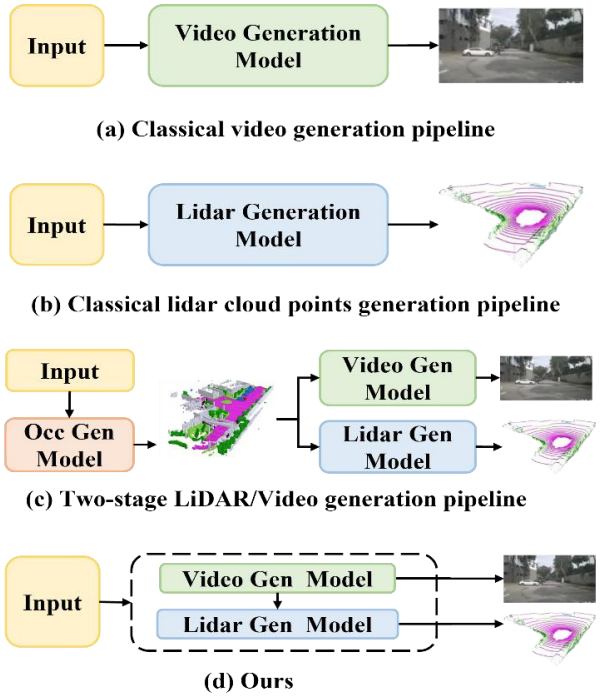
Genesis: Multimodal Driving Scene Generation with Spatio-Temporal and Cross-Modal Consistency

Xiangyu Guo, Zhanqian Wu, Kaixin Xiong, Ziyang Xu, Lijun Zhou, Gangwei Xu, Shaoqing Xu,
Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Wenyu Liu, Xinggang Wang*

Genesis: Multimodal Driving Scene Generation with
Spatio-Temporal and Cross-Modal Consistency. *NeurIPS 2025.*

Background and Motivation



Method	Multi-view	Video	LiDAR
BEVGen	✓	×	×
BEVControl	✓	×	×
DriveDreamer	×	✓	×
Vista	×	✓	×
MagicDrive	✓	✓	×
MagicDriveDiT	✓	✓	×
Drive-WM	✓	✓	×
LiDARGen	×	×	✓
LiDARDiffusion	×	×	✓
LiDARDM	×	×	✓
Copilot4D	×	×	✓
UniScene	✓	✓	✓
Ours	✓	✓	✓

