SimGen:
Simulator-conditioned Driving Scene Generation

(A 60 min Talk)

Online

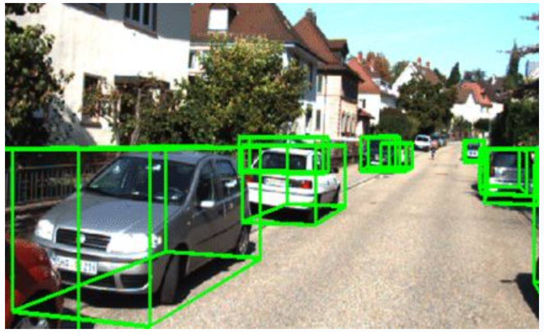
June 2024

Yunsong Zhou

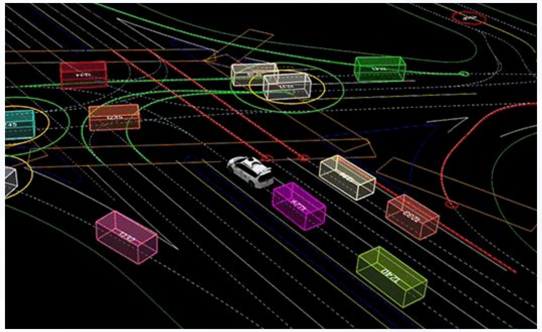
Problem setting | Autonomous Driving (AD) Tasks

Perception → Bounding boxes → Prediction → Waypoints → Planning → Trajectory

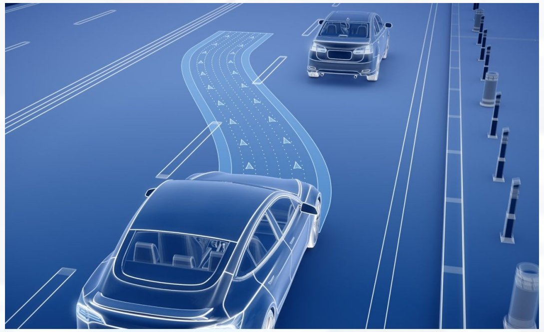
Challenge | Various weathers, illuminations, and scenarios



What are around?

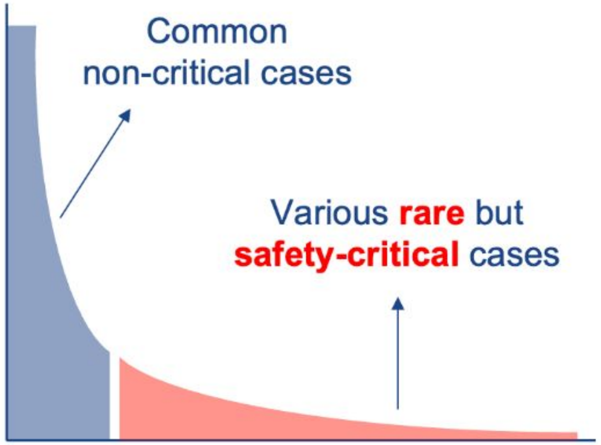


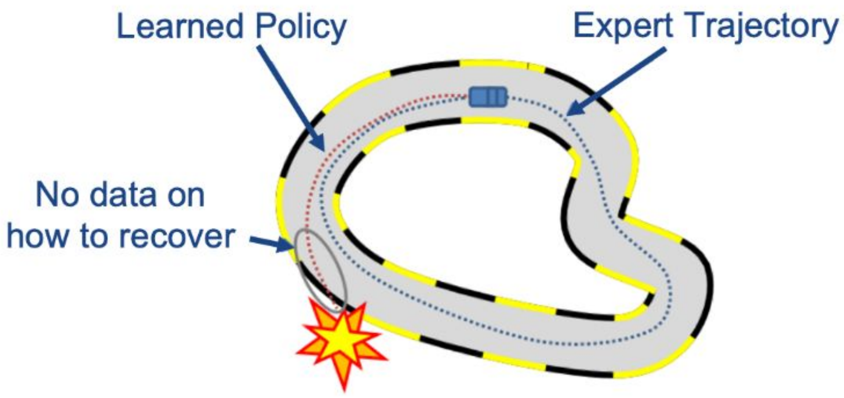
How will they go in the future?



Where should I go?

Challenge - Robustness and Generalization

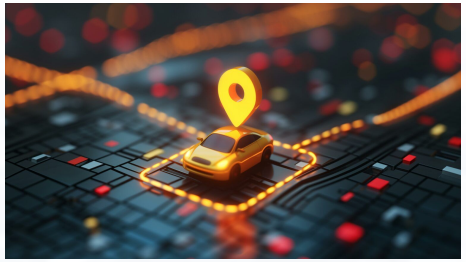




Source Domain	Target Domain
Simulator	Real world
Location A	Location B
Weather A	Weather B
Day	Night
Sensor A	Sensor B

(a) Long-tailed Distribution **(b) Covariate Shift** **(c) Domain Adaptation**

Motivation | Synthetic Data Generation for Driving

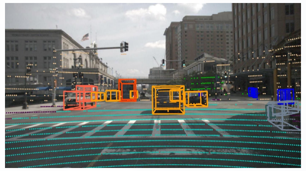


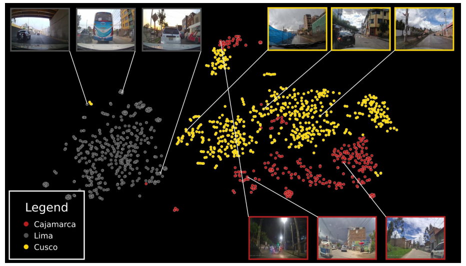
Real Data Collection

- Costly and laborious to collect and annotate the data
- Collecting data on dangerous driving can even pose a risk to life

Small-scale Dataset

Motional ARGO AI lyft





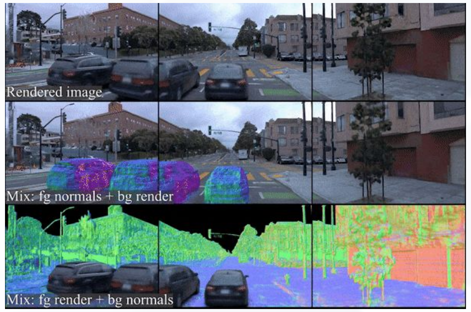
Manual Collection

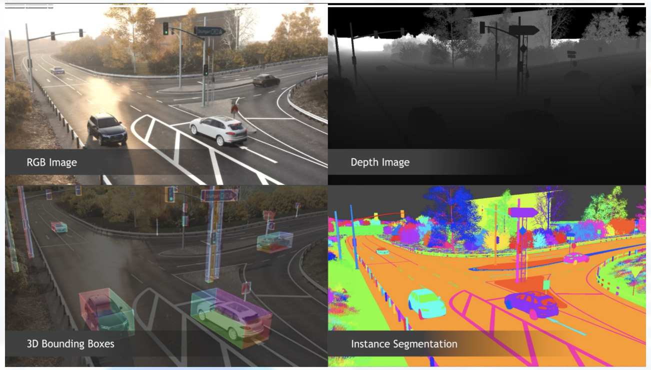
Credit to Seeing Machines

Synthetic Data Generation

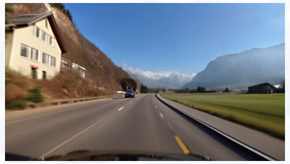
- A promising alternative to harvest annotated training data

