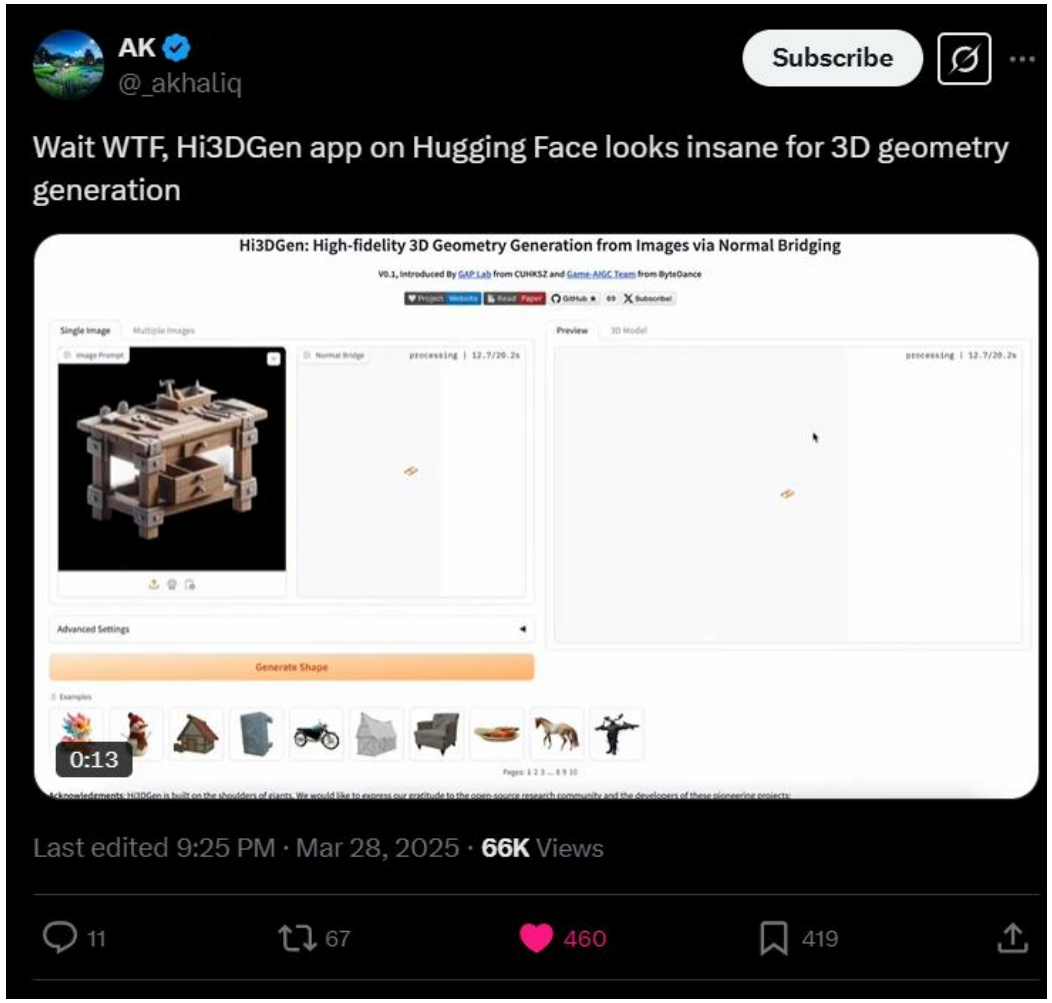
Embodied AI Powered by Physics-aware Generative Simulation

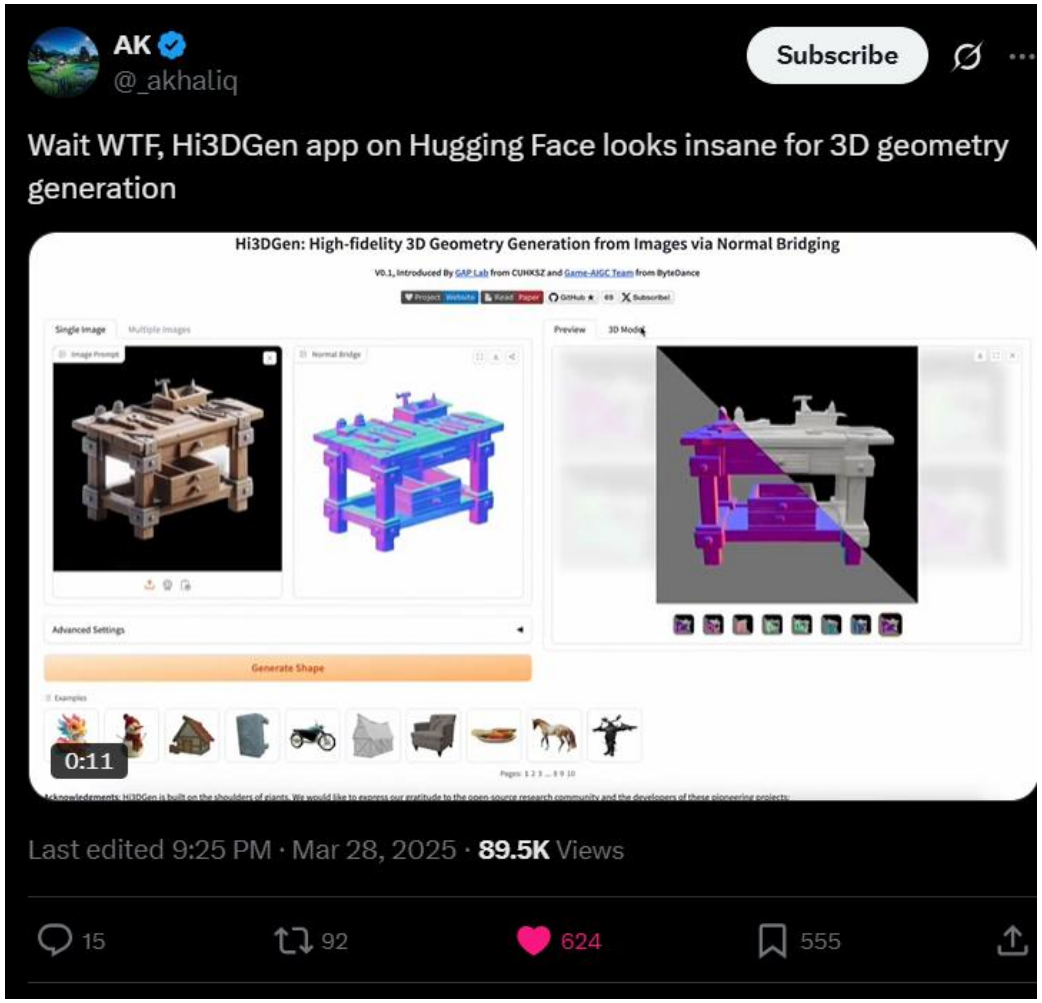
Hao Zhao

https://sites.google.com/view/fromandto

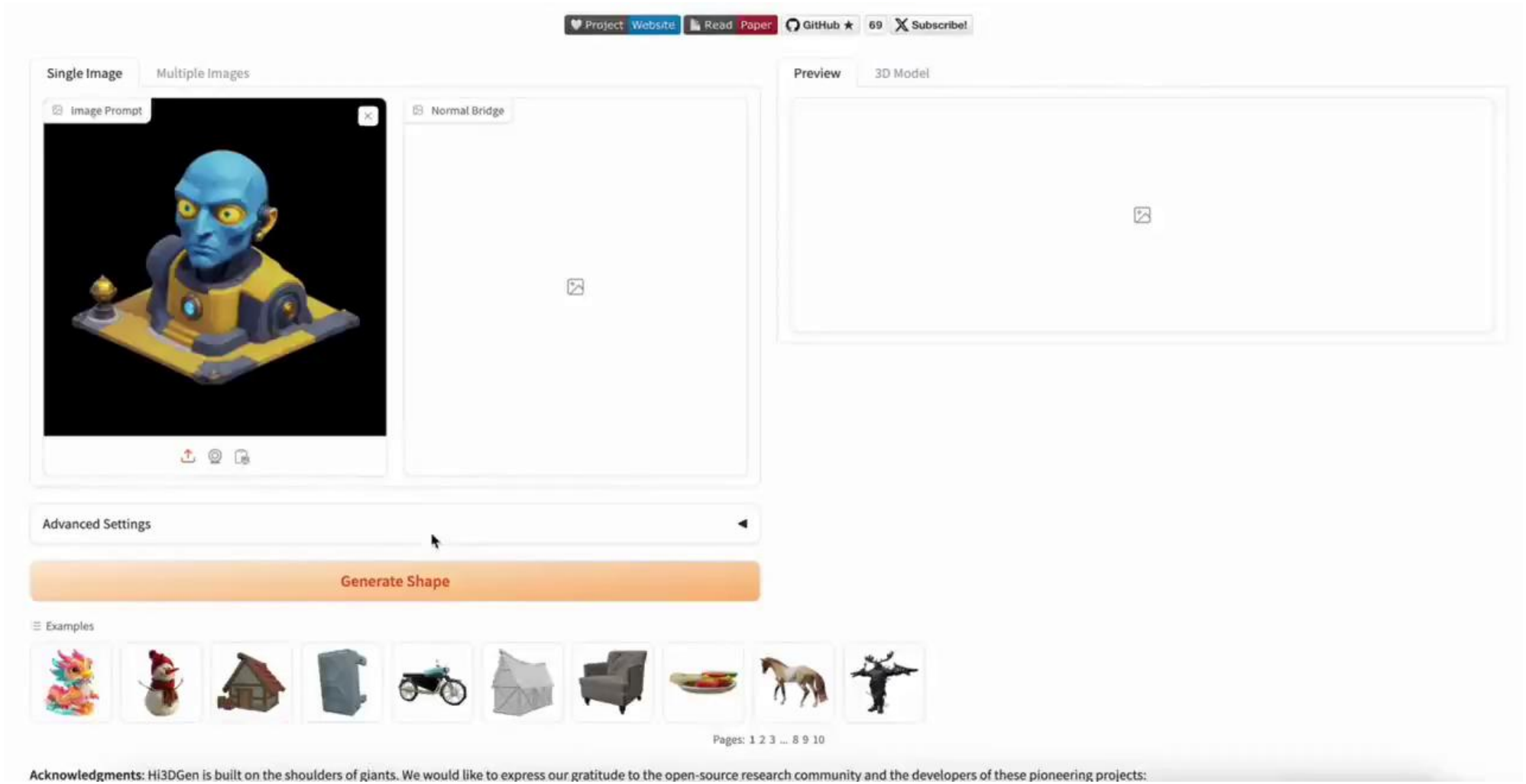
A.K.A. 生成式仿真为具身智能释放无限灵感

Physics-aware AIGC　　Neural Simulation　　Embodied AI

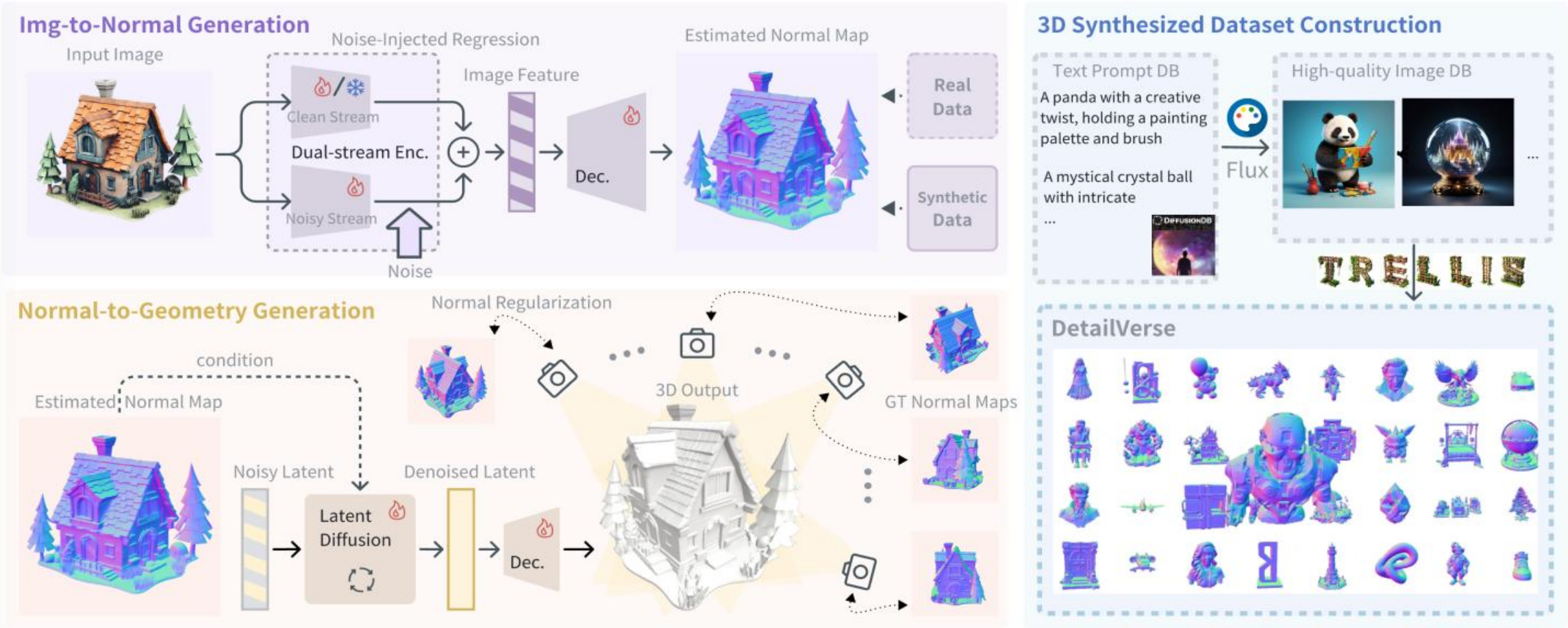




Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging



Hi3DGen: High-fidelity 3D Geometry Generation
from Images via Normal Bridging



Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging

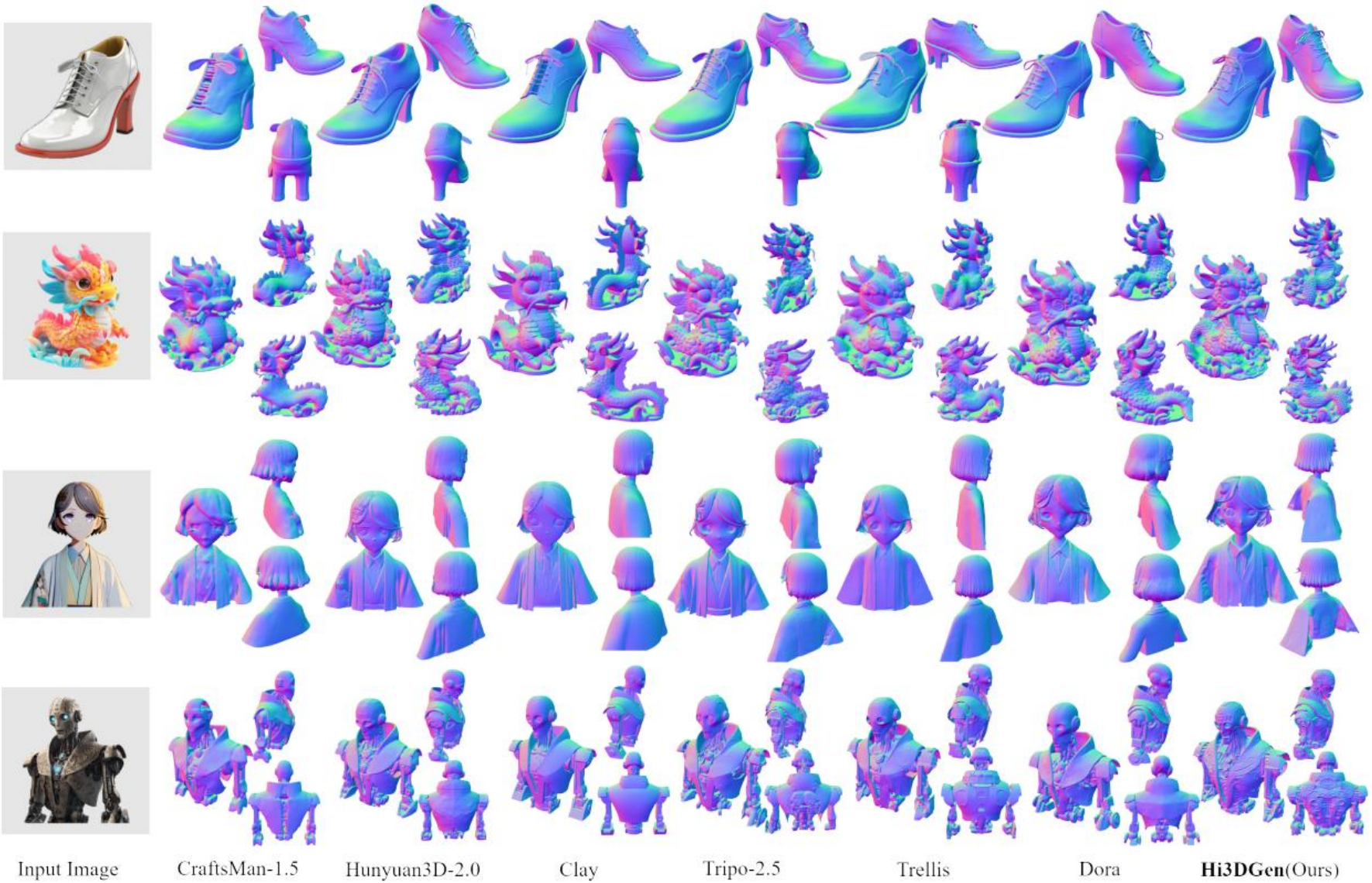
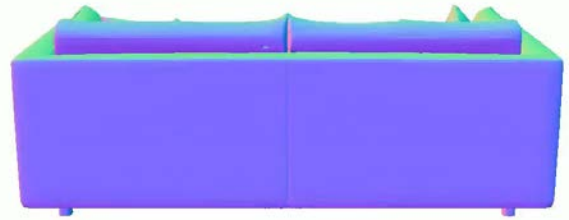
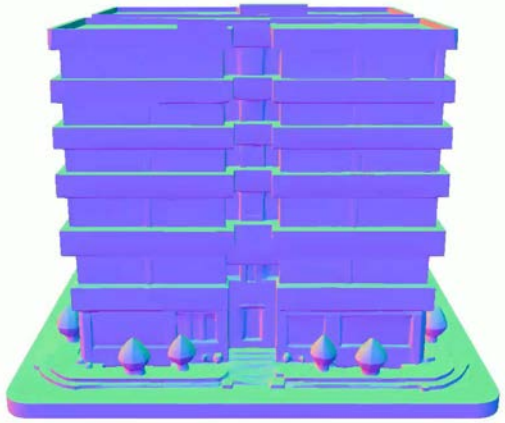
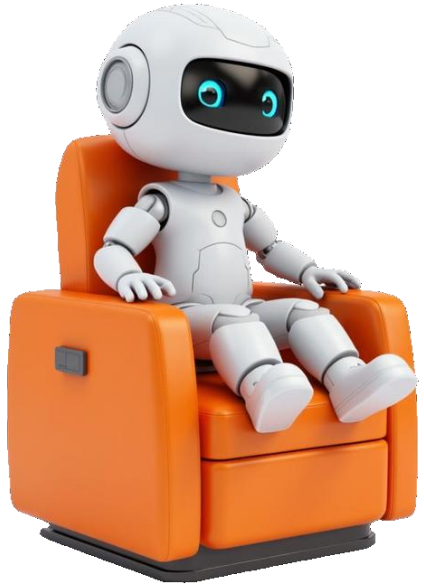


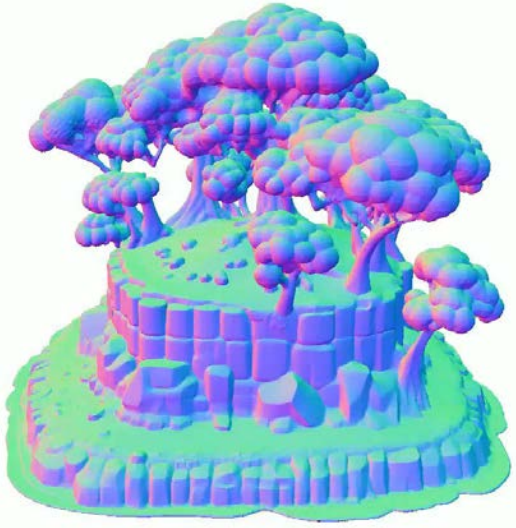
Figure 9. Qualitative 3D generation comparison on samples from Dora's project page [51].

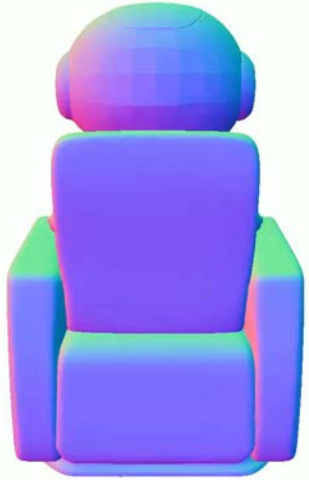
Hi3DGen: High-fidelity 3D Geometry Generation
from Images via Normal Bridging

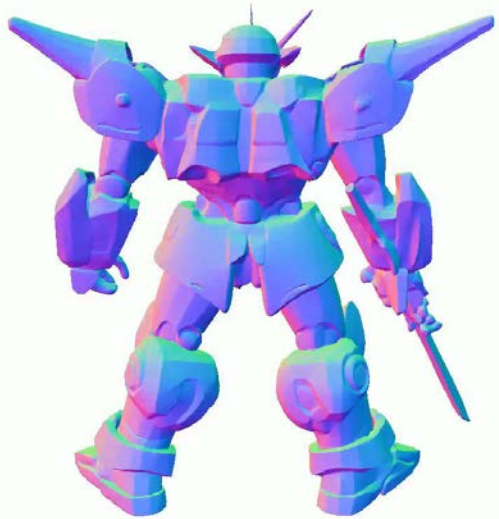


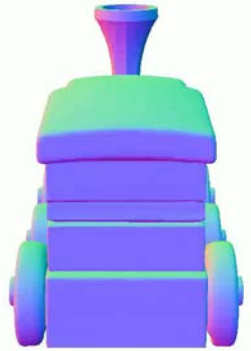


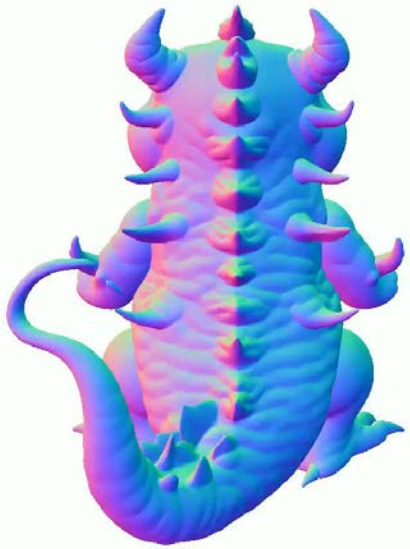










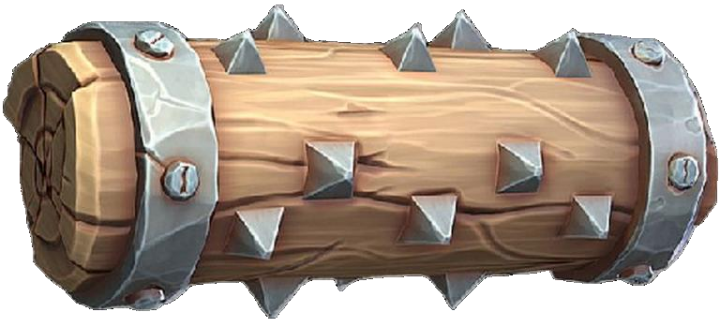


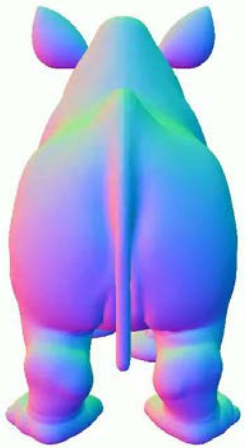


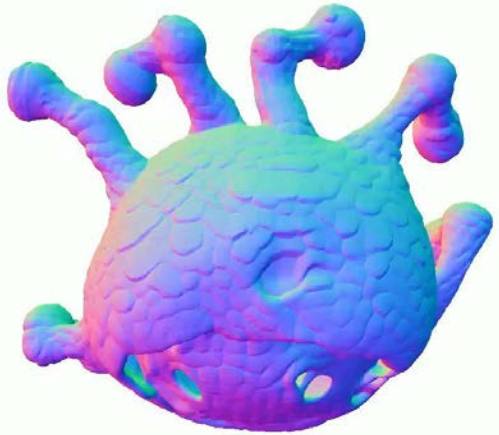


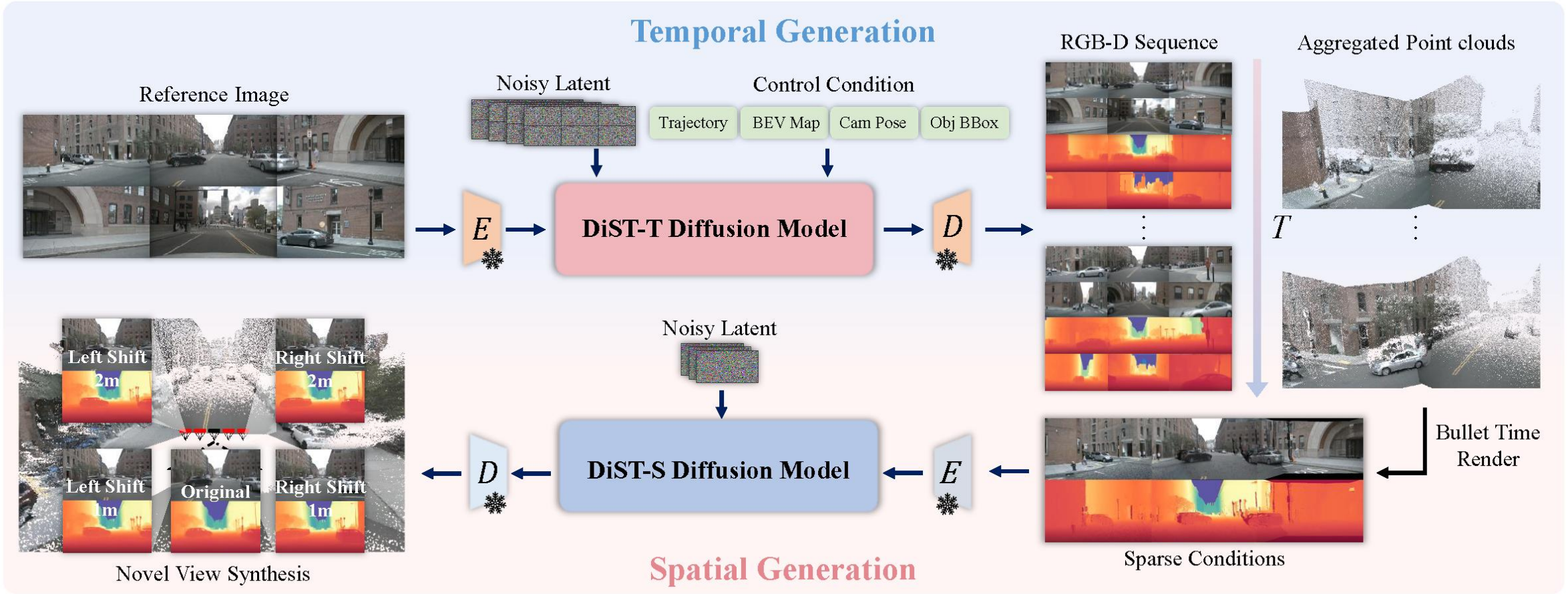


MFA



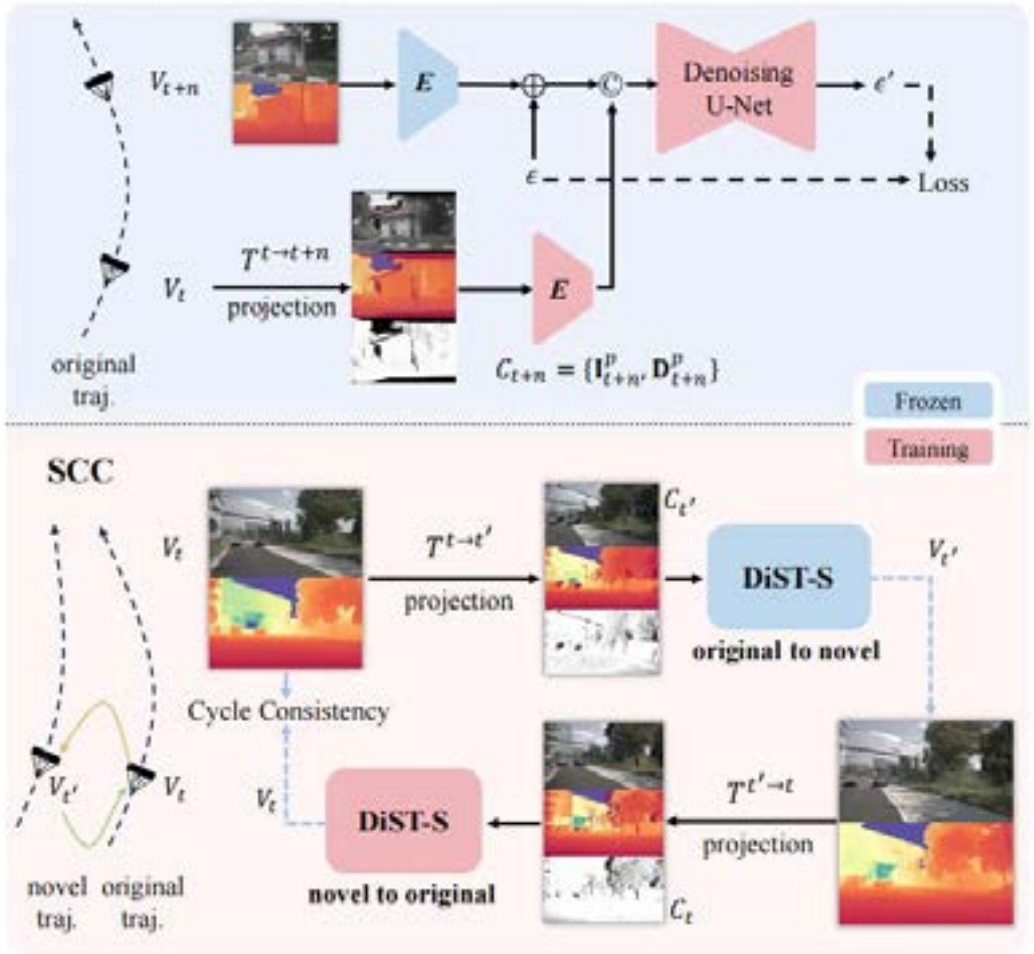






DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

Method	Temporal Gen.	Spatial NVS	Feed-forward
Vista [11]	✓	✗	✓
UniScene [22]	✓	✗	✓
Drive-WM [39]	✓	✗	✓
DriveDreamer [38]	✓	✗	✓
MagicDrive [10]	✓	✗	✓
MagicDriveDiT [9]	✓	✗	✓
3DGS [18]	✗	✓	✗
PVG [3]	✗	✓	✗
UniSim [50]	✗	✓	✗
EmerNeRF [46]	✗	✓	✗
StreetGaussian [44]	✗	✓	✗
DriveDreamer4D [55]	✗	✓	✗
Stag-1 [32]	✗	✓	✓
FreeVS [34]	✗	✓	✓
STORM [47]	✗	✓	✓
DriveForward [31]	✗	✓	✓
StreetCrafter [45]	✗	✓	✓
DreamDrive [26]	✓	✓	✗
MagicDrive3D [8]	✓	✓	✗
DiST-4D (Ours)	✓	✓	✓



DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

Method	Multi-view	Video	Depth	FID ↓	FVD ↓
BEVGen [29]	✓	✗	✗	25.54	-
BEVControl [48]	✓	✗	✗	24.85	-
DriveGAN [19]	✗	✓	✗	73.40	502.30
DriveDreamer [38]	✗	✓	✗	52.60	452.00
Vista [11]	✗	✓	✗	6.90	89.40
WoVoGen [25]	✓	✓	✗	27.60	417.70
Panacea [43]	✓	✓	✗	16.96	139.00
MagicDrive [10]	✓	✓	✗	16.20	217.94
MagicDriveDiT [9]	✓	✓	✗	20.91	94.84
Drive-WM [39]	✓	✓	✗	15.80	122.70
Vista* [11]	✓	✓	✗	13.97	112.65
UniScene [22]	✓	✓	✗	**6.45**	71.94
DiST-T (Ours-D)	✓	✓	✓	7.40	25.55
DiST-T (Ours)	✓	✓	✓	6.83	**22.67**

Table 2. **Quantitative results of RGB Video Generation.** Evaluation on the NuScenes validation set with FID and FVD. Unlike baselines lacking depth modeling, DiST-T generates multi-view RGB videos with depth, achieving state-of-the-art performance. The results of the multi-view variant of Vista* [11] are reported in [22].

