging the limitations of static scenes, newer benchmarks have begun to address dynamics [41,48]. VLM4D [50] treats video as a four-dimensional modality to probe spatiotemporal correlations. However, this shift to dynamic scenes has exposed a critical flaw: multimodal hallucination [20,23,39] has demonstrated that VLMs are highly prone to orientation hallucinations and semantic priors, especially when processing dynamic scenarios or uncommon viewpoints.
- **Driving VQA:** NuScenes-QA [28] targets static perception of the current frame. DriveLM [33] chains perception and planning, but restricts queries to immediate control actions within short horizons. OmniDrive [36] further introduces counterfactual queries, yet outputs geometric waypoints evaluated via L2 displacement error—grounding evaluation in path generation rather than causal scene understanding. `FSU-QA` departs from this paradigm: it targets natural-language causal judgments over multi-agent counterfactual scenarios, a semantic simulation-level capability not addressed by prior nuScenes-based benchmarks (see Tab. 1).

Table 1: Comprehensive comparison between FSU-QA and existing driving VQA benchmarks. FSU-QA uniquely targets long-horizon foresight intelligence, enabling joint evaluation of both VLMs and World Models' semantic consistency.

Feature / Benchmark	**NuScenes-QA**	**DriveLM**	**OmniDrive**	**DriveGPT4**	**FSU-QA (Ours)**
Primary Focus	Perception	Planning	Planning	Action	**Foresight**
Long-horizon Reasoning (**12s**)	×	×	×	×	✓
Counterfactual (What-if) Reasoning	×	×	Partial	×	✓
Hierarchical Cognitive Tasks (**9 Levels**)	×	×	×	×	✓
Multi-Agent-Aware Counterfactual QA	×	Partial	✓	×	✓
Foresight-oriented Training Split	×	×	×	×	✓
World Model Semantic Eval	×	×	×	×	✓
Dataset Source	nuScenes	nuScenes	nuScenes	nuScenes	nuScenes

2.3 Future-Oriented World Models

A core application of WMs is the generation of physically plausible future videos for autonomous driving. Early approaches based on diffusion models [30, 40] suffered from limited controllability and short rollout horizons; subsequent GPT-style autoregressive frameworks [7, 11, 43] improved temporal consistency and enabled extended multi-frame generation under continuous control.

Among the latest advancements, Epona [47] and DrivingWorld [12] represent two representative directions: Epona focuses on large-scale video-to-action pretraining for long-horizon driving imagination, while DrivingWorld incorporates explicit spatial constraints and controllable rollout strategies to stabilize world model predictions. Despite these advancements, the emphasis of current approaches remains centered on prediction. Although the synthesized videos often appear visually realistic, they frequently lack semantic fidelity, raising concerns about whether the predicted futures align with actual causal dynamics and scene semantics [4].

In this work, we introduce FSU-Bench, a unified benchmark designed to address the above limitations and enable a thorough assessment of the foresight capabilities of both VLMs and WMs.

3 FSU-QA

3.1 Foresight Intelligence

We use the term Foresight Intelligenceto encapsulate the advanced predictive capabilities that go beyond simple pattern recognition or short-term forecasting. While "prediction" is often used in machine learning for tasks like next-token generation or immediate trajectory forecasting, "foresight" implies a deeper, more causal understanding. It is the ability to model, simulate, and evaluate complex, long-horizon futures, often involving multiple interacting agents and counterfactual "what-if" scenarios.

Foresight Intelligenceis conceptually distinct from the planning-oriented reasoning studied in prior datasets. Planning-oriented benchmarks and agentic systems, such as DriveLM [33], OmniDrive [36], and One Agent [21], mainly focus

on immediate decisions or short-horizon behaviors under observable contexts. Foresight Intelligence, by contrast, requires anticipating long-horizon semantic consequences under multi-agent uncertainty and counterfactual conditions. This constitutes a qualitatively distinct capability: simulation-level cognition rather than reaction-level control.

Motivated by insights from cognitive neuroscience [31] and the demands of autonomous decision-making [34], we posit that Foresight Intelligenceis supported by three foundational pillars:

- Situational Modeling. The ability to build a structured, semantic understanding of the current world state, including agent relationships and environmental constraints.
- Causal and Dynamic Simulation. The core predictive engine that generates physically plausible and temporally coherent futures, including the ability to run "what-if" counterfactual scenarios based on causality.
- **Goal-Oriented Evaluation.** The capacity to reason over simulated futures to assess risks and opportunities, enabling selection of an optimal action that ensures safety and achieves a specific objective.

