NEURAL INFORMATION PROCESSING SYSTEMS

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

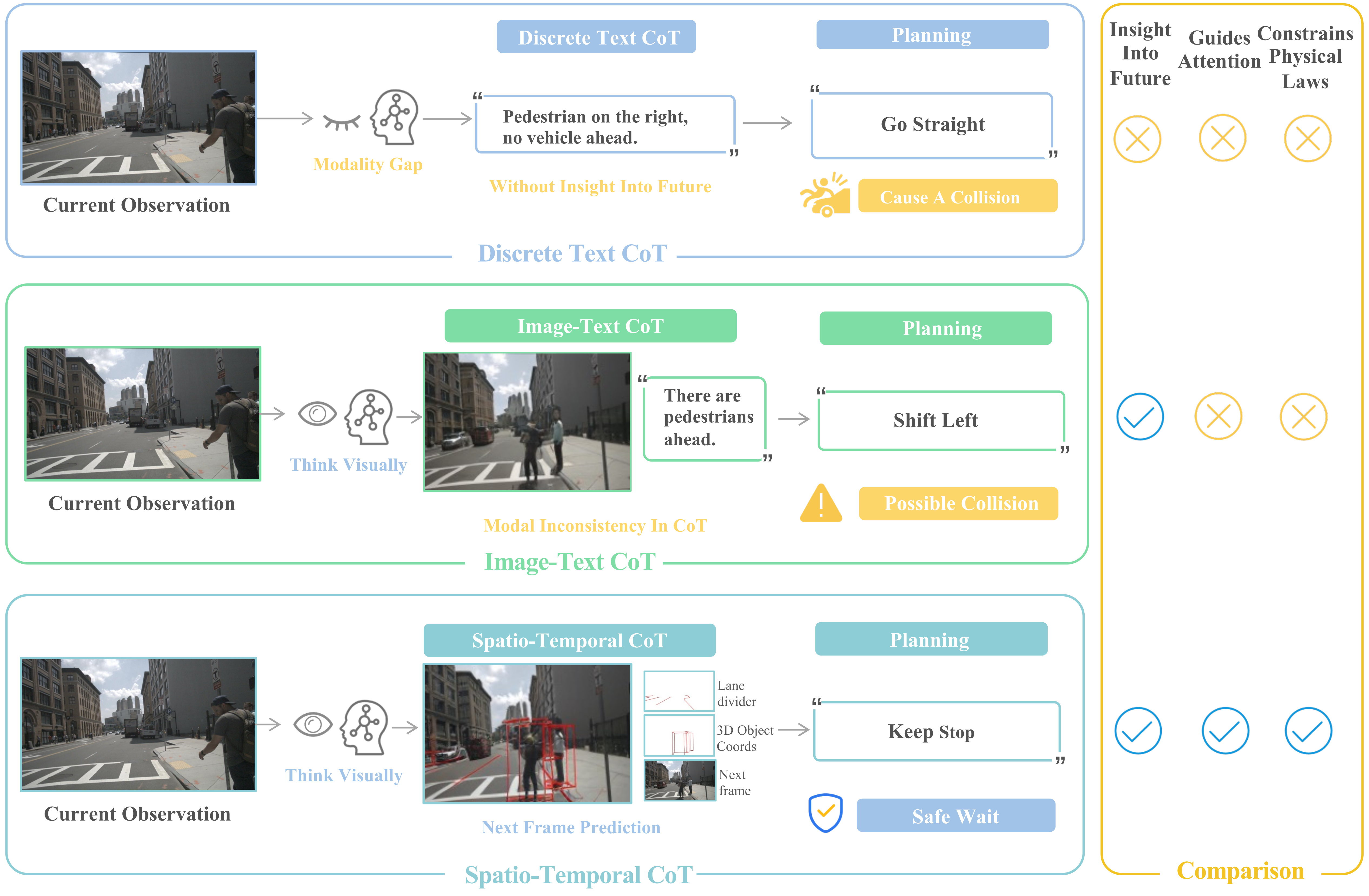
Shuang Zeng[1,2], Xinyuan Chang[2], Mengwei Xie[2], Xinran Liu[2],
Yifan Bai[1,3], Zheng Pan[2], Mu Xu[2], Xing Wei[1]

