# THE CHALLENGE OF RELIABLE VISION–LANGUAGE MODEL RESPONSES IN DRIVING

## **Anonymous authors**

000

001

002003004

006

008 009

010 011

012

013

014

015

016

017

018

019

020

021

022

024

025

026

027

028

029

031 032 033

034

037

040

041

042 043

044

046

047

048

051

052

Paper under double-blind review

#### **ABSTRACT**

Reliable decision-making relies on both prediction and reasoning. In this work, we investigate whether Vision-Language Models (VLMs), when applied as driving assistants, can genuinely understand how present observations shape future outcomes, or whether their outputs merely reflect patterns memorized during training without grounded temporal reasoning. While recent efforts have integrated VLMs into autonomous driving, prior studies typically emphasize scene understanding and instruction generation, implicitly assuming that strong visual interpretation naturally enables future reasoning and thus ensures reliable decision-making—a claim we critically examine. We identify two major challenges limiting VLM reliability in this setting: response inconsistency—where minor input perturbations yield different answers or, in some cases, responses degenerate toward nearrandom guessing—and limited temporal reasoning, in which models fail to reason and align sequential events from current observations, often resulting in incorrect or even contradictory responses. Moreover, we find that models with strong visual understanding do not necessarily perform best on tasks requiring temporal reasoning, indicating a tendency to over-rely on pretrained patterns rather than modeling temporal dynamics. To address these issues, we adopt existing evaluation methods and introduce FutureVQA, a human-annotated benchmark dataset specifically designed to assess future scene reasoning. In addition, we propose a simple yet effective self-supervised tuning approach that improves both consistency and temporal reasoning without requiring temporal labels.

The data and code for our experiments will be released upon acceptance.

# 1 Introduction

Modern Vision-Language Models (VLMs) exhibit human-like perception and reasoning capabilities, enabling more natural and intelligent interactions in everyday applications [Liu et al. (2023b]a); [Wang et al. (2023); [Liu et al. (2024); [Bai et al. (2023); [Young et al. (2024); [Zhang et al. (2023b]). Recent studies [Jiang et al. (2024); [Hwang et al. (2024); [Fu et al. (2025); [Renz et al. (2025)] have explored their potential as driving assistants, applying them to scene analysis and decision-making in complex driving environments. Trained on large-scale visual data, these models demonstrate strong abilities in interpreting visual cues and traffic signs, generating high-level driving instructions that resemble human reasoning and can assist autonomous vehicles.

However, despite these encouraging advancements, most existing approaches implicitly assume that strong visual understanding naturally translates into reliable future scene prediction and reasoning. In this work, we critically examine this assumption by evaluating the consistency and reliability of VLM responses in driving scenarios. Specifically, we investigate whether these responses stem from genuine temporal reasoning or merely reflect memorized knowledge acquired during pretraining Fatemi et al. (2024); Xu et al. (2024).

Specifically, we address the following challenges: (1) Response inconsistency, where identical or nearly identical inputs can lead to divergent or unstable outputs; and (2) limited temporal reasoning, where the model fails to maintain coherent reasoning across events that unfold over time, often producing incorrect predictions or even contradictory answers to follow-up questions requiring temporal understanding. These issues highlight a fundamental limitation: the model's lack of temporal grounding. Unlike humans, VLMs do not experience the flow of time and may over-rely on mem-

Figure 1: Reliability failures in VLMs. The figure illustrates three issues: (i) response inconsistency—identical or very similar prompts yield different answers; (ii) contradiction—correct local interpretation but inconsistent future description; and (iii) temporal misalignment—events predicted at incoherent times despite accurate per-frame cues.

orized patterns from pretraining rather than performing genuine temporal reasoning Fatemi et al. (2024).

Our experiments show that both open-source and commercial VLMs exhibit varying degrees of inconsistency when answering driving-related questions, even under minimal input perturbations—such as shuffling the order of answer options in a visual question answering (VQA) task. Furthermore, we find that models with stronger visual understanding are not necessarily better at reasoning about future scenes or events. In some cases, these models perform worse than others, revealing a disconnect between visual perception and temporal reasoning. These findings underscore a critical concern: the potential risks of deploying VLMs in safety-critical applications such as autonomous driving, where consistent and temporally grounded reasoning is essential.

Alongside standard evaluation methods, we introduce FutureVQA, a fully human-annotated benchmark designed to assess how well VLMs can reason about future scenes based on their understanding of preceding visual observations. In addition, we propose a simple yet effective self-supervised tuning approach that improves the model's ability to perform consistent temporal reasoning and scene prediction—without requiring explicit temporal labels.

In summary, our main contributions include: (1) We identify and analyze key limitations of current Vision-Language Models (VLMs) in driving scenarios, including response inconsistency and lack of temporal reasoning, which pose risks for safety-critical applications. (2) We introduce **FutureVQA**, a human-annotated benchmark designed to evaluate VLMs' ability to reason about future scenes based on prior visual context. (3) We propose a simple yet effective self-supervised tuning method that enhances temporal consistency and future scene prediction without requiring temporal supervision.