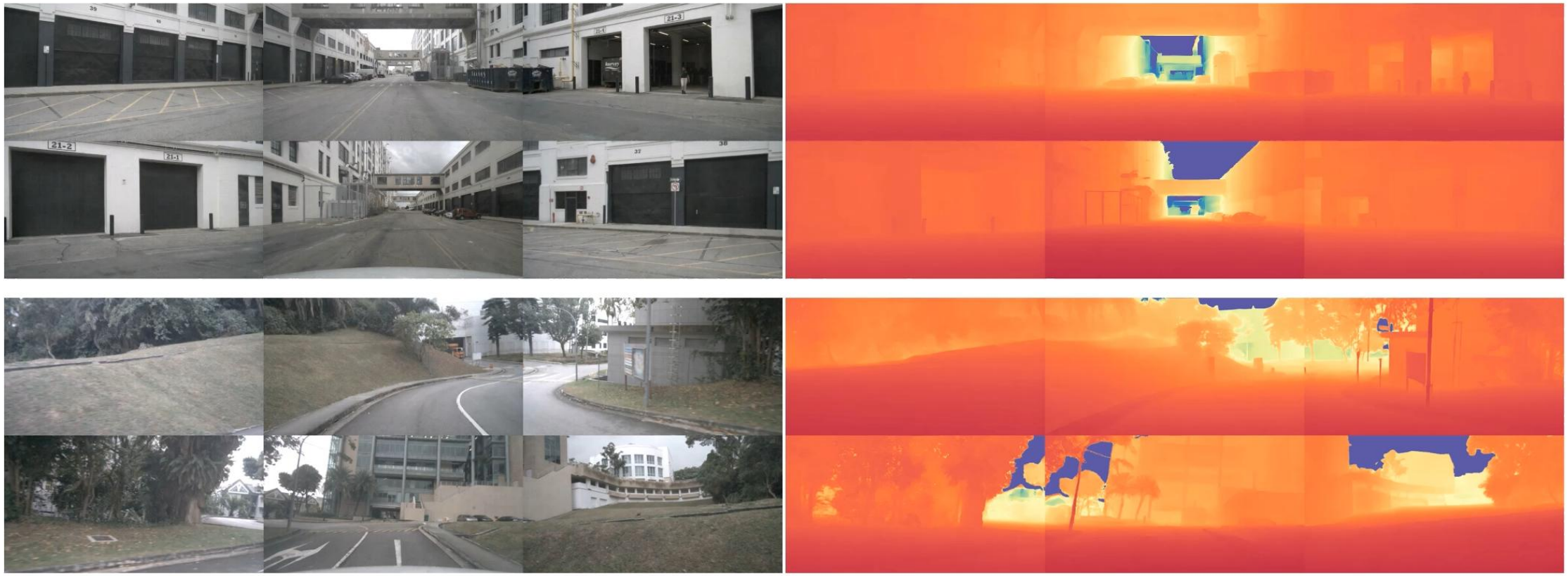
Method	Shift ±1m		Shift ±2m		Shift ±4m	
	FID ↓	FVD ↓	FID ↓	FVD ↓	FID ↓	FVD ↓
PVG [3]	48.15	246.74	60.44	356.23	84.50	501.16
EmerNeRF [46]	37.57	171.47	52.03	294.55	76.11	497.85
StreetGaussian [44]	32.12	153.45	43.24	256.91	67.44	429.98
OmniRe [4]	31.48	152.01	43.31	254.52	67.36	428.20
FreeVS* [34]	51.26	431.99	62.04	497.37	77.14	556.14
DiST-4D (Ours)	20.64	130.98	25.08	156.60	33.56	189.04
DiST-4D (+SCC)	16.40	112.86	19.50	124.97	25.16	146.56
DiST-S (Ours)	12.00	55.80	15.72	89.54	22.54	137.03
DiST-S (+SCC)	**10.12**	**45.14**	**12.97**	**68.80**	**17.57**	**105.29**

Table 5. **Quantitative comparison of novel view synthesis**. We evaluate FID and FVD across different viewpoint shifts (±1m, ±2m, ±4m). FreeVS* denotes we retrain FreeVS [34] using the official code on the nuScenes dataset [2]. DiST-4D uses generated RGB-D videos from DiST-T, while DiST-S leverages real videos with preprocessed metric depth. DiST-S (+SCC) achieves the best performance, demonstrating the effectiveness of self-supervised cycle consistency in improving novel view synthesis quality.

DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

Temporal RGB-D Generation

Temporal Generation by DiST-T

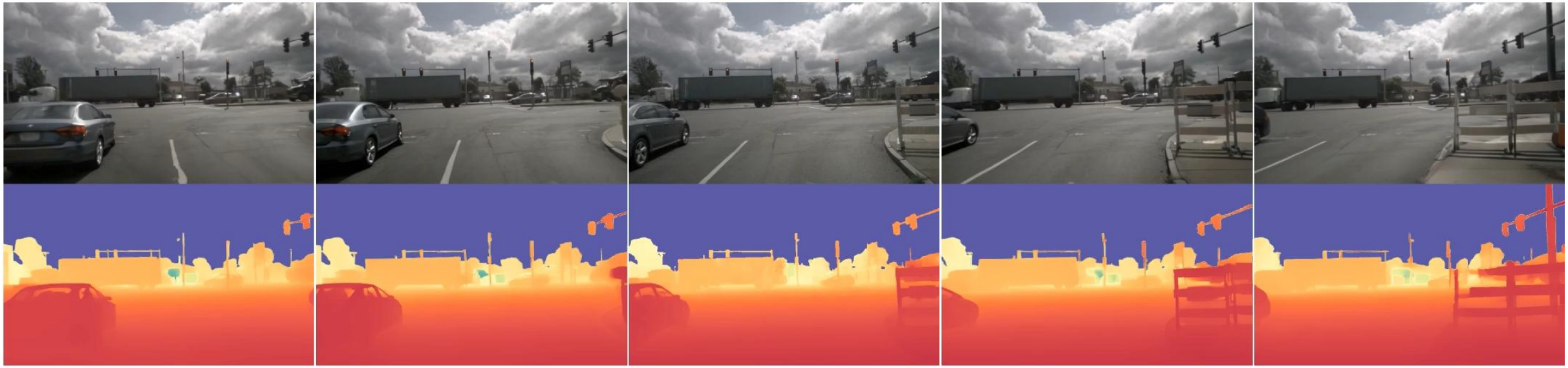


DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

Spatial RGB-D Novel View Synthesis

Spatial Novel View Generated by DiST-S

Left 2m | Left 1m | Original Traj. | Right 1m | Right 2m

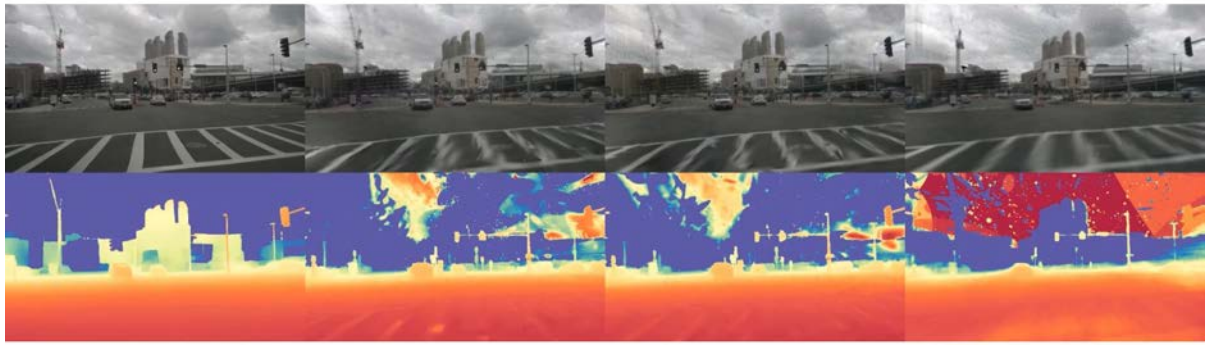


DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

Comparison with other GS works on Spatial RGB-D NVS

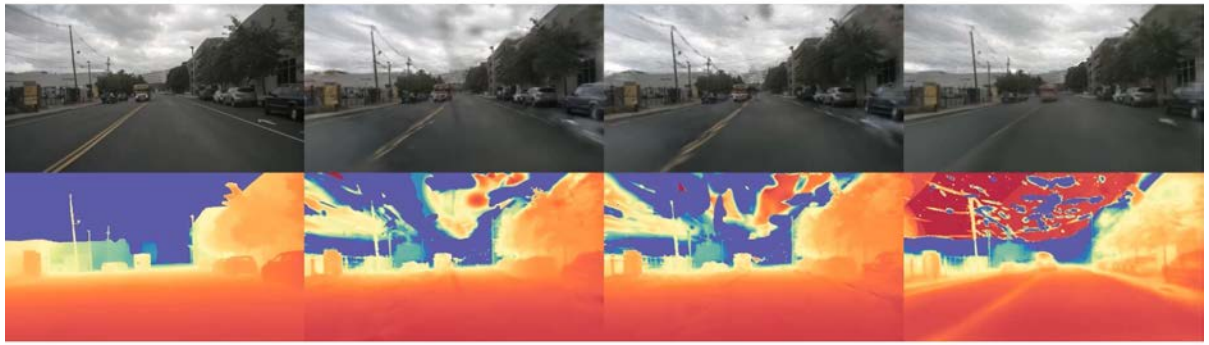
Left 1m

Ours OmniRe StreetGS PVG



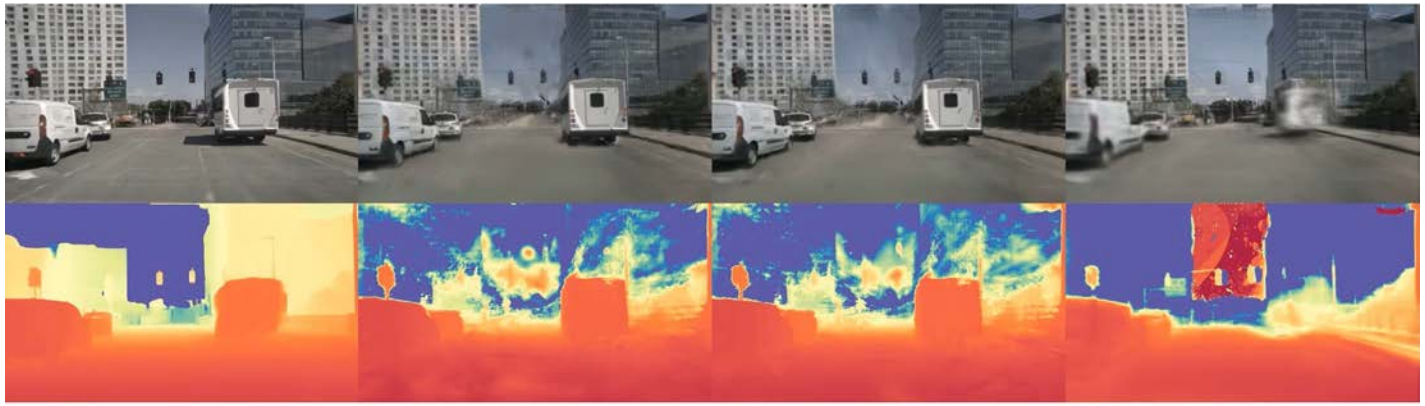
Left 1m

Ours OmniRe StreetGS PVG



Right 1m

Ours OmniRe StreetGS PVG



DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

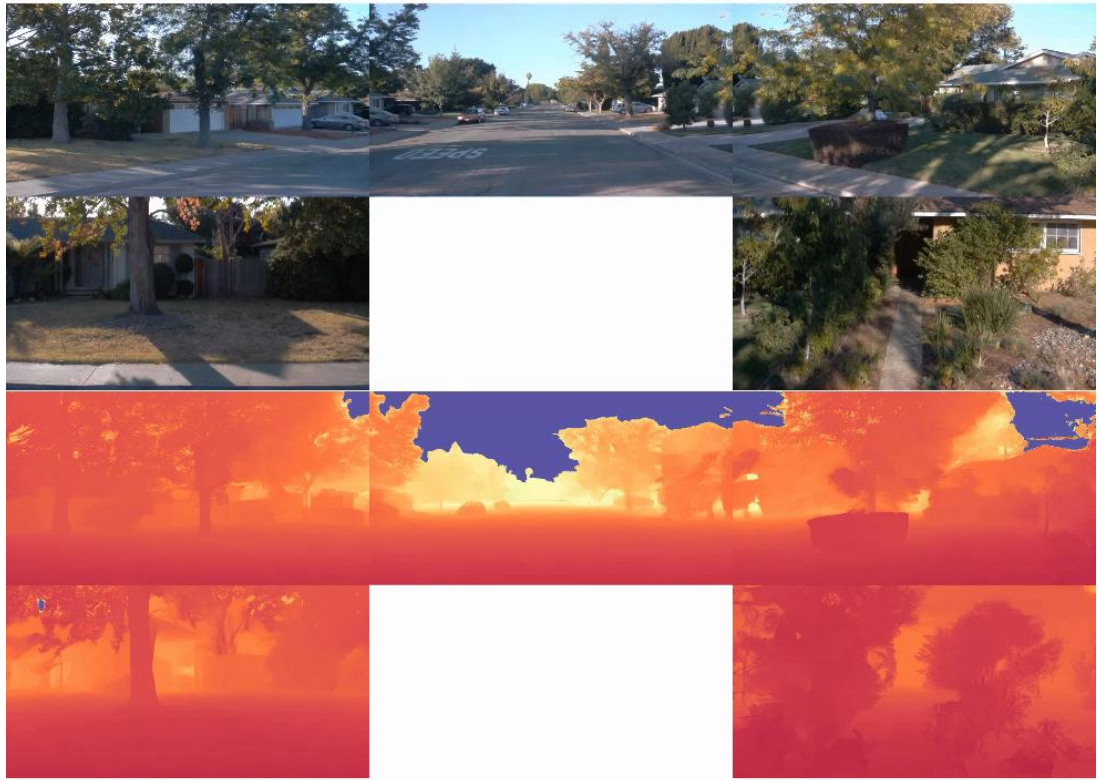
Multi-View Spatial NVS



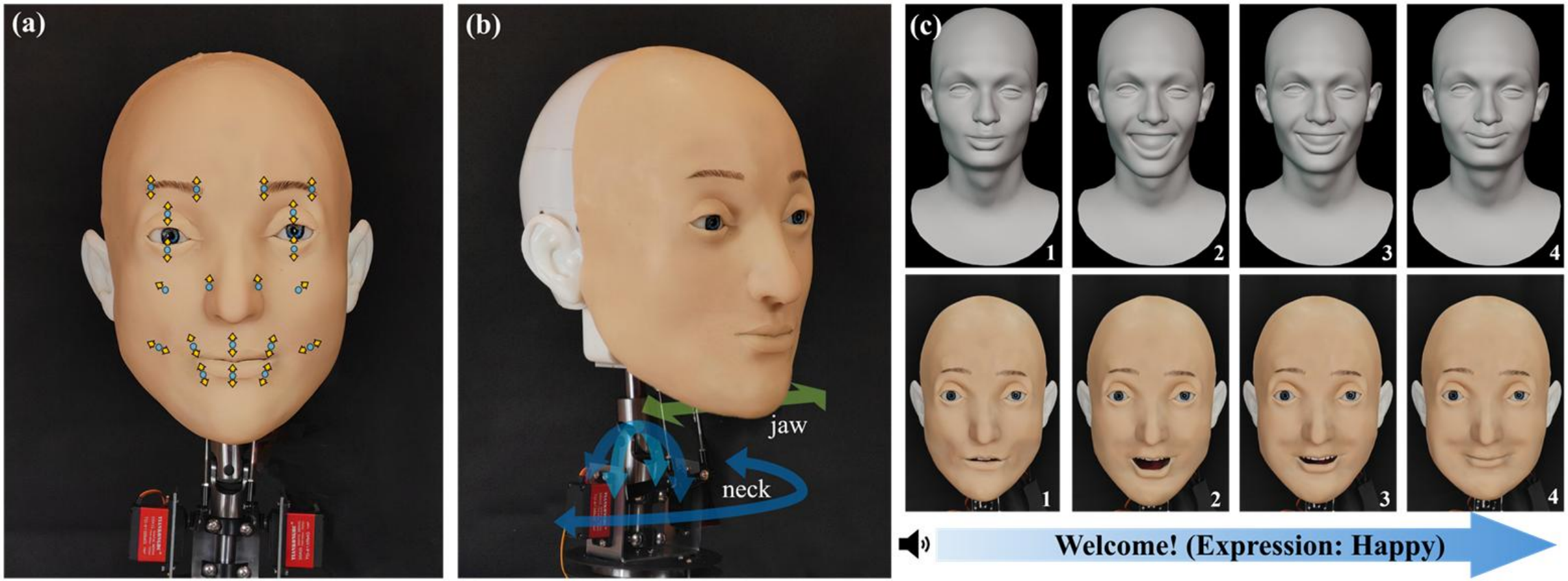



DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

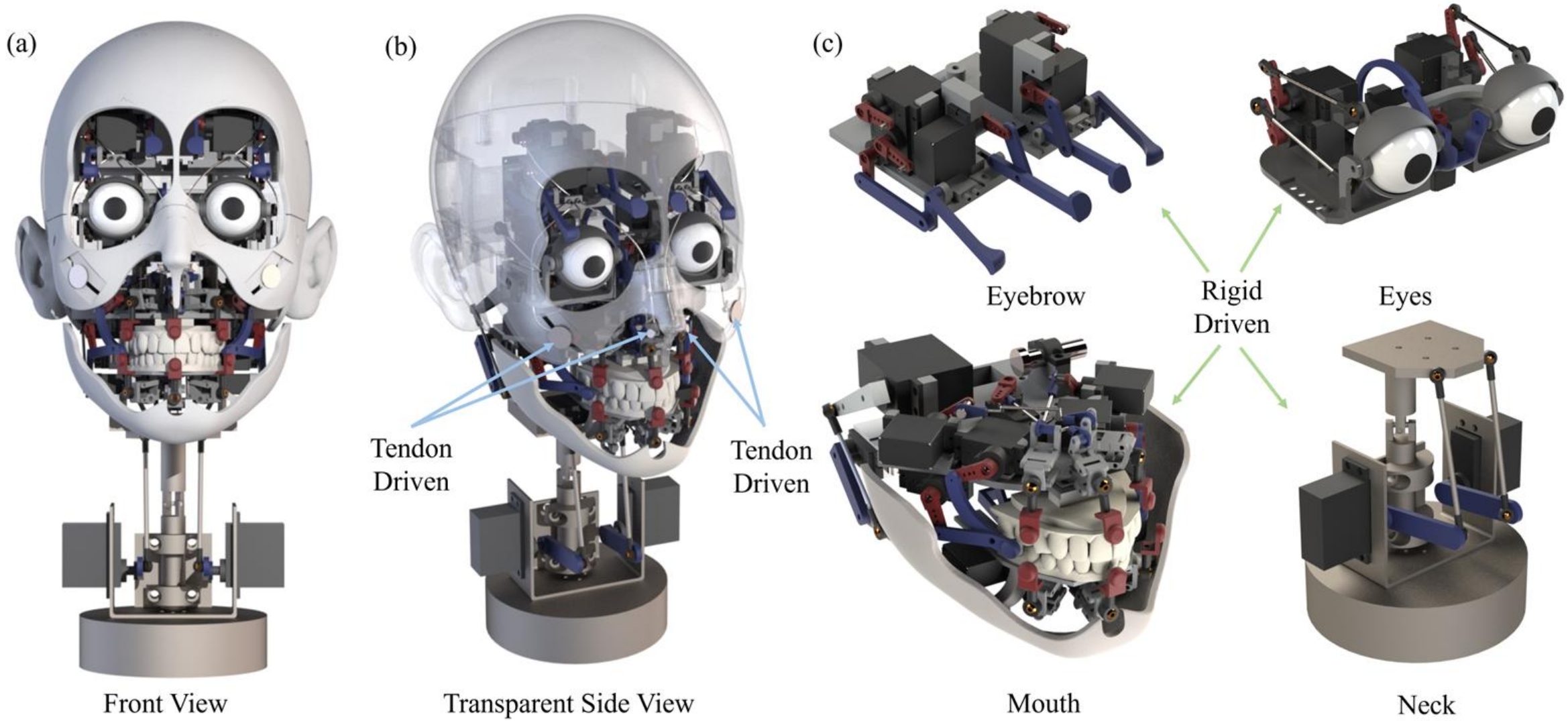
DiST-T: Zero-shot on Waymo



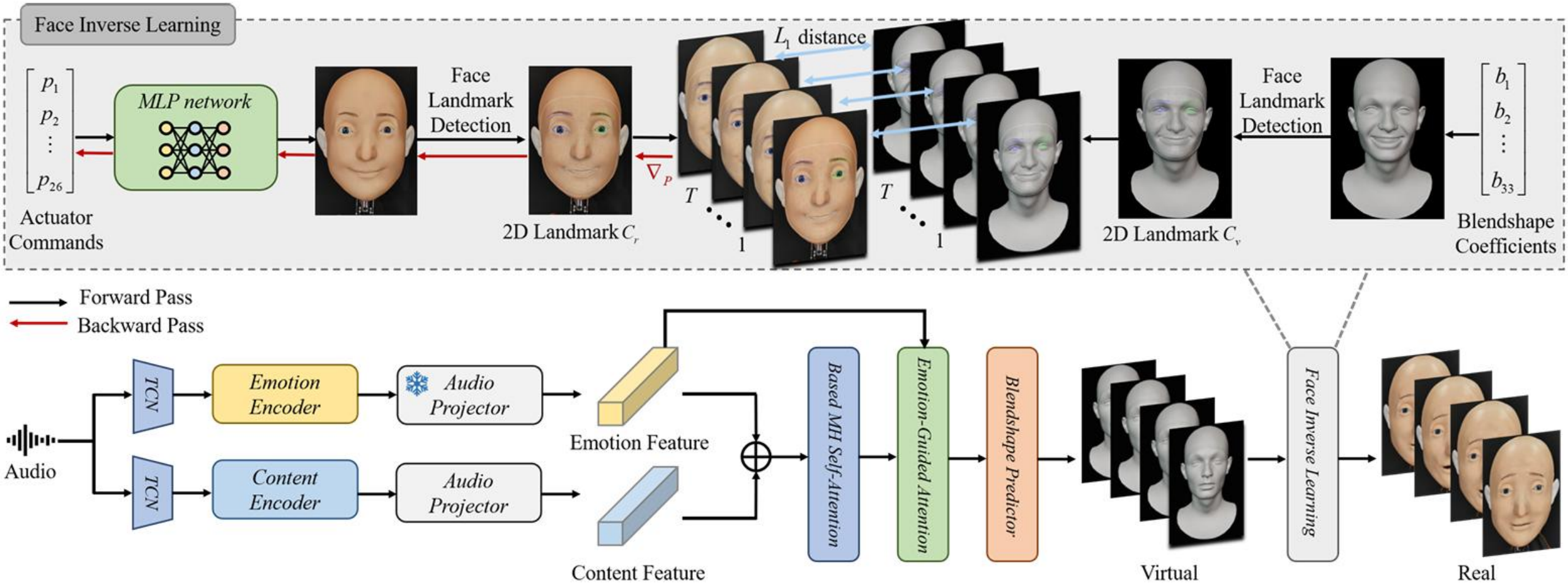
DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation



Morpheus: A Neural-driven Animatronic Face with Hybrid Actuation and Diverse Emotion Control



Morpheus: A Neural-driven Animatronic Face with Hybrid Actuation and Diverse Emotion Control



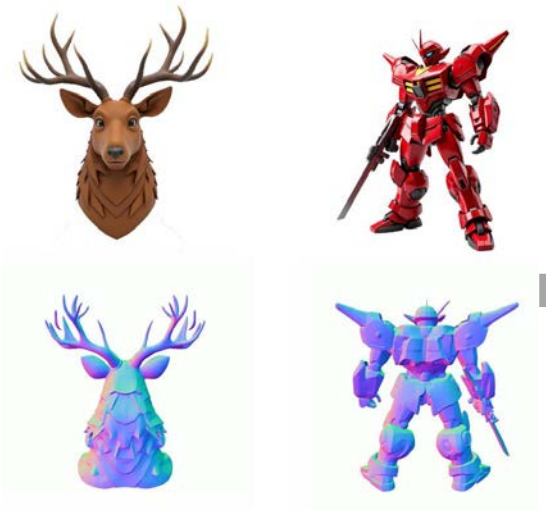
Morpheus: A Neural-driven Animatronic Face with Hybrid Actuation and Diverse Emotion Control

Embodied AI Powered by Physics-aware Generative Simulation

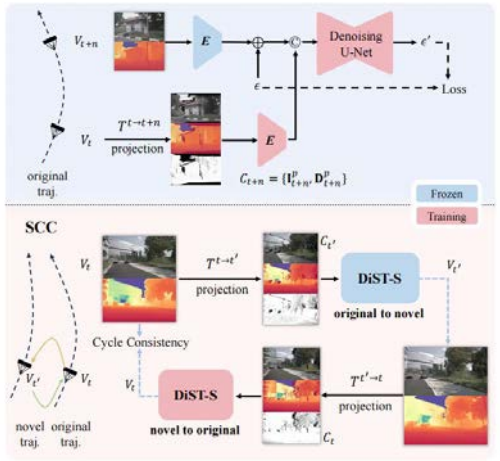
Hao Zhao

https://sites.google.com/view/fromandto

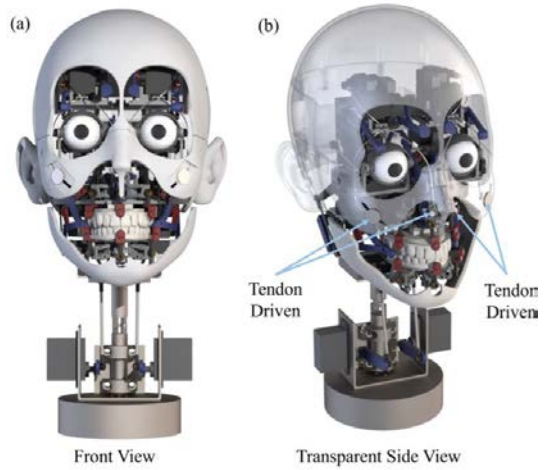
A.K.A. 生成式仿真为具身智能释放无限灵感



Physics-aware AIGC

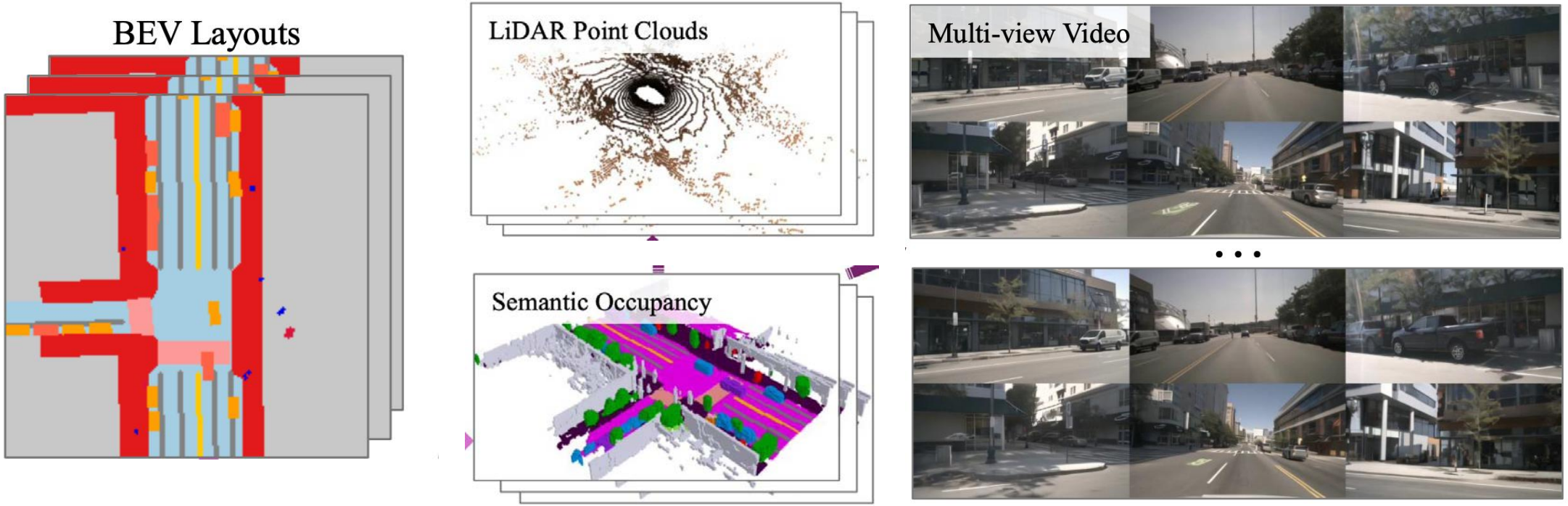


Neural Simulation



Embodied AI

自动驾驶世界模型


BEV Layouts
LiDAR Point Clouds
Semantic Occupancy
Multi-view Video

我们是谁

赵昊

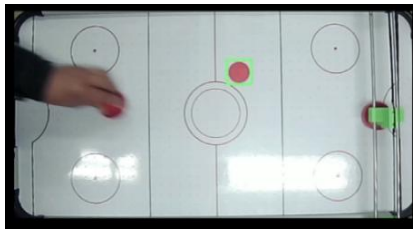
助理教授

- 清华智能产业研究院（AIR） 助理教授
- 北京智源人工智能研究院 可控世界模型创新中心 主任
- 清华大学 电子工程系 博士
- 顶级计算机视觉科学家
- 清华大学“天空工场”机器人社团创始人和负责人
- 专注于几何与认知层面的场景理解及其在机器人中的应用
- 拥有 10 + 项 PCT/US 专利，在 TPAMI/IJCV/CVPR/ICCV/ECCV、ICRA/IROS/RA - L、NeurIPS/ICLR、SIGGRAPH 等顶级期刊和会议发表 40 + 篇论文，多次获得最佳论文奖，如 CICAI 2023 Best Paper Runner - up、3DV 2024 Best Paper
- 个人网站 https://sites.google.com/view/fromandto