Simulators



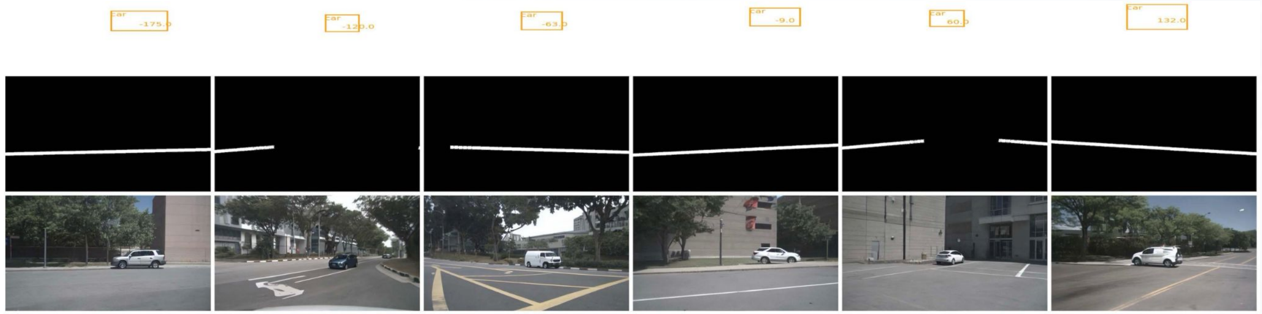


Generative Models



Credit to UniSim, Sora, GenAD

Trending in E2EAD | Synthetic Data Generation



Driving Scene Generation

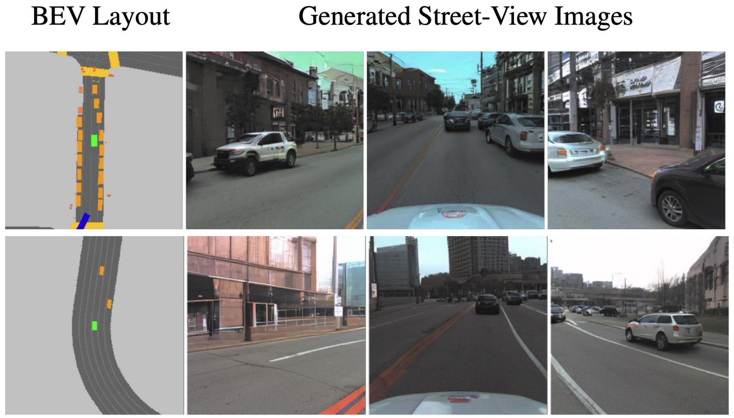
BEVControl — generate images from perspective layouts via diffusion models

BEVGen — generate realistic static images from layouts

2023.6 — Static

2023.12 — Multi-view

2024.6 — Temporal



Trending in E2EAD | Synthetic Data Generation

Panacea — first achieves temporal consistency

DriveDiffusion — best in data augmentation

GenAD — state-of-the-arts with highest video quality

Driving Scene Generation

BEVControl — generate images from perspective layouts via diffusion models

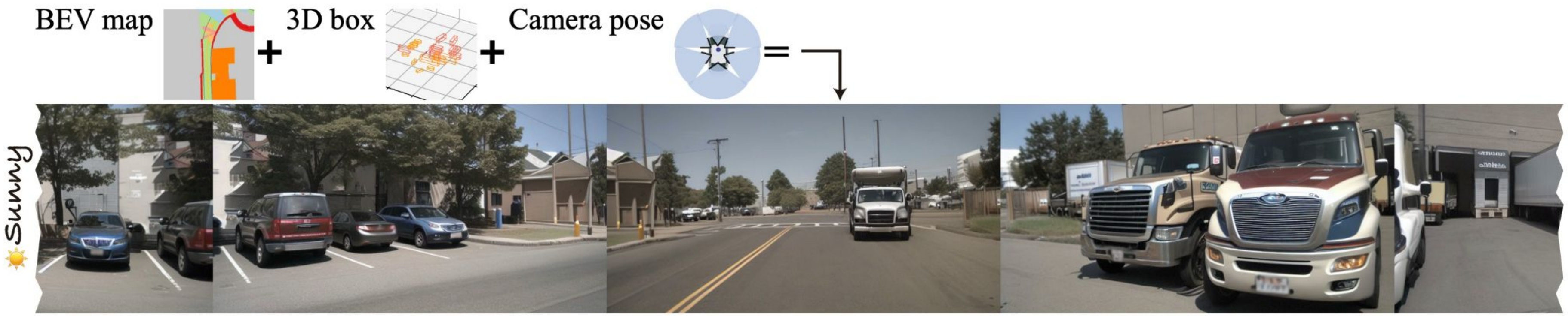
Static — 2023.6

Multi-view — 2023.12

Temporal — 2024.6

BEVGen — generate static images from BEV layouts

MagicDrive — generate multiview images from BEV maps



Trending in E2EAD | Synthetic Data Generation

Panacea — first achieves temporal consistency

DriveDiffusion — best in data augmentation

GenAD — state-of-the-arts with highest video quality

BEVControl — generate images from perspective layouts via diffusion models

Driving Scene Generation

Static | Multi-view | Temporal

BEVGen — generate static images from BEV layouts

2023.6

2023.12

2024.6

MagicDrive — generate multiview images from BEV maps

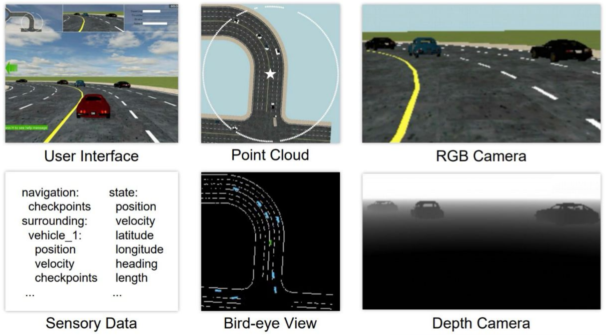
Drawbacks

- *Appearance diversity:* confined to learning on small-scale datasets with limited scenarios (*e.g.*, only urban streets or restricted weather conditions)
- *Layout diversity:* the behaviors are tedious and lack complex or safety-critical situations

