
FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving

Shuang Zeng^{1,2*}, Xinyuan Chang², Mengwei Xie², Xinran Liu²,
Yifan Bai^{1,3}, Zheng Pan², Mu Xu², Xing Wei^{1†}

¹Xi'an Jiaotong University ²Amap, Alibaba Group ³DAMO Academy, Alibaba Group
zengshuang@stu.xjtu.edu.cn, weixing@mail.xjtu.edu.cn,
{changxinyuan.cxy, xiemengwei.xmw, tom.lxr}@alibaba-inc.com,
{baiyifan.byf, panzheng.pan, xumu.xm}@alibaba-inc.com

Abstract

Vision–Language–Action (VLA) models are increasingly used for end-to-end driving due to their world knowledge and reasoning ability. Most prior work, however, inserts *textual* chains-of-thought (CoT) as intermediate steps tailored to the current scene. Such symbolic compressions can blur spatio-temporal relations and discard fine visual cues, creating a cross-modal gap between perception and planning. We propose **FSDrive**, a *visual* spatio-temporal CoT framework that enables VLAs to think in images. The model first acts as a world model to generate a **unified future frame** that overlays coarse but physically-plausible priors—future lane dividers and 3D boxes—on the predicted future image. This unified frame serves as the visual CoT, capturing both spatial structure and temporal evolution. The same VLA then functions as an inverse-dynamics model, planning trajectories from current observations and the visual CoT. To equip VLAs with image generation while preserving understanding, we introduce a **unified pre-training paradigm** that expands the vocabulary to include visual tokens and jointly optimizes VQA (for semantics) and future-frame prediction (for dynamics). A progressive easy-to-hard scheme first predicts lane/box priors to enforce physical constraints, then completes full future frames for fine details. On nuScenes and NAVSIM, FSDrive improves trajectory accuracy and reduces collisions under both ST-P3 and UniAD metrics, and attains competitive FID for future-frame generation despite using lightweight autoregression. It also advances scene understanding on DriveLM. Together, these results indicate that **visual CoT** narrows the cross-modal gap and yields safer, more anticipatory planning. Code is available at <https://github.com/MIV-XJTU/FSDrive>.

1 Introduction

Recently, given the superior capabilities of multimodal large language models (MLLMs) in world knowledge, reasoning ability, and interpretability, they have been widely applied in autonomous driving [20, 43, 87, 31]. One promising direction is the end-to-end vision-language-action (VLA) model, which leverages pre-trained vision-language model (VLM) to directly extract scene features from visual observations and language instructions, subsequently generating vehicle control commands (e.g., speed and trajectory). This paradigm not only simplifies system architecture and minimizes information loss, but also enables the utilization of the model’s world knowledge to analyze driving environments and reason about safe decisions in complex scenarios.

* Work done during the internship at Amap, Alibaba Group.

† Corresponding author.

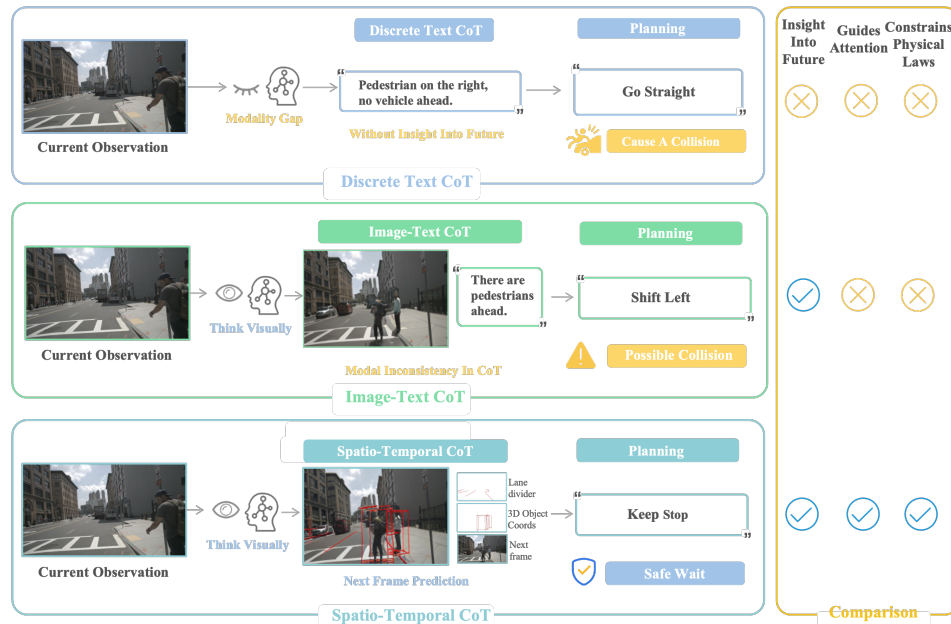


Figure 1: Comparison of different CoT. Textual CoT expression provides insufficient information. The modalities between the image-text CoT are inconsistent. The proposed spatio-temporal CoT captures the temporal and spatial relationships in the future.

In the field of language, Chain-of-Thought (CoT) [67, 50, 15, 52] improves reasoning capabilities and interpretability by encouraging step-by-step thinking. However, existing autonomous driving [27, 44, 14] typically incorporate discrete text CoT (e.g., language descriptions targeting current scenarios and bounding box coordinates) as intermediate reasoning steps. This method is essentially highly abstract and symbolized compression of visual information, which may lead to ambiguous temporal and spatial relationships, loss of fine-grained information, and modality conversion gaps [46, 55, 72], as shown in the top of Figure 1. For autonomous vehicles requiring deep physical-world interaction, should their thinking process more closely resemble simulation and imagination of world, rather than merely relying on logical deduction of language?

Inspired by the human driver's cognitive mechanism of directly constructing visual representations of future scenarios in the mind, rather than converting them into language descriptions for reasoning, we propose a more intuitive spatio-temporal CoT method as shown in the bottom part of Figure 1. This method avoids information loss during text abstraction and enables the model to think visually about trajectory planning. First, the VLA serves as a world model to generate unified image frame for predicting future world states: Inspired by visual prompting engineering [53, 81] that draws red circles on images to guide model attention and by VLIPP [78] first predicts future bounding boxes to introduce physical priors when generating future frames, we represent future world spatial relationships through future red lane dividers and 3D detection boxes on the predicted unified frames [80]. These coarse-grained visual cues direct the model's attention toward drivable areas and critical objects in future scenes while enforcing physically plausible constraints. Meanwhile, the temporal relationships are represented by the ordinary future frame, where the dynamic evolution of visual content intuitively characterizes temporal progression and the inherent laws of scene development. Subsequently, the spatio-temporal CoT acts as an intermediate reasoning step, enabling the VLA to function as an inverse dynamics model for trajectory planning based on current observations and future predictions. Compared to traditional discrete text CoT, and even image-text CoT methods [27, 91, 41] as shown in the middle of the Figure 1, our method unifies both future scene representations and perception outputs in image format, which more effectively conveys the temporal and spatial relationships. This eliminates semantic gaps caused by cross-modal conversions (e.g., converting visual perceptions into textual descriptions for reasoning), establishing an end-to-end visual reasoning pipeline that enables direct visual causal inference by the model.