#### 2 Related Work

Vision Language Models: Recent advances in LLMs have greatly expanded the scope of multimodal research. In the visual domain, models like LLaVA Liu et al. (2023b), QWen Bai et al. (2023), Yi-VL Young et al. (2024), and CogVLM Wang et al. (2023) have made significant strides in image-text reasoning, offering detailed analyses of visual data alongside textual descriptions. LLaVA-Next Liu et al. (2024) further enhances this capability by supporting higher-resolution inputs, enabling more detailed image understanding. For video-text understanding, models such as Video-LLaMA Zhang et al. (2023b) and LLaVA-Video Zhang et al. (2024) have advanced narrative comprehension by incorporating temporal information from dynamic visual content. Beyond generic VLMs, modern VLMs are increasingly integrated into autonomous driving Nie et al. (2024); Chen et al. (2024b); Liao et al. (2024c); Pan et al. (2024c); Gopalkrishnan et al. (2024c), Ehou et al. (2024a); You et al. (2024c); Chen et al. (2024a); Sima et al. (2023c); Wang et al. (2024a), enhancing scene understanding and decision-making.

**Future Scene Reasoning:** Predicting future scenes is a crucial task in robotics and autonomous driving, requiring models to understand the physical world and how scenes evolve over time. Recently, the construction of world models has gained popularity across various modalities, including point

cloud generation, which aims to construct a realistic 3D representation of the world over time Khurana et al. (2023); [Huang et al.] (2024); [Yang et al.] (2024b); [Manivasagam et al.] (2020), and video generation under different environmental conditions and control signals [Wang et al.] (2024b); [Zhao et al.] (2024); [Hu et al.] (2023); [Zhou et al.] (2024b); [Wang et al.] (2024c); [Jia et al.] (2023); [Hassan et al.] (2024).

Visual Question Answering: Early VQA datasets primarily focused on general image-question performance Antol et al. (2015); Zhang et al. (2016); Goyal et al. (2017). Beyond general-purpose VQA, researchers have explored domain-specific applications, such as medical VQA Lau et al. (2018); Bae et al. (2024) and science-driven VQA Kembhavi et al. (2017). To provide a more robust evaluation, some datasets go beyond free-form sentence answers and adopt structured answer formats, such as Yes/No questions Fu et al. (2024) and multiple-choice formats Liu et al. (2023d); Wu et al. (2024); Fu et al. (2024), ensuring a more consistent and objective assessment of model performance. Recently, VQA in autonomous driving has gained attention, aiming to enhance scene understanding in dynamic traffic environments Sachdeva et al. (2024); Malla et al. (2023); Sima et al. (2023); Wang et al. (2024a); Qian et al. (2023); Deruyttere et al. (2019); Vasudevan et al. (2018).

# 3 Problem Formulation and Evaluation

A reliable safe-driving assistant should anticipate how actions and events unfold over time, remain temporally coherent, and respond consistently under semantics-preserving prompt changes. Let  $V_t = \{I_i \mid i \leq t\}$  be the historical frames up to time t. A VLM  $\psi$  is queried to produce a description  $a_{t+\Delta t}$  of the scene at time  $t+\Delta t$ , where  $\Delta t \in \mathbb{Z}^+$  is the prediction horizon. Using the model's response when given the ground-truth future frame  $I_{t+\Delta t}$  as a reference, reliability requires alignment between past-only and future-conditioned predictions:

$$P_{\psi}(a_{t+\Delta t} \mid V_t) \approx P_{\psi}(a_{t+\Delta t} \mid I_{t+\Delta t}). \tag{1}$$

#### 3.1 RESPONSE UNRELIABILITY AND INCONSISTENCY

Existing studies on language models highlight reliability issues, including hallucination Kalai et al. (2025) and sensitivity to input phrasing Ahn & Yin (2025). These concerns are acute in autonomous driving, where decisions must rest on consistent and trustworthy reasoning. For the multiple-choice VQA variant with input x and K options, the VLM  $\psi$  induces a categorical distribution  $P_{\psi}(k \mid x)$  over answers  $k \in \{1, \ldots, K\}$ . We consider semantics-preserving perturbations  $T_{\sigma}(x)$  such as shuffling options by permutation  $\sigma \in S_K$ , and align labels via  $\tilde{P}_{\psi}(k \mid T_{\sigma}(x)) := P_{\psi}(\sigma(k) \mid T_{\sigma}(x))$ .