These three pillars collectively guide the design of `FSU-Bench` and are intrinsically woven into every task.

3.2 Overview

We introduce `FSU-QA` (see Fig. 1), a VQA dataset designed to train and quantitatively evaluate Foresight Intelligence in autonomous driving scenarios, where vehicles must operate in highly dynamic environments and anticipate future situations to ensure safety. `FSU-QA` aims to address two fundamental questions: (1) Whether current VLMs, which are primarily evaluated on present-state understanding, possess the capability for Foresight Intelligence; and (2) Whether the visually realistic data generated by WMs can effectively contribute to developing Foresight Intelligence.

To address these two questions, we design two specialized evaluation paradigms on `FSU-Bench`. Both are cast as a question–answering task in which the model must answer future-related queries based on historical visual observations and motion trajectories.

VLM-oriented evaluation This evaluation is designed to assess the foresight understanding capability of a pretrained VLM ($\mathcal{VLM}$):

$$\hat{a} = \mathcal{VLM}(q, V_{-T_h:0}, Traj_{-T_h:0}) \tag{1}$$

where $\hat{a}$ is the predicted answer, q denotes the textual query, $V_{-T_h:0}$ refers to the historical video observations from $t = -T_h$ to the current time step ($t = 0$), and $Traj_{-T_h:0}$ represents the associated motion trajectories.

WM-oriented evaluation. This evaluation assesses the semantic coherence of WM-generated futures via a proxy-evaluation paradigm: we measure the downstream QA performance gains of VLMs when augmented with WM

Fig. 2: Visualization of FSU-Bench. Yellow boxes represent historical frames, and green boxes represent future frames. The bounding boxes are illustrative only and do not appear in the actual dataset. Content in {} is filled into a sentence template. Note: the questions above are simplified slightly for clarity and brevity.

outputs, treating this gain as a proxy for the semantic utility of the predicted content.

$$\hat{a} = \mathcal{VLM}(q, V_{-T_h:0}, Traj_{-T_h:0}, \hat{V}_{1:T_f}, \hat{Traj}_{1:T_f}) \tag{2}$$

$$\hat{V}_{1:T_f}, \hat{Traj}_{1:T_f} = \mathcal{WM}(V_{-T_h:0}, Traj_{-T_h:0}) \tag{3}$$

where $\hat{V}_{1:T_f}$ and $\hat{Traj}_{1:T_f}$ are predicted future video and trajectory from WM ($\mathcal{WM}$).

4 Evaluation on FSU-Bench

4.1 Dataset Curation & Annotations

Building upon the nuScenes dataset [5], which provides rich sensory and map information, we develop an automatic data annotation pipeline to generate a high-quality dataset enabling both training and systematic evaluation of Foresight Intelligence, as illustrated in Fig. 3.

Scene Pre-processing and Representation. To approximate the driver's perspective, we use videos recorded from the front-facing camera, each approximately 15 seconds long. We segment each video into a 3-second historical observation window and a 12-second future window. Because not every scene is valid for all tasks, we apply filtering to ensure annotation quality, yielding 18 to 28 questions per scene. The HD map is represented using lane lines.

Cognition-Inspired Task Design. To comprehensively train and evaluate Foresight Intelligence, we design nine tasks within FSU-QA (see Fig. 2), grouped

into three major categories that span from low-level perception to high-level reasoning:

*Spatio-temporal Dynamic Reasoning (low-level).*This subset assesses the model's capacity to capture dynamic spatio-temporal relations over long time horizons. While individual sub-tasks such as speed and lane-change prediction are studied in motion forecasting literature, our key distinction is the QA formulation: models must produce interpretable language-grounded answers about ego-vehicle motion states and agent interactions over a 12-second horizon, without access to ground-truth future observations. This formulation bridges motion forecasting with semantic scene understanding, enabling diagnosis of foresight-oriented reasoning capabilities. It involves predicting the ego vehicle's future motion states (both longitudinal and lateral) and understanding its interactions with nearby moving agents, characterized by variations in relative distance and position. This level comprises five tasks: Speed Change, Turn Change, Lane Change, Relative Distance (Rel. Dist.), and Relative Position (Rel. Pos.).