To endow VLAs with image generation capabilities, we propose a pre-training paradigm that simultaneously preserves the semantic understanding of existing MLLM and activates their visual

generation capacity. Specifically, for the semantic understanding preservation part, we follow previous approaches [64, 27, 25] by incorporating visual question answering (VQA) tasks for current scene comprehension. For the activation of visual generation capabilities, we investigate the shared vocabulary space between image and text, directly unleashing the visual generation potential of existing MLLMs in the field of autonomous driving through minimal data (approximately 0.3% of previous methods [70, 73, 24, 35]) without requiring complex model architecture modifications or redesigns. However, directly generating complete detailed future scenes may fail to adhere to physical laws [78, 88]. Thus, we propose a progressive, easy-to-hard generation method. We leverage the world knowledge of VLAs to first infer drivable regions and key object positions in future scenarios, generating coarse-grained future perception images (e.g., lane dividers and 3D detection) to constrain physical laws. Subsequently, full future frames are generated under this constraint to supplement fine-grained details, enabling the model to think visually about accurate future prediction.

Extensive experiments on trajectory planning, future frames generation, and scene understanding tasks demonstrate the effectiveness of pre-training paradigm and spatio-temporal CoT in FSDrive. FSDrive achieves road scene comprehension by establishing pixel-level embodied associations with the environment, rather than relying on human-designed abstract linguistic symbols, advancing autonomous driving towards visual reasoning. In summary, our main contributions are as follows:

- We propose a spatio-temporal CoT reasoning method that allows the model to enhance trajectory planning by thinking visually from future temporal and spatial dimensions.
- We propose a unified pre-training paradigm for visual generation and understanding. Meanwhile, we introduce a progressive generation approach that evolves from imposing physical constraints to supplementing details.
- We conduct comprehensive evaluations across trajectory planning, future frames generation, and scene understanding tasks, demonstrating the effectiveness of our FSDrive.

2 Related work

2.1 Unified multimodal understanding and generation

Recent research efforts [38, 70, 49, 68] have increasingly focused on unifying multimodal understanding and visual generation within a single LLM. On one front, methods like Show-o [74], and VILA-U [73] employ VQ-VAE [61] to transform images into discrete tokens while training LLMs to predict them. However, these methods suffer from insufficient semantic information preservation, often leading to performance degradation in downstream understanding tasks. Alternative methods [57, 11, 48, 9, 82] utilize ViT [12]-based vision encoders (e.g., CLIP [51]) to encode images into continuous feature maps. Nevertheless, such methods typically depend on external diffusion models for image generation or use different training objectives (i.e. diffusion and autoregression) for the two tasks, further complicates the infrastructure design with overall lower efficiency. Moreover, the aforementioned methods usually require massive billion-scale datasets for extensive training from scratch, which results in prohibitively high computational costs when disseminating explorations in this form. In this work, we demonstrate that the visual generative capabilities of existing MLLMs can be directly activated through minimal training costs (approximately 0.3% of previous methods [70, 58, 42, 8]) without requiring sophisticated architectural designs.

2.2 Vision-language models for autonomous driving

Given the superior capabilities of large language models (LLMs) in world knowledge, reasoning, and interpretability, recent researches [2, 83, 39, 85] increasingly integrate Vision-Language Models (VLMs)/LLMs with autonomous driving systems to address limitations in end-to-end approaches. DriveGPT4 [76] employs LLMs through iterative question-answering interactions to explain vehicle behaviors and predict control signals. DriveVLM [60] synergizes LLMs with end-to-end architectures, where LLMs predict low-frequency trajectories that are subsequently refined by the end-to-end model for final planning. Doe-1 [95] reformulates autonomous driving as a next-token prediction task using Lumina-mGPT's [37] multimodal generation capabilities, executing diverse tasks through multimodal token processing. EMMA [27] leverages Gemini's multimodal foundation by encoding all non-sensor inputs (navigation instructions, vehicle status) and outputs (trajectories, 3D positions) as natural language text, fully exploiting pre-trained LLMs' world knowledge. In this work, we propose a

spatio-temporal chain of thought (CoT) reasoning method that unifies the form of images, allowing the model to think visually about trajectory planning.

2.3 World models for autonomous driving

World models [66, 45, 90, 89] aim to infer ego status and dynamic environments from past observations to enable accurate future prediction and planning. Current applications of world models in autonomous driving primarily focus on driving scenario generation [47, 16, 32], planning [66, 41], and representation learning [45, 79, 84]. For driving scenario generation, most prior works are built upon diffusion models, with the exception of GAIA-1 [18] which incorporates a progressive next-token predictor and an additional diffusion image decoder. Recent DrivingGPT [5] leverages existing vision generation LLM LlamaGen [56] while simultaneously outputting predictions for future states and actions. However, such VQ-VAE based visual tokens lack semantic information, often leading to performance degradation in downstream visual understanding tasks [74, 40, 59]. In this work, we propose to directly activate the visual generation capabilities of existing multimodal large language models, enabling VLMs to act as world models and predict future frames.

3 Proposed method: FSDrive

The proposed FSDrive is illustrated in Figure 2. Section 3.1 describes the preliminaries. Section 3.2 presents a unified visual generation and understanding pre-training paradigm and a progressive generation method. Section 3.3 proposes spatio-temporal chain-of-thought methods. Section 3.4 details the training strategy.

3.1 Preliminary

End-to-end trajectory planning. End-to-end autonomous driving directly generates future trajectory from sensor data, convertible to vehicle control actions like acceleration and steering [27]. given N surround-view images $I_t = \{I_t^1, I_t^2, \dots, I_t^N\}$ at timestep t , model $\mathcal{M}$ outputs a BEV trajectory $W_t = \{w_t^1, w_t^2, \dots, w_t^n\}$, where each waypoint $w_t^i = (x_t^i, y_t^i)$. The process is formulated as:

$$W_t = \mathcal{M}(I_t, \text{opt}(T_{com}, T_{ego})), \quad (1)$$

$\text{opt}(T_{com}, T_{ego})$ denotes optional navigation commands and ego status (e.g., velocity, acceleration).