Who am I？科学家/创客/媒体人








- B.S. and Ph.D. from EE @ Tsinghua University
- Postdoc @ Peking University
- Research Scientist @ Intel Labs
- Assistant Professor @ Tsinghua University (now)
- 10+ PCT/US Patents Filed
- 40+ Top-tier Publications in
- TPAMI/IJCV/CVPR/ICCV/ECCV
- ICRA/IROS/RA-L
- NeurIPS/ICLR
- SIGGRAPH
- CICAI 2023 Best Paper Runner-up
- 3DV 2024 Best Paper

各种四旋翼（微型/变桨矩）双旋翼，小车，交互机器人

学生时代：清华最大的机器人社团天空工场创始团队+负责人

顶尖高校/机构的求学/工作经历

横跨视觉/机器人/机器学习/图形学四大领域的科研老兵

多次获得最佳论文奖

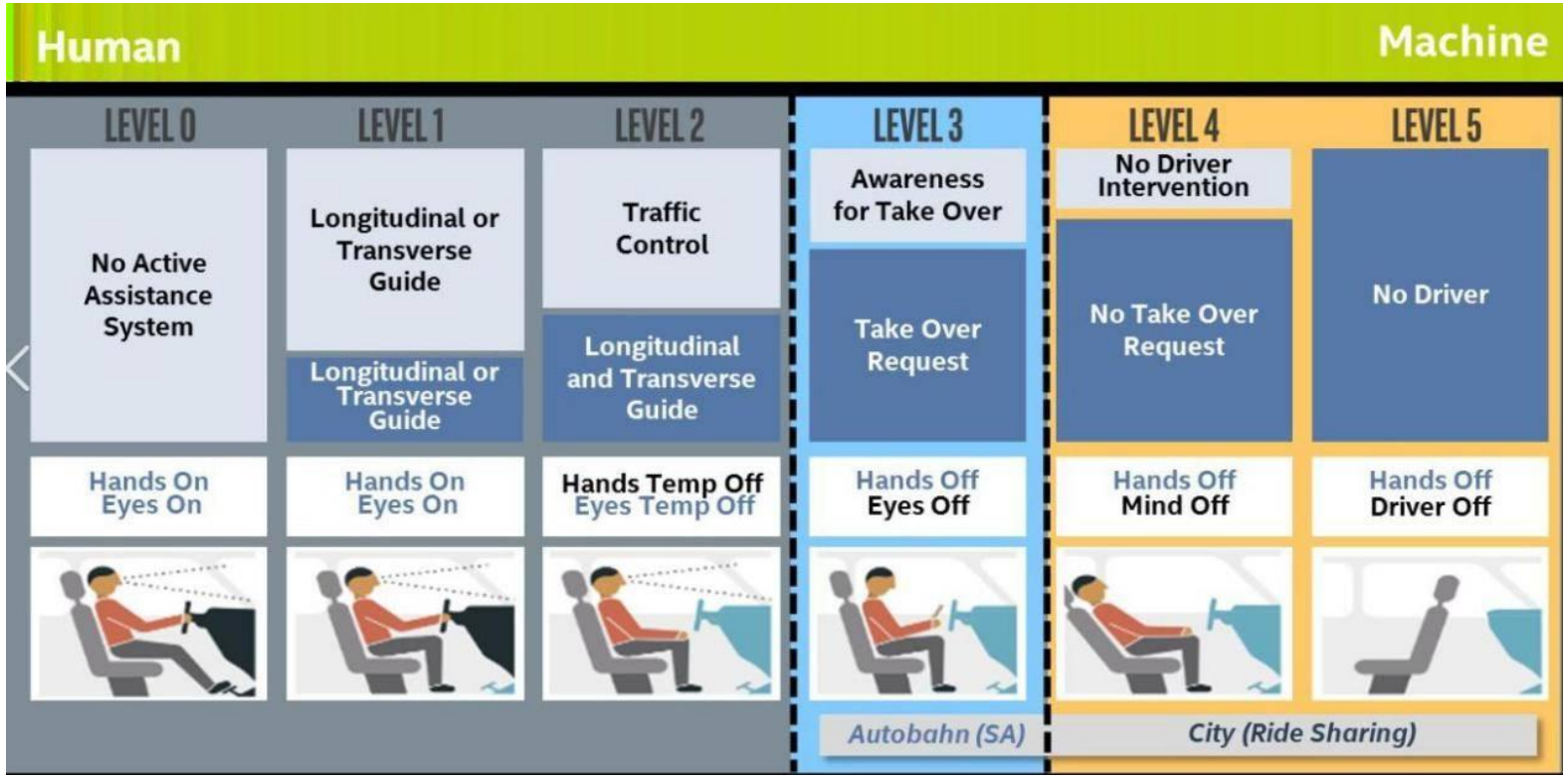
独立音乐人是我的人格底色

多年横跨多平台的全媒体矩阵持续输出科研和产业观点的意见领袖

自动驾驶技术分级与市场渗透现状

1、技术分级：

L2（部分自动化）至L5（完全自动化），当前主流为L2/L3（高速NOA渗透率超10%，城市NOA超3%）47



2、市场数据：

- 2024年中国乘用车L2及以上智驾系统装配量达1001万套，渗透率56.4%6
- 前视一体机（低成本L2方案）渗透率45.7%，成为经济车型主流选择6

全球自动驾驶领军企业及核心产品

1.特斯拉

产品：Autopilot（L2）、FSD（Full Self-Driving，L3测试中）。
技术亮点：端到端技术路线，
BEV+Transformer感知架构，超算Dojo支持数据训练48。

2.Waymo（Google旗下）

服务：Robotaxi（L4级）在美国多城市运营，累计路测超2000万英里。

3.Mobileye（Intel旗下）

产品：EyeQ系列芯片（如EyeQ4H）赋能前视一体机，合作车企包括丰田、宝马等6。

4.Cruise（GM旗下）

服务：旧金山等地的无人驾驶出租车服务，聚焦城市复杂场景。

5.Zoox（Amazon旗下）

产品：全自研L5级无人车，专为共享出行设计。

1.华为

产品：ADS 2.0（高阶智驾系统），支持无图城市NOA，合作车企包括问界、阿维塔78。

2.小鹏汽车

技术：XNGP系统（城市NOA），用户渗透率93.3%，路测数据驱动算法迭代8。

3.百度Apollo

服务：Robotaxi（萝卜快跑）在北上广深等城市运营，累计订单超500万单。

4.小马智行Pony

凭借人工智能技术领域的最新突破，已与丰田、现代、一汽、广汽等车厂建立合作。

5.文远知行

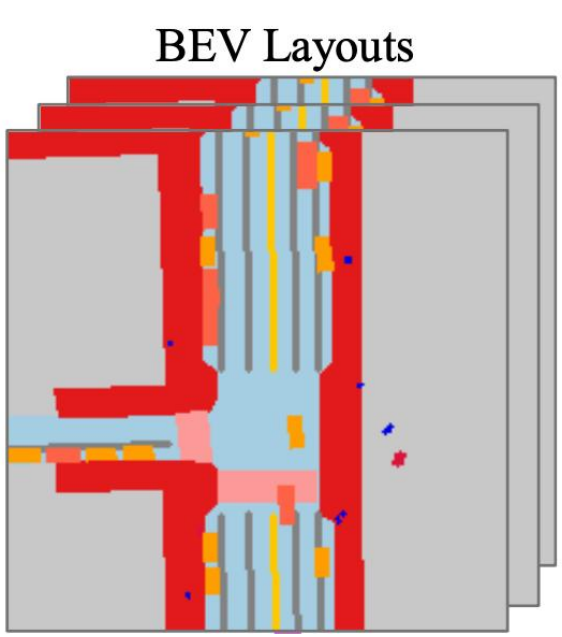
场景：L4级自动驾驶小巴和物流车，中美两地同步运营8。

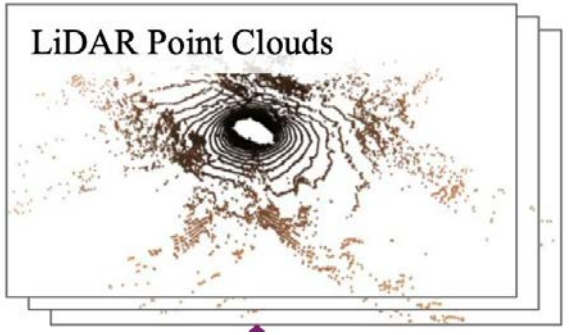
- 世界模型的核心作用：
- 内在环境表示： 利用多模态数据构建车辆的“内心世界”， 对环境进行抽象与建模

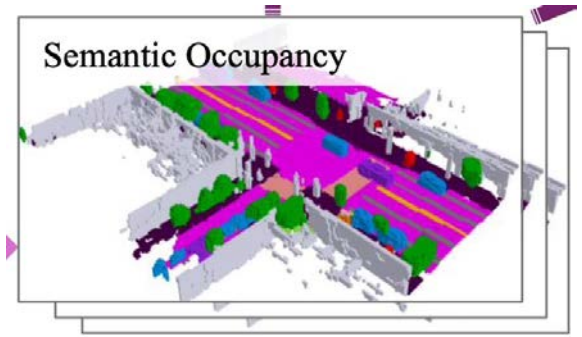
目前自驾世界模型的主要方法：

- Multimodel与统一场景表示： MagicDrive 、UniScene
- 基于鸟瞰视图（BEV）的建模与控制：BEVGen、BEVControl
- 生成式世界模型：DriveDreamer、Vista、Drive-WM
- 基于LiDAR数据的世界模型构建：LiDARGen、LiDARDiffusion、LiDARDM
- 基于Occupancy的空间表示方法：OccSora、OccLLama、OccWorld

Multimodel与统一场景表示



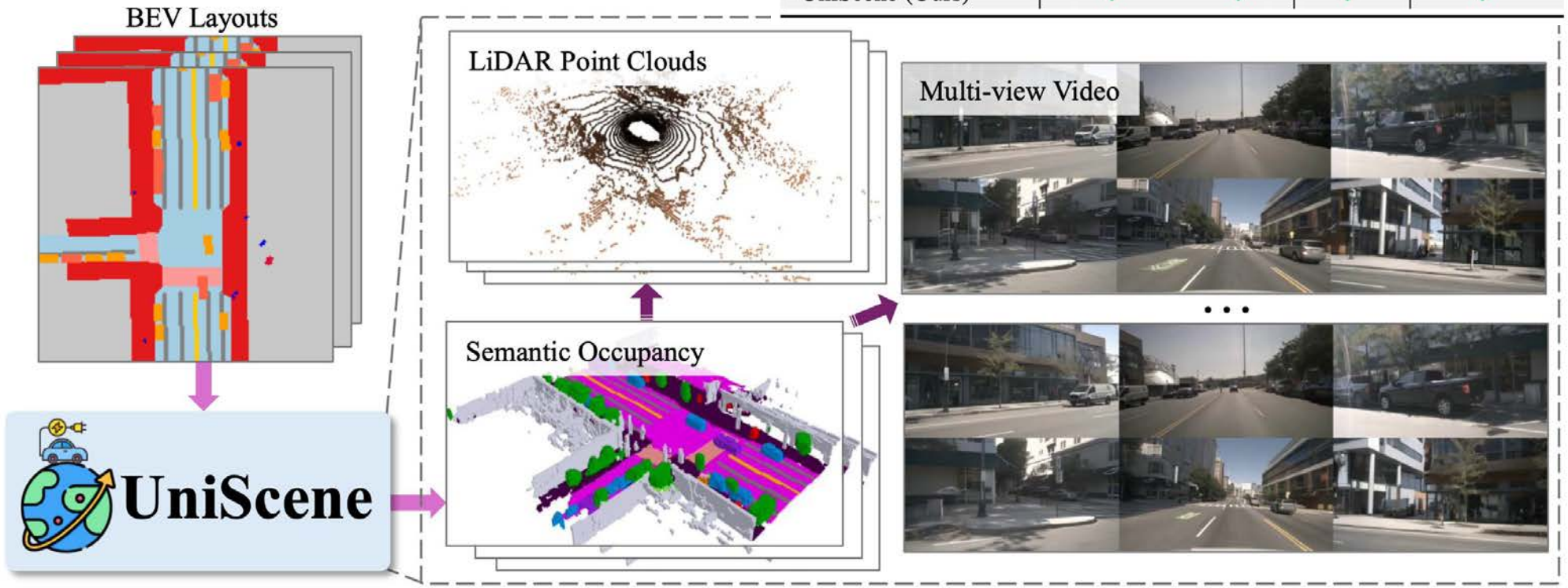




UniScene: Unified Occupancy-centric Driving Scene Generation

鉴于 BEV 布局，UniScene 通过以占用为中心的分层建模方法促进多种数据生成，包括语义占用、多视图视频和 LiDAR 点云。

Method	Multi-view	Video	LiDAR	Occupancy
BEVGen [55]	✓	✗	✗	✗
BEVControl [80]	✓	✗	✗	✗
DriveDreamer [85]	✗	✓	✗	✗
Vista [16]	✗	✓	✗	✗
MagicDrive [15]	✓	✓	✗	✗
Drive-WM [66]	✓	✓	✗	✗
LiDARGen [90]	✗	✗	✓	✗
LiDARDiffusion [50]	✗	✗	✓	✗
LiDARDM [91]	✗	✗	✓	✗
OccSora [59]	✗	✗	✗	✓
OccLLama [68]	✗	✗	✗	✓
OccWorld [86]	✗	✗	✗	✓
UniScene (Ours)	✓	✓	✓	✓

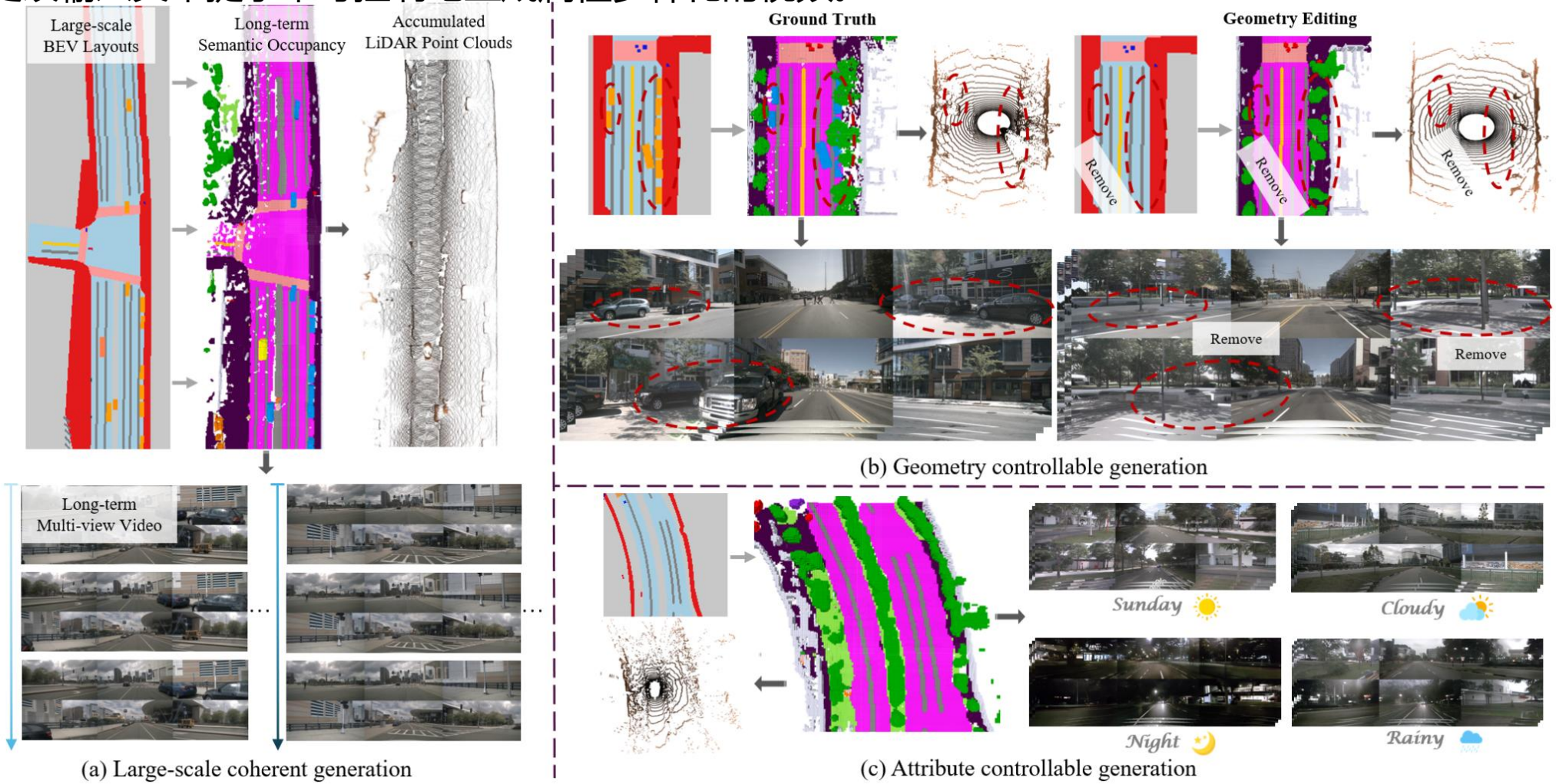


UniScene: Unified Occupancy-centric Driving Scene Generation

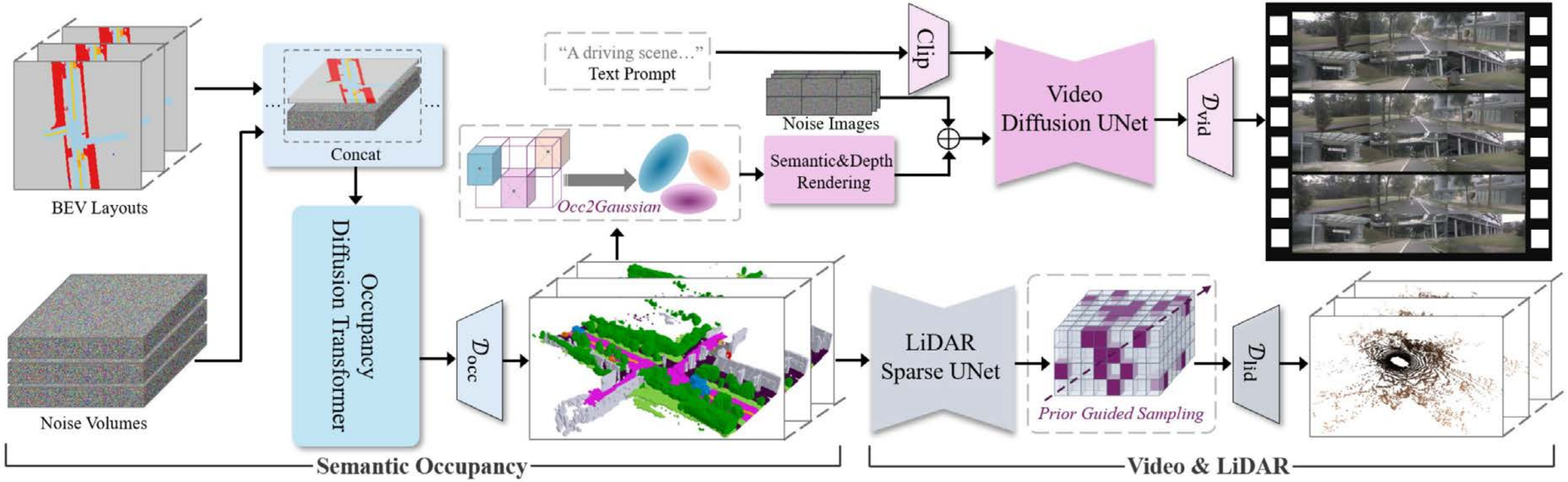
(a) 大规模连贯生成Occupancy、LiDAR 点云和多视角视频。

(b) 通过简单编辑输入的 BEV 布局来传达用户命令，可控制地生成Occupancy 、视频和 LiDAR。

(c) 通过更改输入文本提示来可控制地生成属性多样化的视频。



UniScene: Unified Occupancy-centric Driving Scene Generation



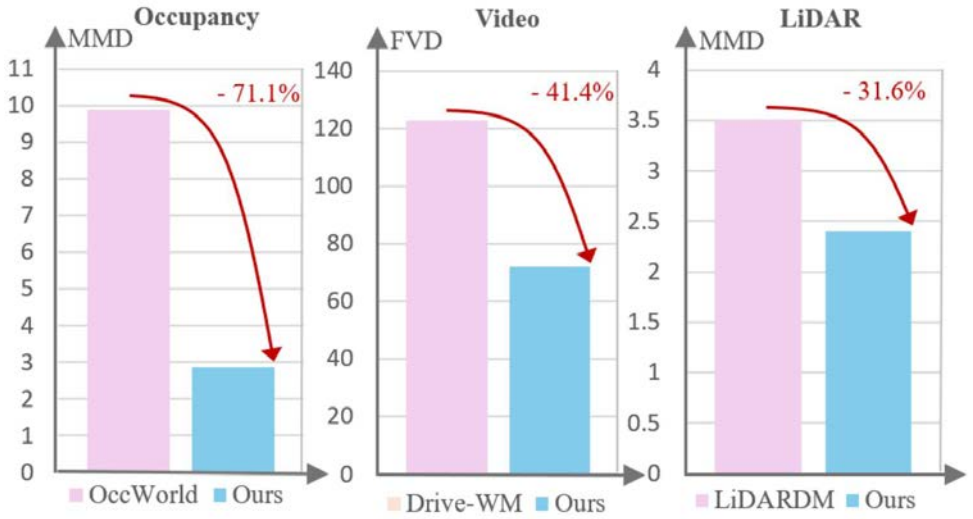
基于 occupancy-centric 层级组织，生成过程分为两大部分：

1.Controllable Occupancy Generation：将 BEV 布局与 noise volumes 拼接后输入 Occupancy Diffusion Transformer，并由 Occupancy VAE Decoder (• _occ) 解码，生成可控的 occupancy。

2.Occupancy-based Video and LiDAR Generation：将生成的 occupancy 转换为 3D Gaussians，再渲染为 semantic 和 depth maps，利用类似 ControlNet 的附加编码器处理后，通过 Video VAE Decoder (• _vid) 输出视频；同时，经过稀疏 UNet 处理并在几何先验引导下采样的 occupancy 被送至 LiDAR head (• _lid) 生成 LiDAR 数据。

UniScene: Unified Occupancy-centric Driving Scene Generation

Quantitative Results



(b) Generation performance comparison on different tasks

Method	MMD (10^{-4})↓	JSD ↓	Time (s)↓
LiDARDM [76]	3.51	0.118	45.12
Open3D [74]	8.15	0.149	2.39
Ours (Gen Occ)	2.40	0.108	0.47
Ours (GT Occ)	**1.53**	**0.072**	**0.25**

Quantitative evaluation for LiDAR generation

Method	Compression Ratio ↑	mIoU ↑	IoU ↑
OccLLama (VQVAE) [57]	8	75.2	63.8
OccWorld (VQVAE) [71]	16	65.7	62.2
OccSora (VQVAE) [48]	**512**	27.4	37.0
Ours (VAE)	32	**92.1**	**87.0**
Ours (VAE)	**512**	72.9	64.1

Quantitative evaluation for occupancy reconstruction

Method	Multi-view	Video	FID ↓	FVD ↓
BEVGen [45]	✓	✗	25.54	-
BEVControl [66]	✓	✗	24.85	-
DriveGAN [21]	✗	✓	73.40	502.30
DriveDreamer [70]	✗	✓	52.60	452.00
Vista [14]	✗	✓	6.90	89.40
WoVoGen [34]	✓	✓	27.60	417.70
Panacea [58]	✓	✓	16.96	139.00
MagicDrive [13]	✓	✓	16.20	-
Drive-WM [55]	✓	✓	15.80	122.70
Vista * [14]	✓	✓	13.97	112.65
Ours (Gen Occ)	✓	✓	6.45	71.94
Ours (GT Occ)	✓	✓	**6.12**	**70.52**

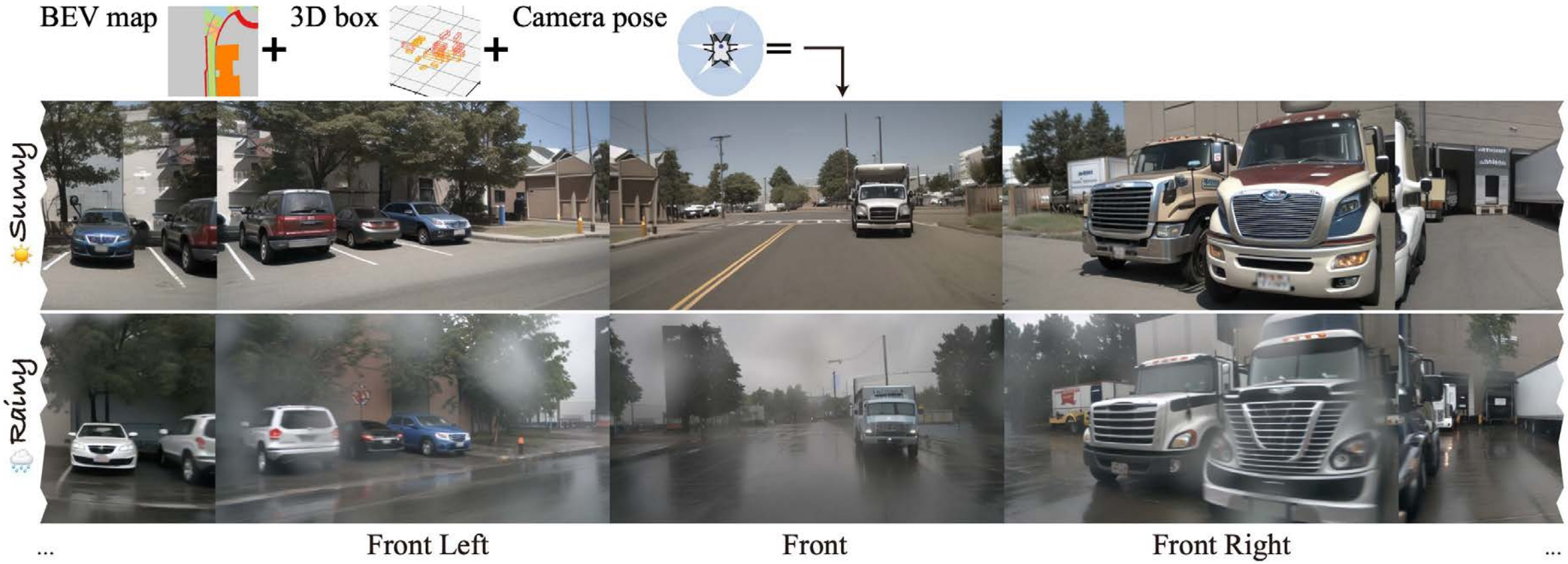
Quantitative evaluation for video generation

MagicDrive: Street View Generation with Diverse 3D Geometry Control

Overview

Multi-camera street view generation from MAGICDRIVE.

