ICLR 2026
ConsisDrive:Identity-Preserving Driving World Models For Video Generation By Instance Mask

Zhuoran Yang, Yanyong Zhang

University of Science and Technology of China

Github: https://shanpoyang654.github.io/ConsisDrive/page.html

Email: shanpoyang@mail.ustc.edu.cn

中国科学技术大学
University of Science and Technology of China

Resume

ConsisDrive

Catalogue

- Motivation
- Method
- Experimental Results
- About Author

Motivation

Motivation: Autonomous Driving Street Scene Video Generation

Necessity of the Task:

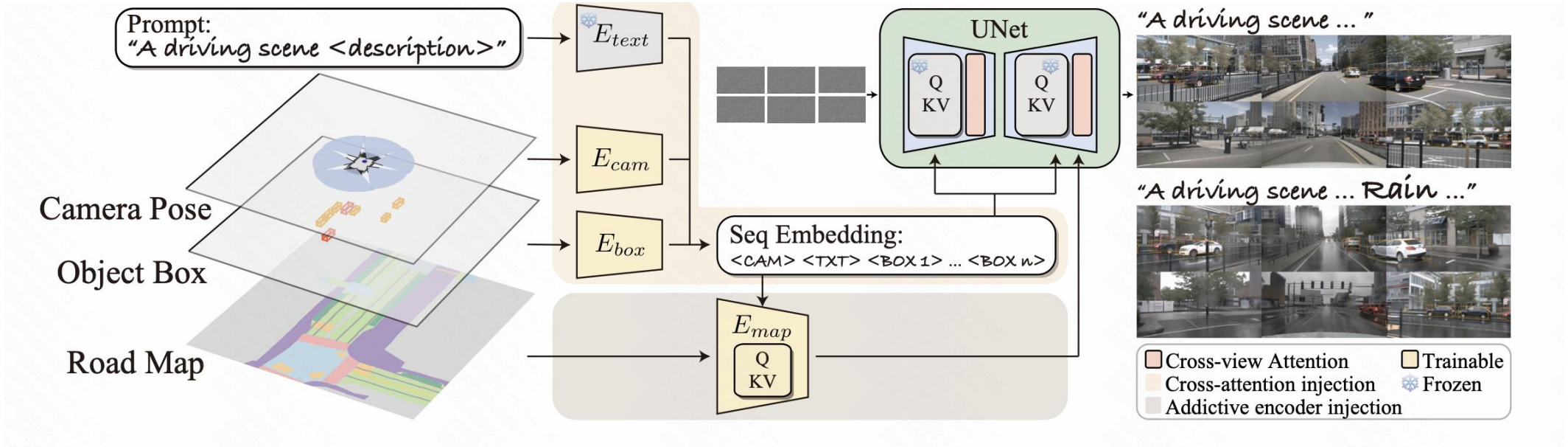
- Autonomous driving demands large-scale, diverse data.
- Data annotation for real-world driving scenes is expensive and time-consuming.
- Generated data can effectively serve as training augmentation for downstream perception and planning models.

Research Motivation: Instance-Level Temporal Consistency

- High realism is essential for synthetic data to be useful in autonomous driving.
  - Fine-grained temporal consistency: critical for training tracking and planning models.
  - Existing works (e.g., MagicDrive-V2, Panacea, DriveDreamer2) still struggle to maintain fine temporal consistency.

Related Works

Instance-Level Temporal Consistency:



MagicDrive-V2 fails to maintain fine-grained temporal consistency:

- the lack of cross-frame instance-level constraints.