Unified visual generation and understanding. Recent works [70, 22] unify multimodal understanding and vision generation in single LLM. While understanding aligns with standard LLMs, generation methods [38, 23] typically use VQ-VAE [61] to encode images into discrete tokens. First, the image tokenizer quantizes image pixels $x \in \mathbb{R}^{H \times W \times 3}$ into discrete tokens $q \in \mathcal{Q}^{h \times w}$, where $h = H/p$, $w = W/p$, p is the downsampling factor, and $q(i, j)$ represents the index of the image codebook. These $h \cdot w$ tokens are arranged in raster order to train a Transformer [62]-based autoregressive model. During image generation, a general language modeling (LM) objective is adopted to autoregressively predict the next token, maximizing the likelihood of each image token:

$$\mathcal{L} = - \sum_{i=1} \log P_{\theta}(q_i | q_{<i}), \quad (2)$$

where q_i denotes the visual token and θ represents the LLM parameters. Finally, the VQ-VAE's detokenizer converts these image tokens back into image pixels.

3.2 Unified pre-training paradigm for visual generation and understanding

To enable unified pre-training, MLLMs require visual generation capabilities. As described in Section 3.1, existing methods (e.g. Lumina-mGPT [37], the visual generation LLM used by Doe-1 [95]) typically employ VQ-VAE to encode images into discrete tokens when extracting visual information. However, these tokens lack semantic information, which hurts downstream understanding performance [74, 97]. Moreover, current methods [70, 96] demand expensive training from scratch on massive billion-scale datasets without leveraging existing LLM knowledge.

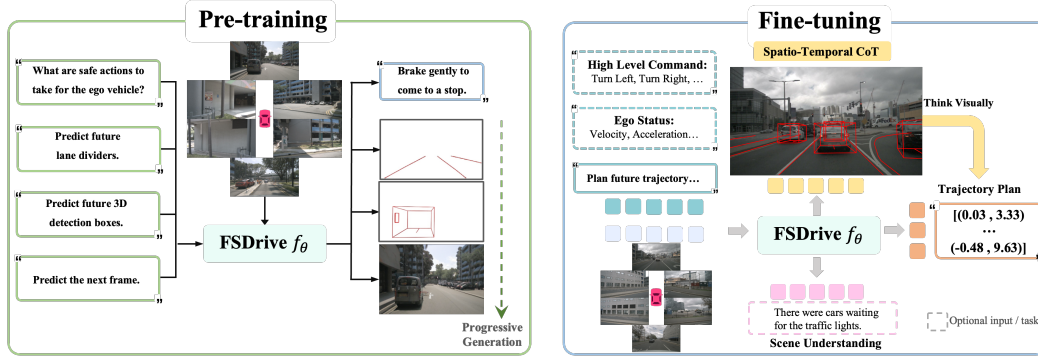


Figure 2: Overview of FSDrive. Taking the currently surround images and task instructions as input, MLLM is trained in the form of next token prediction. MLLM predicts the future spatio-temporal CoT, and then generates trajectory based on the current observation and predicted future.

Our method is directly built upon any existing MLLM that employs ViT-based encoders to convert images into continuous features. We preserve the original MLLM architecture without altering any structural components to maintain compatibility with pretrained weights. The sole modification involves expanding the MLLM’s vocabulary by incorporating image tokens of the VQ-VAE into the text codebook, thereby extending the vocabulary’s scope from language space to a multimodal space encompassing both visual and textual modalities. This enhancement enables the MLLM to predict image tokens, which can then be converted to image pixels through an VQ-VAE’s detokenizer.

Pre-training for visual understanding. To effectively preserve the semantic understanding capabilities of the native MLLM during the pre-training stage, as shown in the left part of Figure 2, we follow previous methods [64, 27] by using a VQA task, which is crucial for autonomous vehicles to analyze complex driving scenarios. Given an image-text question-answer pair (I, L) , where I represents the surround-view images of the current scene and L denotes the instructional question, model $\mathcal{M}$ generates a corresponding answer A :

$$A = \mathcal{M}(I, L). \quad (3)$$

Pre-training for visual generation. Inspired by the world models in autonomous driving [30, 77] that generate future frames to learn physical laws, after activating the visual generation capability, we also enable the VLA to predict future frames, thereby capturing the dynamic evolution of the world. Specifically, given an image-instruction pair (I, L) , the model predicts the next visual token of the future front-view frame through autoregressive generation:

$$P(q_1, q_2, \dots, q_{h \cdot w}) = \prod_{t=1}^{h \cdot w} P_{\theta}(q_i \mid q_{<i}). \quad (4)$$

The predicted visual tokens are then converted back into image pixels by VQ-VAE’s detokenizer. Since future frames naturally exist in video datasets without requiring any labeled data, this approach unlocks the potential to harness abundant video data for improving generation quality.

Progressive image generation. However, directly generating complete detailed future scenes may fail to adhere to physical laws [78]. Therefore, during pre-training stage, we propose a progressive, easy-to-hard generation method, incorporating annotated data containing lane divider and 3D detection. Before generating visual tokens of future frames Q_f , we leverage the world knowledge of VLA to first reason about visual tokens of lane dividers Q_l , which serve as the skeleton of the road scene and define drivable areas to enforce static physical constraints. Subsequently, we reason about visual tokens of 3D bounding boxes Q_d , representing motion patterns of key objects to impose dynamic physical constraints. This progressive method sequence explicitly guides the model to infer structural layouts and geometric details of future scenes while enforcing physical laws. By leveraging these intermediate visual reasoning steps as context, the model learns to think visually about the dynamic evolution of scenes, ultimately enabling accurate future prediction:

$$P(Q_f | Q_l, Q_d) = \Pi_{t=1}^{h \cdot w} P_\theta(q_i | q_{<i}, Q_l, Q_d). \quad (5)$$

3.3 Think visually with spatio-temporal CoT

Autonomous driving planning requires not only understanding the current scene but also envisioning potential future developments to achieve forward-looking comprehension. This thinking process should resemble physical world simulation and imagination rather than purely text symbolic logical deduction. Since our model has already learned physical constraints through the progressive generation during pre-training, and considering efficiency, we no longer separately generate lane dividers, 3D detection, and future frames, but instead integrate all these results into a single unified frame. As shown in the right part of Figure 2, here, VLA serves as a world model to generate a unified image frame predicting the future world state: Inspired by visual prompting engineering [53] that draws red circles on images to guide model attention and by VLIPP [78] first predicts future bounding boxes to introduce physical priors when generating future frames, we represent future world spatial relationships through future red lane dividers and 3D detection boxes on the predicted unified frames. These coarse-grained visual cues direct the model’s attention toward drivable areas and critical objects in future scenes while enforcing physically plausible constraints. Meanwhile, the temporal relationships are represented by the ordinary future frame, where the dynamic evolution of visual content intuitively characterizes temporal progression and the inherent laws of scene development. Subsequently, spatio-temporal CoT Q_{CoT} serves as an intermediate reasoning step, allowing the VLA to function as an inverse dynamics model that plans trajectory based on current observations and future predictions:

$$P(W_t | I_t, Q_{CoT}, opt(T_{com}, T_{ego})) = \Pi_{i=1}^n P_\theta(w_i | w_{<i}, I_t, Q_{CoT}, opt(T_{com}, T_{ego})). \quad (6)$$