Limitations with Current Metrics

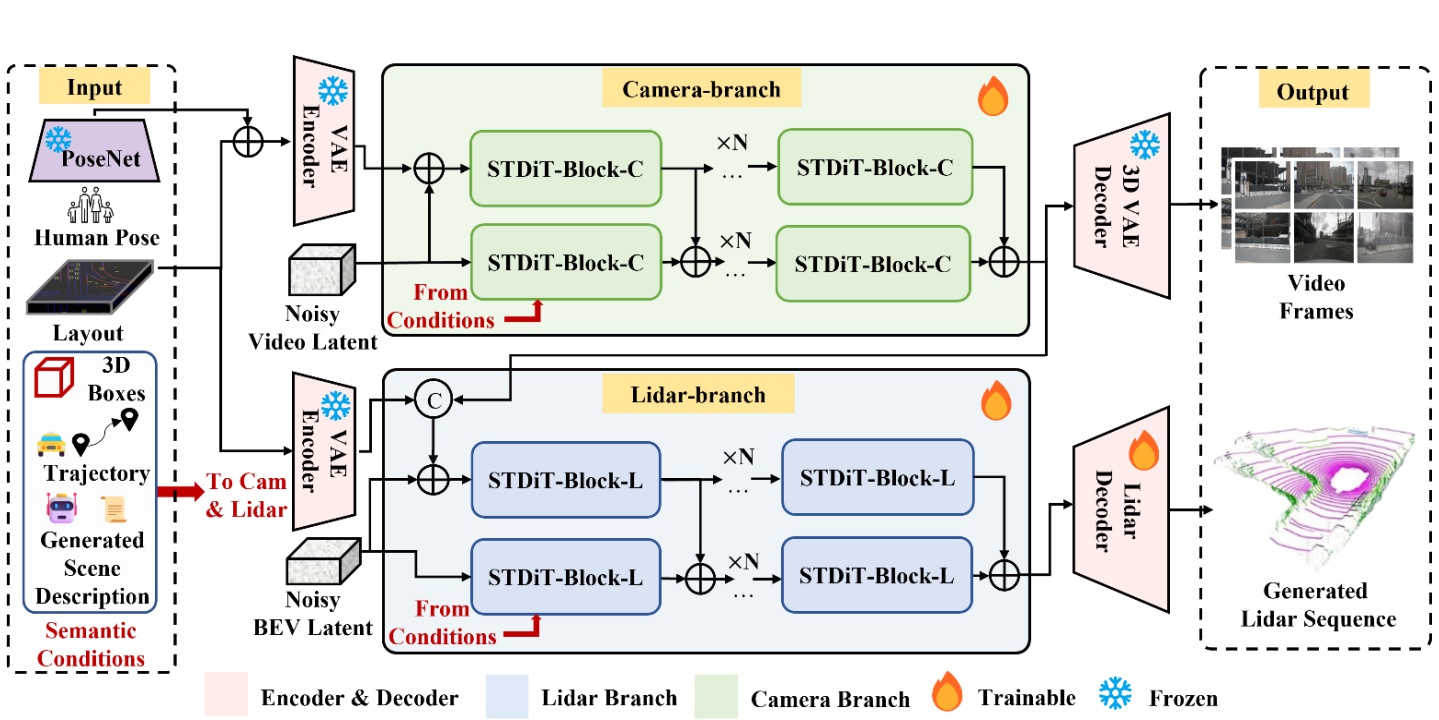
- Support only the unimodal representation generated by video or LiDAR.
- Rely on intermediate representations like occupancy grids that may cause information loss.

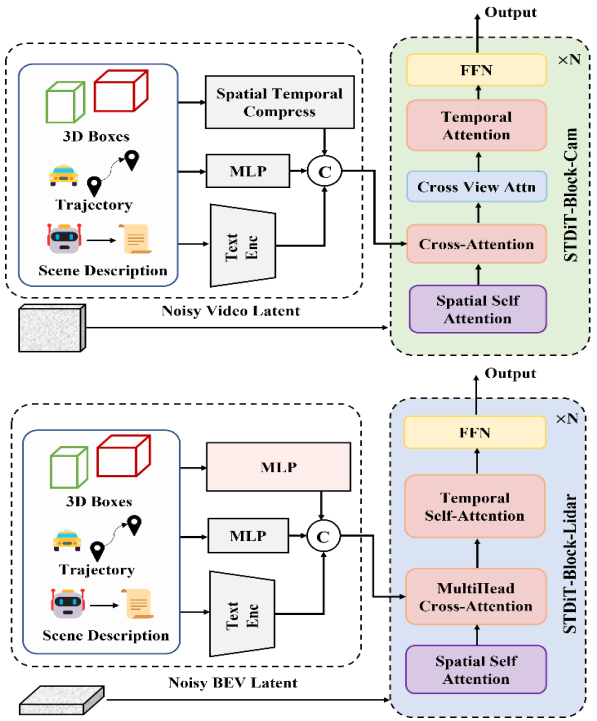
need

✓ A unified joint generation framework tailored for autonomous driving, which synthesizes multi-view RGB videos and LiDAR point clouds in a consistent, semantically grounded manner without reliance on intermediaries such as occupancy grids.