One potential source of inconsistency is *prompt-perturbation sensitivity*, which we describe as a distributional shift under such perturbations, measured by a nonzero total-variation (TV) distance:

$$TV(P_{\psi}(\cdot \mid x), \, \tilde{P}_{\psi}(\cdot \mid T_{\sigma}(x))) = \frac{1}{2} \sum_{k=1}^{K} |P_{\psi}(k \mid x) - \tilde{P}_{\psi}(k \mid T_{\sigma}(x))| > 0.$$
 (2)

Another manifestation is a change in the top-1 prediction under perturbations, denoted as the flip rate (FR) with ties broken by the smallest index:

$$FR(x) := \Pr_{\sigma \sim \text{Unif}(S_K)} \left[ \arg \max_{k} P_{\psi}(k \mid x) \neq \arg \max_{k} \tilde{P}_{\psi}(k \mid T_{\sigma}(x)) \right], \tag{3}$$

Another potential source of inconsistency is  $random\ guessing$ : even when Equation (2) and Equation (3) are (near) zero, repeated runs may differ because the model samples from a near-uniform distribution. In this regime, predictions have accuracy  $\approx 1/K$ , entropy  $\approx \log K$ , self-agreement  $R_2(x) = \sum_{k=1}^K P_{\psi}(k \mid x)^2 \approx 1/K$ , and are invariant to semantics-preserving perturbations (i.e.,  $TV \approx 0$  and FR(x) = 0 at the distribution level with a fixed tie-break). In Section [5], we show that—regardless of model size—both open-source and commercial VLMs exhibit accuracy drops and elevated flip rates under option shuffles with the question fixed, consistent with distributional shifts rather than uniform guessing.

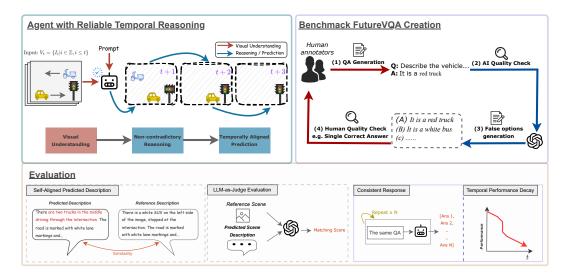


Figure 2: Overview of our framework for evaluating reliable temporal reasoning in VLM driving assistants. Left: The agent consumes past frames  $V_t$  and a prompt to generate temporally aligned predictions over a *variable* future horizon. Right (FutureVQA): Benchmark construction combines human and AI contributions: human experts create natural Q/A pairs, while AI performs quality control to ensure answerability and consistency. Bottom (Evaluation): To thoroughly analyze model reliability, we adopt a self-aligned future description setup, where a model's predicted description is compared to a reference response generated by the same model when the actual future frames are provided. An AI checker is further applied to validate that predictions remain coherent and meaningful. Beyond this, we evaluate consistency under repeated queries and option shuffling, and analyze temporal performance decay to quantify how model reliability changes as the prediction horizon increases.

# 3.2 CONTRADICTION AND TEMPORAL MISALIGNMENT

Despite VLMs' ability to accurately interpret current traffic conditions, they often produce contradictory descriptions when reasoning about future scenes. As shown in Figure [1], a model may correctly identify visual cues and vehicle intentions at the current time based on the input, yet fail to answer follow-up questions consistently. These contradictions suggest that rich and accurate visual interpretation alone does not equip VLMs with the ability to reason about how a scene may evolve over time. In other words, while the model may learn associations between images and their corresponding textual descriptions, it does not genuinely understand their real-world implications or how present actions influence future outcomes. Another issue is temporal misalignment. As shown in Figure [1], VLMs may correctly interpret visual cues and identify individual events, yet they often fail to align these within a coherent temporal structure, as they do not experience time flow as humans do. This limitation is especially critical in driving scenarios, where outcomes such as collisions depend not only on the intentions of surrounding agents but also on the precise timing of their movements.

**Formalization.** Let  $\mathcal{A}$  be the response space and  $\mathcal{R}_{t+\Delta} \subseteq \mathcal{A}$  the set of admissible (reference) responses for time  $t+\Delta$ . Consider two tasks that condition on different information sets:

$$\psi_{\text{pred}}^{\star} := \arg \max_{\psi} \mathbb{E} \big[ P_{\psi}(\mathcal{R}_{t+\Delta} \mid V_t) \big], \qquad \psi_{\text{ref}}^{\star} := \arg \max_{\psi} \mathbb{E} \big[ P_{\psi}(\mathcal{R}_{t+\Delta} \mid V_{t+\Delta}) \big], \qquad (4)$$

where  $V_t$  is the history up to t and  $V_{t+\Delta}$  denotes the future slice at  $t+\Delta$ . In general, these Bayes-optimal solutions are *not necessarily the same*—they are not guaranteed to coincide:

$$\psi_{\text{pred}}^{\star} \neq \psi_{\text{ref}}^{\star}$$
 (5)

Empirically (Section 5), we observe behavior consistent with this non-equivalence: models that perform well when directly shown  $V_{t+\Delta}$  can still contradict themselves across follow-ups and exhibit temporal misalignment when forecasting from  $V_t$  alone.

219220221222

223

224225226

227

228229

230

231

232

235

236

237

238239240

241 242

243

244

245

246 247

249

250

251

253

254

255

256257

258259260

261

262

264

265

266

267

268

269

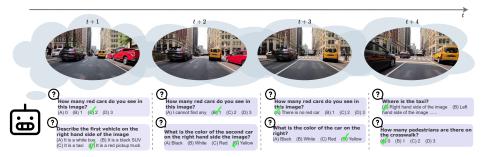


Figure 3: Example of the FutureVQA task. The VLM is asked to answer questions about future scenes based on predictions, without access to the corresponding future frames.