西安交通大學 XI'AN JIAOTONG UNIVERSITY

高德地图

達摩院 DAMO ACADEMY

Background



Problem & Background

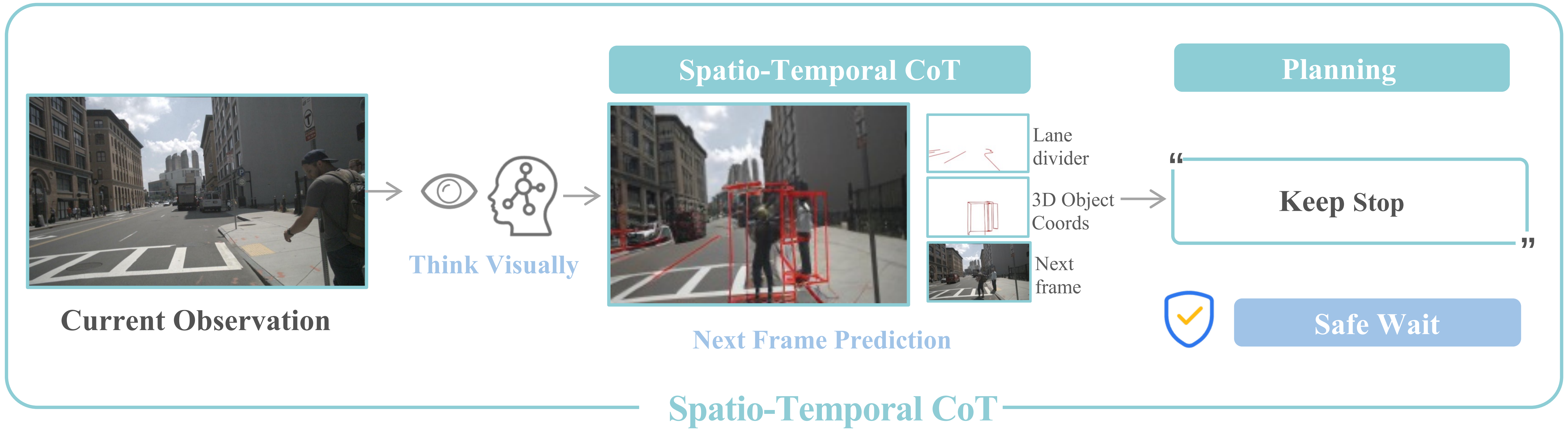
In autonomous driving, when Vision-Language-Action Models (VLAs) are used for reasoning and planning, most methods rely on a textual Chain-of-Thought (CoT). This approach has key limitations:

- High-level abstraction leads to a loss of spatial detail.
- Modality shifts (vision to text) can introduce semantic gaps.
- Difficulty representing spatiotemporal relationships, such as the motion of dynamic objects.

This leads to the core question posed by FSDrive: **Can autonomous vehicles visually imagine the future, instead of relying solely on linguistic logic?**

FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving，https://arxiv.org/abs/2505.17685
Github地址：https://github.com/MIV-XJTU/FSDrive

Background



Motivation

Core Idea: We propose the "Spatio-temporal Chain of Thought":

1）Represent future states as images, capturing both space and time.

2）A VLM acts as a "world model" to predict these future images.

3）It then serves as an "inverse dynamics model" to plan the trajectory.