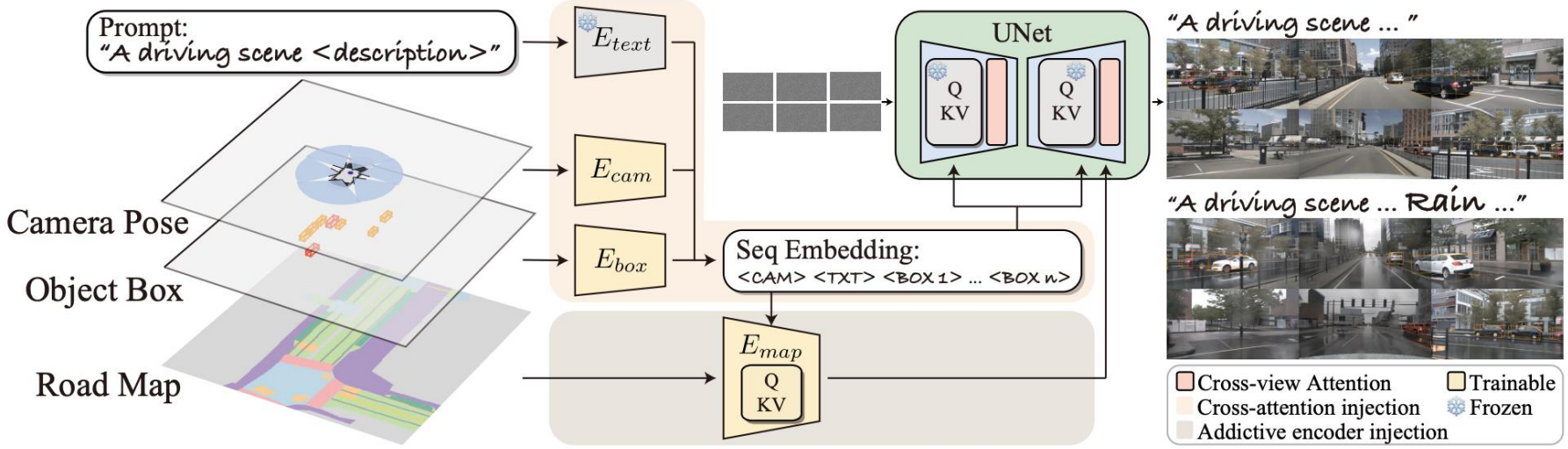
MAGICDRIVE can generate continuous camera views with controls from the road map, object boxes, and text



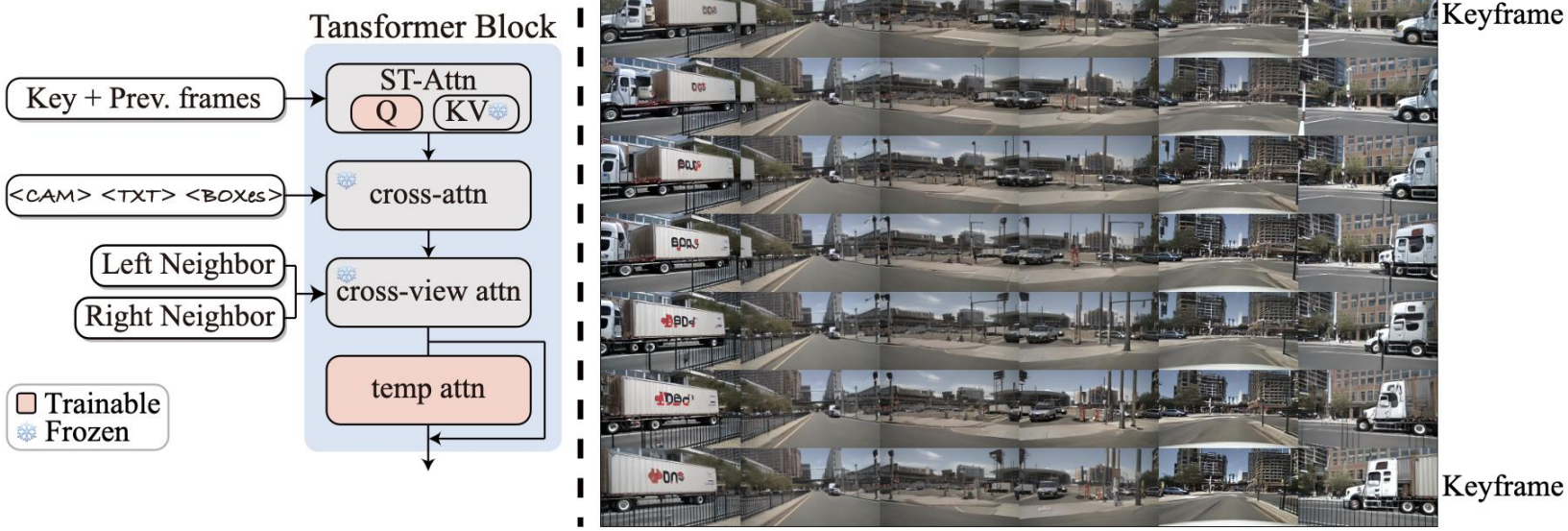
MagicDrive: Street View Generation with Diverse 3D Geometry Control

Overview

MAGICDRIVE 生成高度真实的街景图像，利用 3D annotations 中的几何信息，通过独立编码 road maps、object boxes 和 camera parameters，实现精确的 geometry-guided synthesis；此外，它还能根据 prompt（如 weather）提供生成指导。



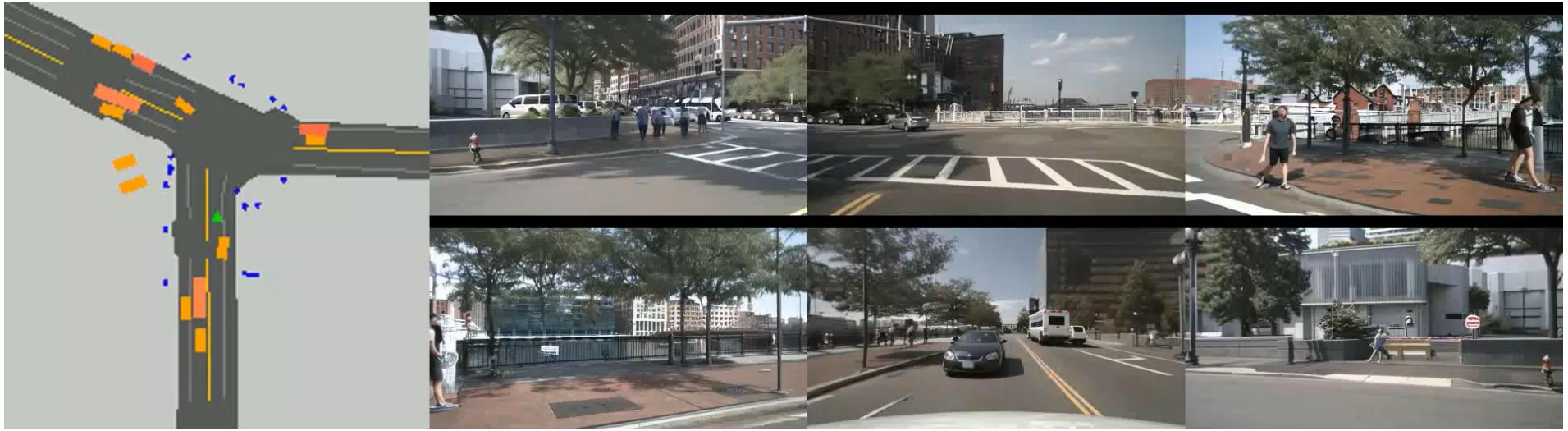
Video Generation



MagicDrive: Street View Generation with Diverse 3D Geometry Control

Diverse Generation

- Latent interpolation with the same geometric conditions
- Image sequences for continuous scenes

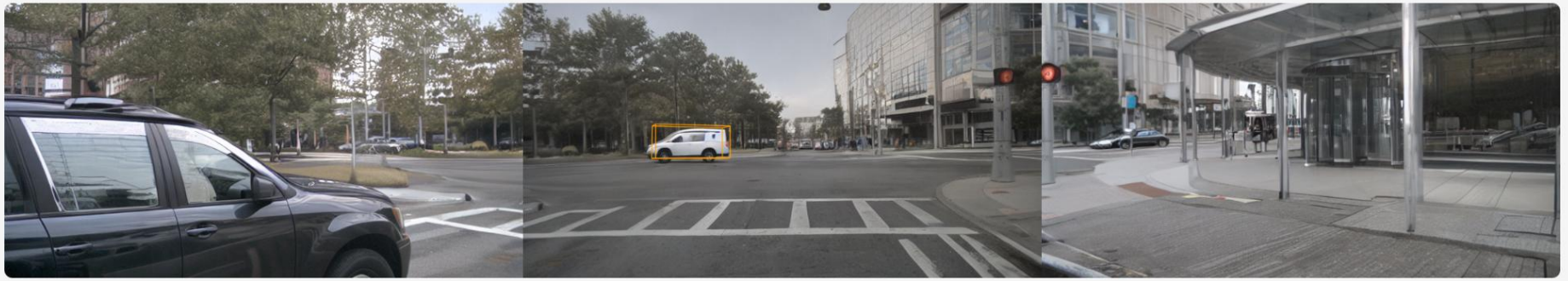


MagicDrive: Street View Generation with Diverse 3D Geometry Control

Diverse 3D Geometry Control

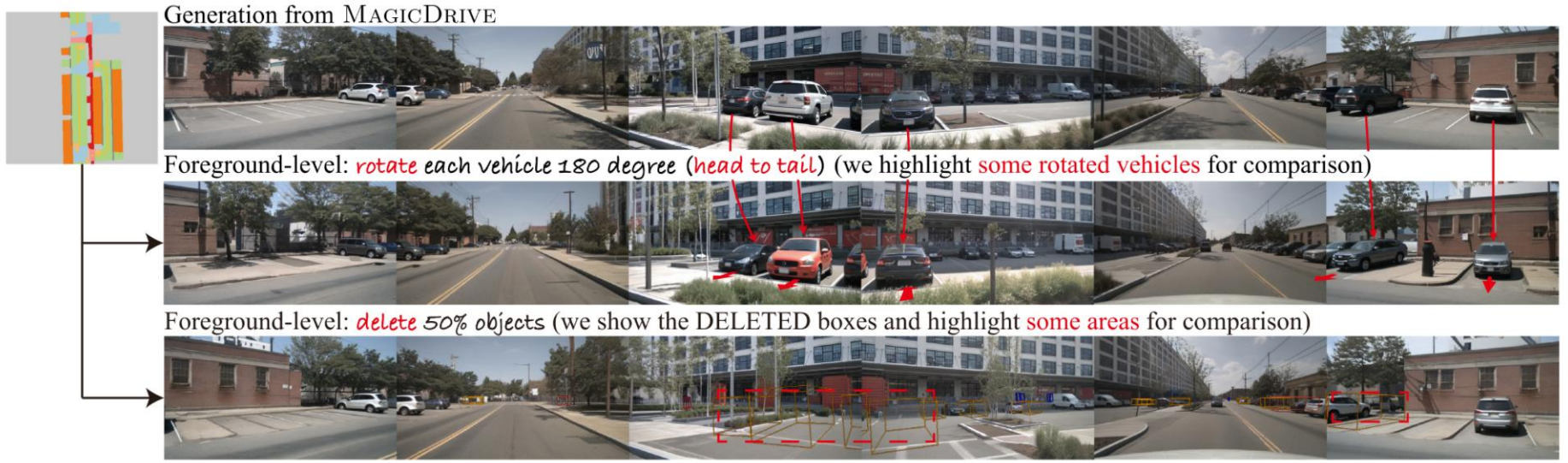
- Object fine-grained control

MagicDrive controls the position of objects precisely, while keeping other objects unchanged. Drag the slider to see how the vehicle (in the bounding box) moves from left to right.



- Multi-level controls

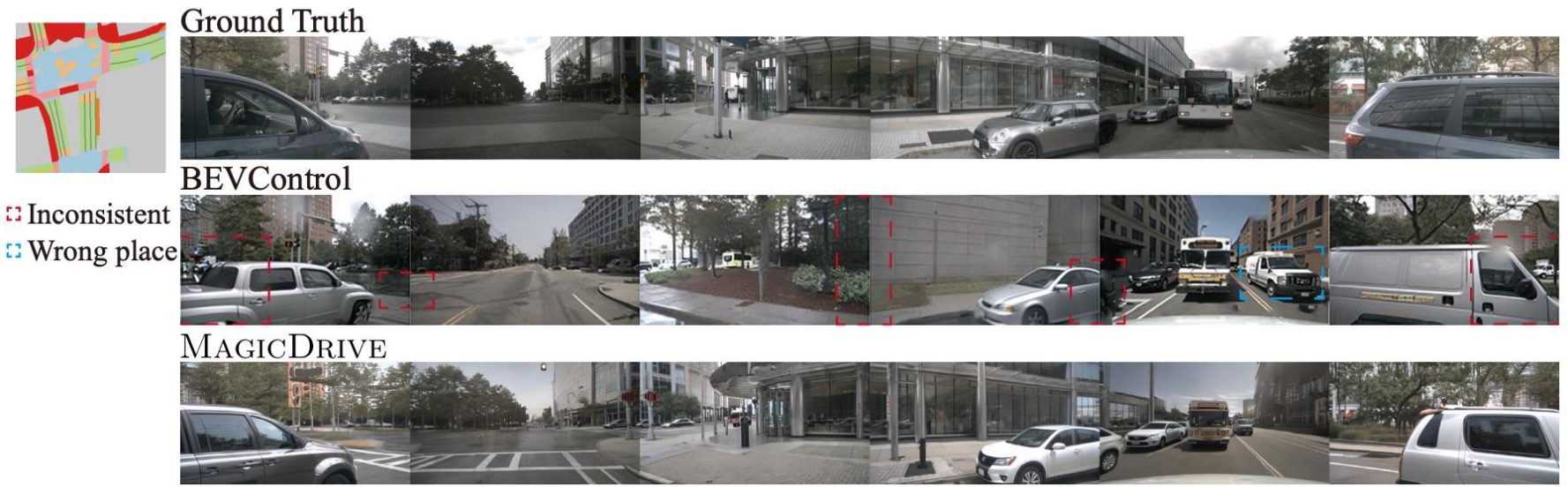
MagicDrive considers controls from road BEV map, object bounding box, camera pose, and textual description.



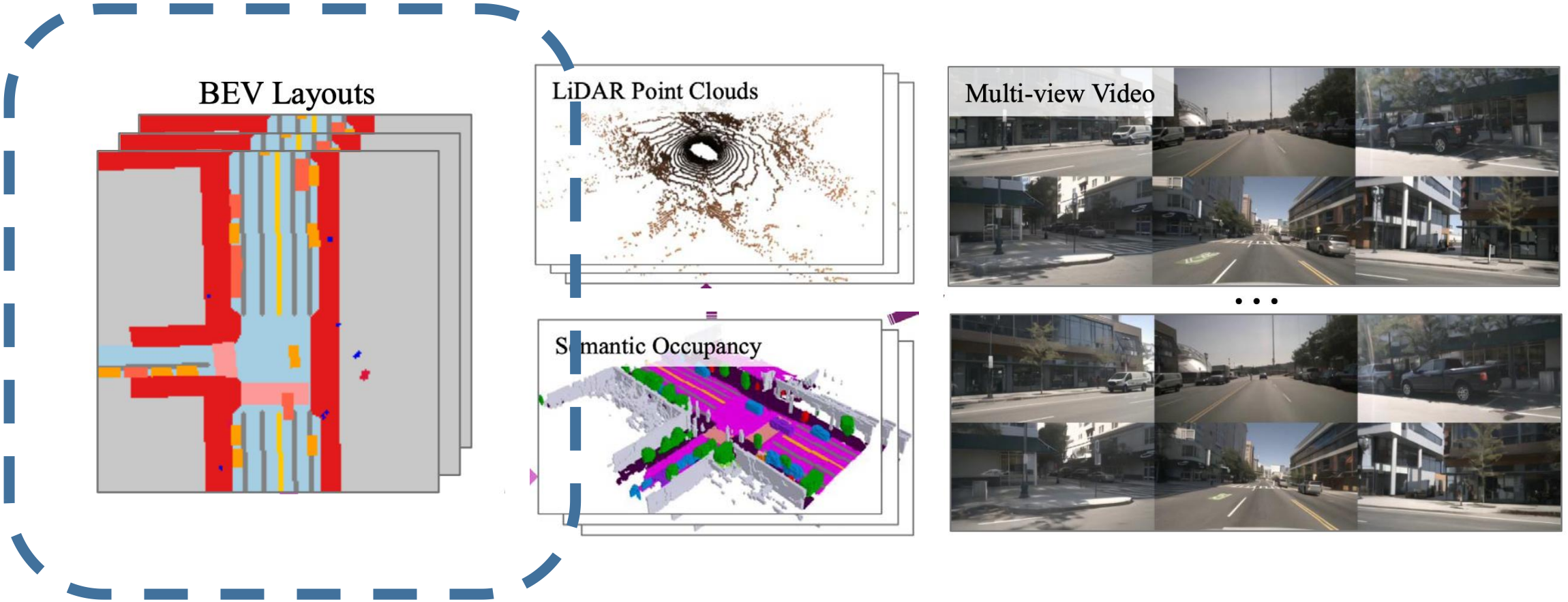
MagicDrive: Street View Generation with Diverse 3D Geometry Control

在 nuScenes 数据集上的实验对比显示，MAGICDRIVE 在街景生成任务中生成的图像保真度超越了所有基线方法。

Method	Synthesis resolution	FID↓	BEV segmentation		3D object detection	
			Road mIoU ↑	Vehicle mIoU ↑	mAP ↑	NDS ↑
Oracle	-	-	72.21	33.66	35.54	41.21
Oracle	224×400	-	72.19	33.61	23.54	31.08
BEVGen	224×400	25.54	50.20	5.89	-	-
BEVControl	-	24.85	*60.80*	26.80	-	-
MAGICDRIVE	224×400	**16.20**	**61.05**	*27.01*	12.30	23.32
MAGICDRIVE	272×736	*16.59*	54.24	**31.05**	**20.85**	**30.26**



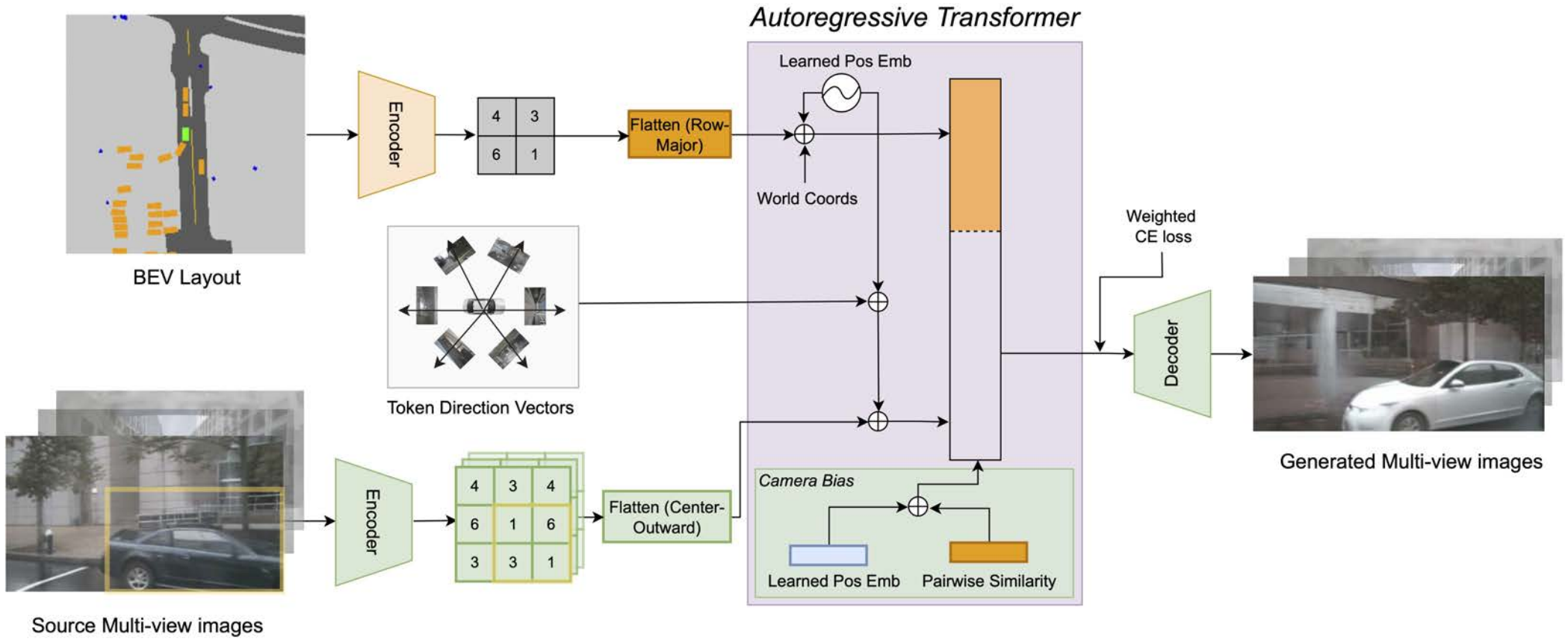
基于鸟瞰视图（BEV）的建模与控制



BEVGen: Street-View Image Generation from a Bird's-Eye View Layout

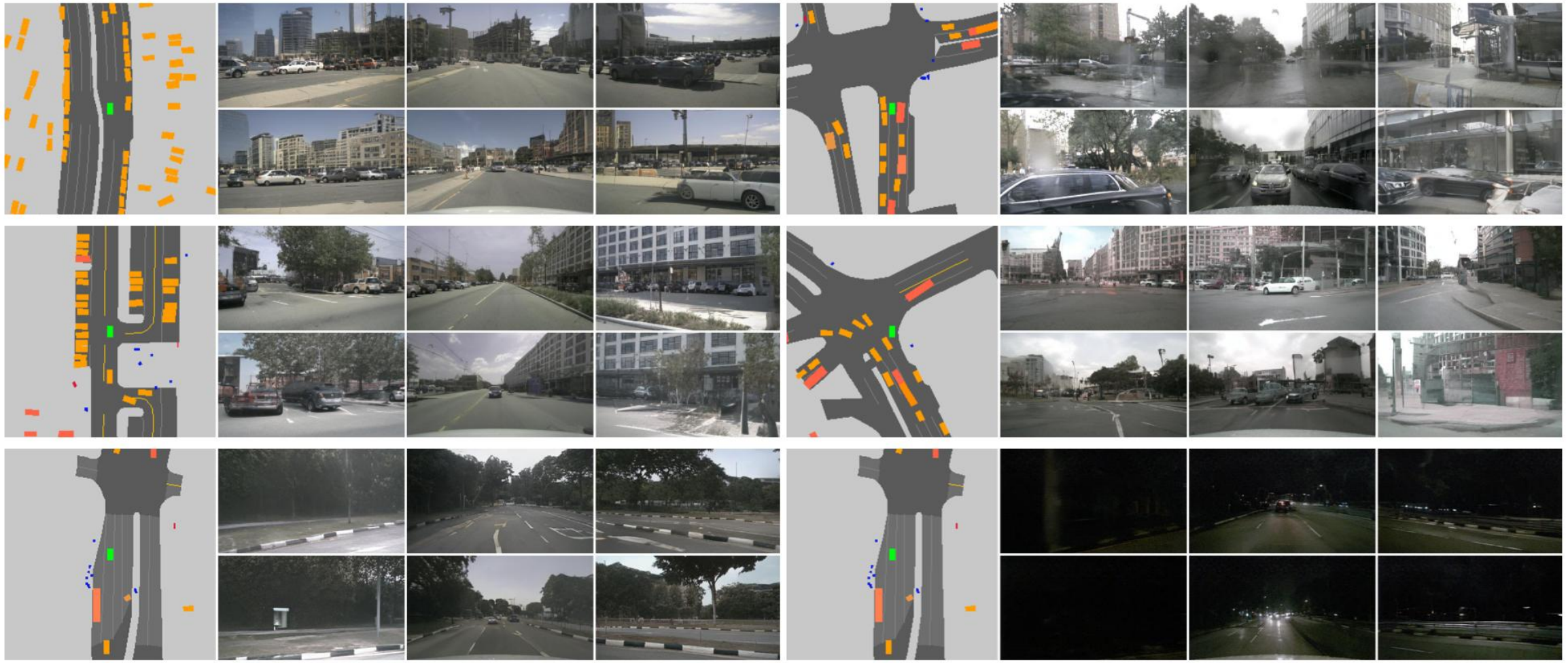
BEVGen 框架：

将鸟瞰图布局和多视角图像编码成离散表示，加入空间嵌入及相机偏置后经过自回归 Transformer 模型，用加权交叉熵进行训练，最终通过解码器生成图像。



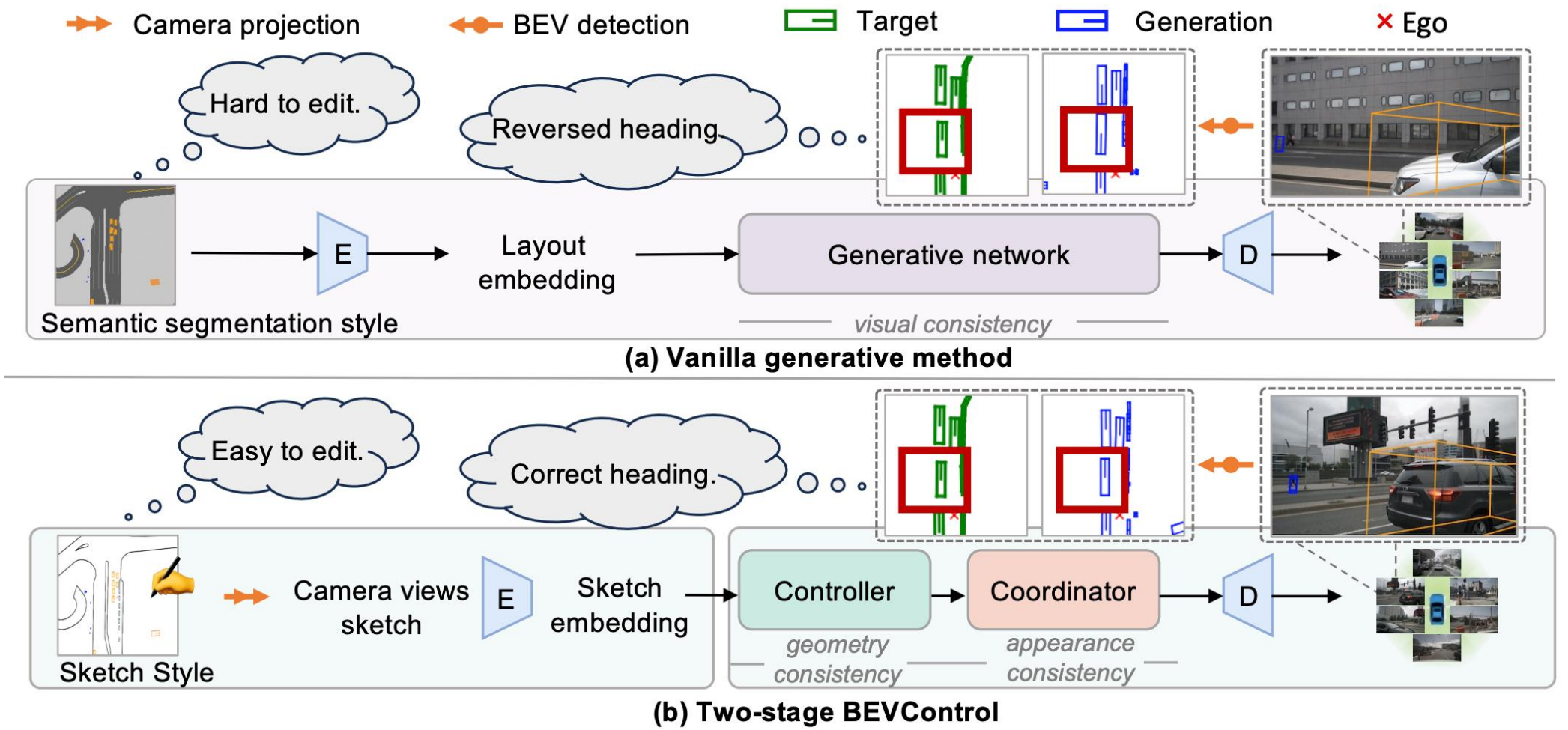
BEVGen: Street-View Image Generation from a Bird's-Eye View Layout

Synthesized multi-view images from BEVGen on nuScenes



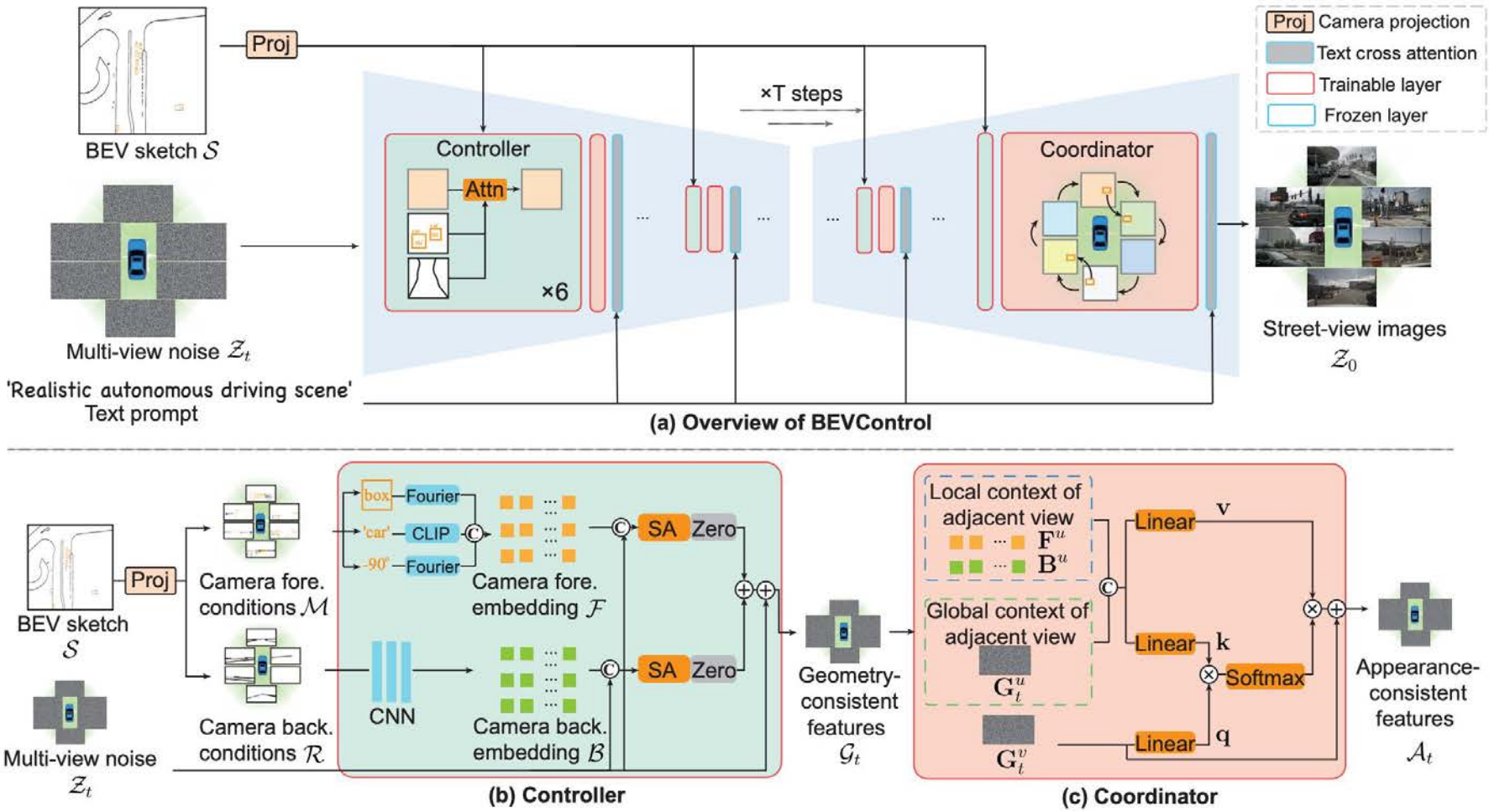
https://youtu.be/8AFKpCFwwOo

BEVControl: Accurately Controlling Street-view Elements with Multi-perspective Consistency via BEV Sketch Layout



Comparison between different generative networks hinted by Bird’ s Eye View (BEV) segmentation layout v.s. sketch layout.

BEVControl: Accurately Controlling Street-view Elements with Multi-perspective Consistency via BEV Sketch Layout



BEVControl 是一个基于 UNet 结构的生成网络，通过由控制器和协调器模块协同工作，从编辑友好的 BEV 草图、多视角噪声图像和文本提示中提取几何与外观一致的特征，从而生成多视角的街景图像。

BEVControl: Accurately Controlling Street-view Elements with Multi-perspective Consistency via BEV Sketch Layout

Method	Real & Diverse	Consistency	Foreground Control				Background Control	$S_{OCS}\uparrow$
	$S_{FID}\downarrow$	$S_{CLIP}\uparrow$	Detection			Segmentation	Segmentation	
			$S_{AP}\uparrow$	$S_{NDS}\uparrow$	$S_{AOE}\downarrow$	$S_{fIoU}\uparrow$	$S_{bIoU}\uparrow$	
Reference-score	0.01	87.96	36.04	44.10	0.42	34.83	74.33	5.00
BEVGen [24]	25.54	-	-	-	-	5.89	50.20	-
LayoutDiffusion [36]	29.64	79.80	3.68	14.68	1.31	15.51	35.31	2.16
GLIGEN [14]	31.34	78.80	15.42	22.35	1.22	22.02	38.12	2.55
BEVControl	**24.85** (↓ **6.49**)	**82.70** (↑ **3.9**)	**19.64** (↑ **4.22**)	**28.68** (↑ **6.33**)	**0.78** (↓ **0.44**)	**26.80** (↑ **4.78**)	**60.80** (↑ **22.68**)	**3.18** (↑ **0.63**)

Table 1. We compare BEVControl with state-of-the-art methods on the validation subset of nuScenes. The results measure the controlling power of different methods. ↓/↑ means a smaller/larger value of the metric represents a better performance.