Benefits

Realistic

Trending in E2EAD | Synthetic Data Generation



Generation via Simulators

2021

MetaDrive — composing driving scenarios for generalizable reinforcement learning

2022

CARLA — supporting development, training, and validation of autonomous driving systems

2023

Benefits

- *Layout diversity:* effortlessly generate scenes with various behaviors and provide accurate control over all objects

Drawbacks

- *Appearance diversity:* only contain a limited amount of 3D assets, and they lack a realistic visual appearance

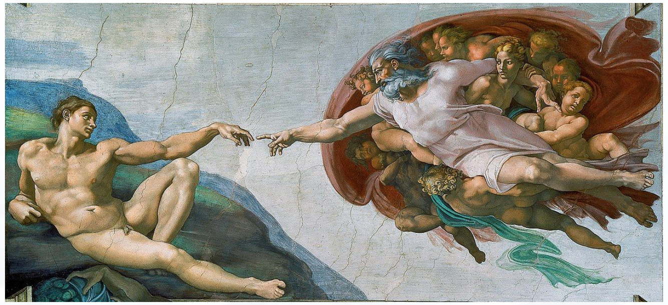
SimGen: Simulator-conditioned Driving Scene Generation

Yunsong Zhou[1,2] Michael Simon[1] Zhenghao Peng[1] Sicheng Mo[1]
Hongzi Zhu[2] Minyi Guo[2] Bolei Zhou[1†]

[1] University of California, Los Angeles [2] Shanghai Jiao Tong University

Credit to metadriverse.github.io/simgen

Insights | Simulator-conditioned Generative Model



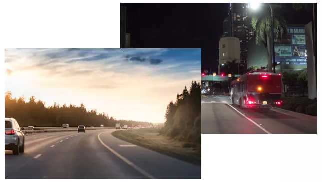
- We propose a *controllable* and *diverse* scene generation paradigm through the simulator-conditioned generative model, SimGen.
- It learns from real-world and simulated data and then generate diverse driving scenes based on the simulator's control conditions and rich text cues.

SimGen - The Big Picture

DIVA Dataset

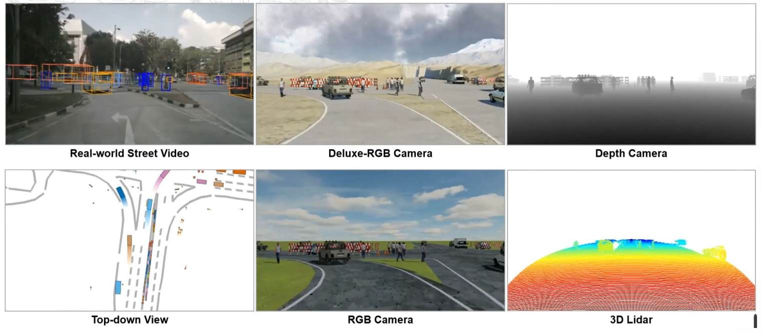
In-the-wild Driving Videos

YouTube

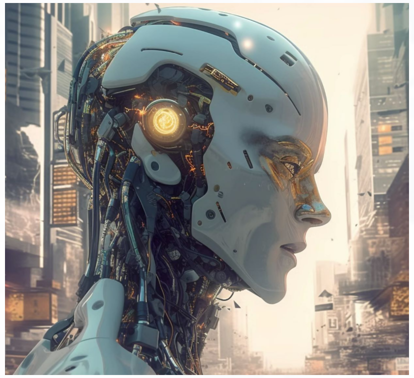


Virtual Data

METADRIVE



Simu-conditioned Model



Cascaded Diffusion Model
for autonomous driving

How to formulate?
Simulation-to-Reality (Sim2Real) Gaps?

Applications

Data Augmentation

Closed-loop Evaluation

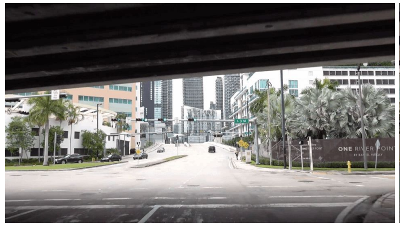
Partial photo by courtesy of online resources.

DIVA Dataset - Appearance and Layout Diversity

Comparisons

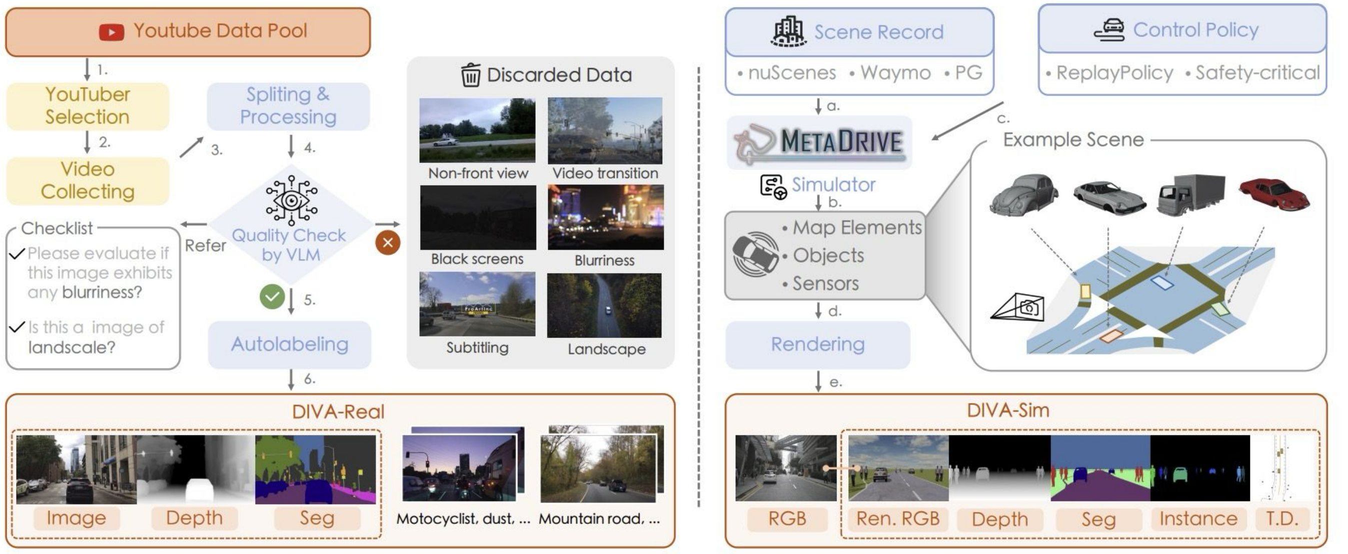
DIVA is the best on **scale**, **diversity**, and **annotations**