Our Approach

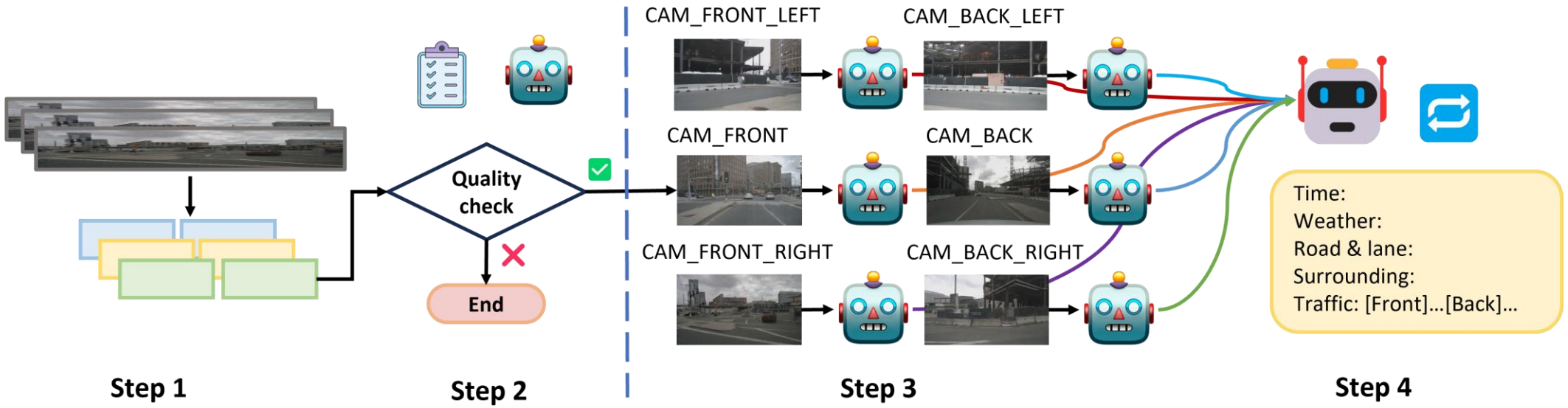
Contribution **Develop a unified two-branch architecture to jointly synthesize multimodal driving scenes.**





- Camera branch leverages a DiT-based spatiotemporal diffusion backbone with a 3D-VAE encoder to capture fine-grained visual dynamics.
- LiDAR branch employs a BEV-aware AE with NeRF-style rendering and adaptive sampling to ensure accurate geometric reconstruction.
- Both branches are jointly conditioned on scene layouts, intermediate video latent features, and structured scene semantics to **maintain strong cross-modal consistency**.

Datacrafter



- Split the multi-view video into several clips, score each clip and only retain the clips with scores exceeding the threshold for training.
- Define three scoring dimensions: Clarity (covering blurriness, distortion, etc.), Structural Rationality (including occlusion, scene integrity, etc.), and Aesthetic Attributes (such as backlighting, abnormal exposure, etc.).
- Generate hierarchical descriptive output for each clip: semantic information that integrates both global-level and instance-level details.

Datacrafter

w/ DataCrafter [RIGHT BACK CAM] A black SUV prominently partially occupying the right side of the frame. A male driver is clearly visible through the window. The background features a brick building under construction, with a blue scaffolding cover and metallic support bars spanning the entrance. One bicycles is parked on the sidewalk beneath the scaffolding.

w/o DataCrafter [RIGHT BACK CAM] Daytime. Driving on the street. A black SUV passing.

- Split the multi-view video into several clips, score each clip and only retain the clips with scores exceeding the threshold for training.
- Define three scoring dimensions: Clarity (covering blurriness, distortion, etc.), Structural Rationality (including occlusion, scene integrity, etc.), and Aesthetic Attributes (such as backlighting, abnormal exposure, etc.).
- Generate hierarchical descriptive output for each clip: semantic information that integrates both global-level and instance-level details.

Evaluation Results–Video Generation

Table: Video Generation Comparison on nuScenes validation set, where green and blue represent the best and the second best values.

Method	Gen. Mode	Multi-view	Video	Sample Num	Frame Num	FVD_{multi} ↓	FID_{multi} ↓
DriveDreamer-2	w/o first cond	✓	✓	–	–	105.10	25.00
MagicDrive-V2	w/o first cond	✓	✓	–	16	94.84	20.91
Drive-WM	w/o first cond	✓	✓	–	–	122.70	15.80
Ours	w/o first cond	✓	✓	5369	16	83.10	14.90
MiLA	w/ first cond	✓	✓	5369	16	18.20	3.00
DriveDreamer-2	w/ first cond	✓	✓	–	–	55.70	11.20
Ours	w/ first cond	✓	✓	5369	16	16.95	4.24
UniScene	w/ noisy latent	✓	✓	6019	–	70.52	6.12
Ours	w/ noisy latent	✓	✓	6019	16	67.87	6.45