Method:ConsisDrive

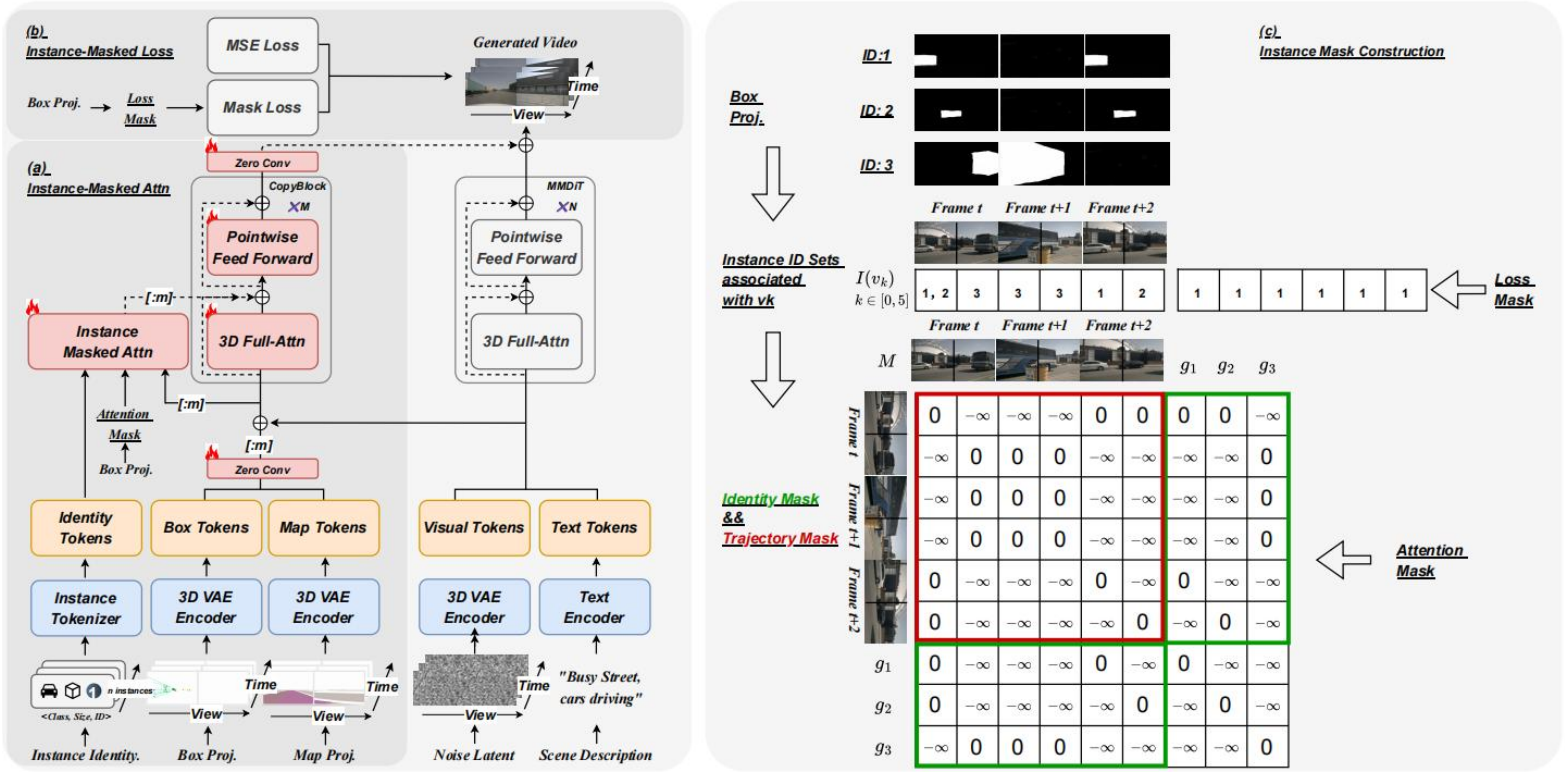
ConsisDrive:Dataset and Model Overview

Dataset: nuScenes open-source dataset
- 320 scenes, each lasting 30 seconds, recorded at 12 FPS

Model:
- Built upon OpenSora v2.0 framework
- Incorporates Multimodal Diffusion Transformer (MM-DiT) for video generation
- Use ControlNet for condition control

Input (Multi-modal Control Conditions):
- Numeric: Object depth information
- Image: 3D bounding box projection, road map projection, multi-object trajectories
- Text: Scene descriptions