Dataset	Time (hours)	Frames	Cts.	Cities	Annotations Text	Depth	Seg.	Virt.
KITTI [18]	1.4	15k	1	1		✓	✓	
CityScapes [11]	0.5	25k	3	50			✓	
Waymo* [58]	11	390k	1	3			✓	
Argoverse 2* [67]	4.2	300k	1	6				
nuPlan* [7]	120	4.0M	2	4				
Honda-HAD [26]	32	1.2M	1	-	✓			
nuScenes [6]	5.5	241k	2	2			✓	
DIVA-Real	120	4.3M	19	71	✓	✓	✓	
DIVA-Sim	27.5^{+}	$998k^{+}$	3	5	✓	✓	✓	✓
DIVA (All)	**147.5**	**5.3M**	**22**	**76**	✓	✓	✓	✓



Construction

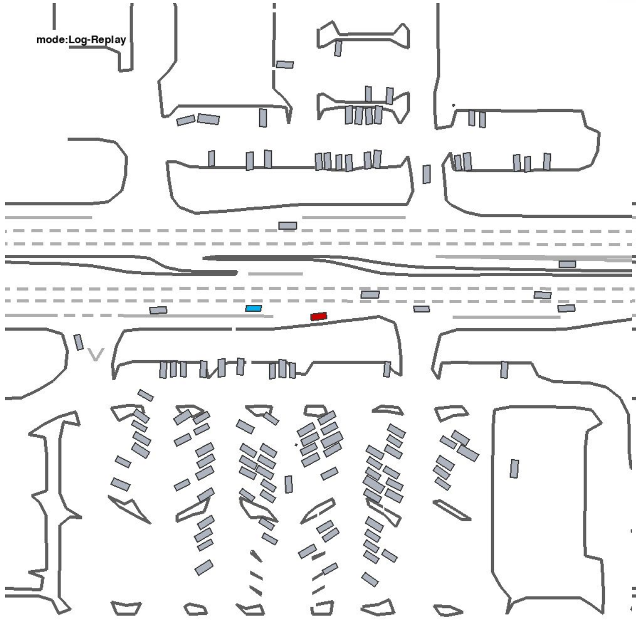
- Including **in-the-wild** and **virtual** driving videos
- Full auto labeling



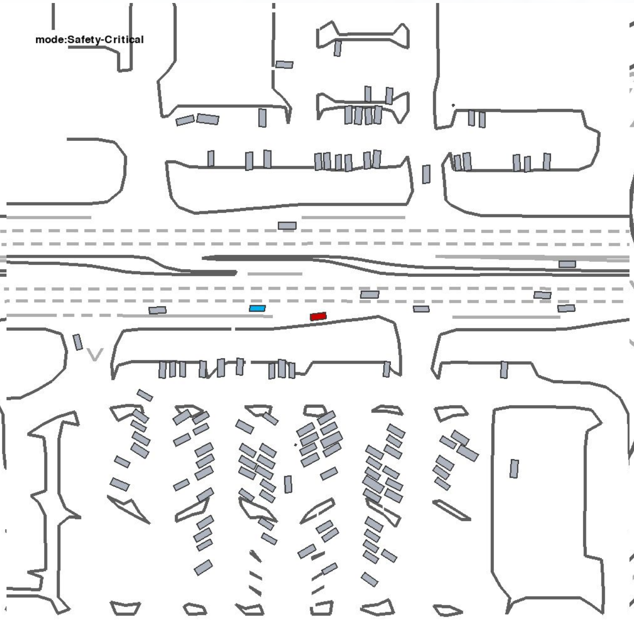
Examples

DIVA Dataset - Appearance and Layout Diversity

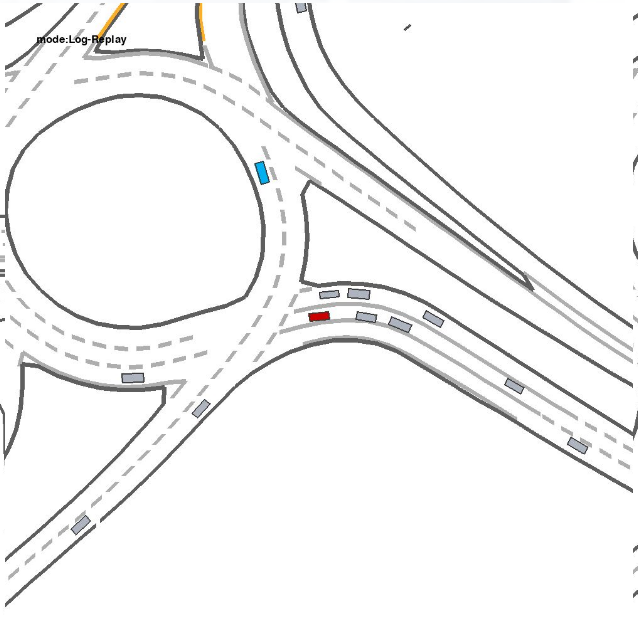
Examples of Generative Adversarial Scenarios