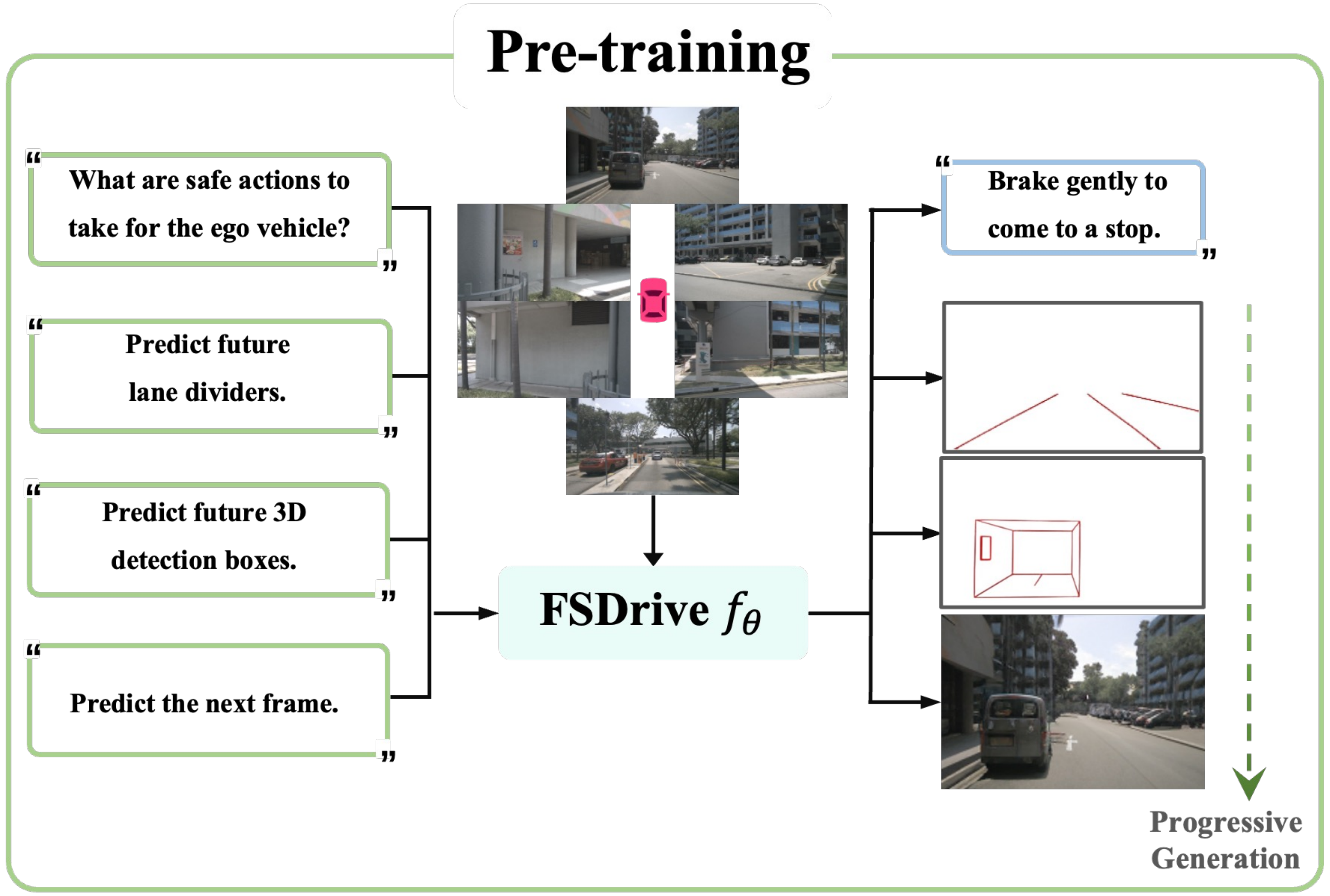
In a nutshell: **Enable the model to "see the future" before deciding how to act in the present.**

Contribution

- Proposed a Spatio-temporal Chain-of-Thought (CoT) method for visual reasoning, enabling the model to visually "think" across future time and space to enhance trajectory planning.

- Proposed a unified pre-training paradigm for visual generation and understanding, featuring a progressive generation method that starts by imposing physical constraints and then gradually adds details.