Output:
- High-quality, multi-view, multi-frame driving scene video



ConsisDrive:Proposed Modules for Instance-level Temporal

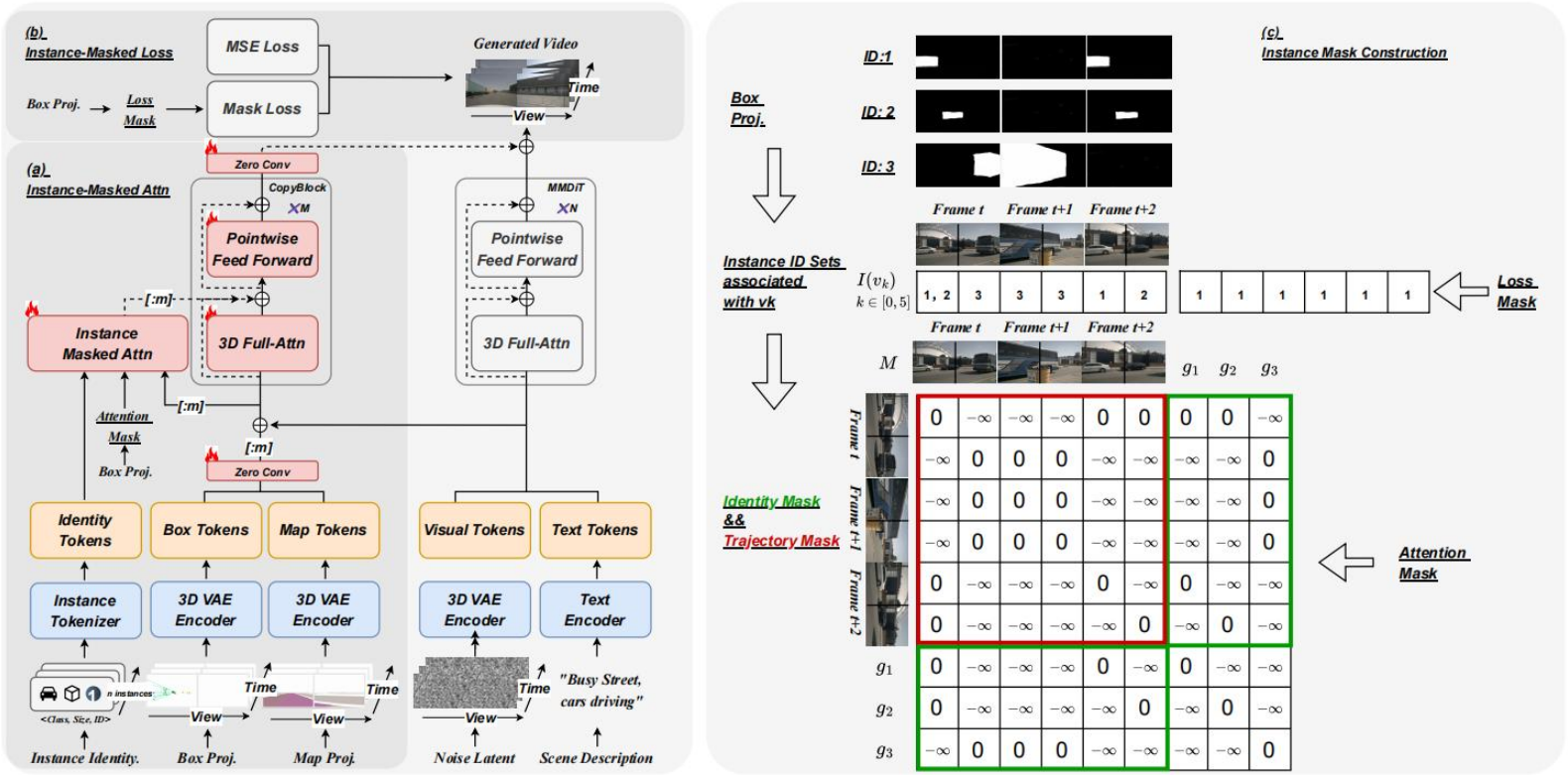
Instance Masked Attention Module (a)

- Ensures instance identity consistency:

Through the Instance Identity Mask, each visual token can only attend to the global identity condition token of the corresponding instance. This prevents identity feature leakage between different instances—for example, avoiding erroneous attention and confusion between features of a "bus" and a "car."