3.4 Training strategy

Our FSDrive can be initialized from any existing MLLM (e.g., Qwen2-VL, LLaVA), avoiding training from scratch and saving significant costs. During training, we fully fine-tune the LLM parameters while freezing all encoders. The training process is divided into two stages:

Stage 1: Unified pre-training. Our objective is to preserve understanding capabilities of MLLMs through VQA tasks and activate their visual generation capabilities to predict future frames. VQA task data originates from OmniDrive-nuScenes [64]. We incorporate a large volume of unlabeled image data from nuScenes [1] for future frame prediction. To implement progressive easy-to-hard CoT, we integrate nuScenes annotated data to teach the model predicting image-formatted future lane dividers and 3D detection. Finally, we add future frame prediction with CoT datas containing intermediate reasoning steps. All the above understanding and generation tasks are trained together.

Stage 2: Supervised fine-tuning. We focus on autonomous driving scene understanding and trajectory planning. Following OmniDrive [64], scene understanding utilizes DriveLM’s GVQA [54] dataset. For trajectory planning, we follow VAD [29, 21] using nuScenes, where our spatio-temporal CoT integrates the holistic future scene, explicit lane dividers, and 3D detection results into a single future frame as intermediate reasoning steps. We train these tasks simultaneously using a single model, enabling task-specific predictions during inference through different task prompts.

4 Experiments

4.1 Experimental settings

Datasets. Following the previous methods [29, 13, 4], we evaluate trajectory planning and future frames generation on the nuScenes [1]. The nuScenes contains 1,000 scenes of approximately 20 seconds each captured by a 32-beam LiDAR and six cameras providing 360-degree field of view. Specifically, The dataset provides 28,130 (train), 6,019 (val), and 193,082 (unannotated) samples. Additionally, we conducted experiments on NAVSIM [10], a realistic scenario dataset designed for real-world planning. This dataset aims to highlight challenging driving scenarios involving dynamic changes in driving intent, while deliberately excluding simple situations such as static scenes or constant-speed driving.

Table 1: End-to-end trajectory planning experiments on nuScenes [1]. We evaluated the L2 and collision metrics based on the distinct computational methodologies of ST-P3 [19] and UniAD [21], respectively. * indicates that the ego status is additionally used. VAD [29] and UniAD [21] results are derived from BEV-Planner [34], while the remaining results are sourced from their respective papers.

Method	ST-P3 metrics								UniAD metrics								LLM
	L2 (m) ↓				Collision (%) ↓				L2 (m) ↓				Collision (%) ↓				
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	
Non-Autoregressive methods																	
ST-P3* [ECCV22] [19]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	-	-	-	-	-	-	-	-	-
VAD [ICCV23] [29]	0.69	1.22	1.83	1.25	0.06	0.68	2.52	1.09	-	-	-	-	-	-	-	-	-
VAD* [ICCV23] [29]	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33	-	-	-	-	-	-	-	-	-
UniAD [CVPR23] [21]	-	-	-	-	-	-	-	-	0.59	1.01	1.48	1.03	0.16	0.51	1.64	0.77	-
UniAD* [CVPR23] [21]	-	-	-	-	-	-	-	-	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37	-
BEV-Planner [CVPR24] [34]	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59	-	-	-	-	-	-	-	-	-
BEV-Planner* [CVPR24] [34]	0.16	0.32	0.57	0.35	0.00	0.29	0.73	0.34	-	-	-	-	-	-	-	-	-
PreWorld [ICLR25] [32]	-	-	-	-	-	-	-	-	0.49	1.22	2.32	1.34	0.19	0.57	2.65	1.14	-
Autoregressive methods																	
ELM [ECCV24] [98]	-	-	-	-	-	-	-	-	0.34	1.23	2.57	1.38	0.12	0.50	2.36	0.99	BLIP2-2.7B
FeD* [CVPR24] [86]	-	-	-	-	-	-	-	-	0.27	0.53	0.94	0.58	0.00	0.04	0.52	0.19	LLaVA-7B
OccWorld [ECCV24] [94]	0.39	0.73	1.18	0.77	0.11	0.19	0.67	0.32	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87	GPT3-like
Doe-1 [arxiv24] [95]	0.37	0.67	1.07	0.70	0.02	0.14	0.47	0.21	0.50	1.18	2.11	1.26	0.04	0.37	1.19	0.53	Lumina-mGPT-7B
RDA-Driver* [ECCV24] [26]	0.17	0.37	0.69	0.40	0.01	0.05	0.26	0.10	0.23	0.73	1.54	0.80	0.00	0.13	0.83	0.32	LLaVA-7B
EMMA* [arxiv24] [27]	0.14	0.29	0.54	0.32	-	-	-	-	-	-	-	-	-	-	-	-	Gemini 1-1.8B
OmniDrive [CVPR25] [64]	0.40	0.80	1.32	0.84	0.04	0.46	2.32	0.94	-	-	-	-	-	-	-	-	LLaVA-7B
OmniDrive* [CVPR25] [64]	0.14	0.29	0.55	0.33	0.00	0.13	0.78	0.30	-	-	-	-	-	-	-	-	LLaVA-7B
FSDrive (ours)	0.28	0.52	0.80	0.53	0.06	0.13	0.32	0.17	0.40	0.89	1.60	0.96	0.07	0.12	1.02	0.40	Qwen2-VL-2B
FSDrive* (ours)	0.14	0.25	0.46	0.28	0.03	0.06	0.21	0.10	0.18	0.39	0.77	0.45	0.00	0.06	0.42	0.16	Qwen2-VL-2B
FSDrive (ours)	0.29	0.57	0.94	0.60	0.04	0.14	0.38	0.19	0.36	1.01	1.90	1.09	0.08	0.34	1.11	0.51	LLaVA-7B
FSDrive* (ours)	0.13	0.28	0.52	0.31	0.03	0.07	0.24	0.12	0.22	0.51	0.94	0.56	0.02	0.07	0.53	0.21	LLaVA-7B

Table 2: Performance comparison on NAVSIM navtest using closed-loop metrics. All the methods only use images as input and do not use lidar.