Method



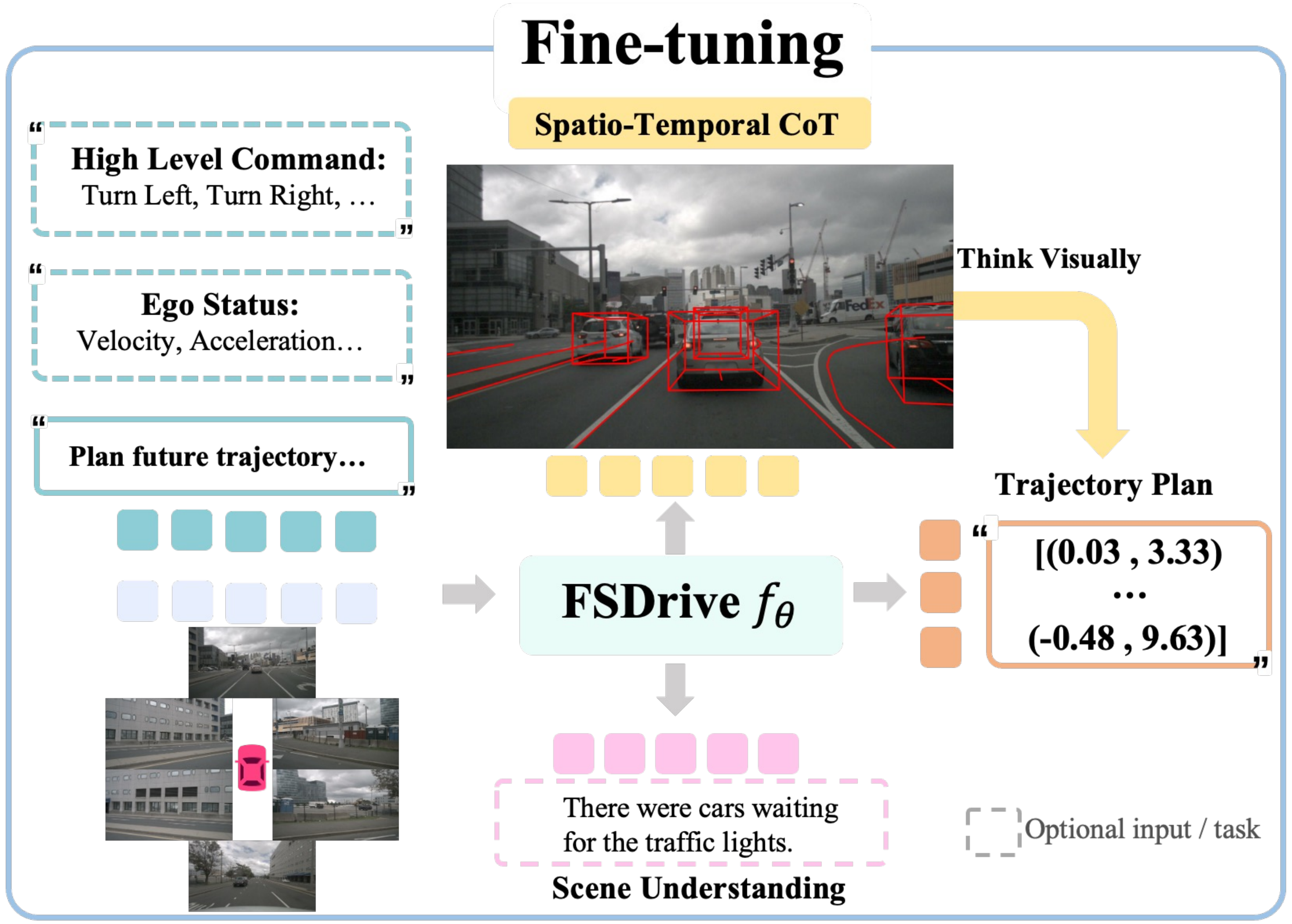
Stage 1: Unified Pre-training for Understanding and Generation

- Visual Understanding: Preserve the MLLM's existing semantic understanding capabilities through VQA tasks.

- Visual Generation: Activate the MLLM's image generation capabilities by leveraging the shared vocabulary space between images and text.

- Progressive Generation: Employ a coarse-to-fine approach: first, generate coarse-grained perceptual maps (e.g., lane lines, 3D detections) to enforce physical constraints; then, render complete future frames to fill in fine-grained details.