#### Algorithm 1 Self-Aligned Future Description Algorithm 2 Multi-trial Evaluation for Consistency **Require:** Model $\psi$ , Visual Input $V_t = \{I_i \mid i \leq$ **Require:** Model $\psi$ , Question Q, Visual Input $V_t$ , Answer A, Number of Trials Nt, horizon $\Delta t \in \mathbb{Z}^+$ , similarity/quality measure $\mathcal{M}(\cdot, \cdot)$ , threshold $\tau$ 1: **for** i = 1 to N **do** 1: $a_{t+\Delta t}^{\text{pred}} \leftarrow \psi(V_t, \Delta t)$ {Predicted response at 2: $Q_i \leftarrow \text{ShuffleOptions}(Q)$ 3: $P_i \leftarrow \psi(V_t, Q_i)$ $t+\Delta t$ from history 4: if $P_i \neq A$ then 2: $a_{t+\Delta t}^{\text{ref}} \leftarrow \psi(\{I_{t+\Delta t}\}, 0)$ {Reference response 5: return False using actual future frame} 6: end if 3: $q \leftarrow \mathcal{M}(a_{t+\Delta t}^{\text{pred}}, a_{t+\Delta t}^{\text{ref}})$ 7: end for 4: return q8: return True

#### 3.3 EVALUATION AND METRICS

This section turns the reliability criteria from Section into practical tests. Since comparing full distributions  $P_{\psi}(\cdot \mid \cdot)$  is impractical, we use paired queries and controlled perturbations as proxies. We evaluate (i) self-alignment between past-only predictions and future-conditioned references, (ii) stability to semantics-preserving prompt changes (paraphrases, option shuffles with label alignment), and (iii) behavior across horizons  $\Delta t$ .

**Self-Aligned Future Description.** As in Algorithm  $\boxed{1}$  we test whether a model's description of the future scene based on past context  $V_t$  aligns with the description it produces when directly given the future slice  $V_{t+\Delta t}$ . We compare the predicted response  $a^{\text{pred}}$  with the reference response  $a^{\text{ref}}$  using a similarity measure  $\mathcal{M}$ . A conventional choice for  $\mathcal{M}$  is to adopt statistical metrics developed for machine translation Papineni et al. (2002); [Lin (2004); Banerjee & Lavie (2005); Vedantam et al. (2014); Anderson et al. (2016). Typical examples include BLEU Papineni et al. (2002) and ROUGE [Lin (2004), which compute n-gram overlaps between sentences. This general family can be expressed as

$$\mathcal{M}_{\text{n-gr}}(a^{\text{pred}}, a^{\text{ref}}) = f\left(\frac{\sum_{n=1}^{N} w_n \cdot g_n(a^{\text{pred}}, a^{\text{ref}})}{\sum_{n=1}^{N} w_n}\right) \cdot BP,\tag{6}$$

where  $g_n$  denotes an n-gram similarity function weighted by  $w_n$ ,  $f(\cdot)$  applies a transformation (e.g., geometric mean), and BP is a brevity penalty to adjust for length differences.

**LLM-as-Judge Evaluation.** While widely used for evaluating language models, statistical methods struggle to capture in-depth spatial relationships Chang et al. (2024a) and the complex semantic meanings Zheng et al. (2023) handled by modern models. An alternative is *model-based evaluation* Liu et al. (2023c); Zheng et al. (2023); Fu et al. (2023b); Yuan et al. (2021); Sellam et al. (2020); Chang et al. (2024ab), which leverages an advanced judge model  $\mathcal{J}_m$  to assess response quality. In this setup,  $\mathcal{J}_m$  is prompted to rate a response based on the visual input  $x_t$ , producing a score in the range  $\mathcal{J}_m(a^{\text{pred}}, x_t) \in \mathbb{Z} \cap [1, 10]$ . In our experiments we use GPT-4o as the judge, with details and prompt templates provided in Appendix D

FutureVQA Benchmark. To complement existing evaluation metrics and address their limitations in capturing temporal reasoning and visual dynamics, we introduce the FutureVQA Benchmark (Figure 2)—a dataset comprising 2.7k manually annotated question-answer pairs. While existing datasets such as DriveLM Sima et al. (2023) contribute to general scene understanding, they do not explicitly challenge VLMs on time-specific future prediction. Moreover, many rely on structured templates or rule-based generation, which limits the diversity and naturalness of question formats. In contrast, our dataset is constructed by human expert annotators based on individual video clips, featuring diverse and naturally phrased questions tailored to each scene. See Figure 3 and Algorithm 2 for the benchmark exampls and the multi-trial protocol. For a detailed comparison and an overview of the dataset's contributions, please refer to Appendix A