VRU-centric Risk Assessment (mid-level). This level assesses the model's capability to detect and interpret Vulnerable Road Users (VRUs) and identify potential risk zones. It consists of three core annotation tasks:

- Pedestrian Intent (Ped. Int.). Annotate the intent of the closest front-view pedestrian within the next 0–3 seconds, based on trajectory and map information.
- Ego–VRU Relative Position Change (E–V Rel. Pos.). Label the final relative position of the closest VRU with respect to the ego vehicle at the end of the 0–3 second horizon.
- Risk Area. Classify the environmental risk level in the upcoming 3-second interval.

High-level Causal Reasoning (high-level). To mitigate nuScenes' inherent "safety bias", the limited presence of unsafe or hazardous driving cases, we introduce a Counterfactual Prediction (CFP) setting that evaluates a model's sensitivity to unsafe behaviors and its capacity for causal understanding. For instance, consider the query: "If the ego vehicle were to turn left with a constant radius in the next 3 seconds, what would be the most likely outcome?" Such prompts allow us to directly assess the model's rule-grounded counterfactual outcome reasoning ability,

whether it can anticipate the safety implications of a prescribed action, such as predicting if a collision would occur.

Counterfactual Annotation Validity. Generating reliable ground truth for counterfactual scenarios (i.e., what would happen under a prescribed but unexecuted action) is inherently challenging. We address this through two complementary mechanisms. First, our rule-based criteria are grounded in the closed-world geometry of nuScenes: given accurate 3D object trajectories, HD-map lane boundaries, and kinematic states, collision outcomes under a prescribed maneuver (e.g., constant-radius left turn) can be determined with high geometric fidelity via spatial intersection checks—without requiring an external simulator. Second, to empirically validate annotation quality, we conducted a human-

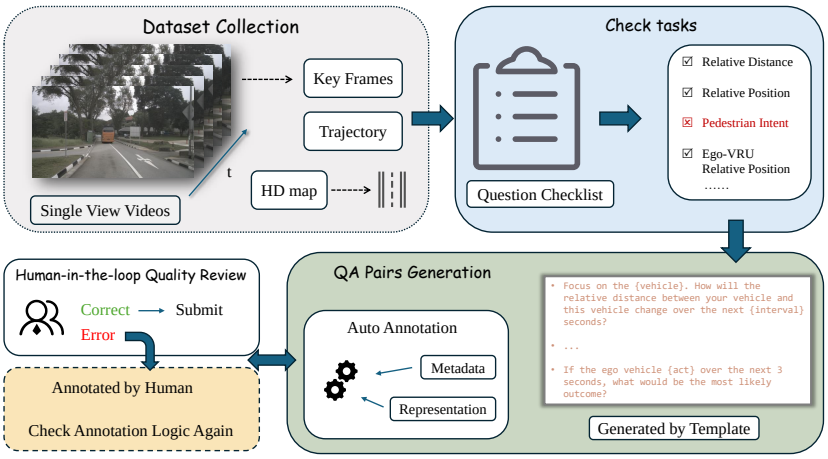


Fig. 3: FSU-QA construction pipeline. Note: the questions above are simplified slightly for clarity and brevity.

in-the-loop verification study in which domain experts independently adjudicated a randomly sampled subset of 150 CFP annotations covering diverse scene types, achieving an agreement rate of 89% with the automated labels (Cohen's $\kappa = 78\%$). These results confirm that rule-based CFP annotation is reliable within the closed geometric constraints of the nuScenes environment.

QA Data Generation. For all nine tasks, we design a comprehensive question checklist using multi-sensor data, 3D object annotations, HD-map representations, and trajectory information. This checklist is applied to each scene to determine the tasks for which annotations are feasible, after which questions are generated via predefined templates. To produce the answers, we establish task-specific label systems and rule-based criteria. We extract and analyze ground-truth information, including kinematic states, spatio-temporal relations, VRU behaviors, environmental risks, and counterfactual verification results, from the dataset and HD-map. Mapping this information to the defined criteria enables systematic derivation of accurate answers for every question.

Specifically, we partition the entire video sequence and trajectory data into N segments of 3 seconds each, denoted as $S_{1:N}$. The first segment S_1 is regarded as the historical observation, whereas the subsequent segments $S_{2:N}$ represent future intervals. For QA generation, we randomly select a timestep i and construct a QA pair conditioned on $S_{1:i}$. This procedure yields QA pairs spanning varying

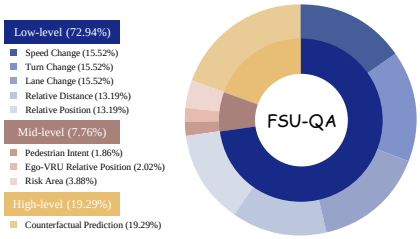


Fig. 4: FSU-QA Statistics. The distribution of tasks across three levels.