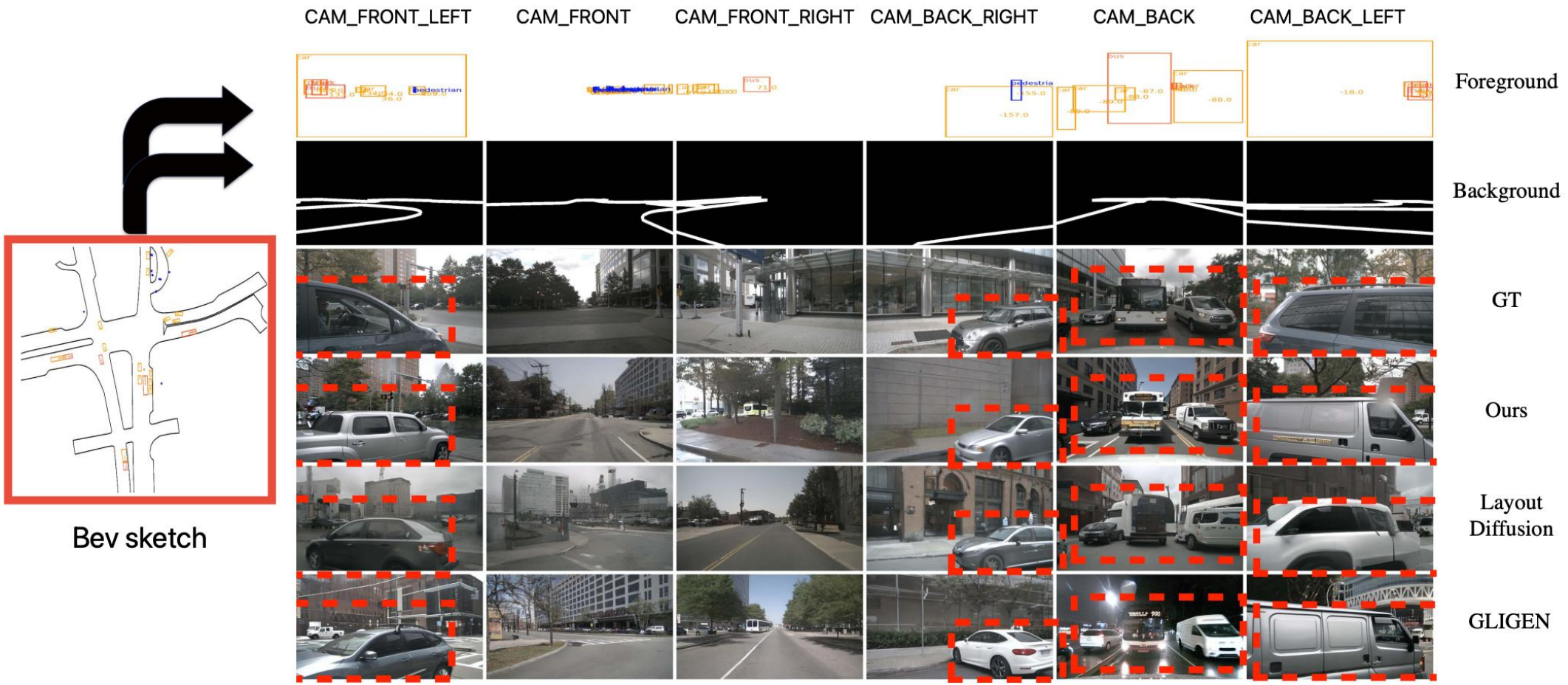
Controller	FC		BC	$S_{OCS}\uparrow$
	$S_{NDS}\uparrow$	$S_{fIoU}\uparrow$	$S_{bIoU}\uparrow$	
Reference-score	44.10	34.83	74.33	5.00
foreground	25.23	22.50	41.70	2.69
background	3.70	3.53	49.71	1.74
both w/o separation	26.87	23.78	52.30	2.90
both w/ separation	**28.68**	**26.80**	**60.80**	**3.18**

Table 2. Different strategies of using the foreground and background hints for controlling the visual elements. The evaluation metrics (i.e., FID, scores of foreground and background control) reported in this table are the same to those in Table 1.

Coordinator	$S_{CLIP}\uparrow$
Reference-score	87.96
w/o coordinator	79.50
w/ CV, w/o CE	82.30
w/ CVCE	**82.70**

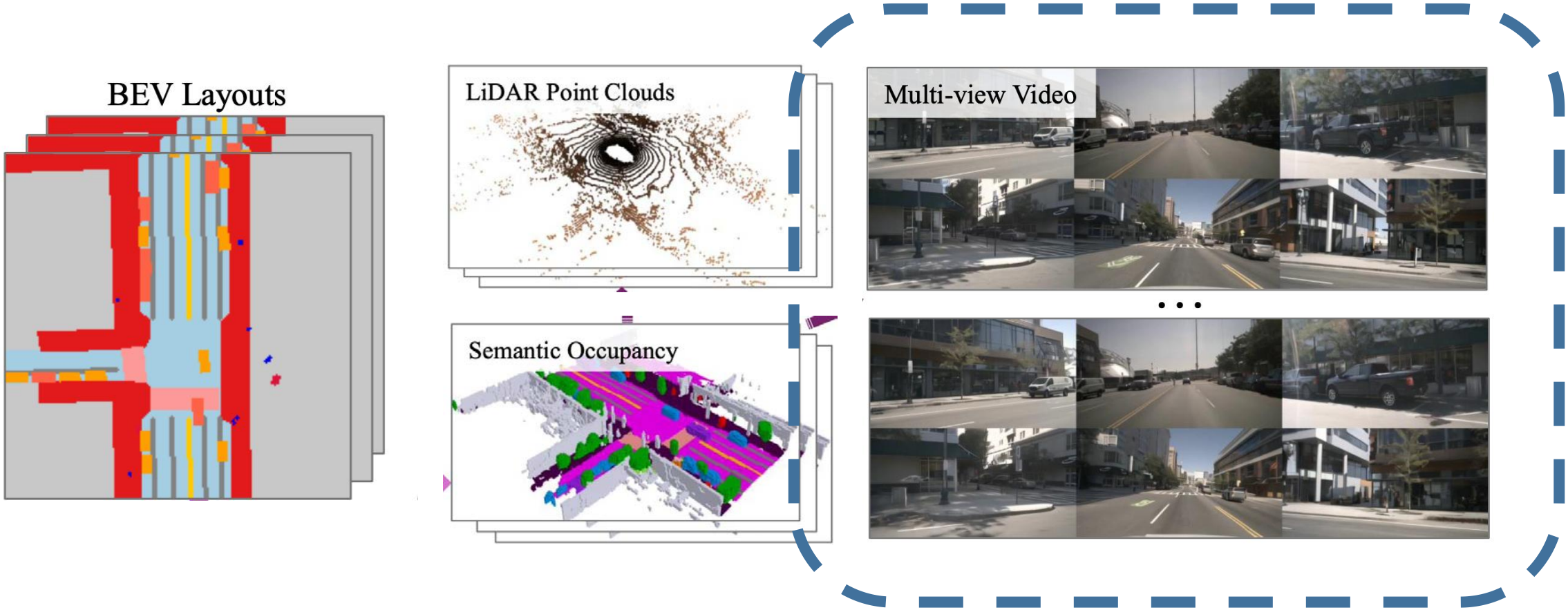
Table 3. Different strategies of using the coordinator for yielding the street-view images from different perspectives. We report the results as CLIP scores, which measure the visual consistency of the street-view images.

BEVControl: Accurately Controlling Street-view Elements with Multi-perspective Consistency via BEV Sketch Layout



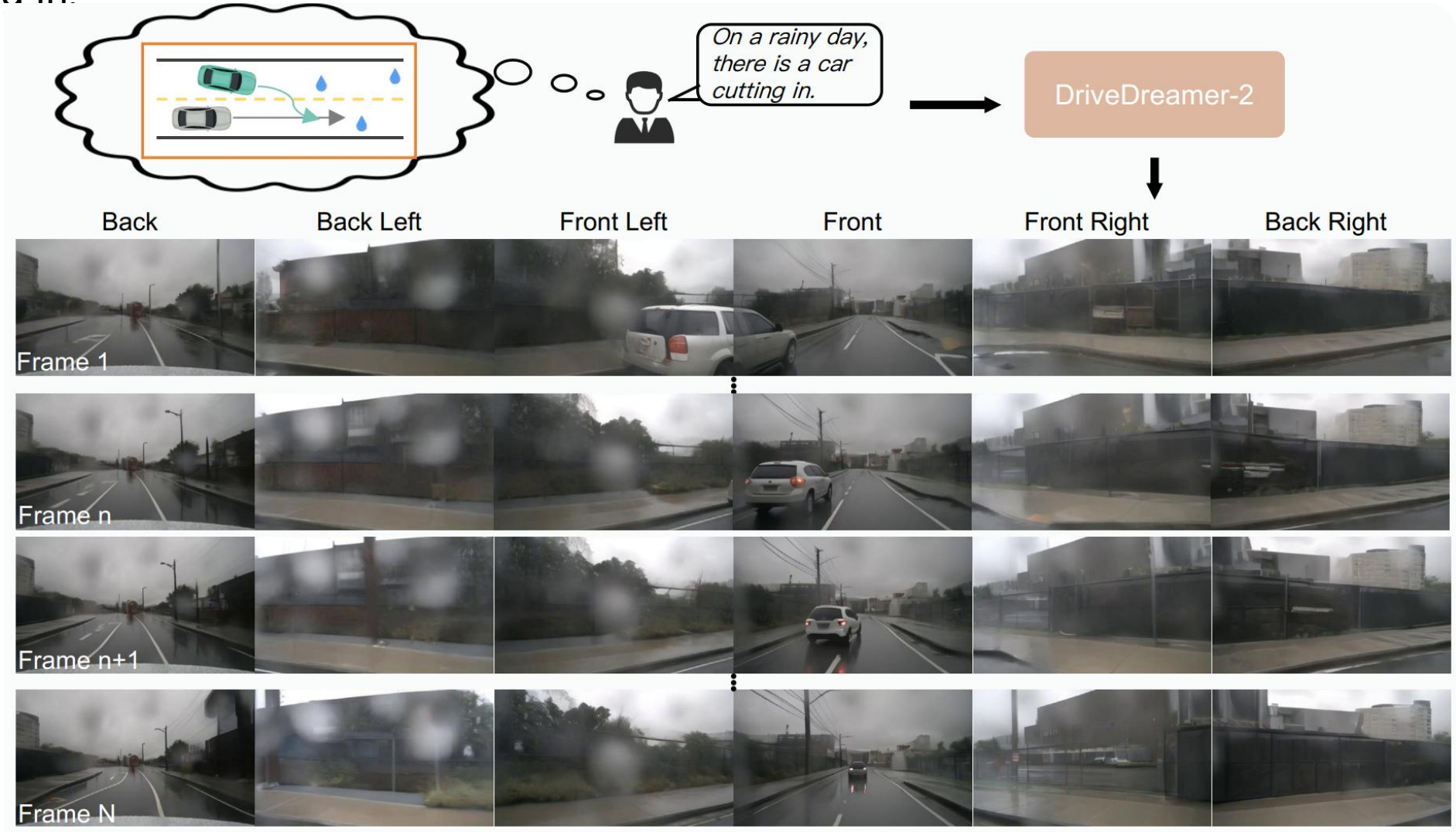
前景控制生成可视化显示，与其他方法相比，这个方法能更精确地生成与 bounding box sketch 条件匹配的物体，尤其在准确定位物体方向上表现更为出色。

生成式世界模型


BEV Layouts
LiDAR Point Clouds
Semantic Occupancy
Multi-view Video

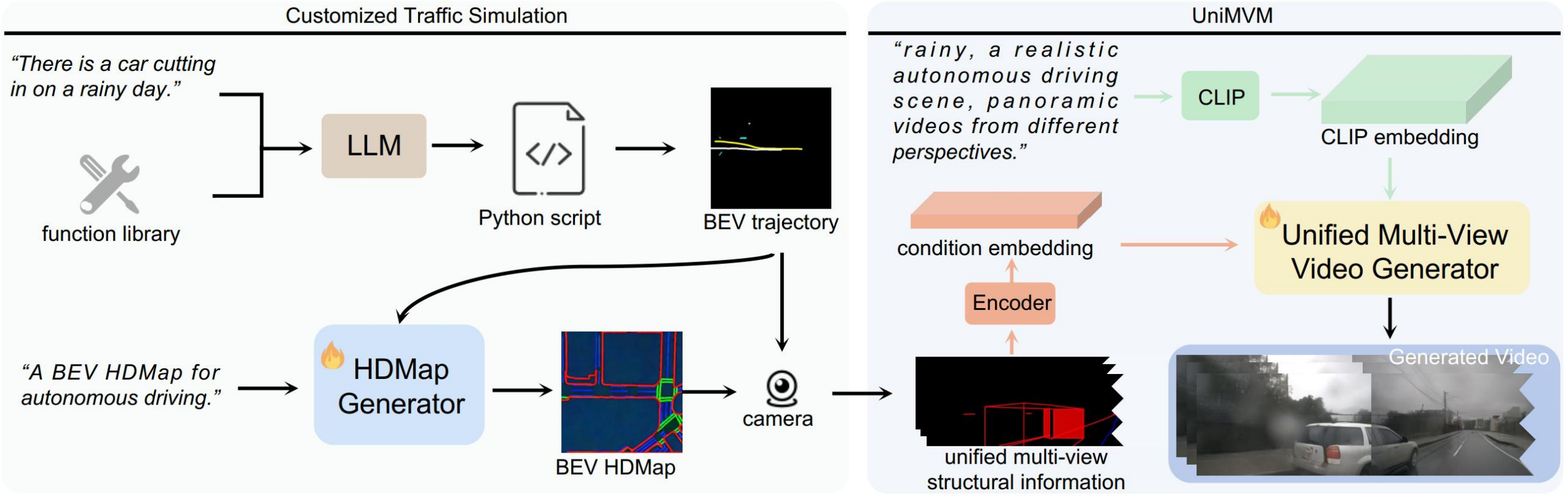
DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation

DriveDreamer-2 can produce multi-view driving videos based on user descriptions: "On a rainy day, there is a car cutting in."



DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation

The overall framework of DriveDreamer-2 involves initially generating agent trajectories according to the user query, followed by producing a realistic HDMap, and finally generating multi-view driving videos.



DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation

Daytime / rainy day / at night, a car abruptly cutting in from the right rear of ego-car.

Rainy day, the ego-car makes a left turn at the traffic signal, with vehicles behind proceeding straight through the intersection.

DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation

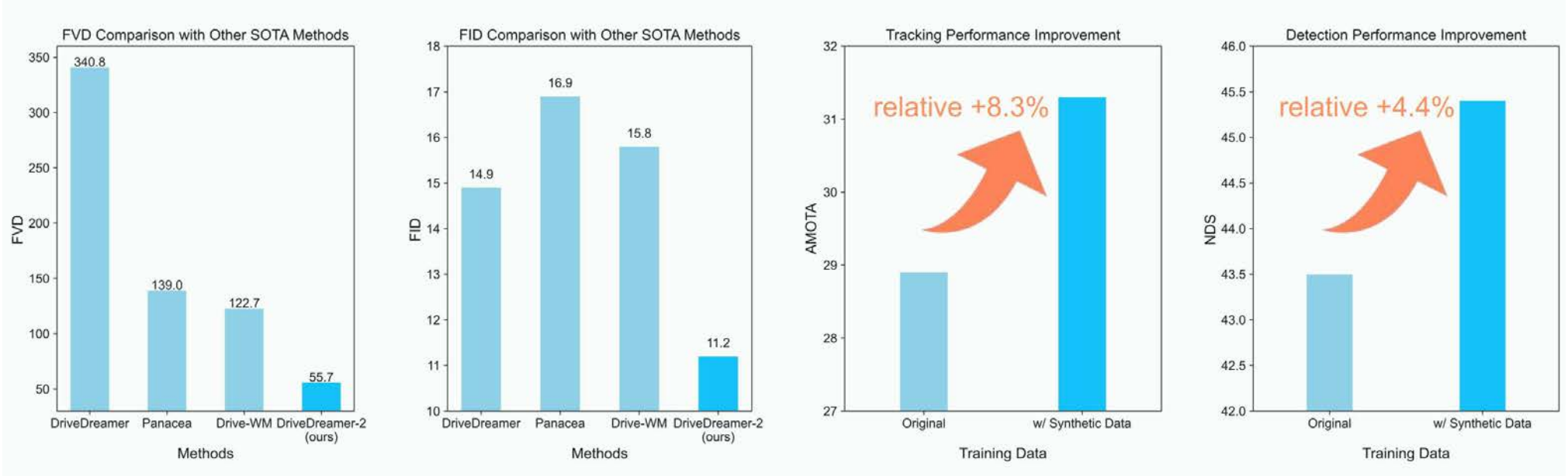
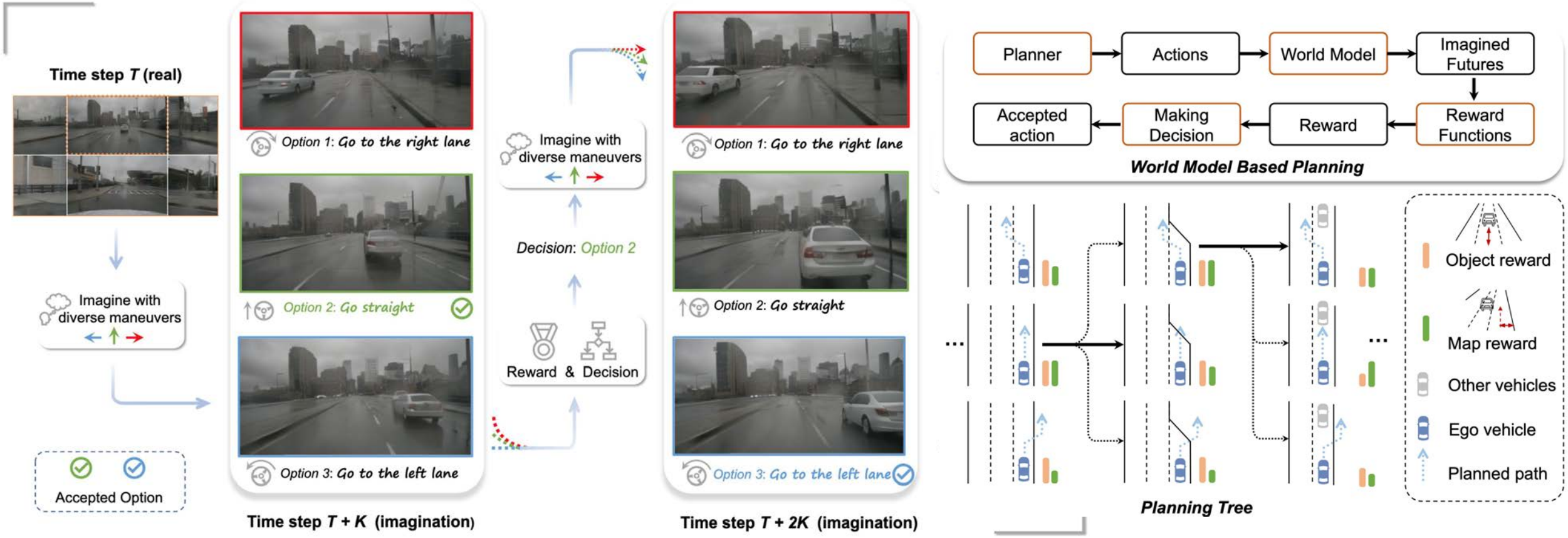


Table 1: Comparison of the generation quality on nuScenes validation set. † denotes that the corresponding conditions are generated.

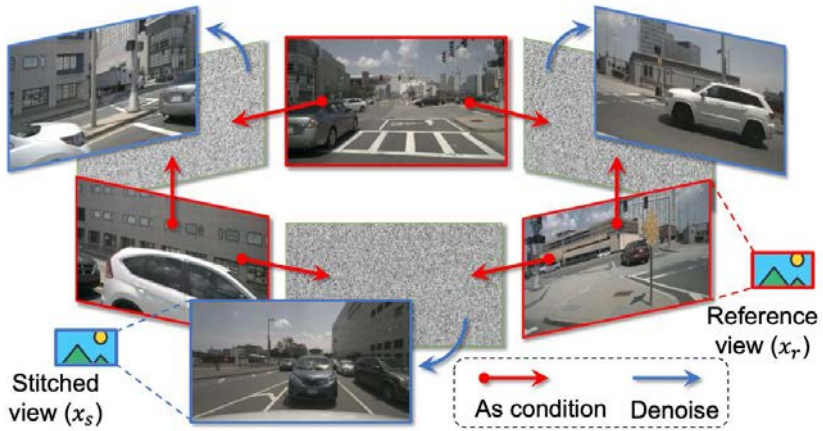
Method	Conditions	FID↓	FVD↓
DriveDreamer [59]	-	26.8	353.2
DriveDreamer-2	-	**25.0**	**105.1**
Drive-WM [62]	3-view videos†	**15.8**	122.7
DriveDreamer-2	1-view video	18.4	**74.9**
DriveDreamer [59]	1st-frame multi-view image	14.9	340.8
Drivingdiffusion [34]	1st-frame multi-view image†	15.8	332.0
Panacea [64]	1st-frame multi-view image†	16.9	139.0
DriveDreamer-2	1st-frame multi-view image	**11.2**	**55.7**

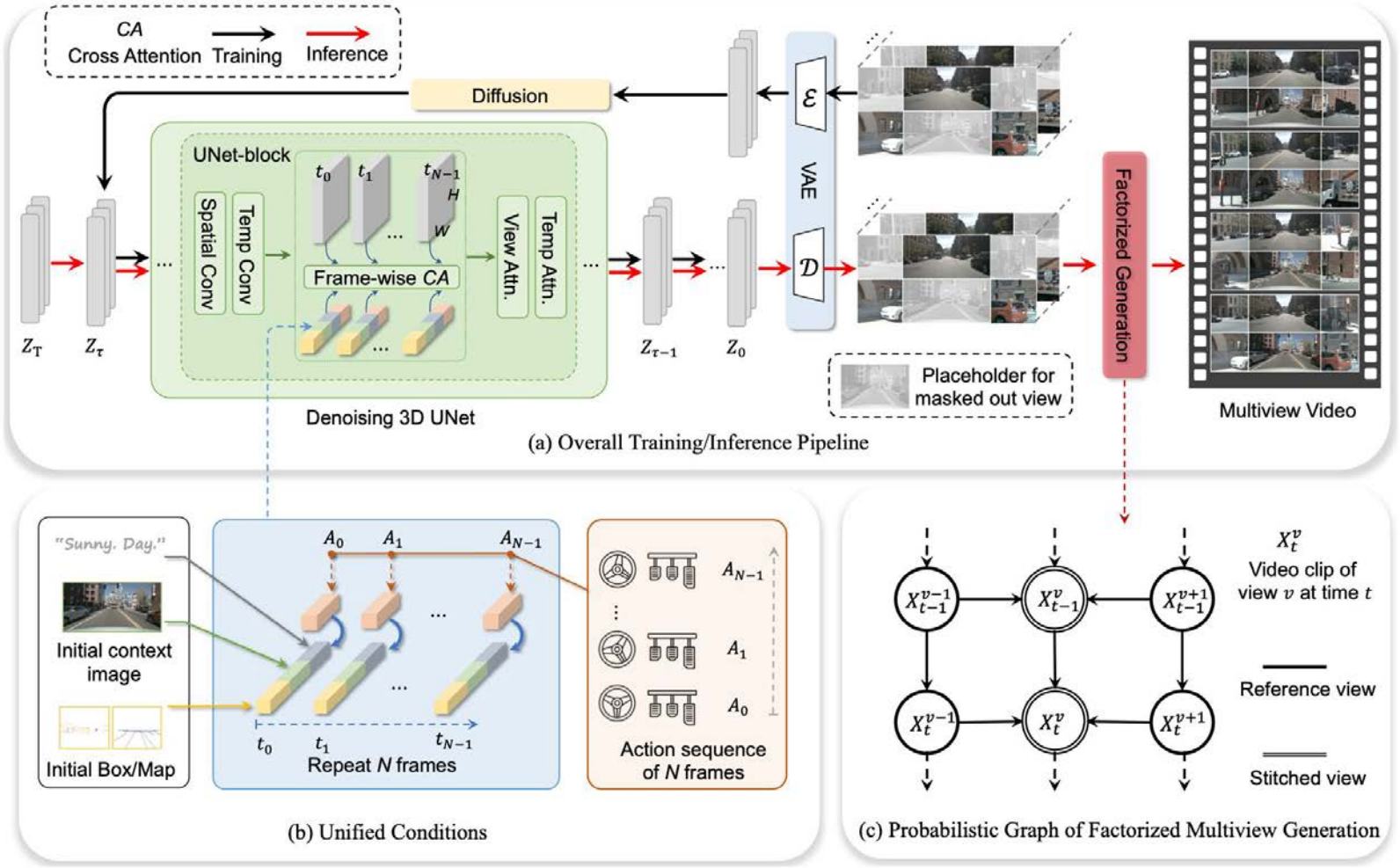
Drive-WM: Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving

通过世界模型进行多视角视觉预测和规划。在时间步骤 T，世界模型设想了 T+K 处的多种未来情况，并发现在 T 处继续直行是安全的。然后，根据时间步骤 T+2K 的想象，模型意识到自车将离前车太近，因此它决定换到左车道以进行安全超车。



Drive-WM: Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving





Framework

（a）说明了所提方法的训练和推理流程。

（b）可视化了用于控制多视图视频生成的统一条件。

（c）表示分解多视图生成的概率图。它将（a）中的 3 视图输出作为输入来生成其他视图，从而增强了多视图一致性。

Drive-WM: Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving

Drive-WM: Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving

Method	Multi-view	Video	FID↓	FVD↓
BEVGen [53]	✓		25.54	-
BEVControl [69]	✓		24.85	-
MagicDrive [17]	✓		16.20	-
Ours	✓		**12.99**	-
DriveGAN [31]		✓	73.4	502.3
DriveDreamer [63]		✓	52.6	452.0
Ours	✓	✓	**15.8**	**122.7**

(a) **Generation quality**.

Method	mAP_{obj} ↑	mAP_{map} ↑	$mIoU_{fg}$ ↑	$mIoU_{bg}$ ↑
GT	37.78	59.30	36.08	72.36
BEVGen [53]	-	-	5.89	50.20
LayoutDiffusion [71]	3.68	-	15.51	35.31
GLIGEN [36]	15.42	-	22.02	38.12
BEVControl [69]	19.64	-	26.80	60.80
MagicDrive [17]	12.30	-	27.01	61.05
Ours	**20.66**	**37.68**	**27.19**	**65.07**

(b) **Generation controllability**.

Table 1. **Multi-view video generation performance on nuScenes.** For each task, we test the corresponding models trained on the nuScenes training set. Our Drive-WM surpasses all other methods in both quality and controllability evaluation.

Method	L2 (m) ↓				Collision (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD (GT cmd)	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
VAD (rand cmd)	0.51	0.97	1.57	1.02	0.34	0.74	1.72	0.93
Ours	**0.43**	**0.77**	**1.20**	**0.80**	**0.10**	**0.21**	**0.48**	**0.26**

Table 3. **Planning performance on nuScenes.** Instead of using the ground truth driving command, we use our tree-based planning to select the best out of three commands.

Map Reward	Object Reward	L2 (m) ↓				Collision (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
		0.51	0.97	1.57	1.02	0.34	0.74	1.72	0.93
✓		0.45	0.82	1.29	0.85	0.12	0.33	0.72	0.39
	✓	0.43	0.77	1.20	0.80	0.12	0.21	0.48	0.27
✓	✓	**0.43**	**0.77**	**1.20**	**0.80**	**0.10**	**0.21**	**0.48**	**0.26**

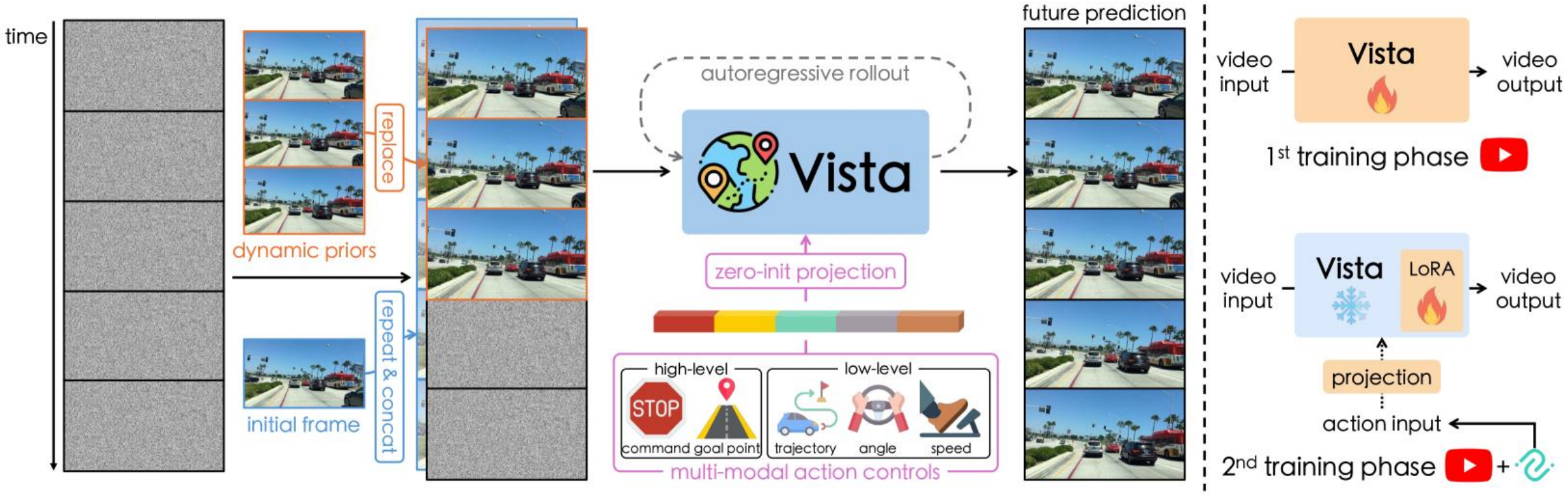
Table 4. **Image-based reward function design.** We use two sub-rewards, map reward and object reward.

Vista: A generalizable driving world model with high fidelity and versatile controllability.

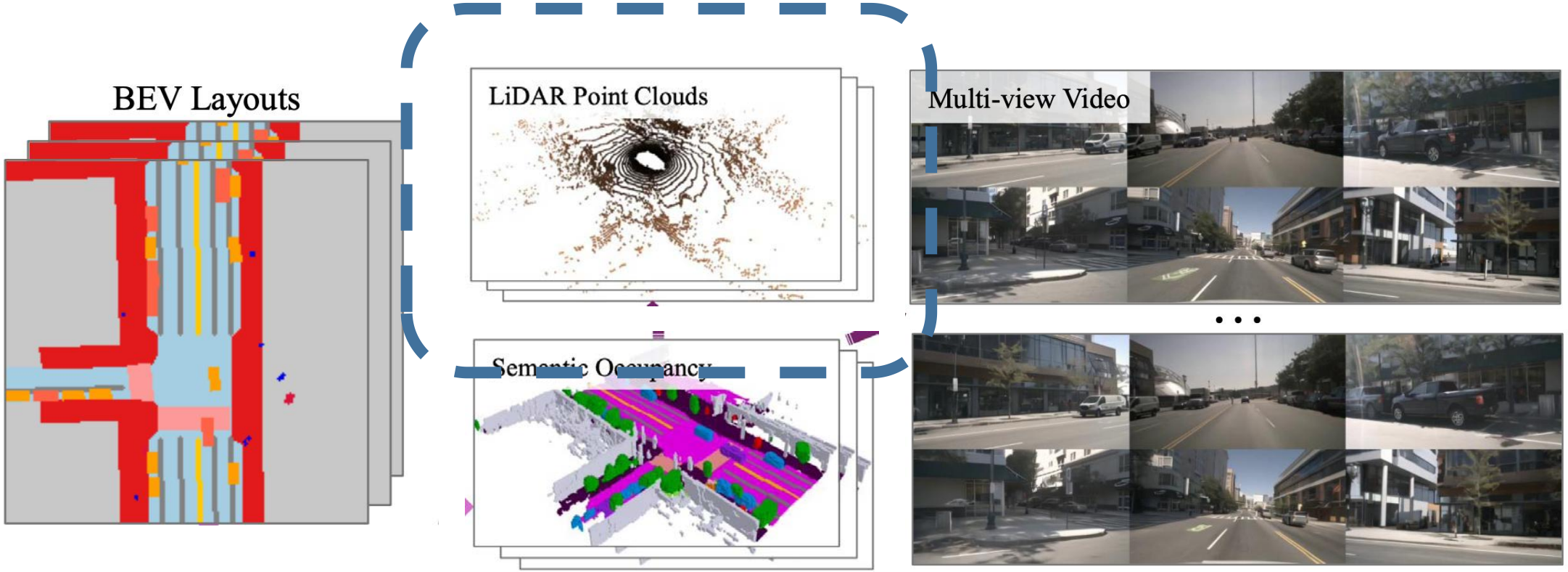
Vista: A generalizable driving world model with high fidelity and versatile controllability.