We evaluate performance across horizons from t+1 to t+12 seconds using accuracy (%). To capture both pointwise and temporal trends, we report: (i)  $\mathbf{Acc@t}$ , accuracy at horizon t, reflecting prediction capability at different time steps; (ii)  $\mathbf{\Delta}\mathbf{Acc_{1s}^{12s}}$ , the accuracy drop between t+1 and t+12, indicating performance decay; (iii)  $\mathbf{mAcc_{(1\to12s)}}$ , mean accuracy over horizons 1-12, summarizing overall performance; (iv)  $\mathbf{Normalized\ Drop\ Ratio\ (NDR)}$ , defined as  $\mathbf{NDR} = \frac{1}{\eta_0} \sum_{t=1}^T (\eta_{t-1} - \eta_t)$ , the cumulative accuracy drop normalized by the initial value  $\eta_0$ , where  $\eta_t$  denotes accuracy at horizon t; and (v)  $\mathbf{Mean\ Relative\ Accuracy\ Retention\ (mRAR)}$ ,  $\mathbf{mRAR} = \frac{1}{T} \sum_{t=1}^T \frac{\eta_t}{\eta_0}$ , the average ratio of accuracy at each horizon relative to the initial value.

# 4 FUTUREAGENT: AN APPROACH FOR ENHANCED TEMPORAL REASONING

To address the limitations in temporally grounded reasoning, we propose a self-supervised fine-tuning approach, as illustrated in Figure 4. The design is motivated by two key challenges: (1) the scarcity of large-scale, high-quality temporal annotations for future scene understanding; and (2) the need to align temporally distributed events based on partial visual context.

Instead of relying on expensive manual labels, we leverage the original pretrained model  $\psi$  to generate pseudo reference descriptions  $a_{t+\Delta t}^{\rm ref}$  using ground-truth future frames  $I_{t+\Delta t}.$  We then fine-tune a new model  $\psi^*,$  initialized from  $\psi,$  to predict these descriptions from past-only inputs  $I_{t-k:t},$  without access to future frames. This encourages the model not only to interpret the current visual input but also to imagine and temporally align possible future events. A time-aware weighting function  $\lambda(\Delta t)$  is applied to modulate the loss contribution from different future steps, allowing the model to focus differently on short-term versus long-term temporal reasoning. In practice, we set k=5, using 5 seconds of past ob-

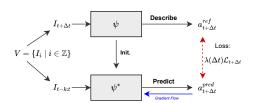


Figure 4: Proposed self-supervised approach to align temporal events and minimize incorrect or contradictory reasoning. Given a video sequence V, we generate detailed descriptions using a pretrained VLM  $\psi$  as pseudo reference labels  $a_{t+\Delta t}^{\rm ref}$ . We then finetune the model  $\psi^*$ , initialized from  $\psi$ , using only past frames as input and training it to predict descriptions of unseen future frames  $a_{t+\Delta t}^{\rm pred}$ . A weighting function  $\lambda(\Delta t)$  adjusts the contribution of each loss term based on the temporal distance  $\Delta t$ .

servations (sampled at 1 frame per second) as input. We observed that increasing the window to 10 seconds did not improve performance but significantly increased computational cost. The weighting function  $\lambda(\Delta t)$  is implemented as an exponential decay:  $\lambda(\Delta t) = 2^{-\Delta t}$ , assigning lower importance to predictions further into the future while still allowing for multi-scale temporal supervision. See Appendix  $\overline{\mathbb{B}}$  for more implementation details.

### 5 EXPERIMENT AND ANALYSIS

In this section, we evaluate how well VLMs can reason about and describe potential future scenes based on preceding visual observations. Specifically, we analyze two key aspects: (1) whether the model can generate consistent responses under minimal input perturbations, which serves as an indicator of genuine understanding versus random guessing; and (2) whether the model can accurately reason about future scenes by interpreting the given history frames

VLM	Evaluation Method		$S-M\downarrow$	$S/M\uparrow$
V LIVI	Single-Trial ↑	Multi-Trial ↑	$S = M \downarrow$	D/W
GPT-40 Hurst et al. (2024)	76.2%	66.1%	11.1%	86.7%
GPT-40-mini Hurst et al. (2024)	66.9%	54.5%	12.4%	81.5%
LLV-v1.5-7b Liu et al. (2023b)	55.1%	33.8%	21.3%	61.3%
LLV-v1.5-13b Liu et al. (2023b)	61.0%	42.3%	18.7%	69.3%
LLV-Next-13b Liu et al. (2024)	41.8%	18.7%	23.1%	44.7%
LLV-Video Zhang et al. (2024)	65.4%	58.1%	7.3%	88.8%
Qwen-VL-7b Bai et al. (2023)	24.6%	4.6%	20.0%	18.7%
Qwen2.5-VL-7b Bai et al. (2025)	79.1%	69.1%	10.0%	87.4%
CogVLM-17b Wang et al. (2023)	53.1%	29.3%	23.8%	44.8%
Yi-VL-34b Young et al. (2024)	60.9%	41.2%	19.7%	67.7%
Vid-LMA2 Zhang et al. (2023a)	67.6%	54.3%	13.3%	80.3%
Baseline <sup>†</sup>	64.5%	51.4%	13.1%	79.7%
FutureAgent <sup>†</sup>	62.7%	52.1%	10.6%	83.1%
Baseline*	73.5%	63.5%	10.5%	85.7%
FutureAgent*	72.3%	64.0%	<u>7.8%</u>	89.2%