Log Replay



Safety-critical Scenarios



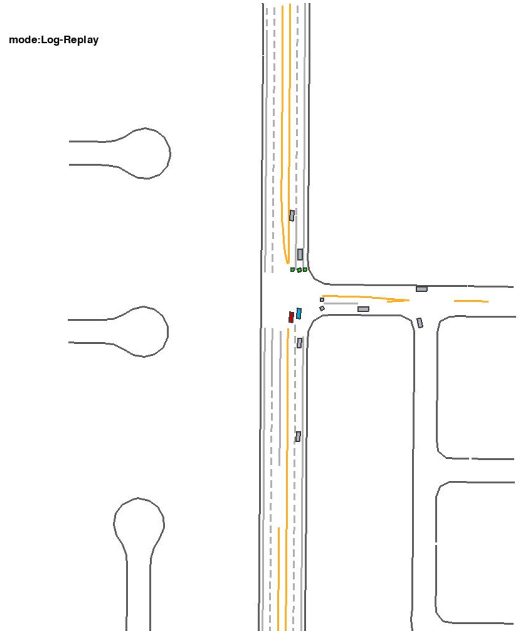
Log Replay

Safety-critical Scenarios

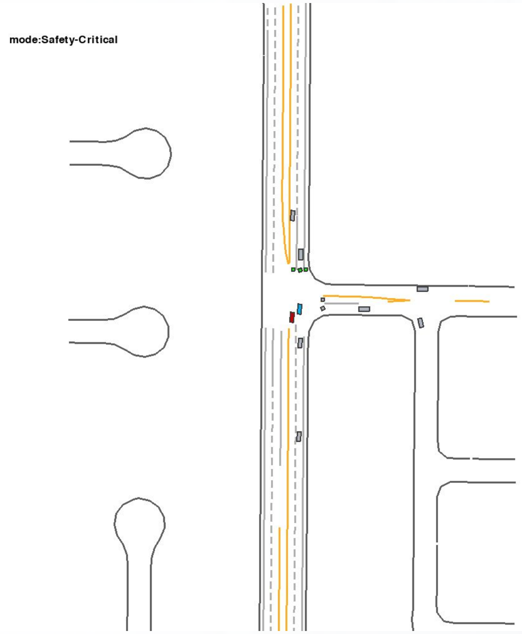
Credit to metadriverse.github.io/cat

DIVA Dataset - Appearance and Layout Diversity

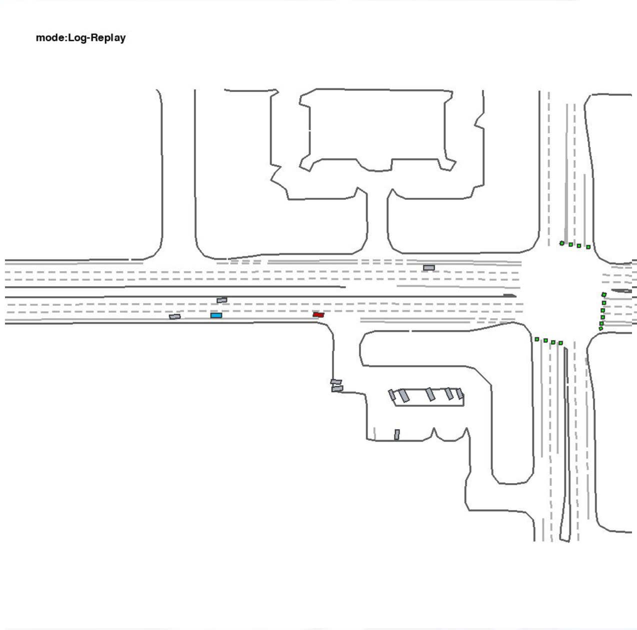
Examples of Generative Adversarial Scenarios