Method



Stage 2: Spatiotemporal CoT: Thinking Visually

VLM as a World Model:

1)Generates unified image frames to predict future world states.

2)Future Spatial Relationships: Represented by predicted lane lines and 3D bounding boxes, guiding the model to focus on drivable areas and key objects.

3)Temporal Evolution: Depicted through standard future frames, intuitively illustrating the dynamic changes in the visual scene.

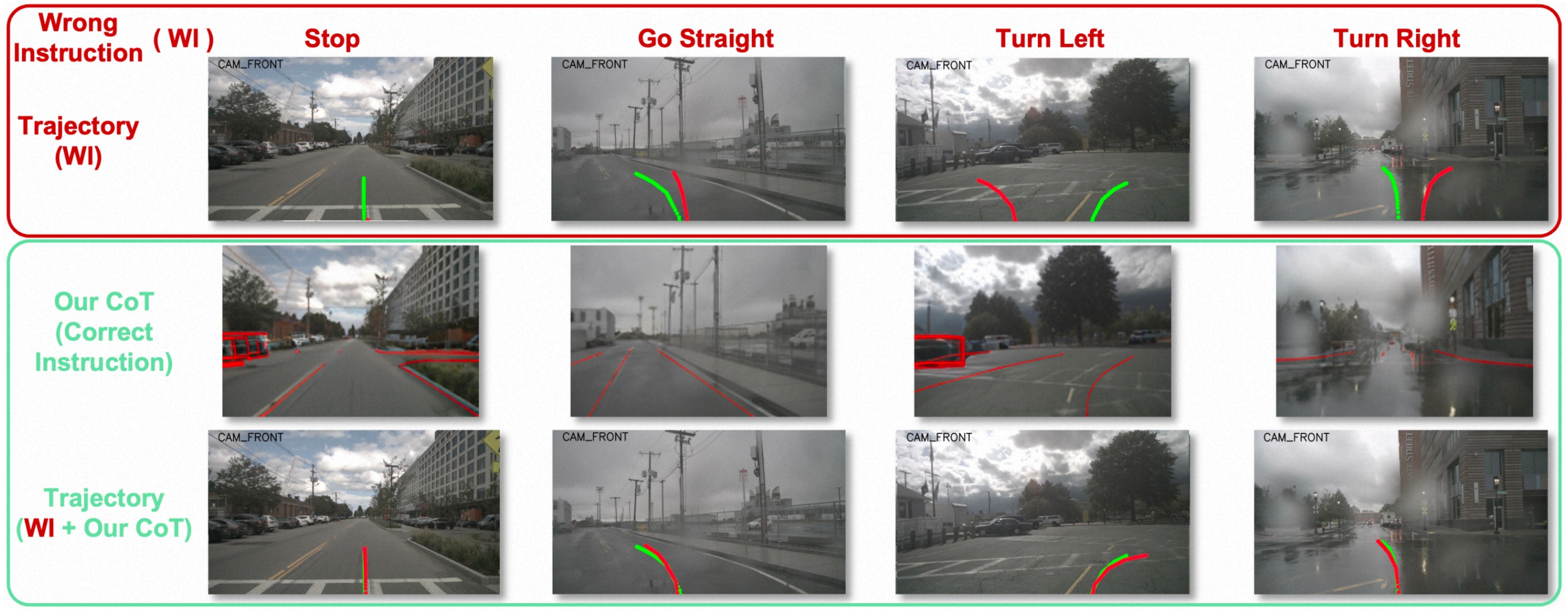
VLM as an Inverse Dynamics Model:

1)Plans trajectories based on current observations and future predictions.

2)The unified image format effectively conveys spatiotemporal relationships, enabling end-to-end visual reasoning.