Vista pipeline

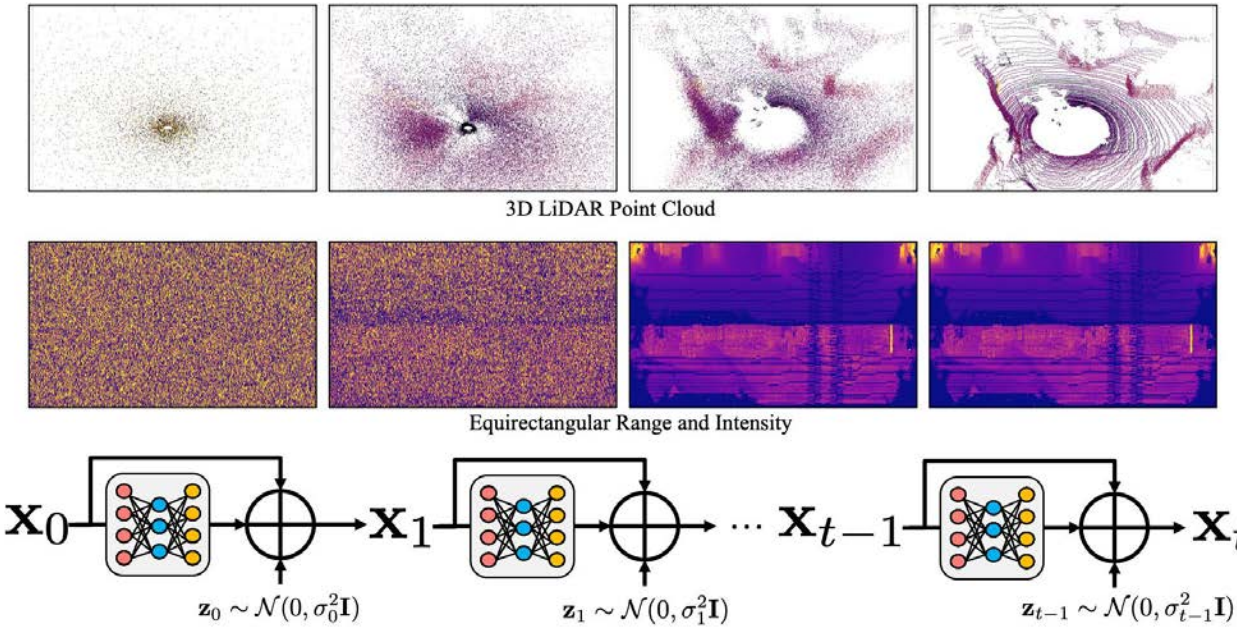
利用 latent replacement 吸收更多关于 future dynamics 的 priors，实现不同 actions 控制下通过 autoregressive rollouts 的长时预测；其训练分为两个阶段，第二阶段冻结预训练权重专注于 action controls 的学习

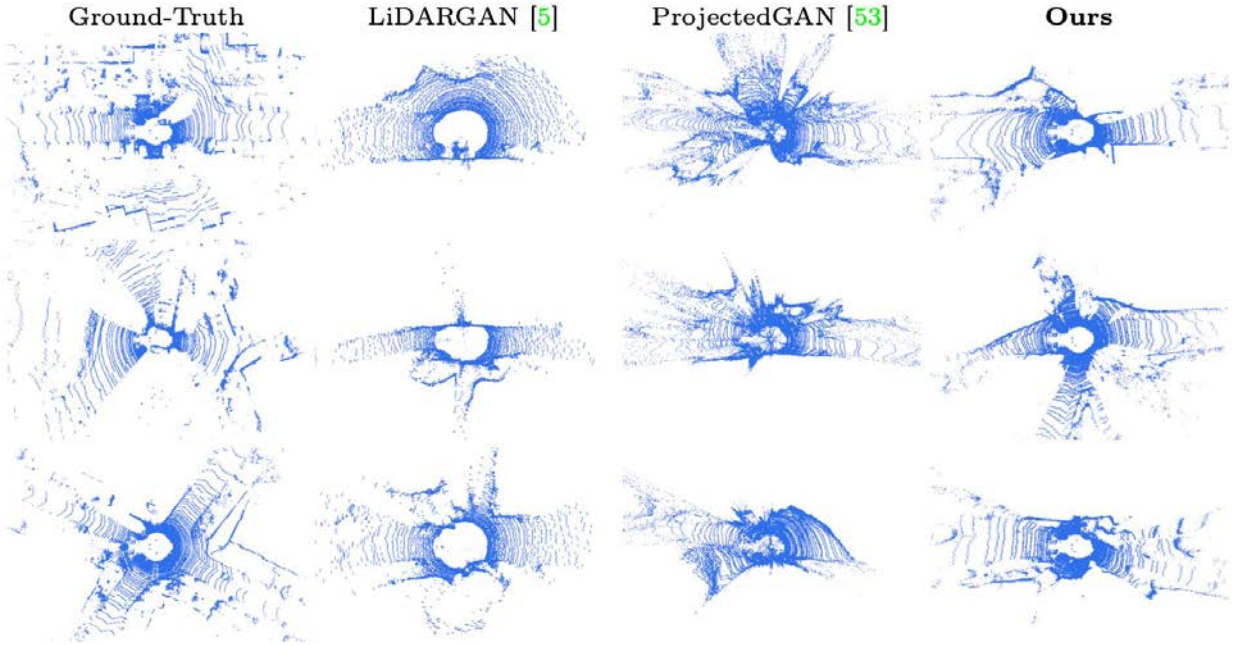


基于LiDAR数据的世界模型构建



LiDARGen: Learning to Generate Realistic LiDAR Point Clouds





	$\text{MMD}_{\text{BEV}} \downarrow$	$\text{FID}_{\text{range}} \downarrow$	$\text{JSD}_{\text{BEV}} \downarrow$
LiDAR GAN [5]	3.06×10^{-3}	3003.8	–
LiDAR VAE [5]	1.00×10^{-3}	2261.5	0.161
Projected GAN [53]	**3.47×10^{-4}**	**2117.2**	**0.085**
Ours	**3.87×10^{-4}**	**2040.1**	**0.067**

Table 3: LiDAR Densification.

	MAE $\downarrow$
PUNet [80]	6.88
NN	2.18
Ours	**1.23**

Table 4: Ablation Study

Coord-aware	CircConv	FRD	MMD
No	No	2422.3	7.60×10^{-4}
Yes	No	2251.1	3.94×10^{-4}
Yes	Yes	**2040.1**	**3.87×10^{-4}**

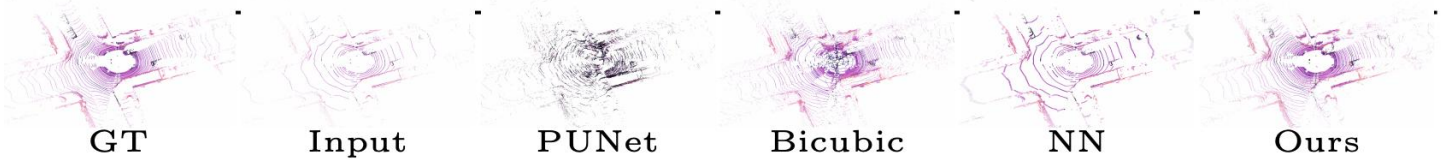
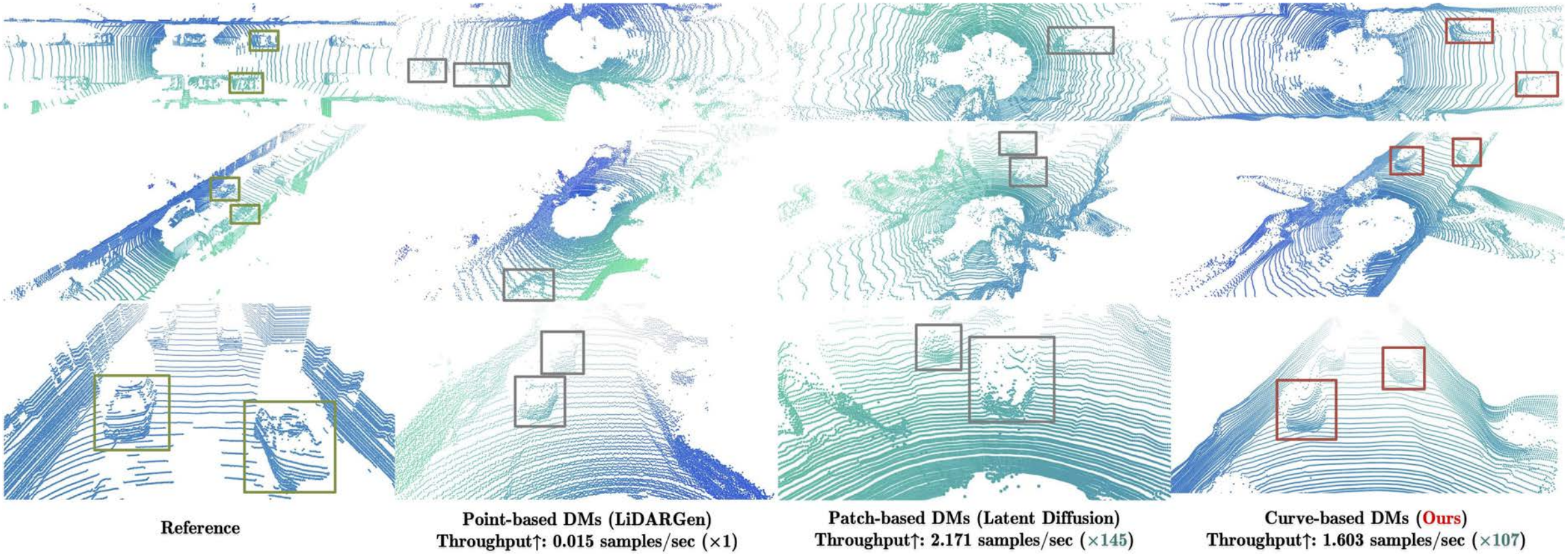


Fig. 6: Densification Results.

LiDAR Diffusion : Towards Realistic Scene Generation with LiDAR Diffusion Models

LiDAR Diffusion allows controllability of semantic maps, cameras, bounding boxes, text, etc.



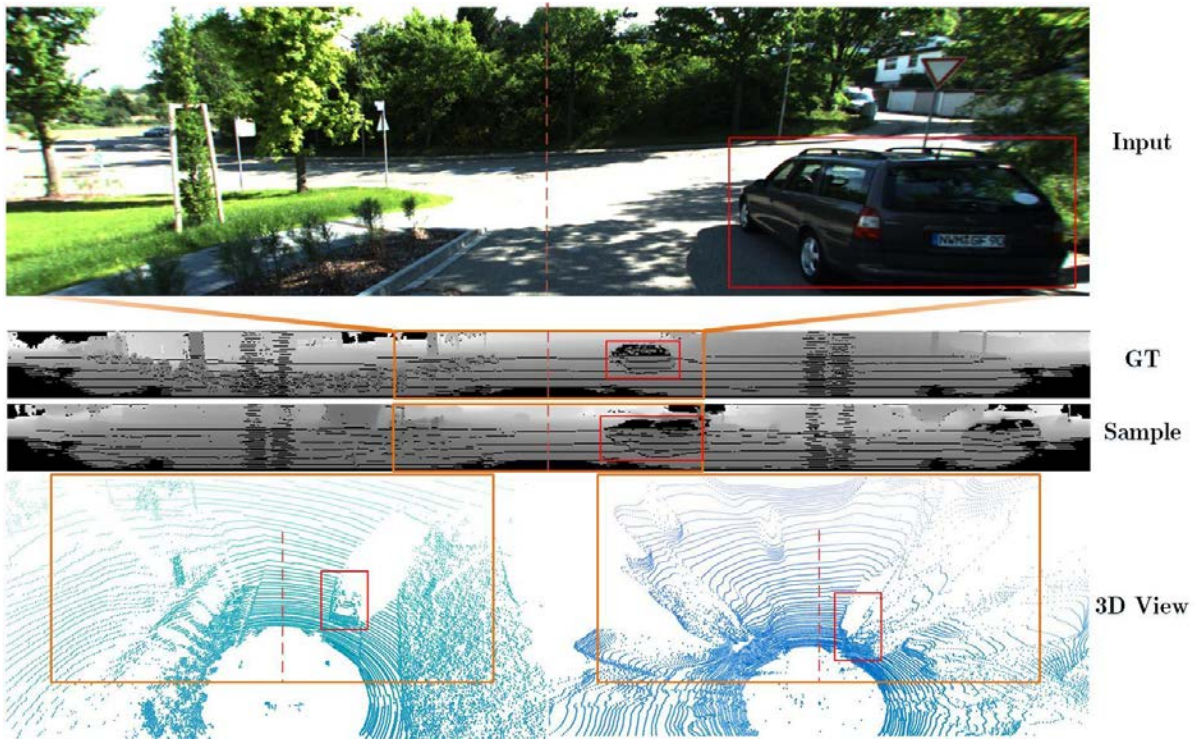
LiDAR Diffusion : Towards Realistic Scene Generation with LiDAR Diffusion Models

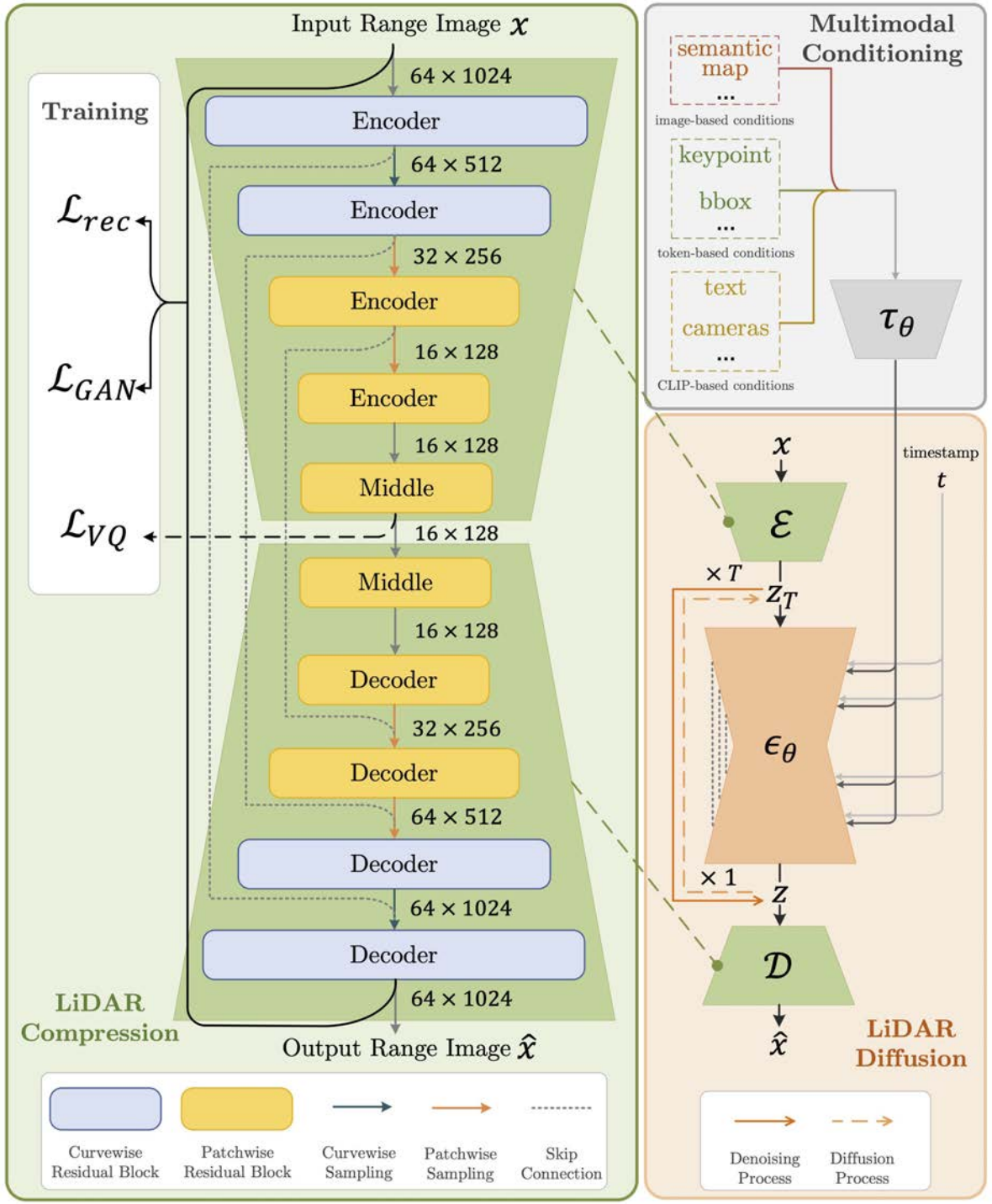
对于64-beam数据的 LiDMs，总体由三部分构成：

(a) LiDAR Compression：将密集的64-beam激光雷达数据压缩成紧凑的表示；

(b) Multimodal Conditioning：融合其他模态（如图像、文本等）信息，为生成过程提供上下文指导；

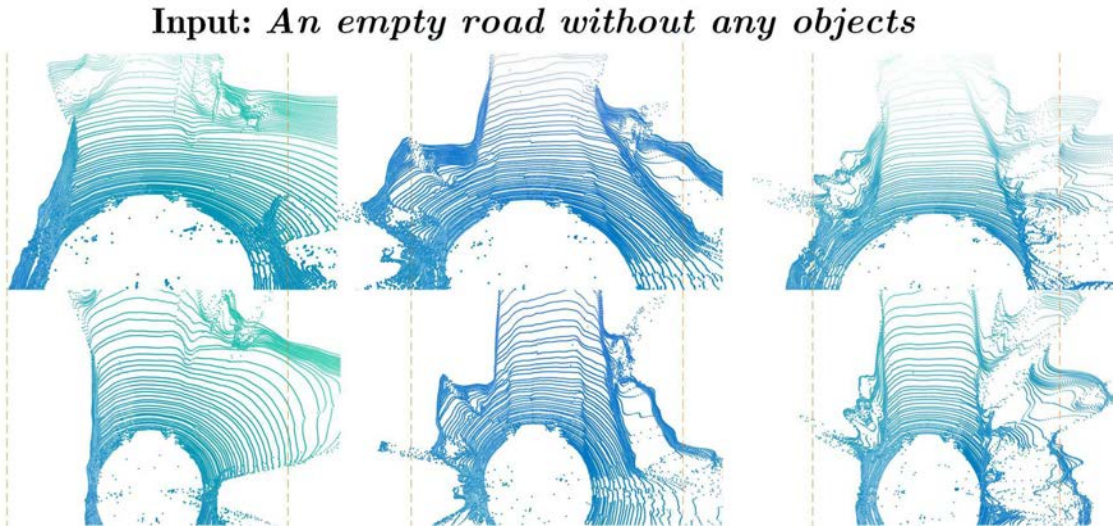
(c) LiDAR Diffusion：利用扩散模型对激光雷达数据进行逐步去噪和生成，实现高质量的点云重建。

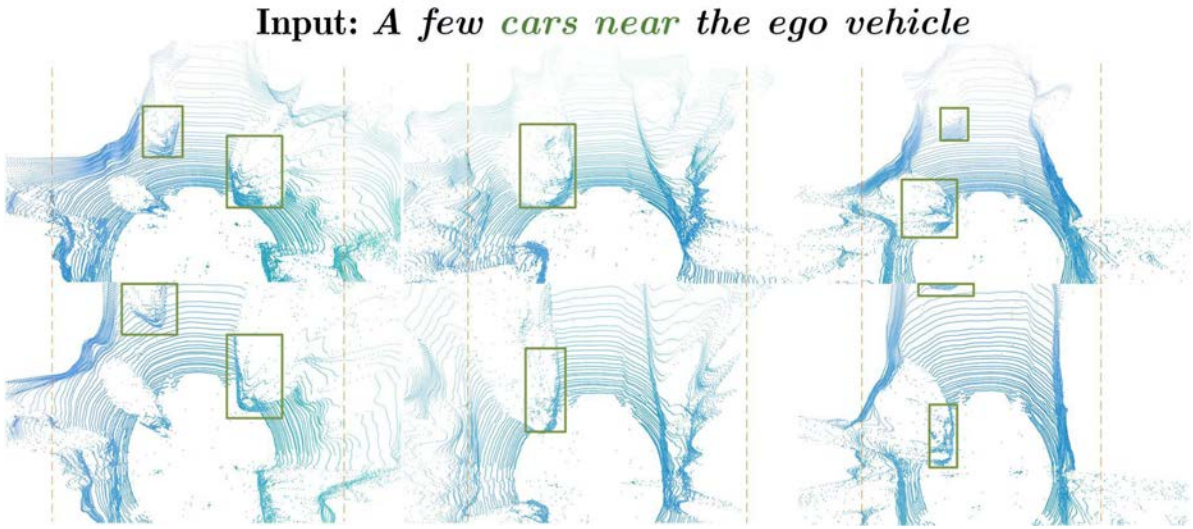




LiDAR Diffusion : Towards Realistic Scene Generation with LiDAR Diffusion Models

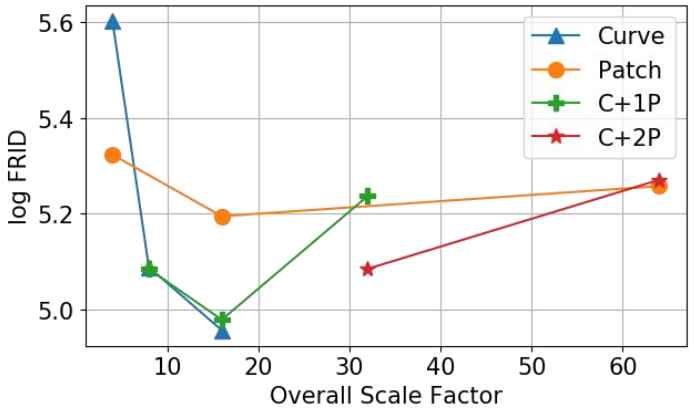
Zero-Shot Text-to-LiDAR

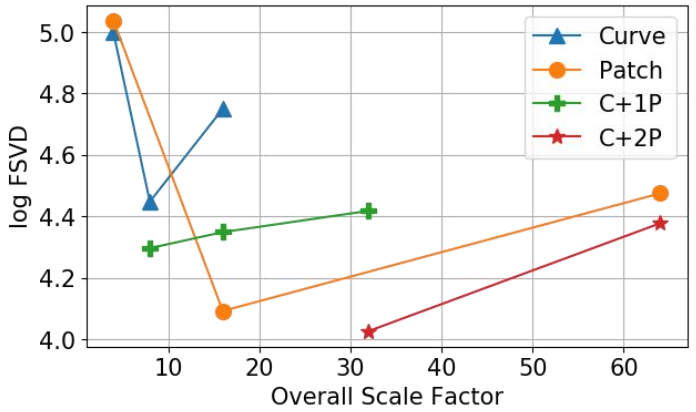




LiDAR Synthesis Quality with Different Scaling Factors

表示整体尺度因子（fc × fp）与采样质量（FRID & FSVD）之间的关系，并在 KITTI-360 数据集上比较了不同的编码策略：单独的 Curve 编码、Patch 编码，以及 Curve 编码结合一阶段（C+1P）或两阶段（C+2P）的 Patch 编码。





LiDAR Diffusion : Towards Realistic Scene Generation with LiDAR Diffusion Models

Method	Perceptual			Statistical	
	FRID ↓	FSVD ↓	FPVD ↓	JSD ↓	MMD ↓ (10^{-4})
Noise	3277	497.1	336.2	0.360	32.09
LiDARGAN [7]	1222	183.4	168.1	0.272	4.74
LiDARVAE [7]	199.1	129.9	105.8	0.237	7.07
ProjectedGAN [54]	149.7	44.7	33.4	0.188	2.88
UltraLiDAR [67]	370.0	72.1	66.6	0.747	17.12
LiDARGen [75] (1160s)	129.1	39.2	33.4	**0.188**	**2.88**
LiDARGen [75] (50s)	2051	480.6	400.7	0.506	9.91
LDM [51] (50s)†	199.5	70.7	61.9	0.236	5.06
LiDM (ours, 50s)†	158.8	53.7	42.7	0.213	4.46
Δ *Improv.*	**20.4%**	**24.0%**	**31.0%**	**9.7%**	**11.9%**
LiDM (ours, 50s)	**125.1**	**38.8**	**29.0**	0.211	3.84
Δ *Improv.*	**37.3%**	**45.1%**	**53.2%**	**10.6%**	**24.1%**

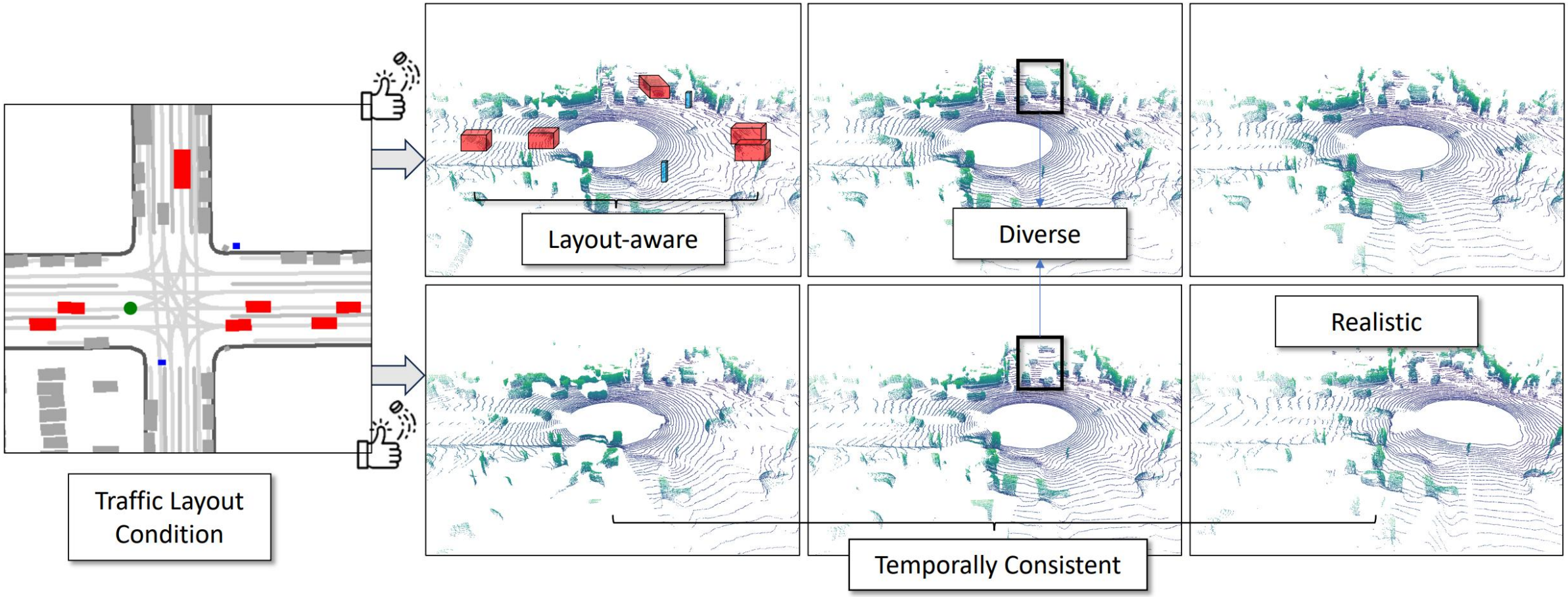
这部分内容比较了无条件 LiDAR 场景生成方法在 64-beam (KITTI-360) 数据集上的表现。其中 “↓” 表示数值越低越好，N-s 表示推理时的采样步数；Δ Improv. 表示相对提升；带 † 的说明表明在其他设置完全相同的前提下

该部分比较了条件式 LiDAR 场景生成方法的表现，分别在 SemanticKITTI 上进行了 Semantic-Map-to-LiDAR 实验和在 KITTI-360 上进行了 Camera-to-LiDAR 实验。

Method	Semantic-Map-to-LiDAR [5]					Camera-to-LiDAR [37]				
	FRID ↓	FSVD ↓	FPVD ↓	JSD ↓	MMD ↓ (10^{-4})	FRID ↓	FSVD ↓	FPVD ↓	JSD ↓	MMD ↓ (10^{-4})
LiDARGen [75]	42.5	31.7	30.1	0.130	5.18	-	-	-	-	-
Latent Diffusion [51]	24.0	21.3	20.3	0.088	3.73	50.2	35.9	26.5	0.256	3.80
LiDAR Diffusion (ours)	**22.9**	**20.2**	**17.7**	**0.072**	**3.16**	**44.9**	**32.5**	**25.8**	**0.205**	**3.69**

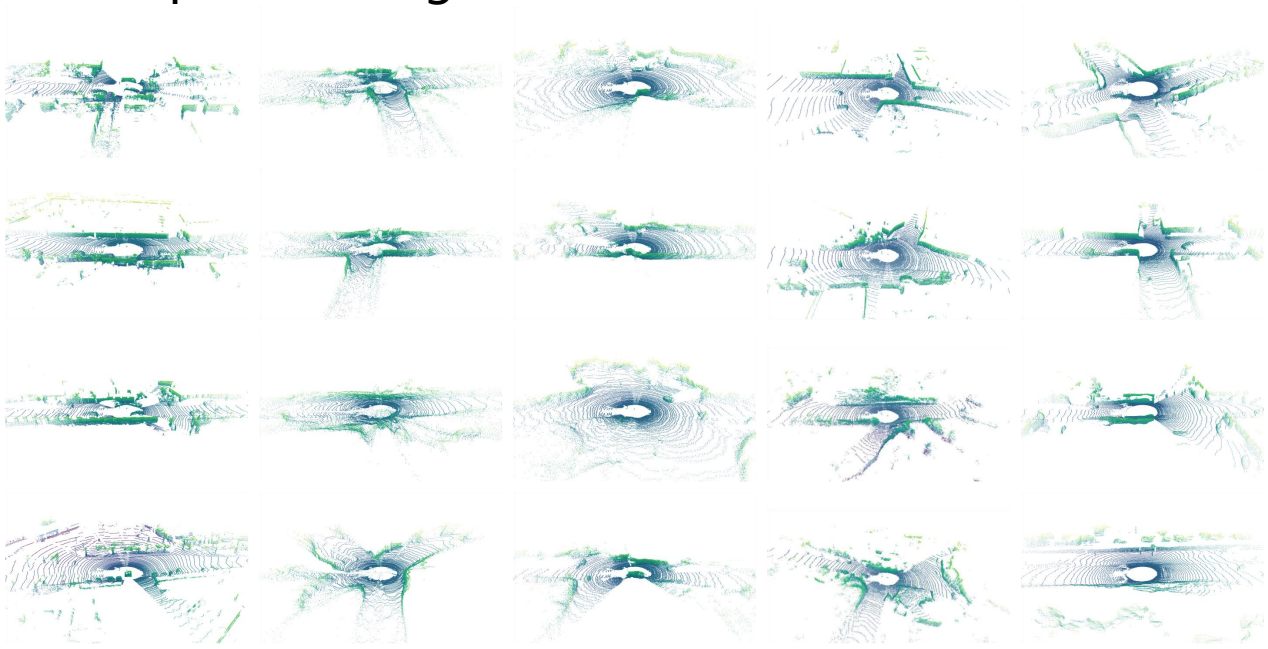
LidarDM: Generative LiDAR Simulation in a Generated World

LidarDM can generate LiDAR videos that are realistic, layout-conditioning, physically plausible, diverse, and temporally coherent, as shown by 2 diverse videos from one map condition.