Methods	Rank	Overall	Speed Change	Turn Change	Lane Change	Rel. Dist.	Rel. Pos.	Ped. Int.	E-V Rel. Pos.	Risk Area	CFP
			Low-level					Mid-level			High-level
Open-source Models											
Qwen2.5-VL-72B	3	45.86	37.00	90.17	97.50	28.66	14.43	21.74	9.88	64.67	10.31
Qwen2.5-VL-7B	6	44.74	34.00	88.33	97.50	27.85	14.43	37.68	3.70	63.33	8.43
Qwen3-VL-32B	6	44.74	33.83	83.00	95.17	38.01	14.43	17.39	6.17	56.00	11.11
Qwen3-VL-8B	12	41.48	31.67	76.50	97.00	32.93	14.43	20.29	3.70	45.33	5.35
Llama 4 Maverick	4	45.47	41.83	66.00	92.00	33.74	18.50	36.23	18.52	42.67	24.36
Llama 4 Scout	11	41.95	41.67	70.33	81.50	30.69	14.43	36.23	3.70	50.67	16.06
Closed-source Models											
GPT-4o Mini	8	44.22	39.17	86.33	93.33	33.94	14.02	18.84	3.70	36.67	13.65
GPT-5	1	48.66	34.67	75.50	93.00	39.63	32.32	36.23	46.91	48.67	20.75
Gemini-2.0 Flash	13	41.43	32.00	74.17	91.17	33.54	15.04	30.43	7.41	29.33	12.45
Gemini-2.5 Flash	9	44.04	38.83	67.33	94.83	30.69	20.73	23.19	27.16	54.67	14.46
Gemini-2.5 Pro	9	44.04	37.83	66.17	91.67	35.37	22.76	30.34	37.04	25.33	18.47
Claude-3.7-Sonnet	5	45.00	31.83	84.67	92.67	35.16	16.67	33.33	19.75	44.00	14.59
Claude-Sonnet-4.5	2	46.38	37.17	76.67	94.50	35.77	20.33	27.54	14.81	49.33	19.54
Finetuned Model											
Qwen3-VL-8B-FI	-	59.59	43.67	92.33	93.17	38.21	39.63	28.99	38.27	66.00	50.20

Table 2: Baseline Evaluation on `FSU-Bench`: only input historical videos and trajectories. Dark gray cells indicate the best result across all models.

temporal horizons, enabling evaluation of
long-term foresight intelligence.

Quality Control. Our quality control follows a structured human-in-the-loop verification framework. We randomly sampled 45 scenes (approximately 30% of `FSU-Bench`), and three domain experts with backgrounds in autonomous driving independently reviewed each QA pair, categorizing flagged issues into: (i) label logic errors, (ii) edge-case misclassification, and (iii) question ambiguity. For each flagged category, we either corrected the annotation rule globally or handled individual corner cases manually. This two-way iteration was run for three rounds until the human-flagged error rate dropped below 5%, at which point the evaluation set was considered finalized. In total, 654 QA pairs were revised or removed as a result of this process.

Additional information about the dataset curation process is provided in the *Supplemental Materials*.

4.2 Statistics & Analysis

`FSU-QA` contains more than 21K question–answer pairs generated from 850 real-world driving videos in the nuScenes dataset [5]. This ensures that `FSU-QA` robustly covers a wide spectrum of real-world driving scenarios, including different times of day, adverse weather conditions, and diverse urban geographies (Boston and Singapore). Among them, 18K QA pairs (700 scenes) are allocated for train-

ing, and 3K+ QA pairs (150 scenes) are held out for evaluation (`FSU-Bench`). An overview of the `FSU-QA` statistics is provided in Fig. 4.

4.3 Evaluation Setup

Benchmark Models. We conduct a comprehensive evaluation of a broad spectrum of Vision–Language Models (VLMs), spanning closed-source and open-source systems of varying parameter sizes and architectural types, to reveal their capabilities in foresight intelligence.

For closed-source models, we assess *Gemini-2.0-Flash*, *Gemini-2.5-Flash*, *Gemini-2.5-Pro* [10], *GPT-4o-Mini* [14], *GPT-5*, *Claude-3.7-Sonnet-Standard* and *Claude-Sonnet-4.5*.