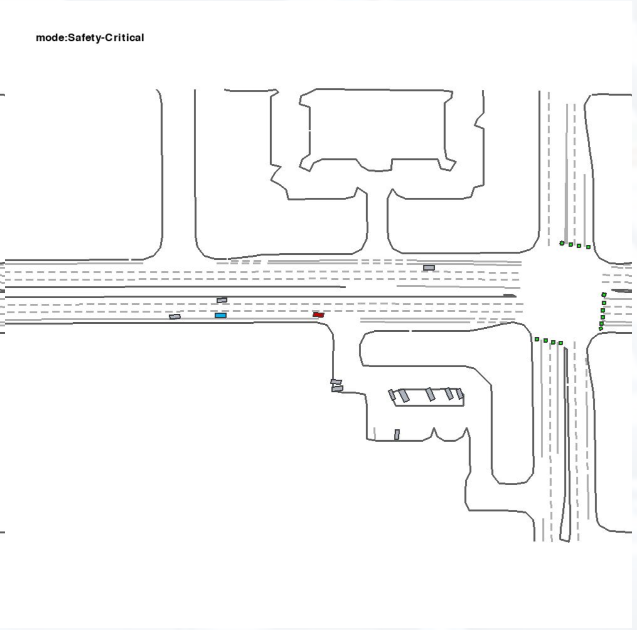
Log Replay



Safety-critical Scenarios



Log Replay



Safety-critical Scenarios

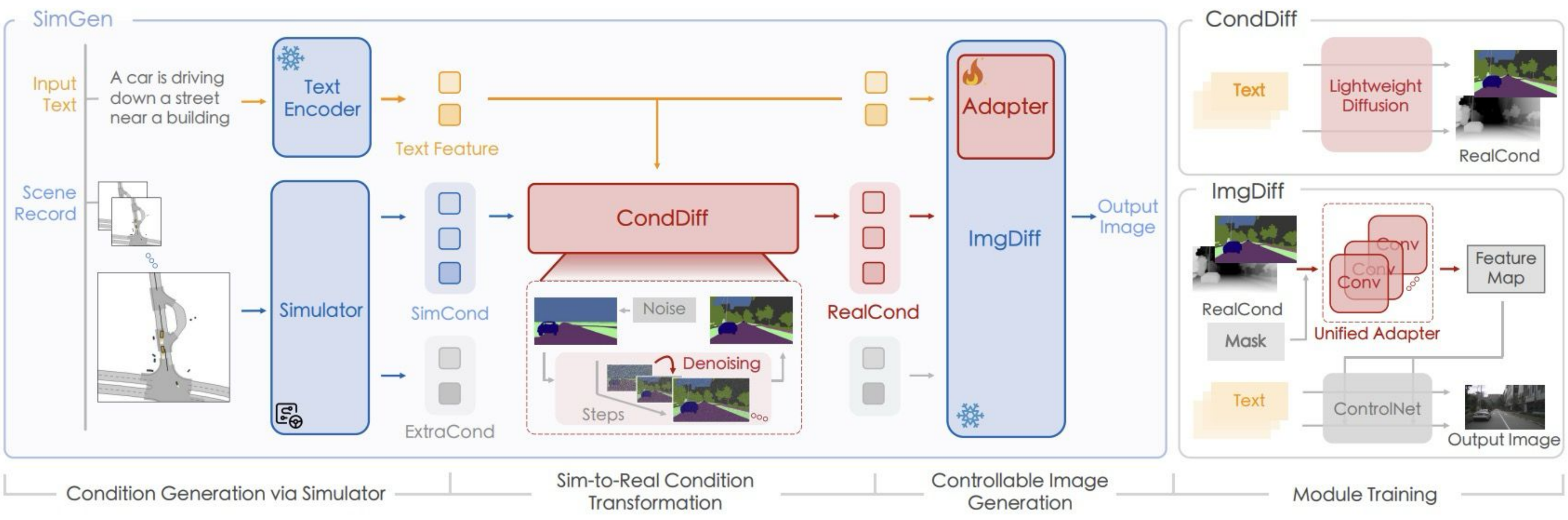
Credit to metadriverse.github.io/cat

SimGen - Overview

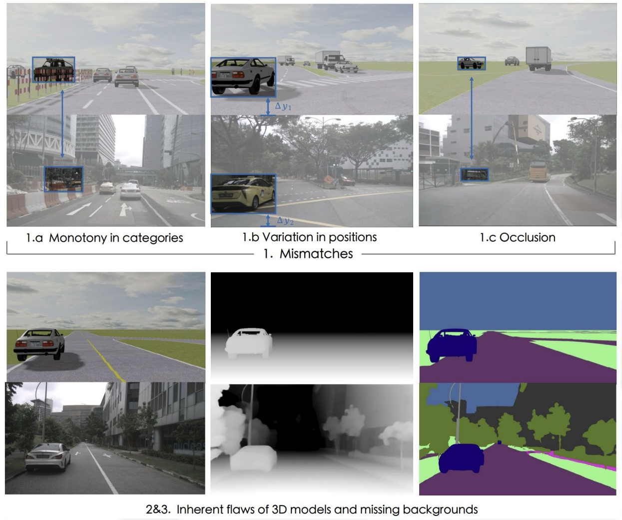
- Input: text and scene record
- Stage 1 (CondDiff): converts SimCond into RealCond, representing real depth and segmentation
- Stage 2 (ImgDiff): an Adapter merges multi-source conditions into a unified control condition and generates driving scene images.

Dataset	RealCond	SimCond	ExtraCond
nuScenes	✓		
DIVA-Real	✓		
DIVA-Sim		✓	✓

Real/SimCond: depth and segmentation; ExtraCond: rendered RGB, instance maps, and top-down views



Empirical Study



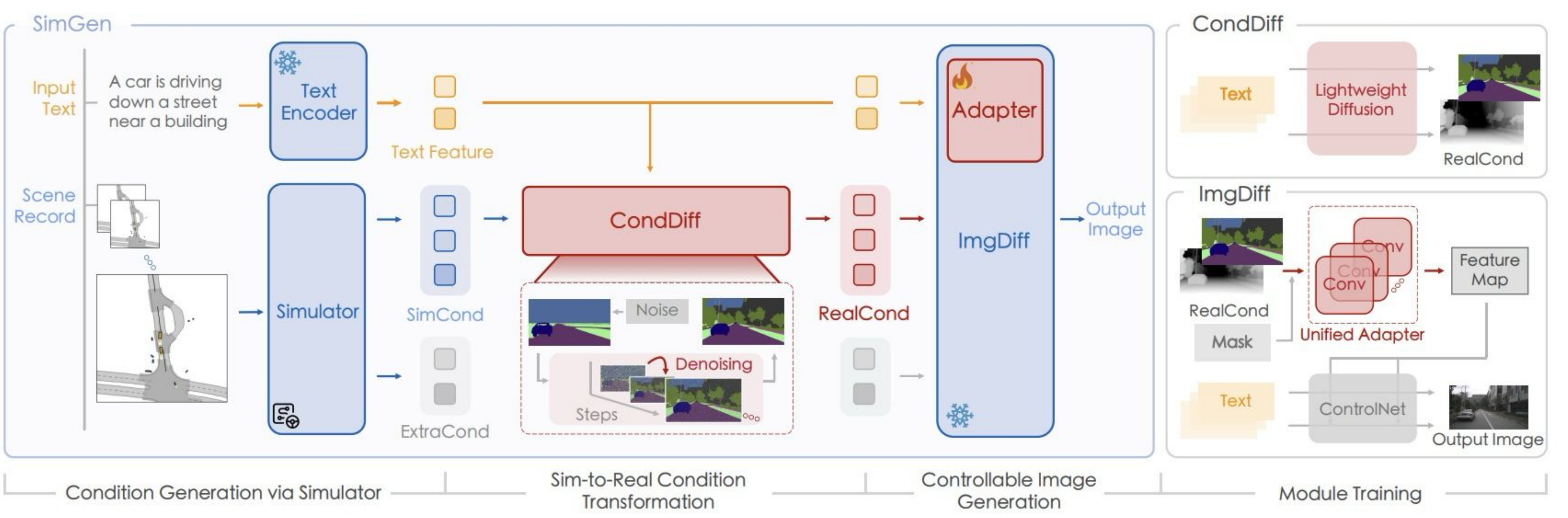
SimGen - Overview

CondDiff

- Naive approach: training a domain transfer model requires paired data far exceeding public datasets
- Ours: CondDiff injects noise-added SimCond into the intermediate sampling process and converts it into realistic conditions via continuous denoising

ImgDiff

- ExtraCond offers additional information but exists condition conflicts
- Ours: mapping variable conditions into fixed-length vectors and enabling a unified control input interface



Experiments

Quantitative Results

Quality — Diversity

Method	Dataset	FID↓	D_{pix} ↑
BEVGen [60]	nuScenes	25.5	17.0
BEVControl [72]		24.9	-
MagicDrive [17]		16.6	19.7
Panacea [65]		17.0	-
DrivingDiffusion [33]		15.9	20.1
SimGen-nuSc	nuScenes	**15.6**	20.5
SimGen	**DIVA**	**15.6**	**26.6**

Controllability

Method	Map Seg		Object Detection	
	$\text{mIoU}_{\text{Road}}$	$\text{mIoU}_{\text{Vehicle}}$	AP_{Car}	AP_{Truck}
Oracle	72.2	34.6	47.0	21.4
BEVGen [60]	50.1 (-21.1)	5.9 (-28.7)	24.7 (-22.3)	9.1 (-15.0)
MagicD. [17]	58.6 (-13.6)	29.5 (-5.1)	37.3 (-9.7)	17.3 (-4.1)
SimGen-nuSc	60.6 (-11.6)	29.9 (-4.7)	39.1 (-7.9)	18.1 (-3.3)
SimGen	**62.9** (-9.3)	**31.2** (-3.4)	**41.0** (-6.0)	**19.6** (-1.8)

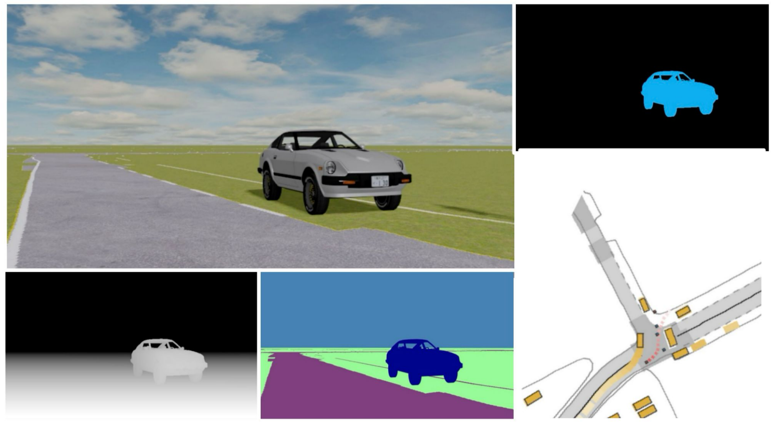
Applications on data augmentation

Method	Map Seg		Object Det	
	$\text{mIoU}_{\text{Road}}$	$\text{mIoU}_{\text{Vehi}}$	AP_{Car}	AP_{Truck}
Baseline	72.2	34.6	47.0	21.4
BEVGen [60]	71.9	34.2	47.3	21.1
MagicD. [17]	77.4	37.7	48.0	22.8
SimGen-nuSc	77.7	38.0	48.3	23.0
SimGen	**78.9**	**39.0**	**49.1**	**23.6**