Table 1: In this evaluation we examine the ability of different VLMs on our evaluation dataset, where multiple answer options are shuffled across several rounds of answering by the VLMs. The accuracy change reflects the difference in performance between single-trial approach and multiple-trial answering, where the LLM must consistently identify the correct option in every round. This method minimizes the influence of random guessing by ensuring that only consistently correct answers are counted. S-M denotes the performance drop from single-trial to multi-trial. The ratio M/S represents the remaining performance.

### 5.1 EVALUATION SETUP AND IMPLEMENTATION DETAILS

All experiments were conducted on a server equipped with  $4\times A100$ -80GB GPUs. For fine-tuning, we utilized all 4 GPUs, while evaluation was performed using a single GPU for all models. In the Future VQA benchmark, each input consists of a 5-second video segment, and the task is to reason about the future scene at time steps t=1 to t=12 seconds. For our fine-tuning method, we sampled training data from the OpenDV-YouTube dataset Yang et al.] (2024a), covering 16 cities across different continents. This subset comprises approximately 84k frames, each with a resolution of  $1280\times720$ , captured at various times of day. Training required approximately 140 GPU hours. Our base model uses Hermes-Yi-34B as the language backbone and CLIP-L Radford et al. (2021) as the visual token encoder. It is pretrained using the LLaVA v1.6 [Liu et al.] (2024) pipeline, and we refer to this model as Baseline\* in our experiments. The fine-tuned version is denoted as Ours\*. We also evaluate a variant using Qwen-VL-32B as the language model, denoted as Baseline† and Ours† after fine-tuning.

#### 5.2 Consistency and Reliability of VLMs Response

In Table [] we evaluate the performance of various VLMs on our proposed FutureVQA benchmark using the corresponding image for each question-answer pair as input—i.e., no future prediction is required. This setup serves both as a baseline for future scene reasoning and as a diagnostic to assess the consistency and reliability of VLM responses. Notably, we observe that all tested VLMs exhibit a significant drop in accuracy when the answer options are simply shuffled, despite the semantic content of the questions remaining unchanged. The most substantial decline occurs with CogVLM [Wang et al.] (2023), which drops by 23.8%, followed by LLaVA-NeXT 13B [Liu et al.] (2024) with a 23.1% decrease.

**Prompt-perturbation sensitivity vs. random guessing.** The performance drop in Table 1 across all models is largely attributable to random guessing, as the decrease scales with the number of options (four in our setup) and the number of trials. In contrast, Figure 5a shows error bars that reflect much smaller shifts when repeating four trials multiple times. These fluctuations (typically 0.5–1.2 points) arise from prompt-perturbation sensitivity: responses are inconsistent across trials but still exhibit a clear preference toward certain answers, rather than uniform randomness.

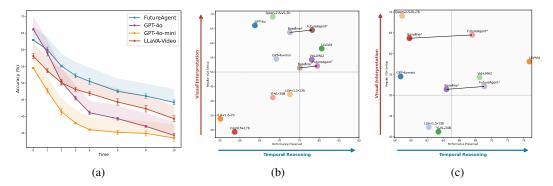


Figure 5: Temporal performance decay analysis on the FutureVQA dataset. (a) Accuracy decay across horizons, where solid lines denote four trials and shaded regions indicate fewer trials (1–3). (b) Relationship between regular VQA performance (y-axis) and relative long-horizon preservation (x-axis: Acc@12 divided by regular VQA accuracy). (c) Relationship between regular VQA performance (y-axis) and relative mean preservation (x-axis:  $\text{mAcc}_{(1\rightarrow 12s)}$  divided by regular VQA accuracy). Together, these plots show how well models retain their performance when extending from immediate perception to future prediction.