For open-source models, we evaluate *Qwen2.5-VL-72B-Instruct*, *Qwen2.5-VL-7B-Instruct* [3], *Qwen3-VL-32B-Instruct*, *Qwen3-VL-8B-Instruct*, *Llama-4-Scout*, and *Llama-4-Maverick* [1].

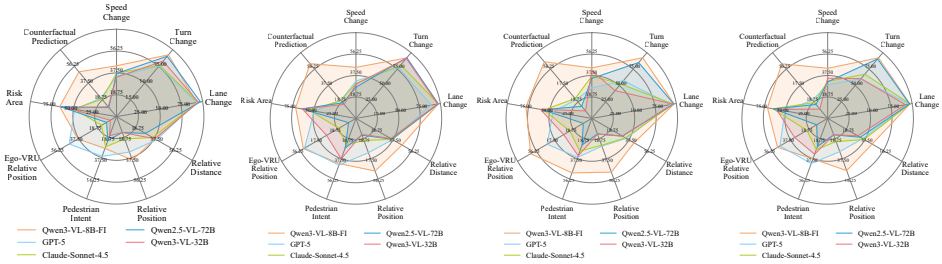
To assess models' capability to generate semantically coherent future data, we employ two representative world models: *Epona* [47], which integrates autoregressive and DiT components, and *DrivingWorld* [12], a purely autoregressive architecture. In the experiments, all previously mentioned VLMs serve as proxy evaluators, enabling a comprehensive evaluation.

Evaluation Protocol. To ensure reproducibility, we detail the full evaluation protocol here and in Appendix A. All VLMs are queried via their official APIs (or local inference for open-source models) using a zero-shot prompt format consisting of a task-specific system instruction, the encoded historical video frames, the trajectory representation, and the multiple-choice question. Frames are sampled at 2 fps from the 3-second historical window, yielding 7 input frames per query. Trajectories are encoded as a sequence of (x, y, θ) waypoints at 2 Hz. For API-based models, including proprietary systems and hosted open-source models, we use the default official API settings with a single-pass query protocol and no retry policy. For locally deployed Qwen models, we use BF16 inference with temperature $= 0.1$ in a single pass, which reduces repetitive outputs observed under strict greedy decoding while keeping generation nearly deterministic. The full prompt template, including system instructions and input formatting, is provided in Appendix.

Metric Design. The questions are categorized into two types: Multiple-Select Questions (MSQ) and Single-Select Questions (SSQ). We adopt *Accuracy* ($\mathcal{ACC}$) as the evaluation metric. An answer is considered correct only when it exactly matches the ground truth; partial matches are not counted as correct. For continuous-valued tasks (e.g., Relative Distance), choices are discretized into labeled bins; we report bin-level accuracy to maintain a unified QA evaluation framework.

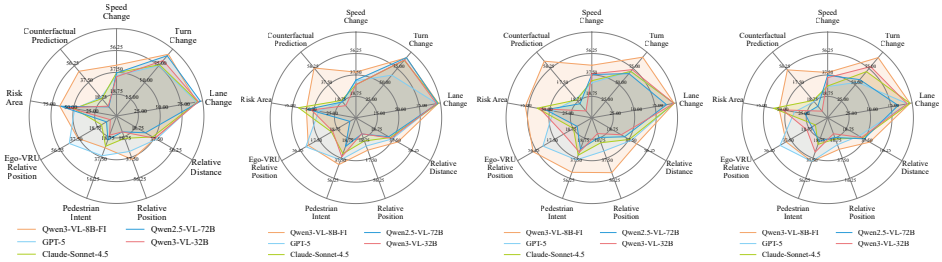
5 Main Results

We follow the joint evaluation pipeline described in Sec. 3.2. Specifically, we first conduct the VLM-oriented evaluation to assess each VLM's foresight intelligence



(a) Baseline **(b)** Base & Pred. Video **(c)** Base & Pred. Traj **(d)** Base & Pred. V+T

Fig. 5: Results with data generated by *Epona*. (a)–(d) correspond to results under four different inputs.



(a) Baseline **(b)** Base & Pred. Video **(c)** Base & Pred. Traj **(d)** Base & Pred. V+T

Fig. 6: Results with data generated by *DrivingWorld*. (a)–(d) correspond to results under four different inputs.

and establish a baseline for subsequent comparisons. We then perform the WM-oriented evaluation, where the semantic coherence of WM-generated future data is quantified by the performance gains over this baseline.