Experiments

Method	ST-P3 metrics L2 (m) ↓ 1s	2s	3s	Avg.	Collision (%) ↓ 1s	2s	3s	Avg.	UniAD metrics L2 (m) ↓ 1s	2s	3s	Avg.	Collision (%) ↓ 1s	2s	3s	Avg.	LLM
Non-Autoregressive methods																	
ST-P3* [ECCV22] [14]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	-	-	-	-	-	-	-	-	-
VAD [ICCV23] [20]	0.69	1.22	1.83	1.25	0.06	0.68	2.52	1.09	-	-	-	-	-	-	-	-	-
VAD* [ICCV23] [20]	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33	-	-	-	-	-	-	-	-	-
UniAD [CVPR23] [16]	-	-	-	-	-	-	-	-	0.59	1.01	1.48	1.03	0.16	0.51	1.64	0.77	-
UniAD* [CVPR23] [16]	-	-	-	-	-	-	-	-	0.20	0.42	**0.75**	0.46	0.02	0.25	0.84	0.37	-
BEV-Planner [CVPR24] [25]	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59	-	-	-	-	-	-	-	-	-
BEV-Planner* [CVPR24] [25]	0.16	0.32	0.57	0.35	**0.00**	0.29	0.73	0.34	-	-	-	-	-	-	-	-	-
PreWorld [ICLR25] [24]	-	-	-	-	-	-	-	-	0.49	1.22	2.32	1.34	0.19	0.57	2.65	1.14	-
Autoregressive methods																	
ELM [ECCV24] [73]	-	-	-	-	-	-	-	-	0.34	1.23	2.57	1.38	0.12	0.50	2.36	0.99	BLIP2-2.7B
FeD* [CVPR24] [65]	-	-	-	-	-	-	-	-	0.27	0.53	0.94	0.58	**0.00**	**0.04**	0.52	0.19	LLaVA-7B
OccWorld [ECCV24] [71]	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] [72]	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] [17]	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] [18]	**0.14**	0.29	0.54	0.32	-	-	-	-	-	-	-	-	-	-	-	-	Gemini 1-1.8B
OminiDrive [CVPR25] [50]	0.40	0.80	1.32	0.84	0.04	0.46	2.32	0.94	-	-	-	-	-	-	-	-	LLaVA-7B
OminiDrive* [CVPR25] [50]	**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



Quantitative Evaluation

On the ST-P3 and UniAD benchmarks, FSDrive, both with and without ego status, outperforms current state-of-the-art (SOTA) methods.

Qualitative Analysis

Even with faulty navigation instructions, FSDrive can correct its plan and avoid collisions by visually predicting the future.

FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving, https://arxiv.org/abs/2505.17685

Github地址： https://github.com/MIV-XJTU/FSDrive

Experiments

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**

Performance comparison on NAVSIM using only images as input

Method	DriveGAN [21] [CVPR21]	DriveDreamer [51] [ECCV24]	Drive-WM [52] [CVPR24]	GenAD [60] [CVPR24]	GEM [10] [CVPR25]	Doe-1 [72] [arxiv24]	**FSDrive (ours)**
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**

Future frames generation results

FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving，https://arxiv.org/abs/2505.17685

Github地址：https://github.com/MIV-XJTU/FSDrive

Demo Video

Demo Video

Predicted Future

Current Observations

Spatio-Temporal CoT

Trajectory Planning

FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving，https://arxiv.org/abs/2505.17685

Github地址：https://github.com/MIV-XJTU/FSDrive

Conclusion

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

Shuang Zeng[1,2*], Xinyuan Chang[2], Mengwei Xie[2], Xinran Liu[2],
Yifan Bai[1,3], Zheng Pan[2], Mu Xu[2], Xing Wei[1†]
[1]Xi'an Jiaotong University [2]Amap, Alibaba Group [3]DAMO 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

⭐Code Available

Conclusion and Future Work

Conclusion: FSDrive introduces the "Visual Chain-of-Thought" (Visual CoT) concept, which uniformly represents intermediate reasoning steps with images. It also proposes a unified pre-training method that unlocks the image generation capabilities of Vision-Language Action Models (VLAs), achieving state-of-the-art (SOTA) results in trajectory planning, image generation, and scene understanding.

Future Work: We plan to extend this work to surround-view future frame generation for more comprehensive environmental perception. We will also explore stronger physical constraints and causal reasoning by leveraging larger-scale training datasets, incorporating closed-loop control, and utilizing more advanced unified architectures for generation and understanding.