Method	NC ↑	DAC ↑	TTC ↑	Comf. ↑	EP ↑	PDMS ↑
VADv2 [arXiv24] [3]	97.2	89.1	91.6	100	76.0	80.9
UniAD [CVPR23] [21]	97.8	91.9	92.9	100	78.8	83.4
DiffusionDrive-Cam [CVPR25] [36]	97.8	92.2	92.6	99.9	78.9	83.6
LTF [TPAMI23] [6]	97.4	92.8	92.4	100	79.0	83.8
PARA-Drive [CVPR24] [69]	97.9	92.4	93.0	99.8	79.3	84.0
LAW [ICLR25] [33]	96.4	95.4	88.7	99.9	81.7	84.6
FSDrive (ours)	98.2	93.8	93.3	99.9	80.1	85.1

Following the previous methods [7, 64], we evaluate scene understanding on DriveLM [54]. This dataset features keyframe descriptions paired with QA annotations covering full-stack autonomous driving (perception, prediction, planning), offering comprehensive language support for development.

Metrics. We evaluate trajectory planning using L2 displacement error and collision rate following previous methods [21, 29, 19]. Notably, UniAD [21] computes L2 metrics and collision rate at each timestep, whereas ST-P3 [19] and VAD [29] considers the average of all previous time-steps. For a fair comparison, we adopted these two different calculation methods. Following existing methods [65, 77, 71], we report Fréchet Inception Distance (FID) [17] to measure the future frames generation quality. DriveLM GVQA [54] metrics include language metrics like BLEU, ROUGE_L, and CIDEr for text generation, the ChatGPT Score for open-ended Q&A and accuracy for multiple-choice questions. For NAVSIM [10], we adopt the official metrics for evaluation, especially PDMS.

Implementation details. We initialize our model with Qwen2-VL-2B [63] and pre-train it for 32 epochs to enable visual generation while preserving semantic understanding. During fine-tuning (12 epochs on 8 NVIDIA RTX A6000), we use 1×10^{-4} learning rate and batch size of 16. We expand the visual codebook of MoVQGAN [92] to the vocabulary of the large language model and use its tokenizer to convert the visual tokens predicted by the large language model to the pixel space.

Table 3: Future frames generation results on the nuScenes [1] dataset.

Method	DriveGAN [CVPR21 [30]]	DriveDreamer [ECCV24 [65]]	Drive-WM [CVPR24 [66]]	GenAD [CVPR24 [77]]	GEM [CVPR25 [16]]	Doe-1 [arxiv24 [95]]	FSDrive
Type	GAN	Diffusion	Diffusion	Diffusion	Diffusion	Autoregressive	Autoregressive
Resolution	256×256	128×192	192×384	256×448	576×1024	384×672	128×192
FID ↓	73.4	52.6	15.8	15.4	10.5	15.9	10.1

Table 4: Results on DriveLM [54] GVQA benchmark.

Method	Accuracy ↑	ChatGPT ↓	BLEU_1 ↑	ROUGE_L ↑	CIDEr ↑	Match ↑	Final Score ↑
DriveLM baseline [54]	0.00	0.65	0.05	0.08	0.10	0.28	0.32
Cube-LLM [7]	0.39	0.89	0.16	0.20	0.31	0.39	0.50
TrackingMeetsLMM [28]	0.60	0.58	0.72	0.72	0.04	0.36	0.52
SimpleLLM4AD [93]	0.66	0.57	0.76	0.73	0.15	0.35	0.53
OmniDrive [64]	0.70	0.65	0.52	0.73	0.13	0.37	0.56
FSDrive (ours)	0.72	0.63	0.76	0.74	0.17	0.39	0.57

Table 5: Ablation results of pre-training.

VQA	Future frames	Future 3D detection	Future lane divider	1s	2s	3s	L2 (m) ↓ Avg.	1s	2s	3s	Collision (%) ↓ Avg.
×	×	×	×	0.45	1.09	2.12	1.22	0.12	0.43	1.45	0.67
✓	×	×	×	0.46	1.07	2.04	1.19	0.12	0.42	1.42	0.65
×	✓	×	×	0.39	0.96	1.71	1.02	0.10	0.38	1.32	0.60
×	×	✓	×	0.46	1.06	1.99	1.17	0.10	0.37	1.35	0.61
×	×	×	✓	0.42	0.97	1.80	1.06	0.13	0.41	1.40	0.65
✓	✓	✓	✓	0.39	0.91	1.63	0.98	0.09	0.36	1.33	0.58

4.2 Main results

End-to-End trajectory planning. We present trajectory planning performance on nuScenes following previous methods [29, 21] in Table 1. When using ego status, FSDrive surpasses previous SOTA methods using ego status in ST-P3 and UniAD metrics. However, following BEV-Planner [34] findings about ego-status’s performance boost, we prioritize non-ego-status evaluations. Compared to non-autoregressive (e.g., UniAD) and autoregressive methods (e.g., OmniDrive), FSDrive demonstrates superior effectiveness. Notably, FSDrive outperforms Doe-1 [95] which also enables vision generation (L2: 0.53 vs. 0.70 and 0.96 vs. 1.26; collision: 0.19 vs. 0.21 and 0.40 vs. 0.53), indicating limitations in their VQ-VAE-based discrete visual features for understanding. For a fair comparison, we also used LLaVA like methods [64, 26, 86, 75]. Under the corresponding settings, FSDrive still has excellent competitiveness, indicating that FSDrive can be widely applied to any existing MLLM.

Results on NAVSIM. Table 2 shows the evaluation results for NAVSIM [10]. All approaches rely exclusively on camera input, with no lidar data being used. Achieving a PDMS score of 85.1, FSDrive outperforms prior camera-only methods like LAW [33] and DiffusionDrive-Cam [36], thus showcasing its efficacy in the pseudo closed-loop setting.

Evaluation of generation results. Although we generate future frames as CoT for trajectory planning, we still validate visual quality via FID in Table 3. To enable rapid generation for real-time driving, we generate frames at 128×192 resolution. Our autoregressive FSDrive achieves competitive performance against specialized diffusion models. Compared to Doe-1 [95] which employs the vision generation MLLM Lumina-mGPT 7B [37], FSDrive 2B maintains superior advantages, indicating that the visual generation capabilities of MLLM can be effectively unlocked even with minimal data.

Results on DriveLM dataset. FSDrive’s scene understanding was evaluated on DriveLM in Table 4, achieving 0.57 and outperforming recent methods like Cube-LLM [7] and OmniDrive [64]. This highlights the effectiveness of FSDrive pre-training paradigm for generation and understanding.



Figure 3: Qualitative analysis of our CoT. The red trajectory is the prediction and the green is the GT.

Table 6: Ablation results of different CoT.