LidarDM: Generative LiDAR Simulation in a Generated World

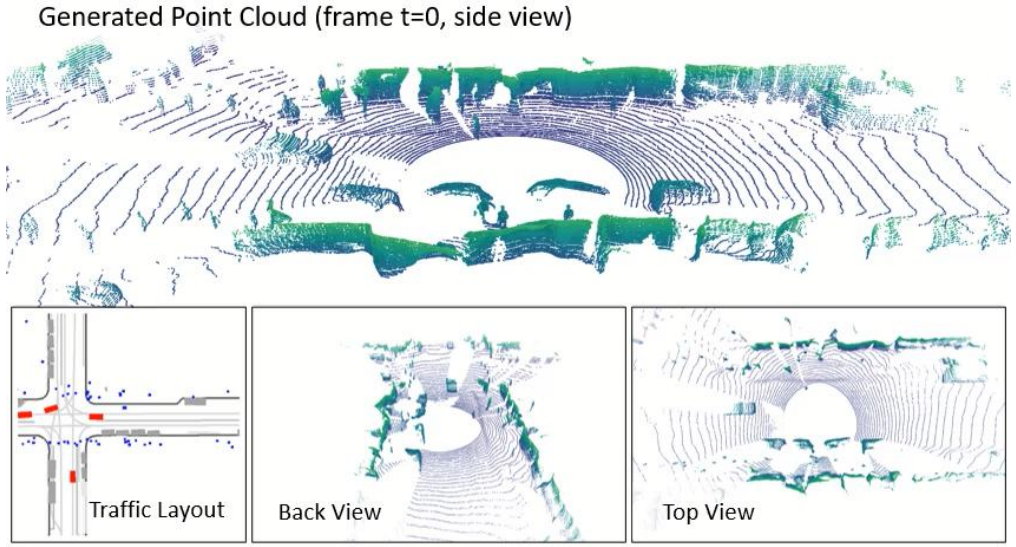
Competitive Single-Frame LiDAR Generation



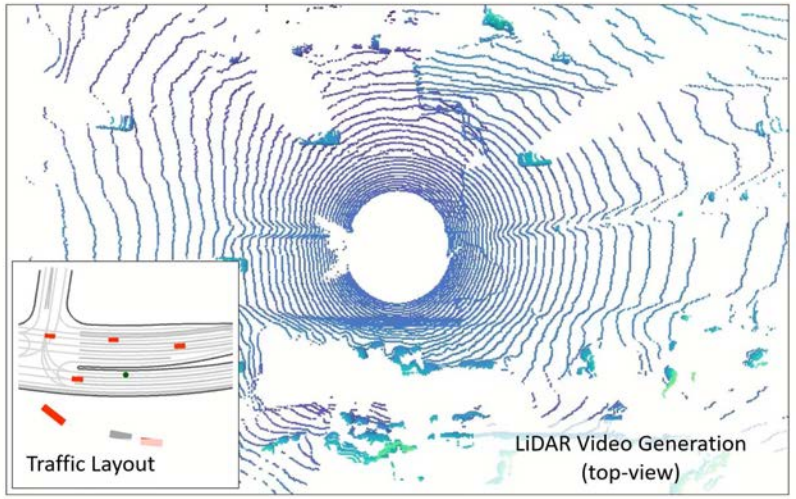
Visualizations of Agent Motions

Consistent Multi-Frame LiDAR Generation

Short Sequence Generation



Long Sequence Generation



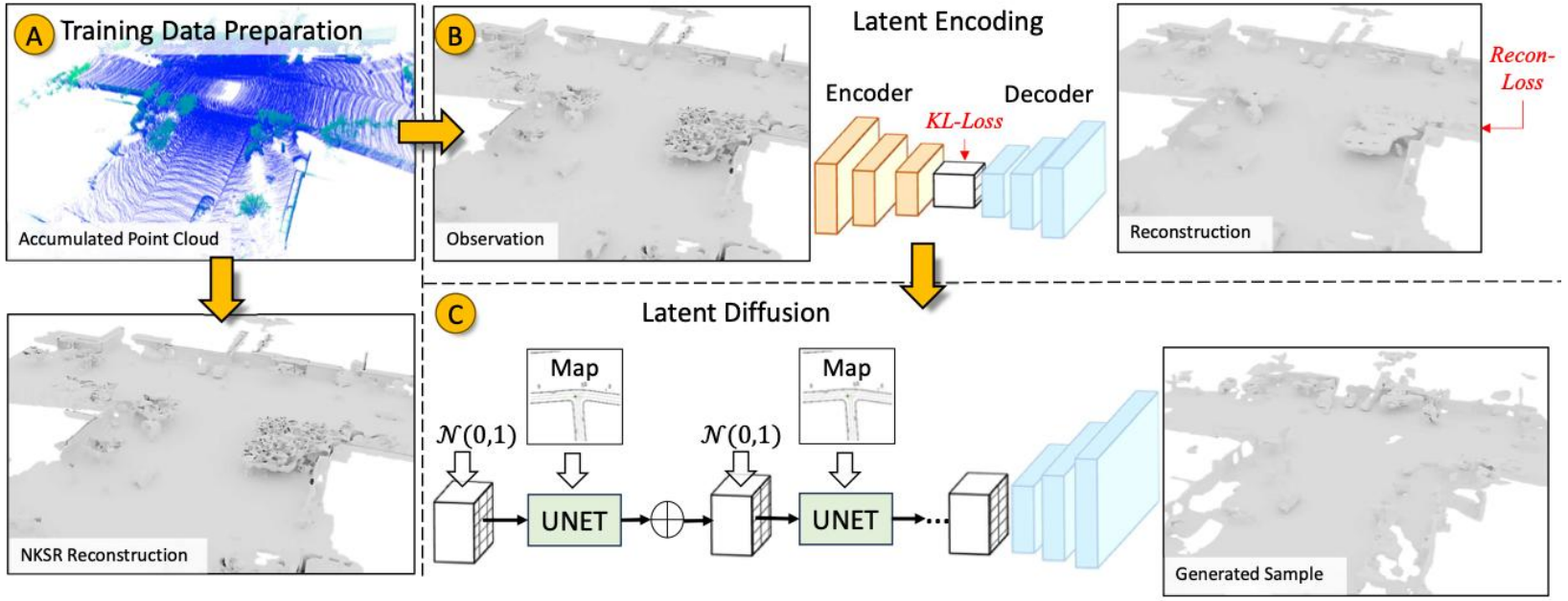
LidarDM: Generative LiDAR Simulation in a Generated World

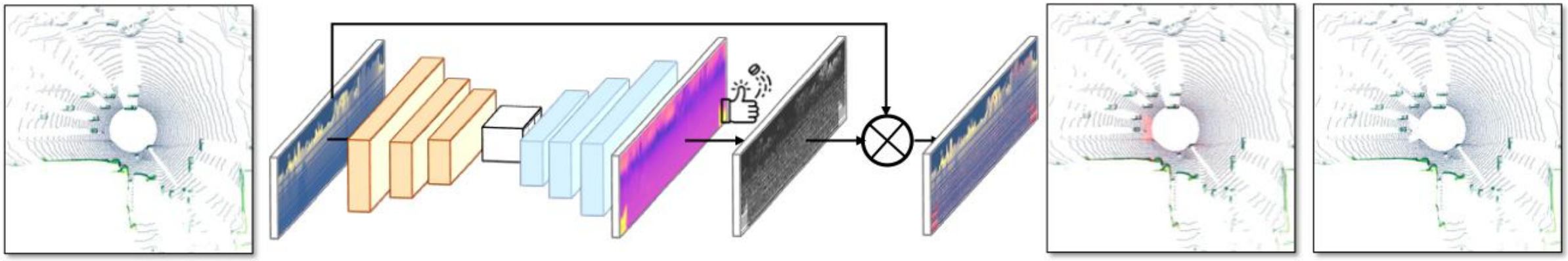
3D scene generation pipeline：

1.Mesh 重建： 利用累计的 point clouds 重建每个 ground truth mesh 样本。

2.VAE 压缩： 训练一个 variational autoencoder 将 mesh 压缩成一个 latent code。

3.扩散采样： 训练一个 map-conditioned diffusion model，在 VAE 的 latent space 中进行采样，生成全新的样本。



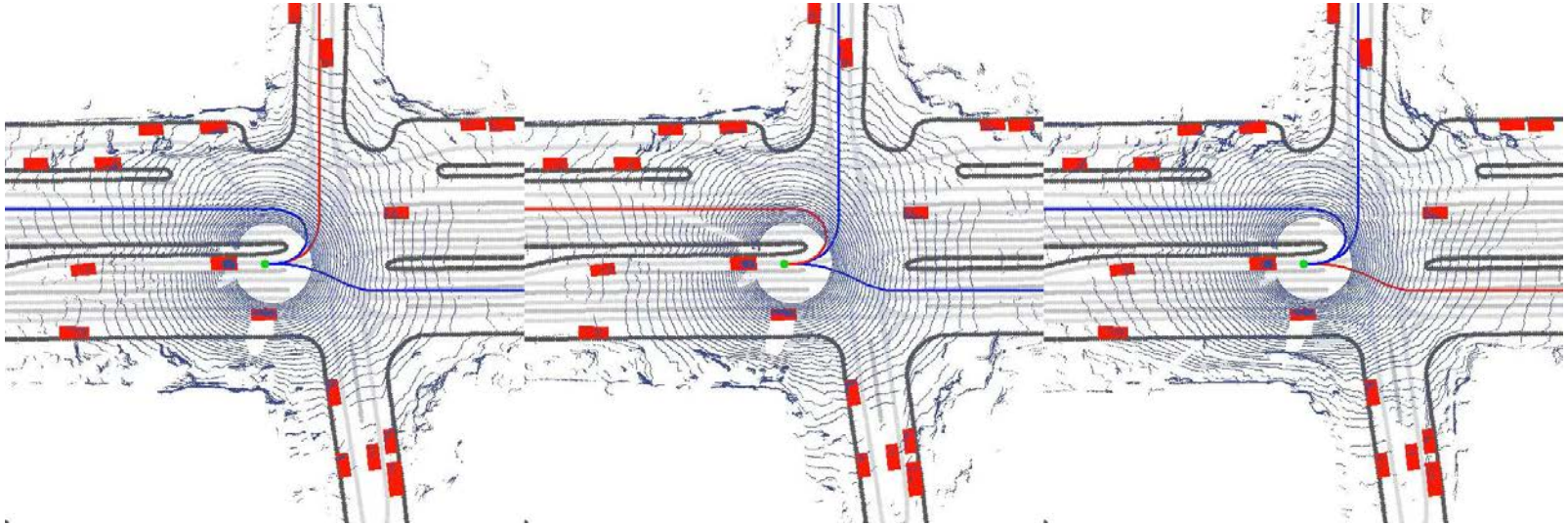


Stochastic raydrop networks 用于模拟传感器噪声，从而进一步提升生成图像的真实感；在上图的 Masked Range 与 Masked LiDAR 图像中，被 raydrop 的点以红色标出。

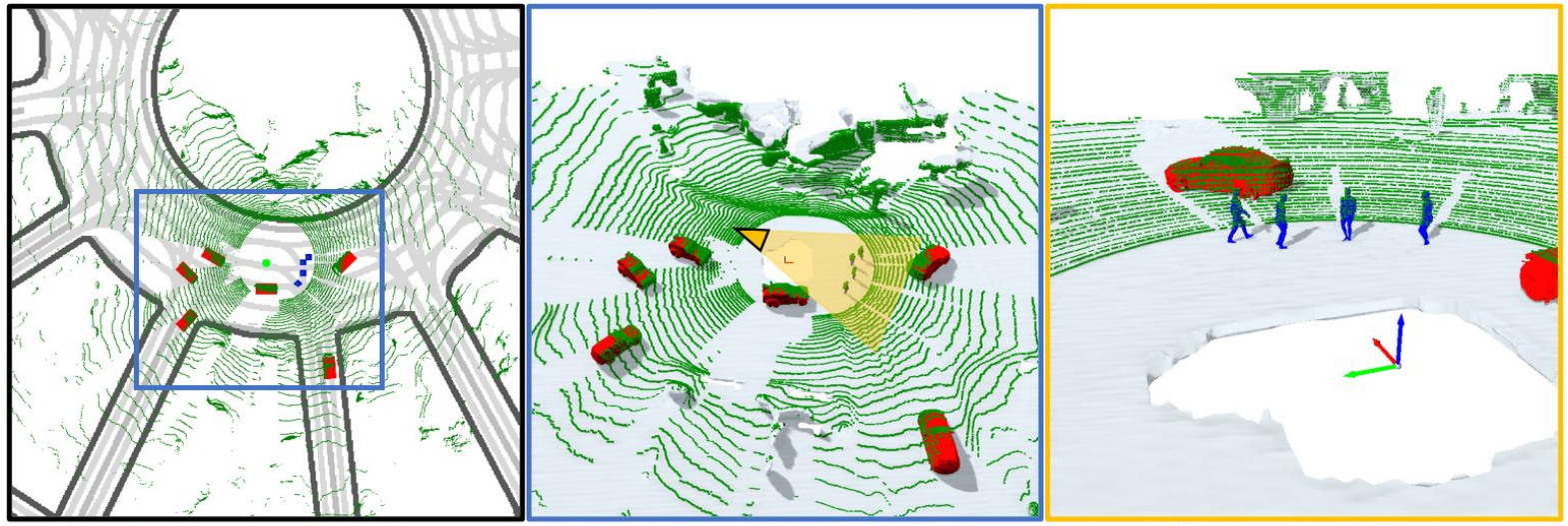
LidarDM: Generative LiDAR Simulation in a Generated World

Applications

End-to-end Traffic Simulation



Out-of-Distribution Scene Generation



Aligned Map-Lidar of Champs-Élysées | Scene Visualization | Closed-up on Pedestrian

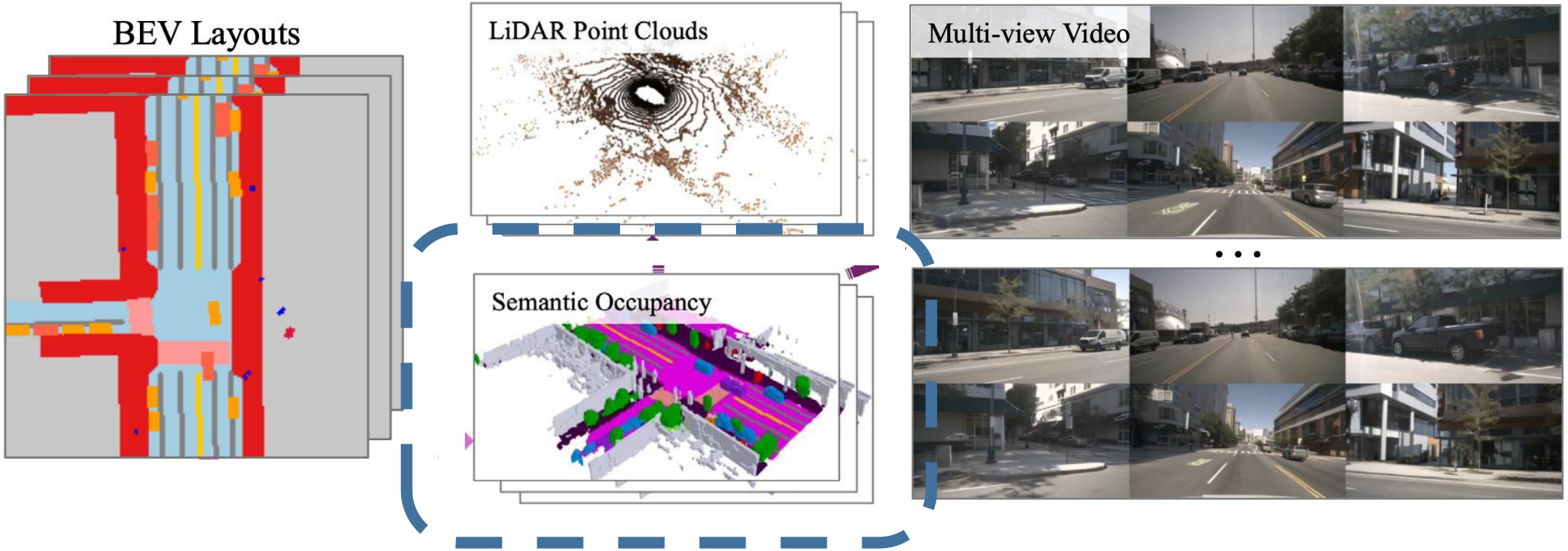
Qualitative results for unconditional generation on KITTI-360 dataset.

Method	MMD_{BEV} (↓)	JSD_{BEV} (↓)
LiDAR VAE [8]	8.53×10^{-4}	0.267
LiDAR GAN [8]	8.95×10^{-4}	0.243
ProjectedGAN [56]	7.07×10^{-4}	0.201
LidarGen [83]	2.95×10^{-4}	0.136
UltraLidar [72]	$\mathbf{9.67 \times 10^{-5}}$	0.132
LidarDM (Ours)	1.67×10^{-4}	**0.119**

	mAP (%)	mAP Agreement
Real	59.7	81.1%
LidarDM	56.4	

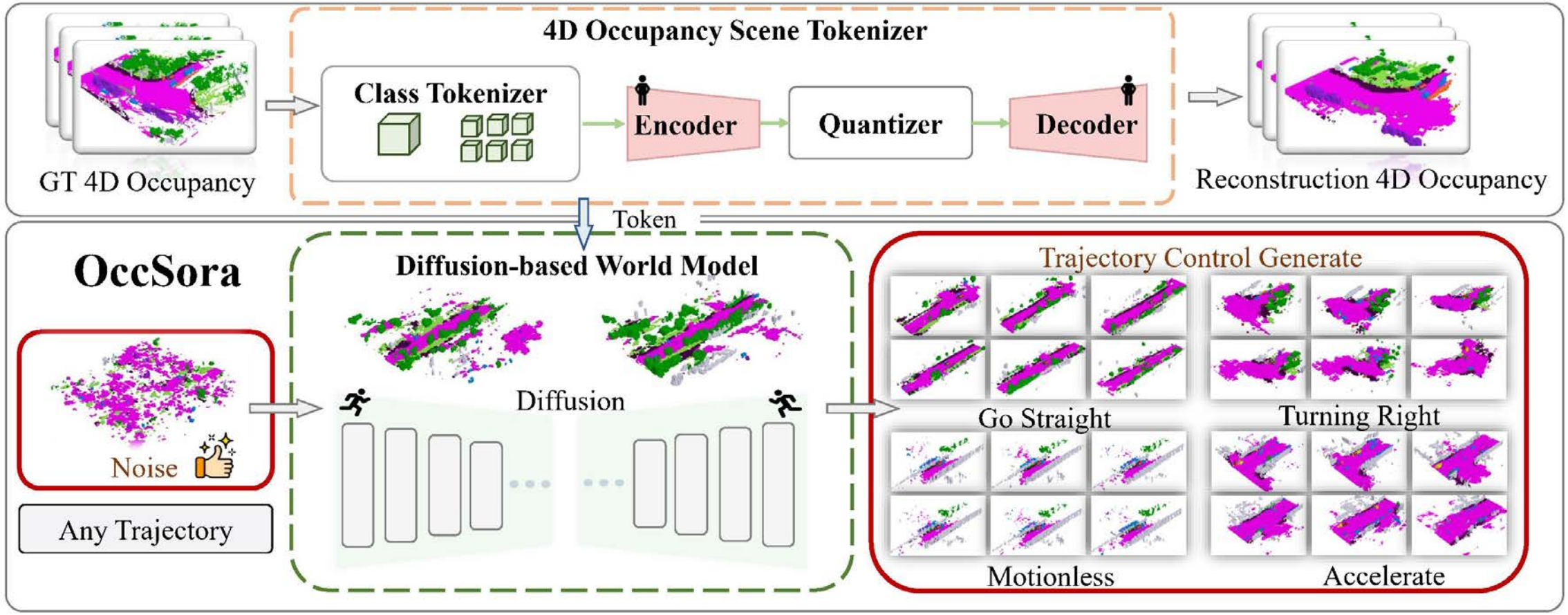
Table 3: Real2Sim: Detector [75] trained on real data can be evaluated on LidarDM-generated data, showing strong agreement with its real counterpart, suggesting its potential for simulation.

基于Occupancy的空间表示



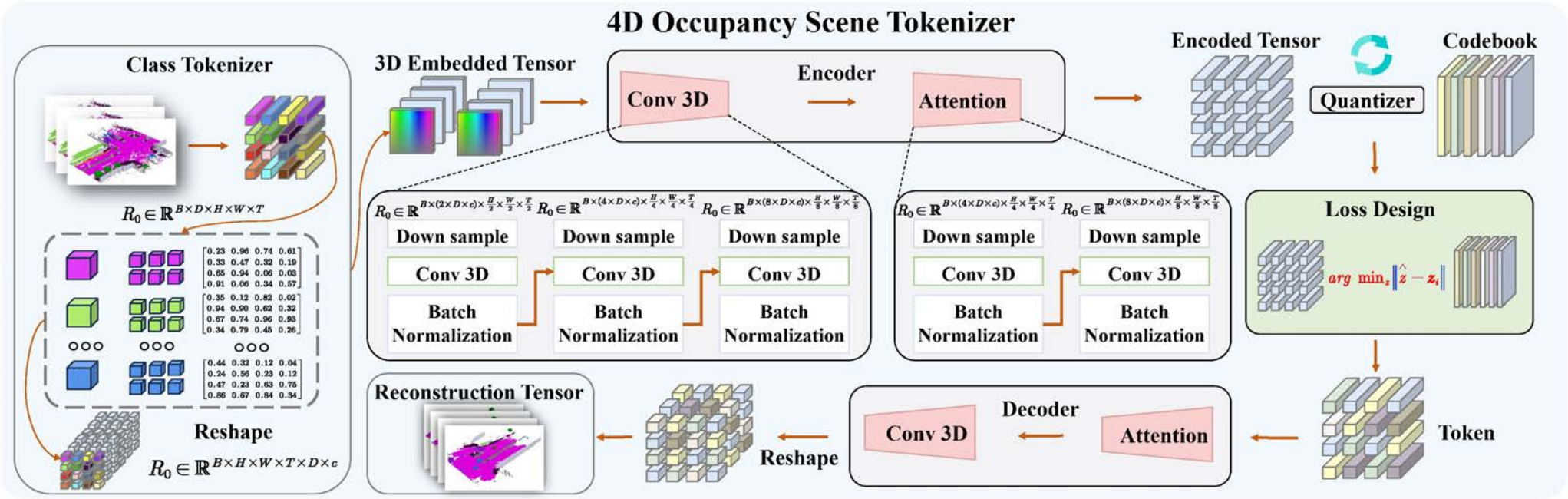
OccSora: 4D Occupancy Generation Models as World Simulators for Autonomous Driving

与大多数现有采用 autoregressive 框架进行 next-token prediction 的世界模型不同，Occsora提出了一种基于 diffusion 的 4D occupancy 生成模型 OccSora，以更高效地建模长期时序演化。
首先，使用 4D scene tokenizer 获取紧凑的离散时空表示，从而为 4D occupancy 输入提供高质量重构长序列 occupancy 视频的基础；接着，在这些时空表示上训练了一个 diffusion transformer，并根据 trajectory prompt 条件生成 4D occupancy。
OccSora 能够生成 16 秒视频，具备真实的 3D 布局和时序一致性，展示了其对驾驶场景空间和时序分布的深刻理解。

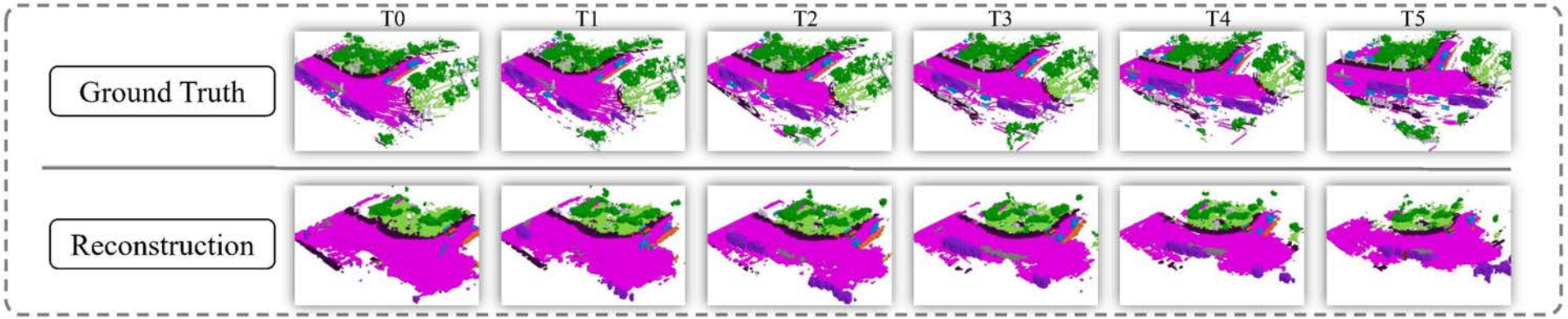


OccSora: 4D Occupancy Generation Models as World Simulators for Autonomous Driving

4D occupancy scene tokenizer 架构通过对 4D 场景进行编码与压缩来提取高维特征，再通过解码器还原出场景的时空物理特性。



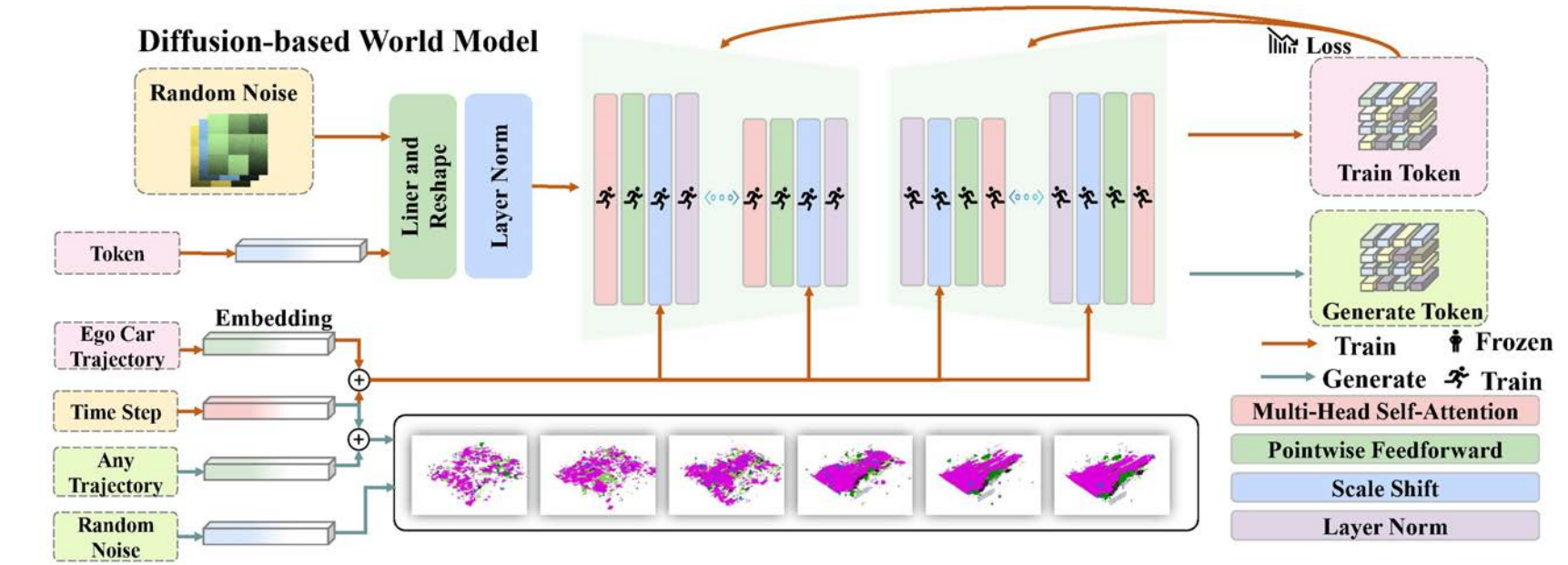
将长视频序列压缩为一个时空场景表示

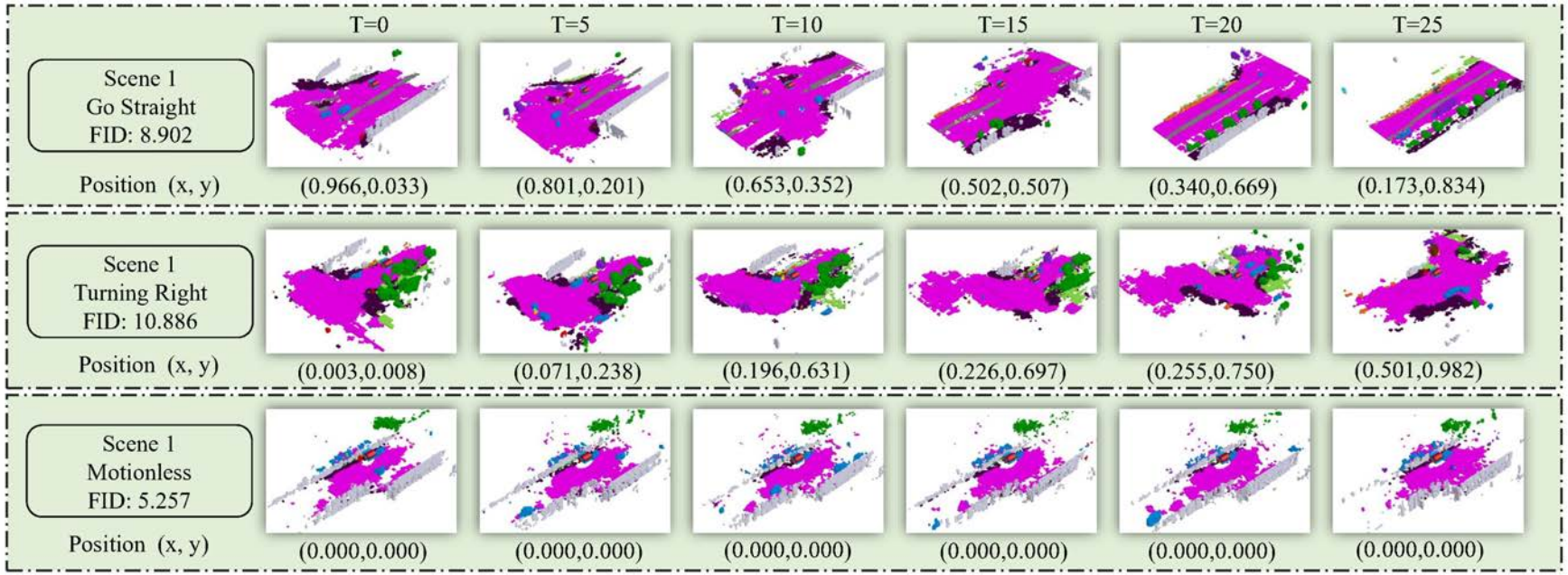


OccSora: 4D Occupancy Generation Models as World Simulators for Autonomous Driving