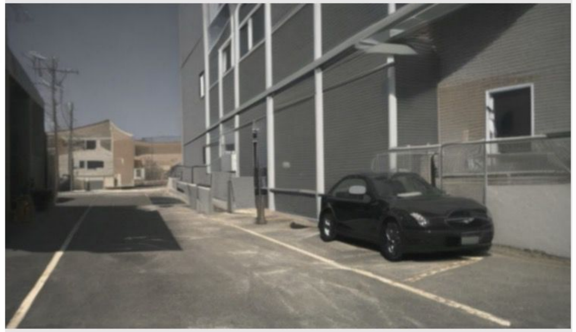
Experiments

Diverse Appearances

Conditions



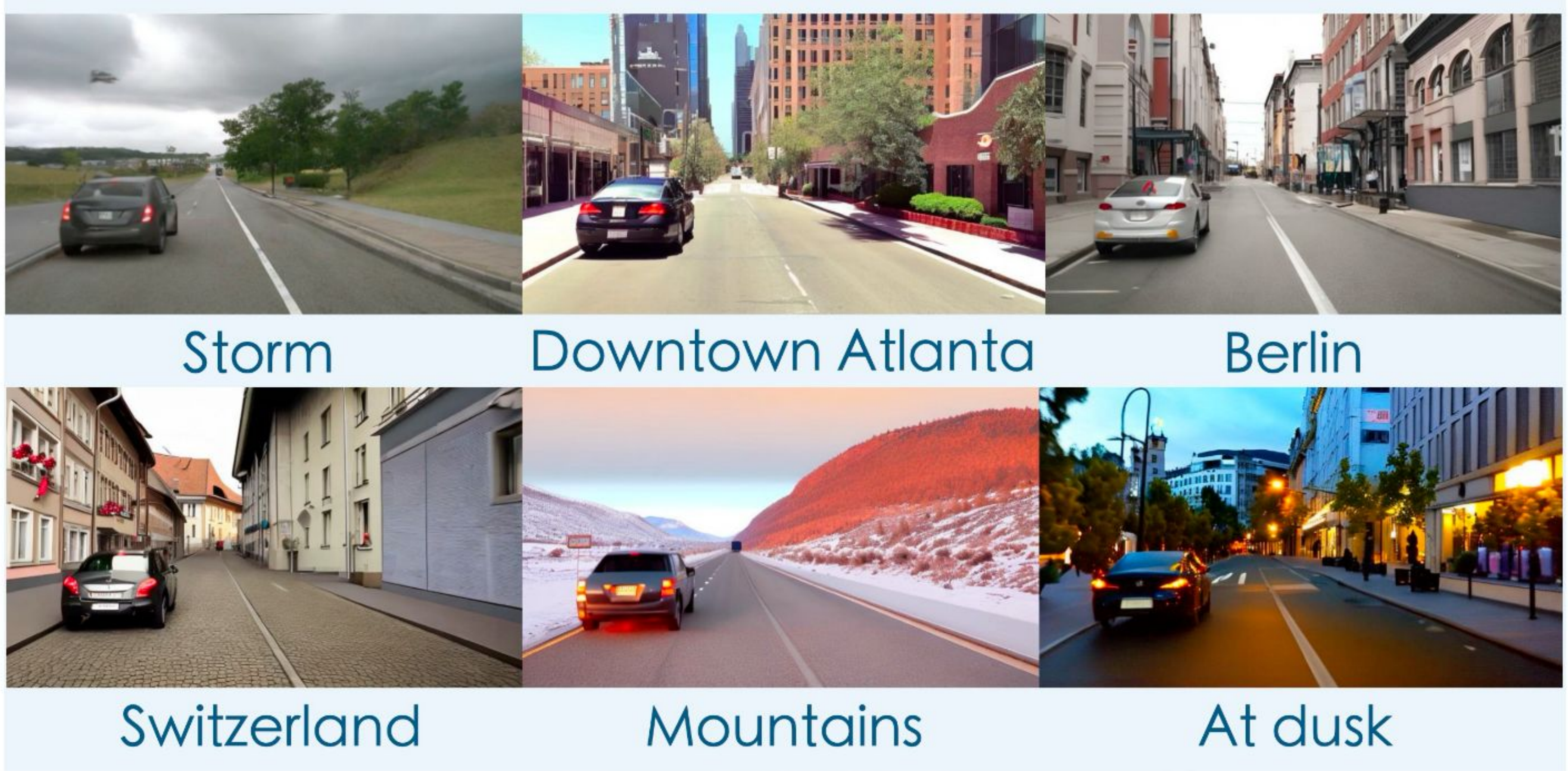
SimGen-nuSc



SimGen

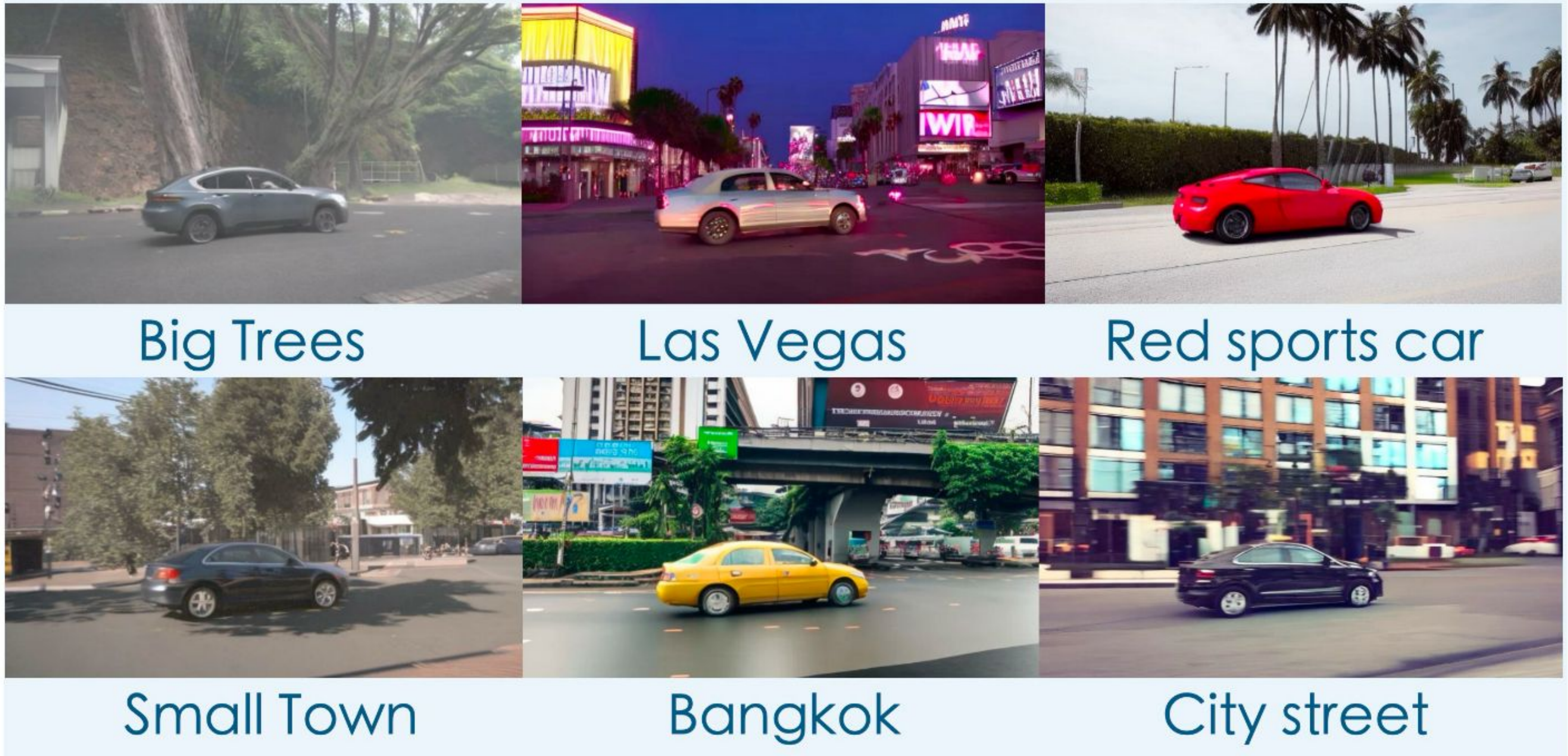
Experiments