Model	Accuracy ↑					- NDR↓	mRAR ↑	
Model	Acc@1s	Acc@4s	Acc@12s	$\Delta Acc_{1s}^{12s}$	$mAcc_{(1 \to 12s)}$	- NDK +	IIIKAK	
GPT-4o	59.1%	41.1%	31.6%	-27.5%	42.2%	0.42	0.64	
GPT-4o-mini	47.7%	36.0%	32.0%	-15.7%	37.7%	0.29	0.69	
LLV-v1.5-7b	24.0%	18.1%	16.0%	-8.0%	18.6%	0.24	0.55	
LLV-v1.5-13b	37.8%	30.9%	26.3%	-11.5%	30.7%	0.27	0.73	
LLV-Next-13b	15.4%	9.3%	4.2%	-11.2%	7.3%	0.60	0.39	
LLV-Video	53.7%	46.5%	43.4%	-10.3%	46.8%	0.18	0.81	
Qwen2.5-VL-7b	61.9%	49.5%	40.7%	-21.2%	47.2%	0.31	0.68	
CogVLM-17b	22.8%	19.4%	14.0%	<u>-8.8%</u>	17.2%	0.30	0.59	
Yi-VL-34b	38.1%	30.0%	26.1%	-12.0%	28.4%	0.29	0.70	
Vid-LMA2	52.4%	41.2%	37.2%	-15.2%	42.4%	0.28	0.78	
Baseline <sup>†</sup>	49.8%	44.1%	33.1%	-16.7%	38.6%	0.33	0.75	
FutureAgent†	49.2%	46.7%	36.0%	-13.2%	41.4%	0.25	0.79	
Baseline*	60.2%	48.2%	38.1%	-22.7%	46.1%	0.36	0.73	
FutureAgent*	60.8%	50.7%	43.6%	16.6%	50.1%	0.21	0.78	
w/o CoT	60.5%	48.4%	41.3%	-19.2%	48.2%	0.30	0.75	

Table 2: Accuracy (Acc) of models on our VQA benchmark at different future time frames. All accuracy values are evaluated across multiple trials to minimize the influence of random chance. The result suggest that models like GPT-40, while showing strong ability in visual understdaning, fail to maintain consistent future scene reasoning across different time interval. †\*Our model is not trained with explicit temporal (video) label.

# 5.3 CAN VLMs "SEE" THE FUTURE?

Effective decision-making in dynamic environments should be grounded in accurate predictions. Here, we investigate whether VLMs are capable of reasoning about future scenes based on their interpretation of present visual cues, and whether they understand how events unfold over time. As shown in Table 2 we evaluate VLMs on our FutureVQA benchmark by asking them to answer questions about unseen future scenes using only the past five seconds of visual input. The task challenges models to make predictions ranging from 1 to 12 seconds into the future. Each question is evaluated using a multi-trial protocol. Interestingly, we find that models that perform best in standard visual understanding tasks do not necessarily excel in future reasoning. For example, while GPT-40 demonstrates strong visual comprehension, its performance drop over time, measured by both  $\Delta Acc_{1s}^{12s}$  and NDR, is significantly higher than that of other models. This suggests that, while equipped with strong visual interpretation capabilities, these models often fail to reason about how a

 scene evolves over time. In particular, they may struggle to understand how present events influence future outcomes, even if they generate accurate responses based on the current image.

In Figure 5b and Figure 5c, we observe that very poor visual interpretation ability typically coincides with weak temporal reasoning—an expected outcome since reliable reasoning requires accurate perception as a foundation. However, models such as GPT-4o Hurst et al. (2024) and Qwen-2.5 Bail et al. (2025), despite strong visual interpretation, experience significant drops when asked to predict the future, suggesting that good perception alone does not guarantee reliable temporal reasoning.

In (Table 3. Table 4.), we compare how closely the predicted future scene descriptions match the model's own descriptions when the actual future image is provided. Ideally, if the prediction is accurate, both descriptions should align, as if the model had seen the future scene. Our results show that, after applying the proposed training method, the predicted descriptions become significantly more accurate and consistent across all time intervals.

Model	Mean $Score_{(0 \rightarrow 12s)} \uparrow$					
Wiodei	mB3	mB4	mRL	mC	mM	
Baseline <sup>†</sup>	10.7	6.0	22.8	2.3	25.4	
FutureAgent†	20.3	19.8	35.2	11.3	34.6	
Baseline*	12.3	7.1	25.2	3.6	28.5	
FutureAgent*	28.8	22.7	37.3	12.3	39.2	
w/o CoT	25.9	20.4	35.9	11.1	38.3	
w/o self-sup.	11.8	6.9	24.7	2.3	26.0	

Table 3: We compare how well our proposed model describes future scenes as if it "sees" them. A higher value indicates greater similarity between the reference description and the predicted description. The mean score, m, is computed over discrete time steps  $t \in \mathbb{Z}_{[1,12]}$  seconds. B3: BLEU-3, B4: BLEU-4, R-L: ROUGE-L, C: CIDEr, M: METEOR.

Model	Score Over Time ↑					
Model	S@1s	S@2s	S@4s	S@8s	S@12s	
LLV-v1.5-7b	2.59	2.67	2.07	2.52	2.25	
LLV-v1.5-13b	2.13	1.92	1.94	2.49	2.40	
LLV-Next-13b	2.11	2.87	2.26	2.57	2.15	
Baseline <sup>†</sup>	4.88	4.01	2.96	2.34	2.41	
FutureAgent†	5.31	5.01	3.98	3.44	.2.46	
Baseline*	5.36	4.23	3.03	3.22	2.98	
FutureAgent*	6.43	6.12	5.33	5.04	4.66	
w/o CoT	5.84	5.44	4.33	4.18	3.92	
(w/o self-sup.	3.72	3.96	3.01	3.19	3.04	

Table 4: Model-based evaluation of predicted caption quality across various time frames using GPT-40, with a specific focus on objective descriptions, such as the accuracy of object appearance and location within the image.