Type	L2 (m) ↓				Collision (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
None	0.39	0.91	1.63	0.98	0.09	0.36	1.33	0.58
Text CoT	0.39	0.92	1.61	0.97	0.10	0.29	1.21	0.53
Image-text CoT	0.38	0.90	1.65	0.98	0.09	0.25	1.15	0.50
Spatio-temporal CoT	0.40	0.89	1.60	0.96	0.07	0.12	1.02	0.40

4.3 Ablation study

In this section, unless otherwise specified, we evaluate the computing metrics of UniAD [21] based on the Qwen2-VL-2B model [63] and do not use the ego status.

Qualitative analysis. We evaluate our CoT’s effectiveness in Figure 3. Without spatial-temporal CoT, erroneous navigation inputs caused significant trajectory deviations and potential collisions. Use correct instruction when reasoning our CoT, while still employing wrong instruction for planning. However, FSDrive mitigated instruction errors through observation-based trajectory planning and future prediction, demonstrating its inverse dynamics modeling capability.

Table 7: Ablation experiments of future frames generation.

(a) Ablations of pre-training data volume.

Data volume	FID↓
None	29.4
~100k	16.2
~200k	12.7

(b) Ablations of progressive method.

Progressive Method	FID↓
×	12.7
✓	10.1

Pre-training ablation study. The impact of pre-training on trajectory planning is summarized in Table 5. Pure VQA tasks show negligible effects. Future frame generation pre-training improves L2 by 16.4% and collisions by 15.8%, validating world-model-based prediction’s effectiveness in capturing physical dynamics. 3D detection and lane divider pre-training yield moderate gains in L2/collision metrics respectively. The combined understanding and generation pre-training achieves better performance, demonstrating our unified paradigm’s capacity to enhance scene representation and effectively learn physical laws, thereby strengthening spatial understanding capabilities.

Results of different CoT. Ablation studies on CoT variants in Table 6 show marginal L2 changes but notable collision rate improvements. Pure text CoT (8.6% improvement) exhibits limited representation capability due to unimodal perception. Compared to text CoT, image-text CoT (combining future frames with textual perception) shows insignificant gains due to the inconsistent modalities

between CoTs. The proposed spatio-temporal CoT achieves 31% improvement, demonstrating that unified image-based reasoning effectively identifies future collision risks.

Ablation study on generation results. We conduct ablation studies on future frames generation in Table 7. Table 7a shows that larger pre-training datasets improve MLLM’s visual generation capability. Despite being much smaller (200K vs. 100M in previous work [70]), our data achieves more robust visual generation. Scaling datasets may further enhance performance. Table 7b confirms our progressive method improves autoregressive image generation.

5 Conclusion

This paper proposes FSDrive, an autonomous driving framework based on spatio-temporal CoT that enables VLAs to think visually. By unifying future scene generation and perception results through intermediate image-form reasoning steps, our FSDrive eliminates the semantic gap caused by cross-modal conversions and establishes an end-to-end visual reasoning pipeline. The VLA serves dual roles: as a world model that predicts future image frames with lane divider and 3D detection, and as an inverse dynamics model that plans trajectory based on both current observations and future predictions. To enable visual generation in VLAs, we present a pretraining paradigm that unifies visual generation and understanding, along with a progressive easy-to-hard visual CoT to enhance autoregressive image generation. Extensive experimental results demonstrate the effectiveness of the proposed FSDrive method, advancing autonomous driving towards visual reasoning.

Limitations and broader impacts. Though autonomous driving requires surrounding environmental awareness, considering real-time efficiency, we currently only generate future frames for the front-view. Future work can attempt to generate Surround views to achieve safer autonomous driving. Moreover, more robust visual quality can be achieved in future work through the use of larger training datasets and a more advanced unified paradigm that integrates generation and understanding. In terms of impact, the ethical challenges posed by LLMs extend to autonomous driving. Advances in technology and regulation will drive development of safer, more efficient systems.

Acknowledgments. This work was support by the National Natural Science Foundation of China No. 62572385, the Fundamental Research Funds for the Central Universities No. xxj032023020, and CAAI-CANN Open Fund, developed on OpenI Community.

References

- [1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. *CVPR*, 2020.
- [2] X. Chang, M. Xue, X. Liu, Z. Pan, and X. Wei. Driving by the rules: A benchmark for integrating traffic sign regulations into vectorized hd map. *CVPR*, 2025.
- [3] S. Chen, B. Jiang, H. Gao, B. Liao, Q. Xu, Q. Zhang, C. Huang, W. Liu, and X. Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024.
- [4] Y. Chen and R. Greer. Technical report for argoverse2 scenario mining challenges on iterative error correction and spatially-aware prompting. *arXiv preprint arXiv:2506.11124*, 2025.
- [5] Y. Chen, Y.-Q. Wang, and Z. Zhang. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. *ICCV*, 2025.
- [6] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *TPAMI*, 2023.
- [7] J. H. Cho, B. Ivanovic, Y. Cao, E. Schmerling, Y. Wang, X. Weng, B. Li, Y. You, P. Krähenbühl, Y. Wang, et al. Language-image models with 3d understanding. *ICLR*, 2025.
- [8] M. Dai, S. Liu, Z. Zhao, J. Gao, H. Sun, and X. Li. Secure tug-of-war (sectow): Iterative defense-attack training with reinforcement learning for multimodal model security. *arXiv preprint arXiv:2507.22037*, 2025.
- [9] M. Dai, J. Sun, Z. Zhao, S. Liu, R. Li, J. Gao, and X. Li. From captions to rewards (carevl): Leveraging large language model experts for enhanced reward modeling in large vision-language models. *arXiv preprint arXiv:2503.06260*, 2025.
- [10] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, A. Geiger, and K. Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *NeurIPS*, 2024.