- Maintains temporal coherence:

Using the Instance Trajectory Mask, visual tokens of the same instance are allowed to attend to each other across different frames, while cross-instance attention between visual tokens is strictly prohibited. This enables smooth propagation of the instance's appearance features (color, texture, shape) along its motion trajectory over time, thereby avoiding frame-to-frame flickering or abrupt changes.

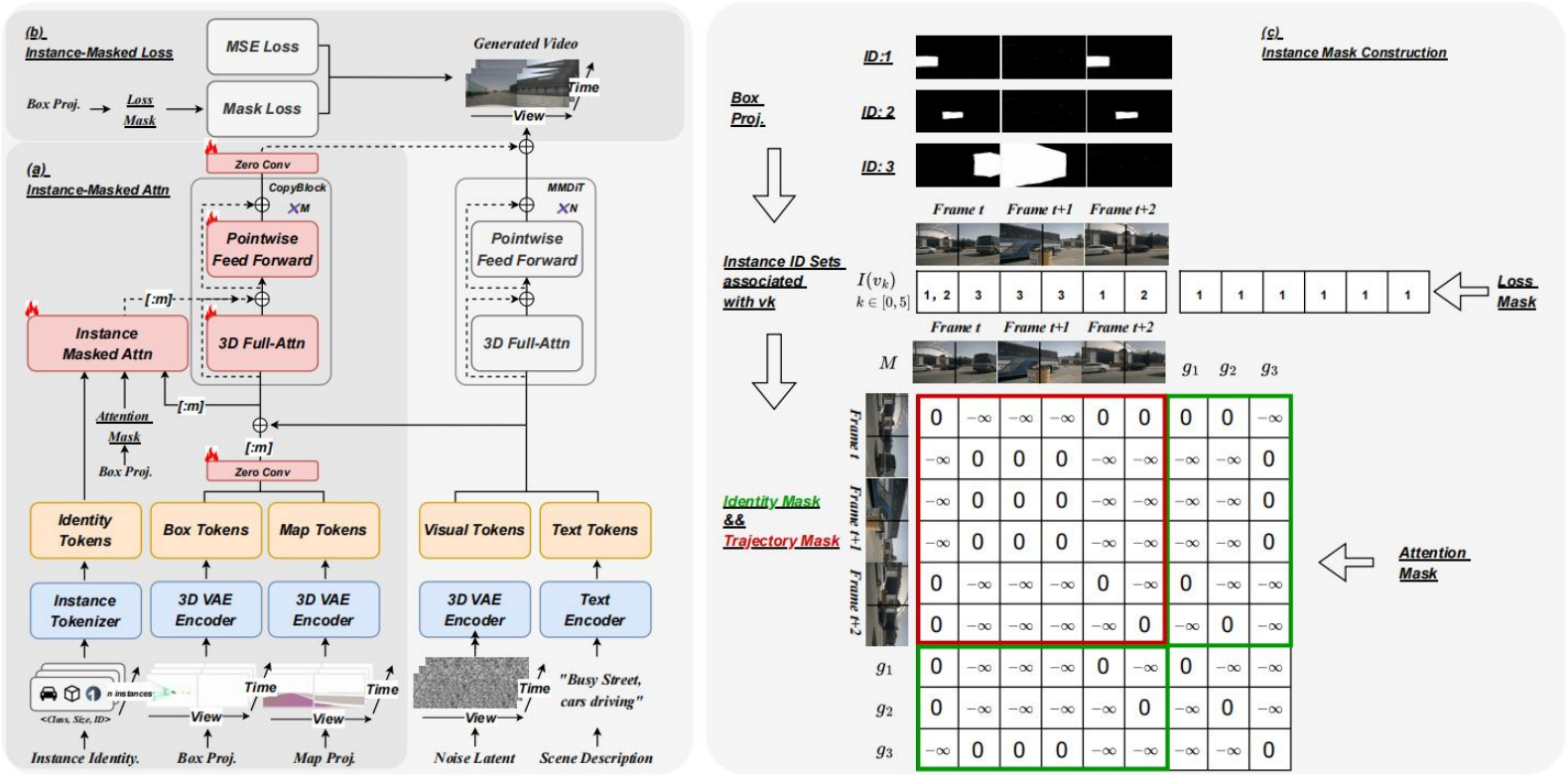


ConsisDrive:Proposed Modules for Instance-level Temporal

Instance Masked Loss Module (b)

- Ensures long-term consistency:

We propose the Instance Masked Loss module, which leverages a dynamic mask to emphasize supervision on foreground regions. This balances foreground temporal coherence with global scene fidelity.