5.1 VLM-oriented Evaluation

As shown in Eq. 1, the historical video sequence and trajectory are used as inputs to the VLM to generate future-related answers.

Closed-source Models. Closed-source models demonstrate overall stronger foresight reasoning capabilities compared to most open-source counterparts, with GPT-5 achieving the highest overall score (48.66) and securing Rank 1 among all evaluated models. Notably, GPT-5 shows clear advantages in mid-level spatio-temporal reasoning, indicating superior ability to understand dynamic interactions between agents.

Claude-Sonnet-4.5 ranks 2nd overall, with a notable advantage on Risk Area (49.33)—the highest among closed-source models—suggesting heightened sensitivity to safety-critical scene cues. Meanwhile, Claude-3.7-Sonnet shows consistent mid-level performance but lags slightly behind Claude-Sonnet-4.5 in high-level reasoning. In contrast, Gemini models (2.0 Flash, 2.5 Flash, 2.5 Pro) ex-

hibit more conservative performance, with overall scores around 41–44. Gemini models score competitively on geometric tasks such as Lane Change (91–95) and Turn Change (66–74), but consistently underperform on Risk Area (25–55) and high-level CFP tasks (12–18), indicating limitations in safety-critical and counterfactual reasoning. Finally, GPT-4o Mini, despite being a lightweight model, achieves competitive overall accuracy (44.22), surpassing some open-source models. To better understand these performance differences, we analyze model-specific error patterns. GPT-5 leads on both Rel. Pos. (32.32) and E-V Rel. Pos. (46.91), suggesting superior capacity for multi-agent interaction modeling and understanding. These patterns suggest that foresight reasoning is not merely a function of model scale, but is sensitive to the richness of semantic scene understanding in the pretraining corpus.

Open-source Models. Among open-source models, Qwen2.5-VL-72B achieves the best overall performance (45.86), ranking 3rd across all models. It exhibits strong capabilities on Low-level foresight tasks, particularly in Turn Change (90.17) and Lane Change (97.50), indicating that large Qwen models excel at recognizing imminent maneuver intentions from short-term visual cues. However, its performance drops noticeably on mid-level tasks, suggesting limited understanding of multi-agent spatial dynamics. Medium-sized models such as Qwen2.5-VL-7B and Qwen3-VL-32B deliver competitive overall scores (44.74), showing that open-source models can remain effective even at smaller scales. Nevertheless, they generally exhibit weaker high-level reasoning, revealing challenges in long-term, safety-critical forecasting. The Llama 4 series shows different behavior: Llama 4 Maverick performs well on mid-level tasks, but its high-level reasoning ability remains lower than closed-source models. Llama 4 Scout, despite being lightweight, maintains stable performance on foundational tasks, though its accuracy decreases on high-level prediction tasks.

Across the board, open-source models tend to excel in low-level maneuver prediction (likely benefiting from large-scale video pretraining) but systematically fall short on complex multi-agent interaction and counterfactual prediction tasks. This performance gap is likely attributable to the scarcity of safety-critical and counterfactual scenarios in standard pretraining corpora, highlighting the value of `FSU-QA` as a targeted training resource for foresight-oriented supervision.

Finetuning with `FSU-QA`. We further finetune Qwen3-VL-8B using `FSU-QA`, yielding the model Qwen3-VL-8B-FI.

Finetuning on `FSU-QA` significantly boosts Qwen3-VL-8B's performance within the nuScenes domain, allowing it to outperform all baselines— demonstrating that `FSU-QA` provides targeted supervision for the foresight reasoning gap identified in zero-shot evaluations. Cross-domain generalization (e.g., to Waymo or nuPlan [6]) remains an important direction for future work.

5.2 WM-oriented Evaluation

We use VLM zero-shot performance as our baseline. The semantic coherence of WM-generated futures is then quantified by measuring downstream QA performance gains when VLMs are augmented with predicted video and/or trajectory.

Crucially, to validate that observed improvements genuinely stem from semantically meaningful WM content rather than incidental input augmentation, we additionally conduct a **shuffled control experiment**: for each test scene, the WM-predicted video and trajectory are replaced with predictions from a randomly selected different scene. Shuffled predictions consistently decrease VLM accuracy across all task categories relative to the baseline, confirming that the gains under ground-truth WM predictions arise from semantically informative future content, not from the mere presence of additional visual tokens.