- [11] R. Dong, C. Han, Y. Peng, Z. Qi, Z. Ge, J. Yang, L. Zhao, J. Sun, H. Zhou, H. Wei, X. Kong, X. Zhang, K. Ma, and L. Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Mindecker, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
- [13] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *NeurIPS*, 2024.
- [14] Y. Gao, C. Li, Z. You, J. Liu, Z. Li, P. Chen, Q. Chen, Z. Tang, L. Wang, P. Yang, et al. Openfly: A comprehensive platform for aerial vision-language navigation. *CoRR*, 2025.
- [15] J. Guo, Z. Li, J. Wu, Q. Wang, Y. Li, L. Zhang, hai zhao, and Y. Yang. Tom: Leveraging tree-oriented mapreduce for long-context reasoning in large language models. In *EMNLP*, 2025.
- [16] M. Hassan, S. Stapf, A. Rahimi, P. M. B. Rezende, Y. Haghighi, D. Brüggemann, I. Katircioglu, L. Zhang, X. Chen, S. Saha, M. Cannici, E. Aljalbout, B. Ye, X. Wang, A. Davtyan, M. Salzmann, D. Scaramuzza, M. Pollefeys, P. Favaro, and A. Alahi. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. *CVPR*, 2025.
- [17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.
- [18] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.
- [19] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, 2022.
- [20] Y. Hu, Q. Li, D. Zhang, J. Yan, and Y. Chen. Context-alignment: Activating and enhancing LLMs capabilities in time series. In *ICLR*, 2025.
- [21] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li. Planning-oriented autonomous driving. In *CVPR*, 2023.
- [22] J. Huang, M. Yan, S. Chen, Y. Huang, and S. Chen. Magicfight: Personalized martial arts combat video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.
- [23] J. Huang, G. Zhang, Z. Jie, S. Jiao, Y. Qian, L. Chen, Y. Wei, and L. Ma. M4v: Multi-modal mamba for text-to-video generation. *arXiv preprint arXiv:2506.10915*, 2025.
- [24] Z. Huang, H. Qian, Z. Cai, A. Wang, J. Wang, and F. Xiong. Intelligent recognition method for urban road grid patterns by fusing mesh and road features. *International Journal of Digital Earth*, 2024.
- [25] Z. Huang, H. Qian, Z. Cai, X. Wang, L. Xie, and X. Niu. An intelligent multilane roadway recognition method based on pseudo-tagging. *Cartography and Geographic Information Science*, 2025.
- [26] Z. Huang, T. Tang, S. Chen, S. Lin, and Z. e. a. Jie. Making large language models better planners with reasoning-decision alignment. *ECCV*, 2024.
- [27] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, J. Guo, D. Anguelov, and M. Tan. Emma: End-to-end multimodal model for autonomous driving. *TMLR*, 2025.
- [28] A. Ishaq, J. Lahoud, F. S. Khan, S. Khan, H. Cholakkal, and R. M. Anwer. Tracking meets large multimodal models for driving scenario understanding. *ArXiv preprint arXiv:2503.14498*, 2025.
- [29] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang. Vad: Vectorized scene representation for efficient autonomous driving. *ICCV*, 2023.
- [30] S. W. Kim, J. Phillion, A. Torralba, and S. Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, 2021.
- [31] B. Li, Y. Wang, J. Mao, B. Ivanovic, S. Veer, K. Leung, and M. Pavone. Driving everywhere with large language model policy adaptation. In *CVPR*, 2024.
- [32] X. Li, P. Li, Y. Zheng, W. Sun, Y. Wang, and Y. Chen. Semi-supervised vision-centric 3d occupancy world model for autonomous driving. *ICLR*, 2025.
- [33] Y. Li, L. Fan, J. He, Y. Wang, Y. Chen, Z. Zhang, and T. Tan. Enhancing end-to-end autonomous driving with latent world model. *ICLR*, 2025.
- [34] Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. M. Álvarez. Is ego status all you need for open-loop end-to-end autonomous driving? *CVPR*, 2024.
- [35] S. Liang, X. Chang, C. Wu, H. Yan, Y. Bai, X. Liu, H. Zhang, Y. Yuan, S. Zeng, M. Xu, et al. Persistent autoregressive mapping with traffic rules for autonomous driving. *arXiv preprint arXiv:2509.22756*, 2025.
- [36] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, and X. Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *CVPR*, 2025.
- [37] D. Liu, S. Zhao, L. Zhuo, W. Lin, Y. Qiao, H. Li, and P. Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.
- [38] H. Liu, W. Yan, M. Zaharia, and P. Abbeel. World model on million-length video and language with ringattention. *ICLR*, 2025.
- [39] J. Liu, F. Shang, Y. Liu, H. Liu, Y. Li, and Y. Gong. Fedbcgd: Communication-efficient accelerated block coordinate gradient descent for federated learning. In *ACM MM*, 2024.

- [40] W. Liu, J. Chen, K. Ji, L. Zhou, W. Chen, and B. Wang. Rag-instruct: Boosting llms with diverse retrieval-augmented instructions. *EMNLP*, 2025.
- [41] W. Liu, J. Xu, F. Yu, Y. Lin, K. Ji, W. Chen, Y. Xu, Y. Wang, L. Shang, and B. Wang. Qfft, question-free fine-tuning for adaptive reasoning. *NeurIPS*, 2025.
- [42] W. Lu, Y. Tong, and Z. Ye. Dammfnd: Domain-aware multimodal multi-view fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- [43] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao. Dolphins: Multimodal language model for driving. *ECCV*, 2024.
- [44] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang. A language agent for autonomous driving. *COLM*, 2024.
- [45] C. Min, D. Zhao, L. Xiao, J. Zhao, X. Xu, Z. Zhu, L. Jin, J. Li, Y. Guo, J. Xing, L. Jing, Y. Nie, and B. Dai. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. *CVPR*, 2024.
- [46] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.
- [47] J. Ni, Y. Guo, Y. Liu, R. Chen, L. Lu, and Z. Wu. Maskgwm: A generalizable driving world model with video mask reconstruction. *CVPR*, 2025.
- [48] Y. Qian, X. Li, J. Zhang, X. Meng, Y. Li, H. Ding, and M. Wang. A diffusion-tgan framework for spatio-temporal speed imputation and trajectory reconstruction. *IEEE T-ITS*, 2025.
- [49] K. Qiu, Z. Gao, Z. Zhou, M. Sun, and Y. Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *CVPR*, 2025.
- [50] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *RSS*, 2025.
- [51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [52] A. Sarkar, M. Y. I. Idris, and Z. Yu. Reasoning in computer vision: Taxonomy, models, tasks, and methodologies. *arXiv preprint arXiv:2508.10523*, 2025.
- [53] A. Shtedritski, C. Rupprecht, and A. Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *ICCV*, 2023.
- [54] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li. Drivelm: Driving with graph visual question answering. *ECCV*, 2024.
- [55] H. Song, D. Qu, Y. Yao, Q. Chen, Q. Lv, Y. Tang, M. Shi, G. Ren, M. Yao, B. Zhao, et al. Hume: Introducing system-2 thinking in visual-language-action model. *arXiv preprint arXiv:2505.21432*, 2025.
- [56] P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.
- [57] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang. Generative pretraining in multimodality. *ICLR*, 2024.
- [58] S. Sun, W. Yu, Y. Ren, W. Du, L. Liu, X. Zhang, Y. Hu, and C. Ma. Gdiffretro: Retrosynthesis prediction with dual graph enhanced molecular representation and diffusion generation. *AAAI*, 2025.
- [59] W. Tan, D. Chen, J. Xue, Z. Wang, and T. Chen. Teaching-inspired integrated prompting framework: A novel approach for enhancing reasoning in large language models. In *COLING: Industry Track*, 2025.
- [60] X. Tian, J. Gu, B. Li, Y. Liu, Z. Zhao, Y. Wang, K. Zhan, P. Jia, X. Lang, and H. Zhao. Drivelm: The convergence of autonomous driving and large vision-language models. *CoRL*, 2024.
- [61] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.
- [62] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
- [63] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [64] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Álvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *CVPR*, 2025.
- [65] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *ECCV*, 2024.
- [66] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *CVPR*, 2024.
- [67] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.
- [68] Q. Wei, P. Dai, W. Li, B. Liu, and X. Wu. Copeft: Fast adaptation framework for multi-agent collaborative perception with parameter-efficient fine-tuning. In *AAAI*, 2025.
- [69] X. Weng, B. Ivanovic, Y. Wang, Y. Wang, and M. Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *CVPR*, 2024.