Experimental Results

Results-Qualitative

Video Generation Quality:

- Instance-level

  Temporal Consistency:

  Class

- Instance-level

  Temporal Consistency:

  Color

Results-Quantitative

- FID / FVD :

Method	Multi-View	Multi-Frame	FVD↓	FID↓
BEVControl Yang et al. (2023)	✓		-	24.85
DrivingDiffusion Li et al. (2023)	✓	✓	332	15.83
Panacea Wen et al. (2024)	✓	✓	139	16.96
MagicDrive-V2 Gao et al. (2024b)	✓	✓	94.84	20.91
DriveScape Wu et al. (2024)	✓	✓	76.39	8.34
DriveDreamer2 Zhao et al. (2024)	✓	✓	55.7	11.2
DiVE Jiang et al. (2024)	✓	✓	94.6	-
DrivingSphere Yan et al. (2024)	✓	✓	103.42	-
UniScene Li et al. (2025)	✓	✓	70.52	6.12
InstaDrive Yang et al. (2025a)	✓	✓	38.06	3.96
UniMLVG Chen et al. (2025)	✓	✓	60.1	8.8
ConsisDrive	✓	✓	37.23	3.88

- Downstream Task Validation:
  - Multi-Object Tracking (MOT) — Evaluating Fine-Grained Temporal ConsistencyUsing the trained StreamPETR MOT model, we perform training on: Real-nuScenes train Dataset && Real + Generated-nuScenes train Dataset
    - Comparison is based on the IDS (Identity Switches) metric — fewer switches indicate stronger instance-level temporal consistency.

Method	Real	Gen.	AMOTA↑	AMOTP↓	IDS↓
Oracle	✓	-	0.289	1.419	687
DriveDreamer2	✓	✓	0.313	1.387	593 (-94)
InstaDrive	✓	✓	0.496	1.376	532 (-155)
ConsisDrive (Ours)	✓	✓	0.498	1.350	525 (-162)

  - Perception Task — Evaluating Fidelity
    - Using the pre-trained StreamPETR perception model, we compare: Real-nuScenes Val Dataset && Generated-nuScenes Val Dataset
      - Performance is assessed using the NDS (NuScenes Detection Score), reflecting improvements in fidelity.

Method	Real	Gen.	mAP↑	mAOE↓	mAVE↓	NDS↑
Oracle	✓	-	34.5	59.4	29.1	46.9
Panacea	-	✓	22.5 (65.22%)	72.7	46.9	36.1 (76.97%)
Panacea	✓	✓	37.1 (+2.6%)	54.2	27.3	49.2 (+2.3%)
ConsisDrive (Ours)	-	✓	31.5 (91.3%)	63.0	33.1	42.06 (89.68%)
ConsisDrive (Ours)	✓	✓	43.2 (+8.7%)	39.8	25.2	54.6 (+7.7%)

Results-Ablation Study

- Qualitative Analysis:
  - We conduct controlled experiments by selectively removing or keeping key modules
  - Visual comparisons show that removing any of these modules leads to noticeable degradation in temporal coherence or fidelity, confirming their complementary effects.

(a) w/o Identity Mask: Category Shift (b) w/o Trajectory Mask: Color Shift

Figure 3: Ablation study of the three key modules. **(a)** Removing the Identity Mask leads to incorrect instance category rendering, e.g., a traffic cone turns into a crouching pedestrian. **(b)** Removing the Trajectory Mask results in color shifts of the car.

About Author

Looking for a Job in 2026 Fall!

University of Science and Technology of China

Get Computer Science Master Degree in 26 Fall

Researcher & Engineer in Video Generation

Demo && Resume:

https://github.com/shanpoyang654/ConsisDrive

https://github.com/shanpoyang654

Resume

ConsisDrive

Thanks !