- ✓ **Without first-frame conditioning:** Our method outperforms MagicDrive-v2 by 12.38% on FVD_{multi} and DriveWM by 5.69% on FID_{multi}.
- ✓ **With first-frame conditioning:** Our method further improves to 16.95 FVD_{multi} and 4.24 FID_{multi}, showing competitive results compared to MiLA.
- ✓ **With the noisy latent:** Our method achieves 67.87 FVD on 6,019 samples, surpassing the previous best results reported by UniScene.

Evaluation Results–Lidar Generation

Table: Lidar Generation Comparison on nuScenes validation set, where green and blue represent the best and the second best values. “gt_img“ and “gen_img” indicate using ground-truth or generated images as BEV condition input, respectively.

Method	Condition Type	Chamfer@1s ↓	Chamfer@2s ↓	Chamfer@3s ↓
4D-Occ	gt_img	1.13	1.53	2.11
ViDAR	gt_img	1.12	1.38	1.73
HERMES	gt_img	0.78	0.95	1.17
Ours	gt_img	0.611	0.625	0.633
Ours	gen_img	0.634	0.638	0.641

✓ At 1 second, our method achieves a Chamfer Distance of 0.611, surpassing the previous best (0.78 by HERMES) by 21%. At 3 seconds, the advantage widens to a 45% relative reduction (from 1.17 to 0.633).

Ablation studies

Table: Ablation in the video generation model.

Method	Gen. Mode	Sample Num	Frame Num	FVD_{multi} ↓	FID_{multi} ↓
baseline	w/o first cond	5369	16	117.49	22.32
w/ DataCrafter	w/o first cond	5369	16	85.91	15.20
w/ DataCrafter and PoseNet	w/o first cond	5369	16	83.10	14.90

✓ Video generation ablation study verifying the proposed DataCrafter and PoseNet modules (which introduce structured semantic captions to video generation) shows removing both significantly degrades quality: FVD_{multi} rises from 85.91 to 117.49, FID_{multi} from 15.20 to 22.32.

Table: Ablation in the Lidar generation model.

Method	Chamfer@1s ↓	Chamfer@2s ↓	Chamfer@3s ↓
w/o Img BEV Latent + w/o Ref. Frame	0.661	0.669	0.673
w/ Img BEV Latent + w/o Ref. Frame	0.668	0.672	0.677
w/ Img BEV Latent + w/ Ref. Frame	**0.634**	**0.638**	**0.641**

✓ LiDAR generation ablation study evaluates BEV latent features and first-frame conditioning, demonstrating both components are critical for geometrically consistent long-range LiDAR synthesis.

Downstream task

Table: Domain gap on bev segmentation.

Method	mIoU↑	mAP↑
MagicDrive [10]	18.34	11.86
MagicDrive3D [8]	18.27	12.05
MagicDriveDiT [9]	20.40	18.17
DiVE [16]	35.96	24.55
Cogen [15]	37.80	27.88
Ours	**38.01**	**27.90**

Table: Effect of Multimodal Data Generation on 3D Object Detection. Inputs of the methods in the table are both camera and lidar modalities.

Method	mAP↑	NDS↑
Baseline	66.87	69.65
Ours(+cam_gen)	67.09 (+0.22)	70.12 (+0.47)
Ours(+lidar_gen)	67.69 (+0.82)	70.58 (+0.93)
Ours(+cam&lidar_gen)	67.78 (+0.91)	71.13 (+1.48)

- ✓ **Domain gap on bev segmentation**: our method outperforms prior methods (e.g., DiVE, Cogen) in BEV segmentation, achieving the best mIoU (38.01) and mAP (27.90).
- ✓ **3D Object Detection**: joint camera-LiDAR generation achieves the highest gains (+0.91 mAP / +1.48 NDS), confirming multimodal complementarity and high-quality synthetic data's value for downstream perception.

Visualizations

Joint Generation on nuScenes

Visualizations

Joint Generation on Private Data