- [70] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *CVPR*, 2024.
- [71] C. Wu, H. Huang, L. Zhang, J. Chen, Y. Tong, and M. Zhou. Towards automated 3d evaluation of water leakage on a tunnel face via improved gan and self-attention dl model. *Tunn Undergr Space Technol*, 2023.
- [72] J. Wu, H. Li, X. Zhang, X. Liu, Y. Huang, J. Luo, Y. Zhang, Z. Li, R. Chu, Y. Yang, and S. Li. Teaching your models to understand code via focal preference alignment. In *EMNLP*, 2025.
- [73] Y. Wu, Z. Zhang, J. Chen, H. Tang, D. Li, Y. Fang, L. Zhu, E. Xie, H. Yin, L. Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *ICLR*, 2025.
- [74] J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z. Shou. Show-o: One single transformer to unify multimodal understanding and generation. *ICLR*, 2025.
- [75] M. Xie, S. Zeng, X. Chang, X. Liu, Z. Pan, M. Xu, and X. Wei. Seggrowgraph: Learning lane topology as a chain of graph expansions. In *ICCV*, 2025.
- [76] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.
- [77] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo, J. Zhang, A. Geiger, Y. Qiao, and H. Li. Generalized predictive model for autonomous driving. In *CVPR*, 2024.
- [78] X. Yang, B. Li, Y. Zhang, Z. Yin, L. Bai, L. Ma, Z. Wang, J. Cai, T.-T. Wong, H. Lu, and X. Jia. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. *ICCV*, 2025.
- [79] Z. Yang, L. Chen, Y. Sun, and H. Li. Visual point cloud forecasting enables scalable autonomous driving. In *CVPR*, 2024.
- [80] Z. Yu, M. Y. I. Idris, H. Wang, P. Wang, J. Chen, and K. Wang. From physics to foundation models: A review of ai-driven quantitative remote sensing inversion. *arXiv preprint arXiv:2507.09081*, 2025.
- [81] Z. Yu, M. Y. I. Idris, P. Wang, Y. Xia, and Y. Xiang. Forgetme: Benchmarking the selective forgetting capabilities of generative models. *Engineering Applications of Artificial Intelligence*, 2025.
- [82] Y. Yuan, C. Wu, X. Chang, S. Wang, H. Zhang, S. Liang, S. Zeng, and M. Xu. Unimapgen: A generative framework for large-scale map construction from multi-modal data. *arXiv preprint arXiv:2509.22262*, 2025.
- [83] B. Yue, S. Guo, K. Hu, C. Wang, B. Wang, K. Jia, and G. Liu. Real-time verification of embodied reasoning for generative skill acquisition. *arXiv preprint arXiv:2505.11175*, 2025.
- [84] S. Zeng, X. Chang, X. Liu, Z. Pan, and X. Wei. Driving with prior maps: Unified vector prior encoding for autonomous vehicle mapping. *arXiv preprint arXiv:2409.05352*, 2024.
- [85] S. Zeng, D. Qi, X. Chang, F. Xiong, S. Xie, X. Wu, S. Liang, M. Xu, and X. Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*, 2025.
- [86] J. Zhang, Z. Huang, A. Ray, and E. Ohn-Bar. Feedback-guided autonomous driving. In *CVPR*, 2024.
- [87] J. Zhang, C. Xu, and B. Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *CVPR*, 2024.
- [88] L. Zhang, B. Wang, X. Qiu, S. Reddy, and A. Agrawal. Rearank: Reasoning re-ranking agent via reinforcement learning. *EMNLP*, 2025.
- [89] Y. Zhang, X. Liu, R. Tao, Q. Chen, H. Fei, W. Che, and L. Qin. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. *ACM MM*, 2025.
- [90] Y. Zhang, X. Liu, R. Zhou, Q. Chen, H. Fei, W. Lu, and L. Qin. Cchallenge: A novel benchmark for joint cross-lingual and cross-modal hallucinations detection in large language models. *ACL*, 2025.
- [91] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, M.-Y. Liu, D. Xiang, G. Wetzstein, and T.-Y. Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, 2025.
- [92] C. Zheng, L. T. Vuong, J. Cai, and D. Q. Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *NeurIPS*, 2022.
- [93] P. Zheng, Y. Zhao, Z. Gong, H. Zhu, and S. Wu. Simplellm4ad: An end-to-end vision-language model with graph visual question answering for autonomous driving. *ArXiv preprint arXiv:2407.21293*, 2024.
- [94] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *ECCV*, 2024.
- [95] W. Zheng, Z. Xia, Y. Huang, S. Zuo, J. Zhou, and J. Lu. Doe-1: Closed-loop autonomous driving with large world model. *arXiv preprint arXiv: 2412.09627*, 2024.
- [96] P. Zhou, W. Min, C. Fu, Y. Jin, M. Huang, X. Li, S. Mei, and S. Jiang. Foodsky: A food-oriented large language model that can pass the chef and dietetic examinations. *Patterns*, 2025.
- [97] P. Zhou, X. Peng, J. Song, C. Li, Z. Xu, Y. Yang, Z. Guo, H. Zhang, Y. Lin, Y. He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *CVPR*, 2025.
- [98] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li. Embodied understanding of driving scenarios. *ECCV*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims in the abstract and introduction accurately reflect the contribution and scope of the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of the work in Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper focuses on experimental verification and does not involve strict mathematical proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclose information that reproduces the main experimental results of the paper in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code upon completion of the approval process.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We illustrate all the training and testing details in the Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report information about the statistical significance of the experiments in the Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide information on the computer resources needed to reproduce the experiments in the Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper complies in all respects with the NeurIPS ethical guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the impacts of the work in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators of the data and models used in the paper were appropriately credited and in compliance with the license and terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our new assets (code and model) will be made publicly available upon completion of the approval process.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only use LLM to polish the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.