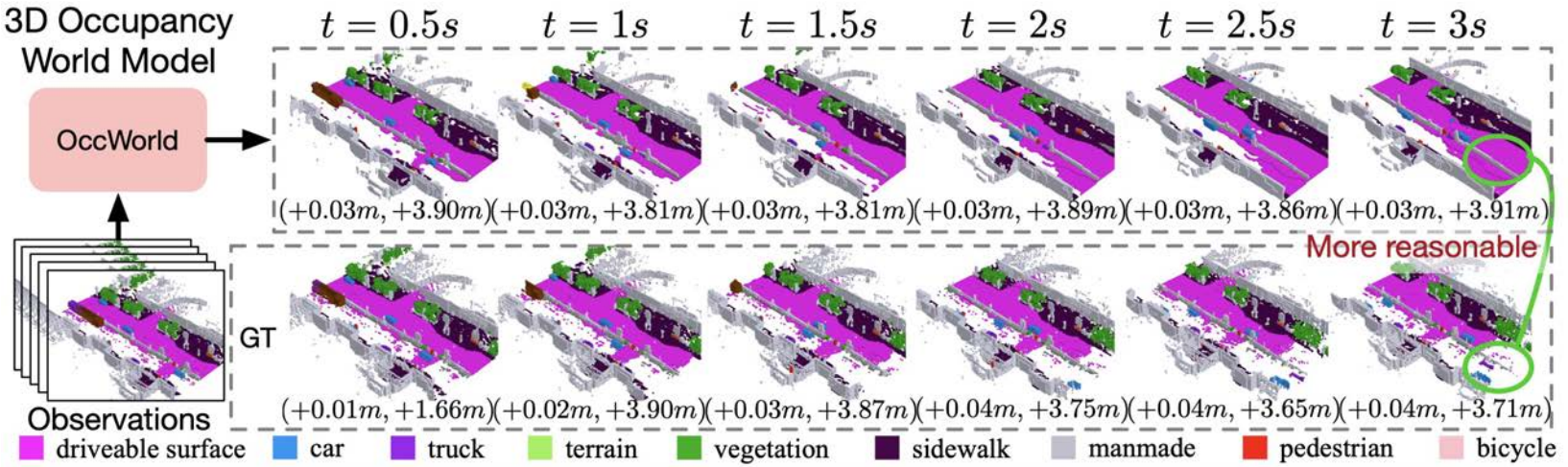
Diffusion based World Model 利用从 4D occupancy scene tokenizer 训练获得的最优 codebook，将 4D occupancy 转换为 token 序列；这些 token 再与 ego vehicle trajectory 和随机噪声组合，作为 denoising 训练的输入，最终生成目标 token。

在不同输入轨迹下生成的 4D occupancy 场景，从上到下依次对应直行、右转和静止状态，每个场景的生成都与所给轨迹匹配，确保逻辑连贯和时序连续。



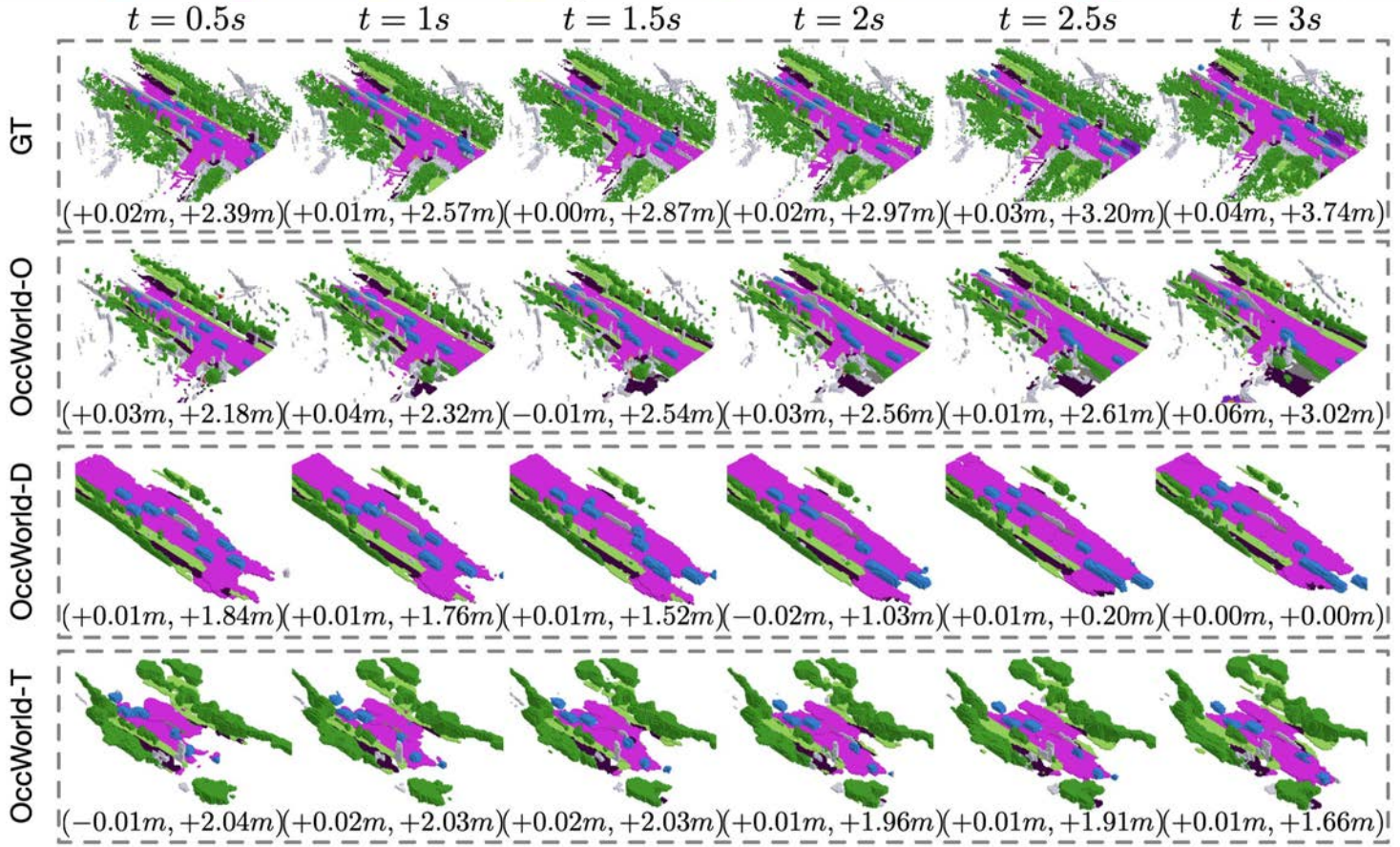


OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving



通过对过去的 3D occupancy 观测进行自监督训练，OccWorld 能够同时预测未来场景演化和自车运动。

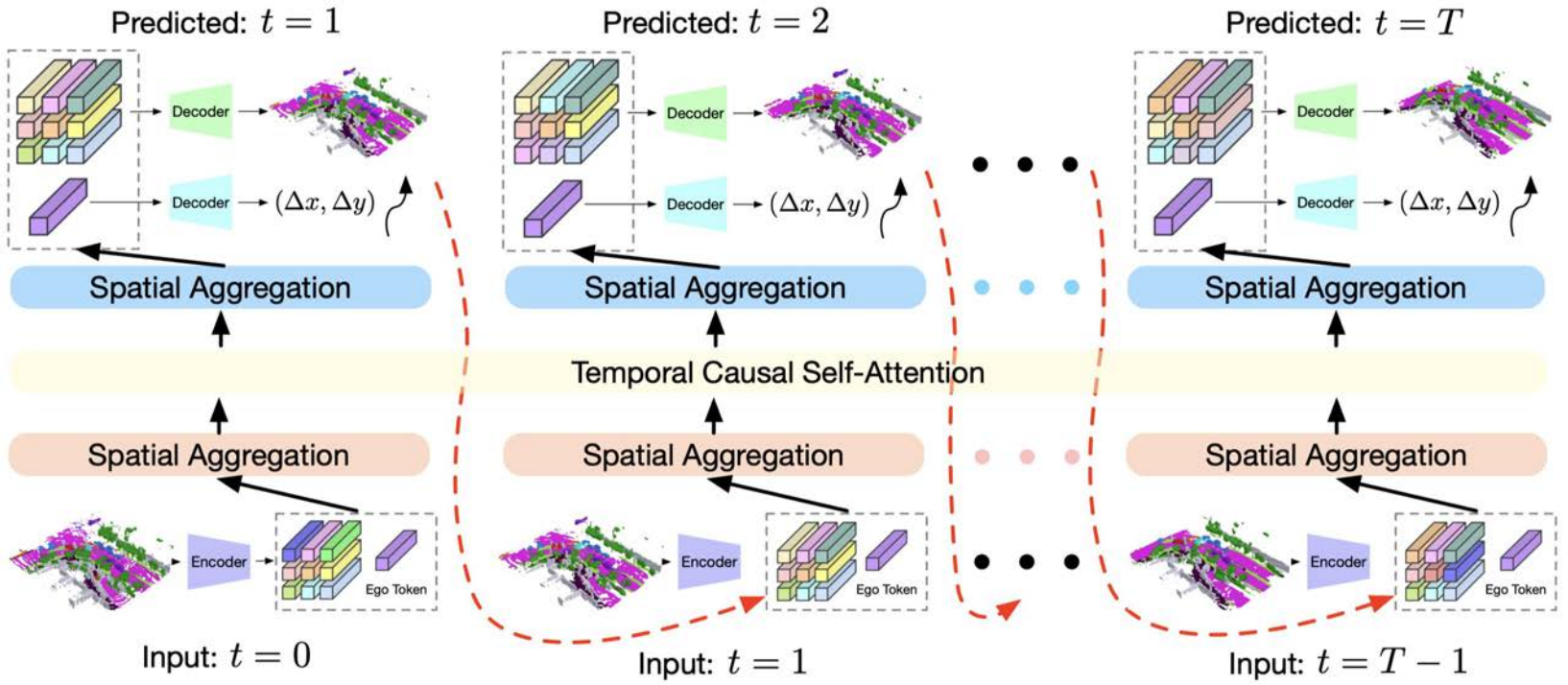
OccWorld 能成功预测周围agents的运动以及未来地图元素（如可行驶区域），甚至在某些情况下生成的可行驶区域比真实标注更合理，展示了模型对场景的理解能力而非单纯记忆。

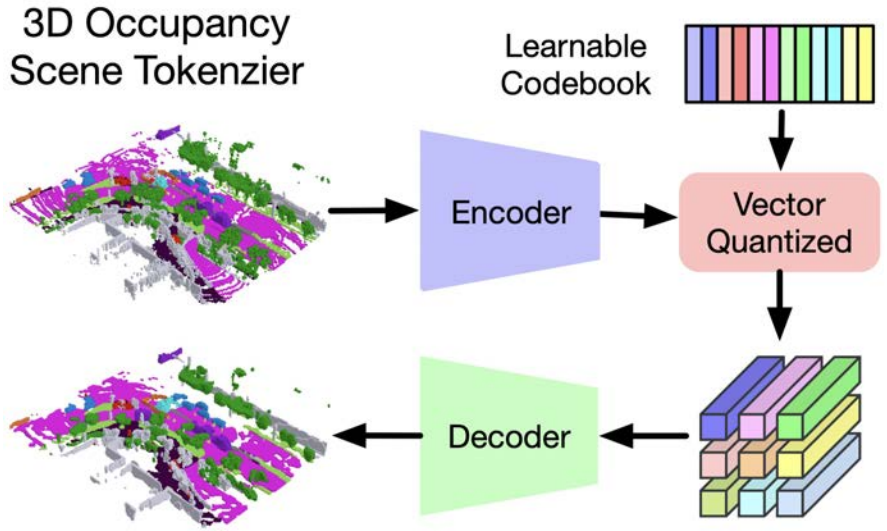


OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving

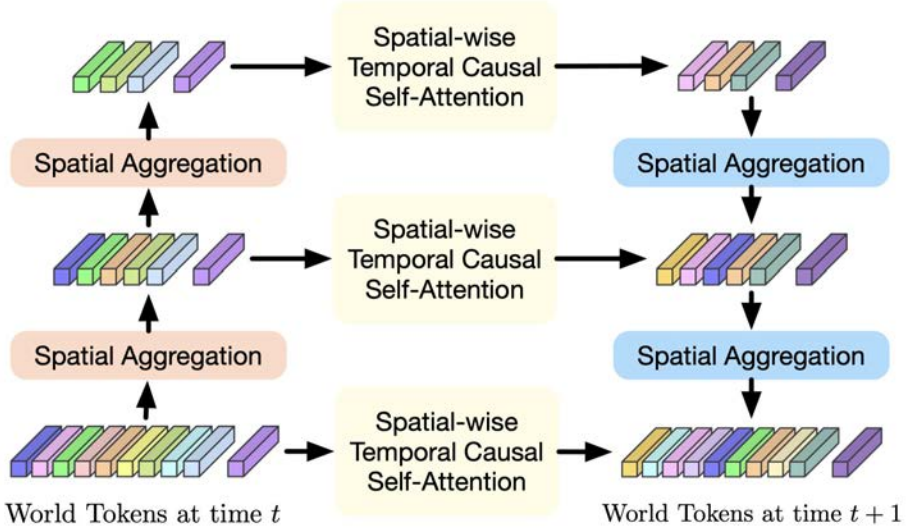
将 GPT 模型应用于自动驾驶场景：
1.训练 3D occupancy scene tokenizer，以生成 3D 场景的离散化高层表示；
2.在 spatial-wise temporal causal self-attention 模块前后进行 spatial mixing，以高效产生全局一致的场景预测。
在训练时，使用 ground-truth scene tokens 作为未来生成的输入，而在推理时则使用预测得到的 scene tokens。

使用 CNNs 对 3D occupancy 进行编码，并通过可学习的 codebook 进行 vector quantization，从而获得离散的 tokens。接着，利用 decoder 通过这些量化 tokens 重构输入的 3D occupancy，并采用重构目标函数同时训练 autoencoder 和 codebook。





每个场景由大量的 world tokens 组成，通过 spatial mixing modules 建模它们的内在依赖关系，从而获得捕捉多级别信息的多尺度 world tokens；随后，在各层级上采用 spatial-wise temporal causal self-attention 来预测下一个场景，最终利用 U-net 结构聚合多尺度预测结果。



OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving

四种设置下评估了 OccWorld：

•OccWorld-O：使用 ground-truth 3D occupancy，生成的未来 3D occupancy 表现远优于 Copy&Paste，证明模型学到了场景演化规律；

•OccWorld-D/T/S：分别使用 TPVFormer（dense ground-truth 和 sparse semantic LiDAR 训练）和 SelfOcc（自监督训练）的预测结果作为输入，实现了端到端的基于视觉的 4D occupancy 预测。

在不同设置下对比了 OccWorld 与最先进端到端自动驾驶方法的 motion planning 表现。

尽管 UniAD 表现强劲，但其需要额外的 3D 空间标注，这在大规模驾驶数据中难以获取；

作为替代，OccWorld 利用通过累计 LiDAR 扫描高效获得的 3D occupancy 作为场景表示。

Table 1. **4D occupancy forecasting performance.** Aux. Sup. denotes auxiliary supervision apart from the ego trajectory. Avg. denotes the average performance of that in 1s, 2s, and 3s. We use bold numbers to denote the best results.

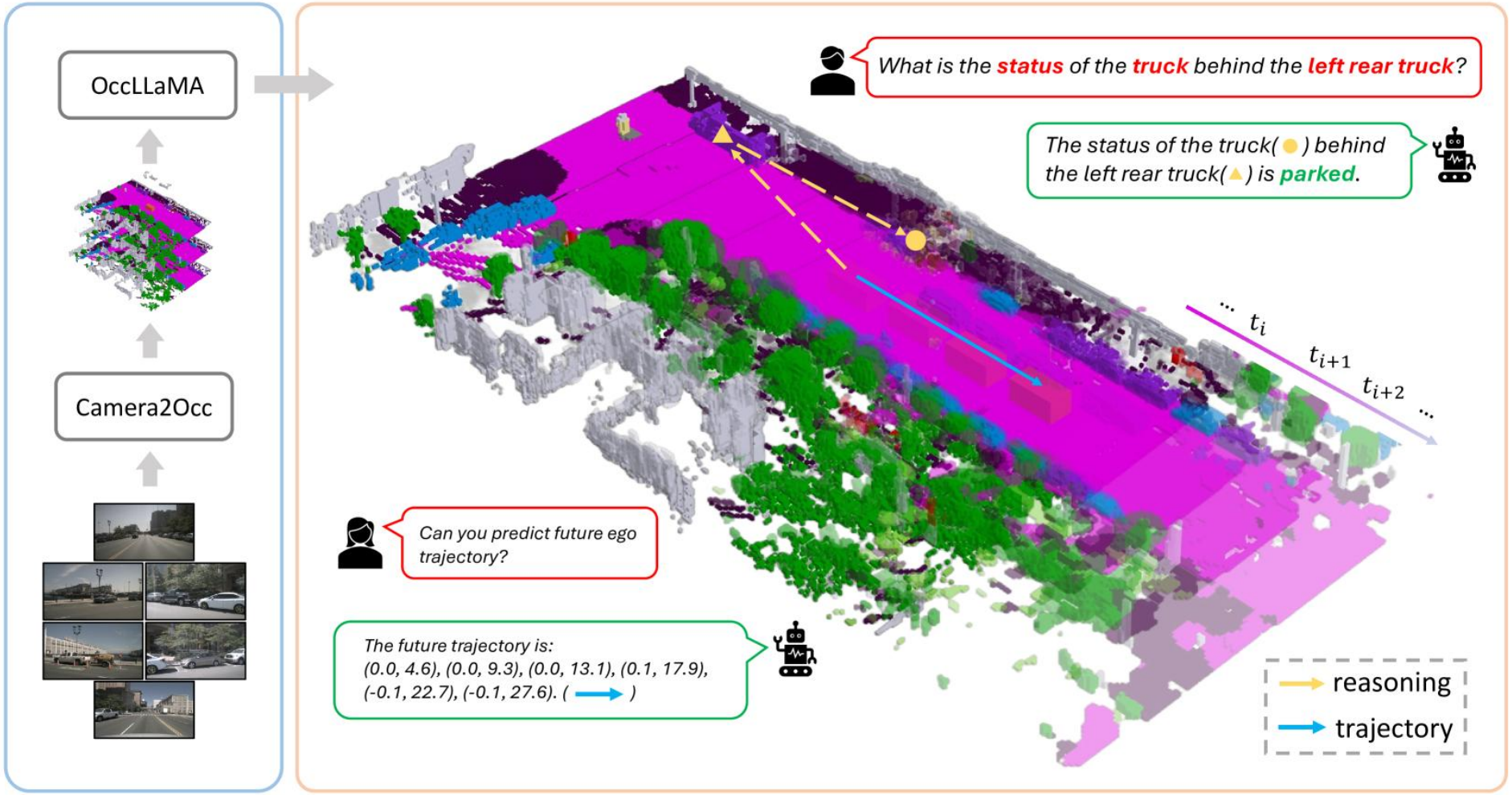
Method	Input	Aux. Sup.	mIoU (%) ↑					IoU (%) ↑					FPS
			0s	1s	2s	3s	Avg.	0s	1s	2s	3s	Avg.	
Copy&Paste	3D-Occ	None	66.38	14.91	10.54	8.52	11.33	62.29	24.47	19.77	17.31	20.52	-
OccWorld-O	3D-Occ	None	**66.38**	**25.78**	**15.14**	**10.51**	**17.14**	**62.29**	**34.63**	**25.07**	**20.18**	**26.63**	**18.0**
OccWorld-D	Camera	3D-Occ	18.63	11.55	8.10	6.22	8.62	22.88	18.90	16.26	14.43	16.53	2.8
OccWorld-T	Camera	Semantic LiDAR	7.21	4.68	3.36	2.63	3.56	10.66	9.32	8.23	7.47	8.34	2.8
OccWorld-S	Camera	None	0.27	0.28	0.26	0.24	0.26	4.32	5.05	5.01	4.95	5.00	2.8

Table 2. **Motion planning performance.** Aux. Sup. denotes auxiliary supervision apart from the ego trajectory. We use bold and underlined numbers to denote the best and second-best results, respectively. † denotes using the metric computation adopted in VAD [25].

Method	Input	Aux. Sup.	L2 (m) ↓				Collision Rate (%) ↓				FPS
			1s	2s	3s	Avg.	1s	2s	3s	Avg.	
IL [43]	LiDAR	None	0.44	1.15	2.47	1.35	0.08	0.27	1.95	0.77	-
NMP [64]	LiDAR	Box & Motion	0.53	1.25	2.67	1.48	**0.04**	0.12	0.87	0.34	-
FF [16]	LiDAR	Freespace	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43	-
EO [26]	LiDAR	Freespace	0.67	1.36	2.78	1.60	**0.04**	**0.09**	0.88	0.33	-
ST-P3 [17]	Camera	Map & Box & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	1.6
UniAD [18]	Camera	Map & Box & Motion & Tracklets & Occ	0.48	**0.96**	**1.65**	**1.03**	0.05	0.17	**0.71**	**0.31**	1.8
VAD-Tiny [25]	Camera	Map & Box & Motion	0.60	1.23	2.06	1.30	0.31	0.53	1.33	0.72	16.8
VAD-Base [25]	Camera	Map & Box & Motion	0.54	1.15	1.98	1.22	**0.04**	0.39	1.17	0.53	4.5
OccNet [53]	Camera	3D-Occ & Map & Box	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72	2.6
OccNet [53]	3D-Occ	Map & Box	1.29	2.31	2.98	2.25	0.20	0.56	1.30	0.69	-
OccWorld-O	3D-Occ	None	**0.43**	1.08	1.99	1.17	0.07	0.38	1.35	0.60	**18.0**
OccWorld-D	Camera	3D-Occ	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87	2.8
OccWorld-T	Camera	Semantic LiDAR	0.54	1.36	2.66	1.52	0.12	0.40	1.59	0.70	2.8
OccWorld-S	Camera	None	0.67	1.69	3.13	1.83	0.19	1.28	4.59	2.02	2.8
VAD-Tiny† [25]	Camera	Map & Box & Motion	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38	16.8
VAD-Base† [25]	Camera	Map & Box & Motion	0.41	0.70	1.05	0.72	0.07	**0.17**	**0.41**	**0.22**	4.5
OccWorld-O†	3D-Occ	None	**0.32**	**0.61**	**0.98**	**0.64**	**0.06**	0.21	0.47	0.24	**18.0**
OccWorld-D†	Camera	3D-Occ	0.39	0.73	1.18	0.77	0.11	0.19	0.67	0.32	2.8
OccWorld-T†	Camera	Semantic LiDAR	0.40	0.77	1.28	0.82	0.12	0.22	0.56	0.30	2.8
OccWorld-S†	Camera	None	0.49	0.95	1.55	0.99	0.19	0.56	1.54	0.76	2.8

OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving

OccLLaMA accepts occupancy data from existing occupancy prediction algorithms, and performs a series of tasks, including scene understanding and reasoning, 4D occupancy prediction, and motion planning.

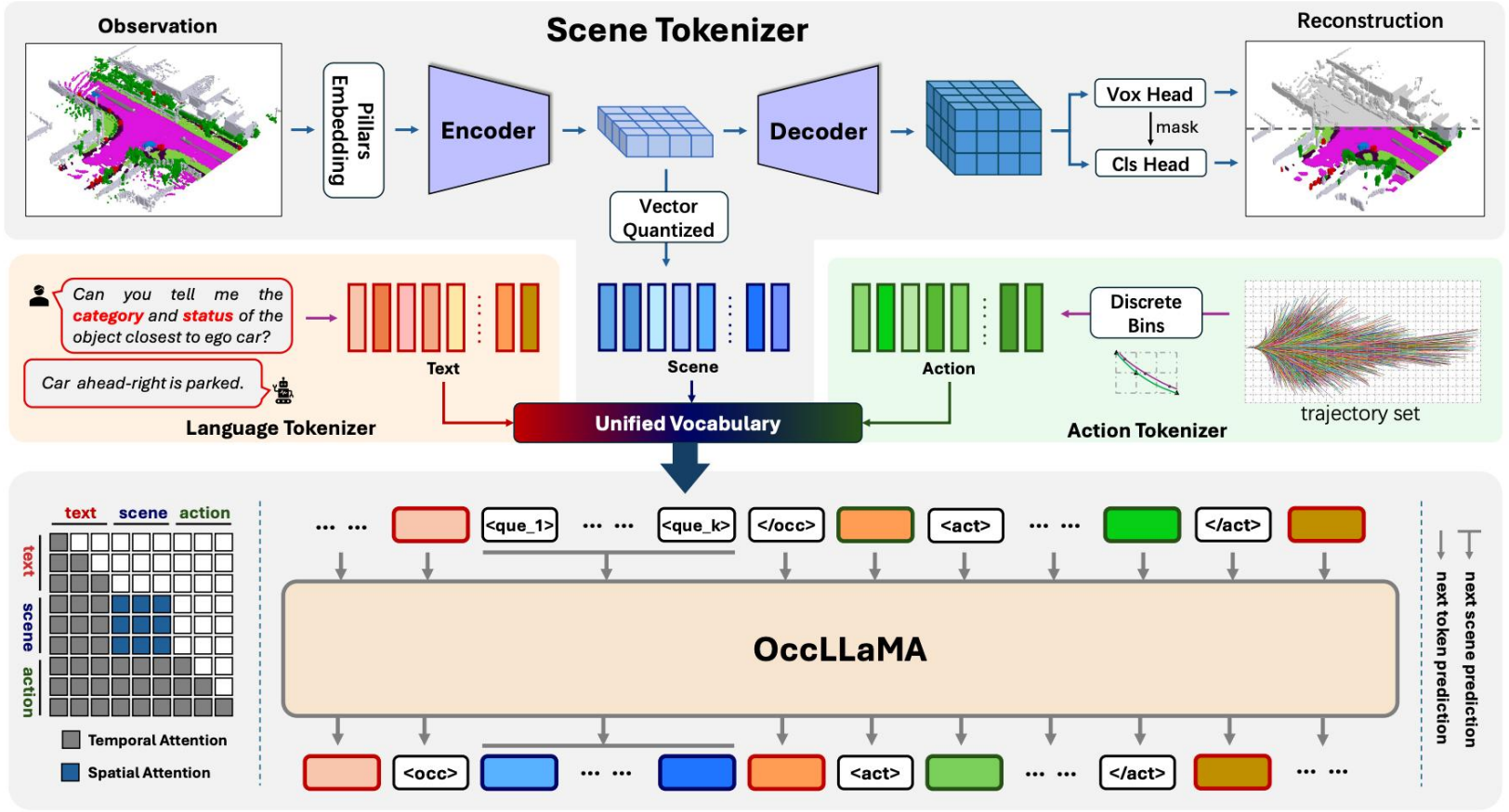


OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving

OccLLaMA 架构：

Scene Tokenizer：将 Occ 表示离散化为场景词汇，并与 language 和 action vocabularies 结合，构成一个统一的词汇空间。

Generative World Model：在统一词汇空间内执行 next token/scene 预测，实现对场景的理解、推理和动作预测。



OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving

Method	input	exist(%)↑			count(%)↑			object(%)↑			status(%)↑			comparison(%)↑			acc(%)↑
		h0	h1	all	h0	h1	all	h0	h1	all	h0	h1	all	h0	h1	all	
LLaVA-D	Depth	38.9	51.9	45.8	7.7	7.6	7.7	10.5	7.4	7.8	7.0	9.9	9.0	64.5	50.8	52.1	26.2
LLaVA	Camera	74.8	72.9	73.8	14.9	14.3	14.6	57.7	34.5	37.9	48.6	44.5	45.9	65.9	52.1	53.3	47.4
LiDAR-LLM	LiDAR	79.1	70.6	74.5	15.3	14.7	15.0	59.6	34.1	37.8	**53.4**	42.0	45.9	67.0	57.0	57.8	48.6
Ours-LLaMA2	Occ	80.6	**79.3**	79.9	18.6	19.1	18.9	**64.9**	39.0	42.8	48.0	**49.6**	**49.1**	**80.6**	63.7	65.2	53.4
Ours-LLaMA3.1	Occ	**82.9**	79.2	**80.9**	**19.2**	**19.2**	**19.2**	64.8	**43.1**	**46.3**	51.0	46.1	47.8	76.5	**65.6**	**66.6**	**54.5**

Table 3: Quantitative results on NuScenes-QA. LLaVA-D input depth generated from point cloud.

Method	Input	Sup.	L2(m)↓				Coll.(%)↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
IL	LiDAR	None	0.44	1.15	2.47	1.35	0.08	0.27	1.95	0.77
NMP	LiDAR	Box & Motion	0.53	1.25	2.67	1.48	0.04	**0.12**	0.87	0.34
FF	LiDAR	Freespace	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
ST-P3	Camera	Map & Box & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD	Camera	Map & Box & Motion & Track & Occ	0.48	**0.96**	**1.65**	**1.03**	0.05	0.17	**0.71**	**0.31**
VAD	Camera	Map & Box & Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
OccWorld-F	Camera	Occ	0.45	1.33	2.25	1.34	0.08	0.42	1.71	0.73
Ours-F	Camera	Occ	0.38	1.07	2.15	1.20	0.06	0.39	1.65	0.70
OccNet	Occ	Map & Box	1.29	2.31	2.98	2.25	0.20	0.56	1.30	0.69
OccWorld-O	Occ	None	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60
Ours-O†	Occ	None	0.38	1.06	2.08	1.18	**0.02**	0.17	1.39	0.53
Ours-O	Occ	None	**0.37**	1.02	2.03	1.14	0.04	0.24	1.20	0.49

Table 2: Quantitative results of motion planning. Ours-O† refers to output trajectories without scene predictions.

未来技术突破方向与核心挑战

1.感知系统极限突破

1. 极端环境可靠性：激光雷达与摄像头在暴雨、浓雾等场景下的性能不足（误检率超30%），需开发多模态冗余感知架构。
2. 成本控制：激光雷达价格仍高于500美元/台，制约L4级车型量产。

2. 决策算法的长尾难题

1. 复杂场景泛化：现有AI模型对0.1%的罕见场景（如无标识乡村道路、动物突然穿行）处理失败率超50%。
2. 伦理决策机制：事故中“电车难题”尚无国际标准算法框架。

3. 算力与能效瓶颈

1. 车载芯片算力需求：L5级需1000+TOPS算力，当前最高芯片（如英伟达Thor）仅2000TOPS但功耗达1000W。

4. 车路协同基础设施

1. 全局数据融合：需覆盖90%以上道路的V2X通信网络，当前中国智能道路改造率不足5%。

自动驾驶重构未来交通的五大场景

1.城市出行革命

1. Robotaxi规模化：成本降至0.3元/公里（传统出租车1.5元），2030年或占共享出行50%份额。
2. 私人车辆订阅化：用户按需购买“驾驶里程包”，车辆空闲时自动加入共享车队创收。

2. 物流运输智能化

1. 无人重卡干线网络：L4级卡车编队行驶降低油耗15%，物流成本下降30%。
2. 最后一公里无人配送：社区级无人车枢纽实现30分钟达，人力成本减少70%。

3. 城市空间重塑

1. 停车场用地释放：自动驾驶车辆利用率提升至60%（传统10%），城市可转化30%停车位为绿地或商业区。
2. 动态道路管理：AI实时调整车道功能（如早高峰增设为潮汐车道），通行效率提升40%。

4. 新型服务经济

1. 移动商业空间：车辆变身移动咖啡厅/会议室，用户通勤时间转化为消费场景。
2. 数据驱动的保险模式：按自动驾驶系统评分动态定价保费，事故率下降80%。

5. 特殊场景渗透

1. 矿区/港口自动驾驶：7×24小时无人化作业，安全事故减少95%。
2. 老年/残障专属出行：语音/手势交互车辆解决特殊群体出行刚需

谢谢！

提问文档