The results using data generated by *Epona* as input are shown in Fig. 5, while the results using data generated by *DrivingWorld* as input are shown in Fig. 6.

Future Augmentation with Epona. When predicted video is incorporated (b), all models show noticeable gains in tasks requiring short-term motion anticipation. In particular, GPT-5 benefits most from video cues, especially in Relative Distance and Ego–VRU Relative Position, indicating that video sequences effectively support mid-level dynamic understanding. With only predicted trajectories as additional input (c), the improvements shift toward relational reasoning. Models such as Claude-Sonnet-4.5 show larger gains in Pedestrian Intent and Risk Area, suggesting trajectory inputs provide clearer signals for future agent–agent interactions. Meanwhile, Qwen models exhibit more conservative improvements, implying a stronger reliance on visual cues. Finally, combining both predicted video and trajectories (d) yields the most comprehensive benefits. The finetuned Qwen3-VL-8B-FI achieves the most significant and uniform boosts across nearly all tasks, surpassing larger models in several categories such as Counterfactual Prediction and Ego–VRU Relative Position. This indicates that multi-modal predictive inputs effectively complement its finetuned foresight capability. GPT-5 also reaches its strongest performance in this setting, particularly in dynamic maneuver understanding (Turn Change, Lane Change). Overall, these results demonstrate that (1) predictive modalities enhance foresight reasoning in complementary ways—video for motion cues, trajectories for relational cues; and (2) models trained on `FSU-QA`, even with smaller parameter sizes, can substantially benefit from predictive inputs and outperform much larger closed-source systems.

Future Augmentation with DrivingWorld. DrivingWorld-generated predictions yield larger and more consistent gains across VLMs than Epona, particularly on tasks requiring temporal consistency and agent–agent relational reasoning (e.g., Relative Position, Pedestrian Intent). The video–trajectory fusion setting again achieves the strongest overall performance, with improvements in Risk Area and E-V Rel. Pos. being especially pronounced—suggesting that DrivingWorld's autoregressive architecture produces more geometrically coherent futures that better complement VLMs' semantic reasoning. In contrast, Epona's gains are more modest and task-selective, suggesting its predicted futures are less complementary to VLMs' existing visual reasoning, possibly due to differences in motion modeling fidelity or temporal resolution.

6 Conclusion

We presented `FSU-QA` and `FSU-Bench` for training and evaluating Foresight Intelligence in autonomous driving. Our evaluation across 13 VLMs reveals three findings: (1) current models systematically underperform on high-level counterfactual tasks, with the gap not explained by model scale alone; (2) world-model-generated futures provide architecture-dependent semantic gains, with DrivingWorld outperforming Epona on relational reasoning tasks; (3) fine-tuning on `FSU-QA` substantially closes the foresight gap within the nuScenes domain. Future work should address cross-domain generalization and domain adaptation across driving datasets [27], as well as tighter integration between world model generation and VLM-based semantic reasoning.

Acknowledgments

This work is supported by the Agency for Science, Technology and Research (A*STAR) under its Career Development Fund (Project No. H26-KSR0066) and is also supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

The authors would like to sincerely thank the Program Chairs, Area Chairs, and Reviewers for the time and efforts devoted during the review process.

References

1. AI, M.: The llama 4 herd: The beginning of a new era of natively multimodal ai innovation (April 2025)
2. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19107–19117 (2022)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025)
4. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A Multimodal Dataset for Autonomous Driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA (Jun 2020)
6. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810 (2021)
7. Chen, J., Zhu, H., He, X., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Fu, Z., Pang, J., He, T.: Deepverse: 4d autoregressive video generation as a world model (2025)

8. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: SpatialRGPT: Grounded spatial reasoning in vision-language models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
9. Cheng, Y., Wang, R., Yang, X., Prakash, A., Rus, D., Ang Jr, M.H., Li, S.: Perception-aware multimodal spatial reasoning from monocular images. In: IROS (2026)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
11. Cui, Y., Huang, S., Zhong, J., Liu, Z., Wang, Y., Sun, C., Li, B., Wang, X., Khajepour, A.: Drivellm: Charting the path toward full autonomous driving with large language models. IEEE Transactions on Intelligent Vehicles **9**(1), 1450–1464 (2024)
12. Hu, X., Yin, W., Jia, M., Deng, J., Guo, X., Zhang, Q., Long, X., Tan, P.: DrivingWorld: Constructing World Model for Autonomous Driving via Video GPT. arXiv preprint arXiv:2412.19505 (Dec 2024)
13. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6693–6702 (2019)
14. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card (2024)
15. Jain, A., Swerdlow, A., Wang, Y., Arnaud, S., Martin, A., Sax, A., Meier, F., Fragkiadaki, K.: Unifying 2d and 3d vision-language understanding. In: Forty-second International Conference on Machine Learning (2025)
16. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning . In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1988–1997. IEEE Computer Society, Los Alamitos, CA, USA (Jul 2017)
17. Li, R., Li, S., Kong, L., Yang, X., Liang, J.: Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In: CVPR. pp. 3707–3717 (2025)
18. Li, S., Liu, C., Xu, X., Yeo, S.Y., Yang, X.: Future-aware interaction network for motion forecasting. In: ICCV. pp. 7505–7515 (2025)
19. Li, S., Zhou, Y., Yi, J., Gall, J.: Spatial-temporal consistency network for low-latency trajectory forecasting. In: ICCV. pp. 1940–1949 (2021)
20. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Association for Computational Linguistics, Singapore (Dec 2023)
21. Li, Z., Zheng, H., Zhao, F., Chan, A., Zhou, J., Lin, S., Li, S., Wu, Q.: One agent to guide them all: Empowering mllms for vision-and-language navigation via explicit world representation. arXiv preprint arXiv:2602.15400 (2026)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)
23. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6924–6934 (2025)

24. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-ChatGPT: Towards detailed video understanding via large vision and language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024)
25. Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt. arXiv preprint arXiv:2310.01415 (2023)
26. Mao, J., Ye, J., Qian, Y., Pavone, M., Wang, Y.: A language agent for autonomous driving. arXiv preprint arXiv:2311.10813 (2023)
27. Pan, Y., Li, S., Wu, Y., Yang, X., Zhao, N.: Panda: Unsupervised domain adaptation for multimodal 3d panoptic segmentation in autonomous driving. In: CVPR. pp. 33057–33067 (2026)
28. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4542–4550 (2024)
29. Renz, K., Chen, L., Arani, E., Sinavski, O.: Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11993–12003 (2025)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 (2022)
31. Schacter, D.L., Addis, D.R., Buckner, R.L.: Remembering the past to imagine the future: the prospective brain. Nature Reviews Neuroscience **8**(9), 657–661 (Sep 2007)
32. Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., Zhao, B.: Video-xl: Extra-long vision language model for hour-scale video understanding. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26160–26169 (2025)
33. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: European conference on computer vision. pp. 256–274. Springer (2024)
34. Sutton, R., Barto, A.: Reinforcement learning: An introduction. IEEE Transactions on Neural Networks **9**(5), 1054–1054 (1998)
35. Wang, Q., He, J., Pan, Y., Yeo, S.Y., Yang, X., Li, S.: Monosr: Open-vocabulary spatial reasoning from monocular images. In: ECCV (2026)
36. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22442–22452 (Jun 2025), iSSN: 2575-7075
37. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W.,

Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
38. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y.: Internvideo: General video foundation models via generative and discriminative learning (2022)
39. Wang, Z., Zhang, Z., Pang, T., Du, C., Zhao, H., Zhao, Z.: Orient anything: Learning robust object orientation estimation from rendering 3d models. In: Forty-second International Conference on Machine Learning (2025)
40. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6902–6912 (2024)
41. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9772–9781 (2021). https://doi.org/10.1109/CVPR46437.2021.00965
42. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. IEEE Robotics and Automation Letters **9**(10), 8186–8193 (Oct 2024)
43. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. 2104.10157 (2021)
44. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10632–10643. IEEE, Nashville, TN, USA (Jun 2025)
45. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., Lin, D., Wang, T., Pang, J.: Mmsi-bench: A benchmark for multi-image spatial intelligence (2025)
46. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Feng, Y., Lefever, E. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553. Association for Computational Linguistics, Singapore (Dec 2023)
47. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., Cao, X., Yin, W.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 27220–27230 (October 2025)
48. Zhang, Z., Wang, Z., Zhang, G., Dai, W., Xia, Y., Yan, Z., Hong, M., Zhao, Z.: DSI-Bench: A Benchmark for Dynamic Spatial Intelligence. arXiv preprint arXiv:2510.18873 (Oct 2025)
49. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Zhang, K., Xu, S., Chen, D., Timofte, R., Van Gool, L.: Equivariant multi-modality image fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25912–25921 (2024)
50. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models (2025)