# 6 LIMITATION AND DISCUSSION

While our approach offers data efficiency and improved temporal reasoning, it also presents tradeoffs. The self-supervised fine-tuning relies on the quality of the baseline model; its limitations may propagate through pseudo labels. Future work could explore alternative forms of supervision such as constructing high-quality, large-scale training data. Similarly, although CoT prompting enhances reasoning without additional training, its step-by-step nature increases inference time. This may be a concern in real-time settings. A promising direction is to distill multi-step reasoning into a singlestep model for faster inference. Despite these challenges, our framework provides a practical and extensible foundation for enhancing temporal understanding in VLMs.

#### 7 CONCLUSION

We investigated the foresight capabilities of VLMs and found that, despite strong visual understanding, they struggle with consistent future scene reasoning. To address this, we introduced the FutureVQA Benchmark, a human-annotated dataset designed to evaluate VLMs' perception and prediction across different time intervals. Our experiments demonstrate that conventional models fail to maintain consistency in future predictions, while our self-supervised training pipeline improves temporal reasoning without requiring annotated temporal data. Notably, our model outperforms video-based VLMs despite lacking explicit temporal supervision. These findings highlight the need for better integration of visual perception and temporal reasoning in VLMs.

### REFERENCES

- Jihyun Janice Ahn and Wenpeng Yin. Prompt-reverse inconsistency: Llm self-inconsistency beyond generative randomness and prompt paraphrasing. *arXiv preprint arXiv:2504.01282*, 2025.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. *CoRR*, abs/1607.08822, 2016. URL <a href="http://arxiv.org/abs/1607.08822">http://arxiv.org/abs/1607.08822</a>.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36, 2024.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL <a href="https://aclanthology.org/W05-0909">https://aclanthology.org/W05-0909</a>
- Chun-Peng Chang, Alain Pagani, and Didier Stricker. 3d spatial understanding in mllms: Disambiguation and evaluation, 2024a. URL <a href="https://arxiv.org/abs/2412.06613">https://arxiv.org/abs/2412.06613</a>.
- Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14131–14140, 2024b.
- Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024a.
- Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 14093–14100. IEEE, 2024b.
- Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019.
- Bahare Fatemi et al. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv*, 2024.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023a.
- Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang,
Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. arXiv preprint arXiv:2503.19755, 2025.

- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv* preprint arXiv:2302.04166, 2023b.
- Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- Akshay Gopalkrishnan, Ross Greer, and Mohan Trivedi. Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving. *arXiv preprint arXiv:2403.19838*, 2024.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. *arXiv preprint arXiv:2412.11198*, 2024.
- Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv* preprint arXiv:2309.17080, 2023.
- Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructingworld model for autonomous driving via video gpt. *arXiv* preprint arXiv:2412.19505, 2024.
- Zanming Huang, Jimuyang Zhang, and Eshed Ohn-Bar. Neural volumetric world models for autonomous driving. In *European Conference on Computer Vision*, pp. 195–213. Springer, 2024.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. arXiv preprint arXiv:2410.23262, 2024.
- Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.
- Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.
- Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.
- Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 4999–5007, 2017.
  - Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1116–1124, 2023.

- Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578, 2018.
  - Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
  - Guibiao Liao, Jiankun Li, and Xiaoqing Ye. Vlm2scene: Self-supervised image-text-lidar learning with foundation models for autonomous driving scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3351–3359, 2024.
  - Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.
  - Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv*:2310.03744, 2023a.
  - Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023b.
  - Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <a href="https://llava-vl.github.io/blog/2024-01-30-llava-next/">https://llava-vl.github.io/blog/2024-01-30-llava-next/</a>
  - Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023c.
  - Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023d.
  - Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1043–1052, 2023.
  - Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Lidarsim: Realistic lidar simulation by leveraging the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11167–11176, 2020.
  - Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pp. 292–308. Springer, 2024.
  - Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14760–14769, 2024.
  - Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.
  - Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2023.
  - Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

- Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. 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.
- Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7513–7522, 2024.
- Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.
- Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv* preprint arXiv:2312.14150, 2023.
- Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4129–4138, 2018.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. URL http://arxiv.org/abs/1411.5726.
- Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv*:2405.01533, 2024a.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.
- Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024b.
- Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024c.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.
- Rongwu Xu et al. Knowledge conflicts for llms: A survey. arXiv, 2024.
- Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.
- Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14673–14684, 2024b.
- Junwei You, Haotian Shi, Zhuoyu Jiang, Zilin Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *arXiv preprint arXiv:2408.09251*, 2024.
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

- Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.
- Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023b. URL <a href="https://arxiv.org/abs/2306.02858">https://arxiv.org/abs/2306.02858</a>.
- Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL <a href="https://llava-vl.github.io/blog/2024-04-30-llava-next-video/">https://llava-vl.github.io/blog/2024-04-30-llava-next-video/</a>
- Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv* preprint arXiv:2403.06845, 2024.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024a.
- Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, et al. Learning flow fields in attention for controllable person image generation. *arXiv preprint arXiv:2412.08486*, 2024b.