Diverse Appearances



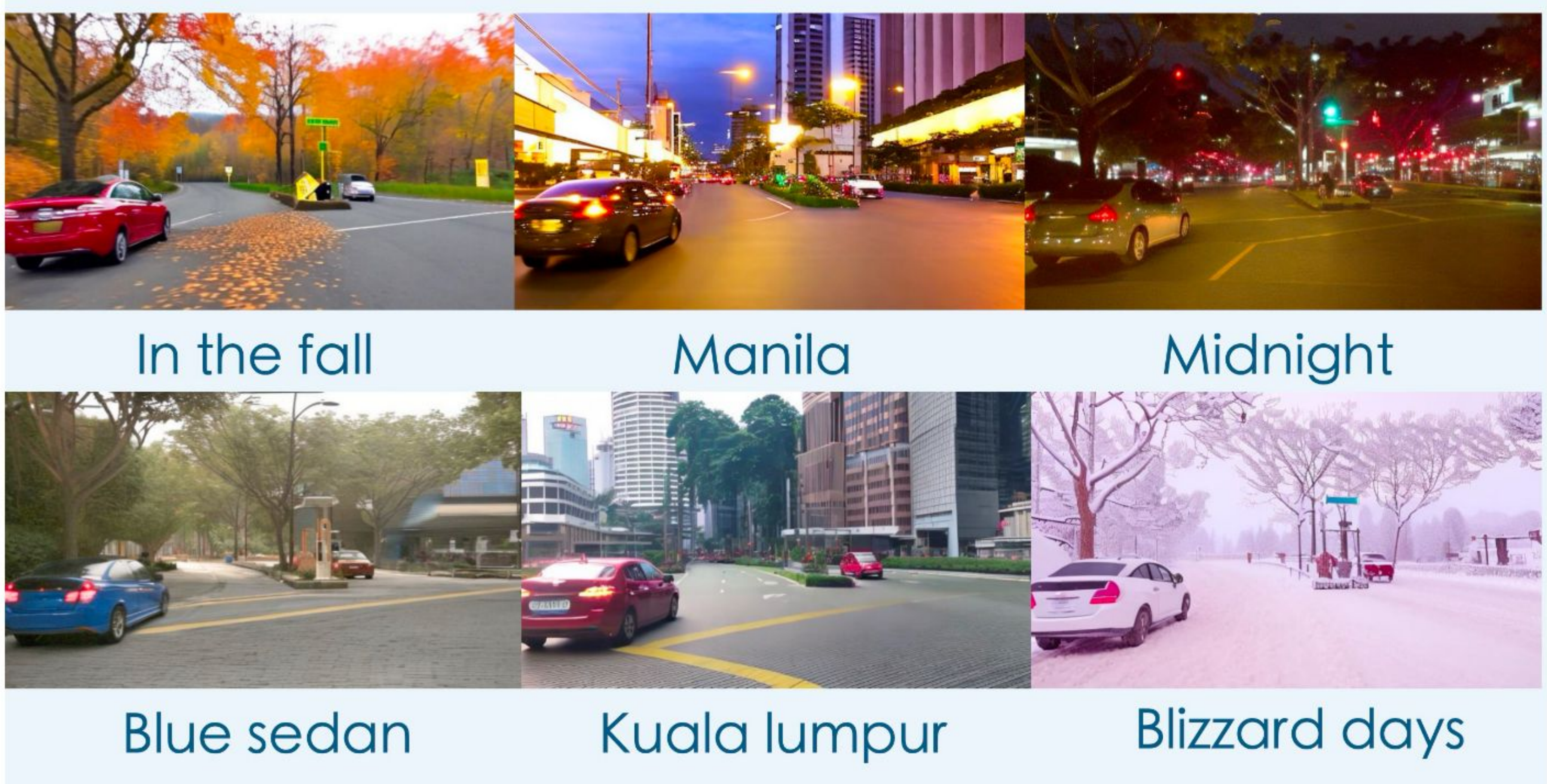
Experiments

Diverse Appearances



Experiments

Diverse Appearances



Experiments

Diverse Appearances

LEGO | Ukiyo-e | Minecraft | Super Mario

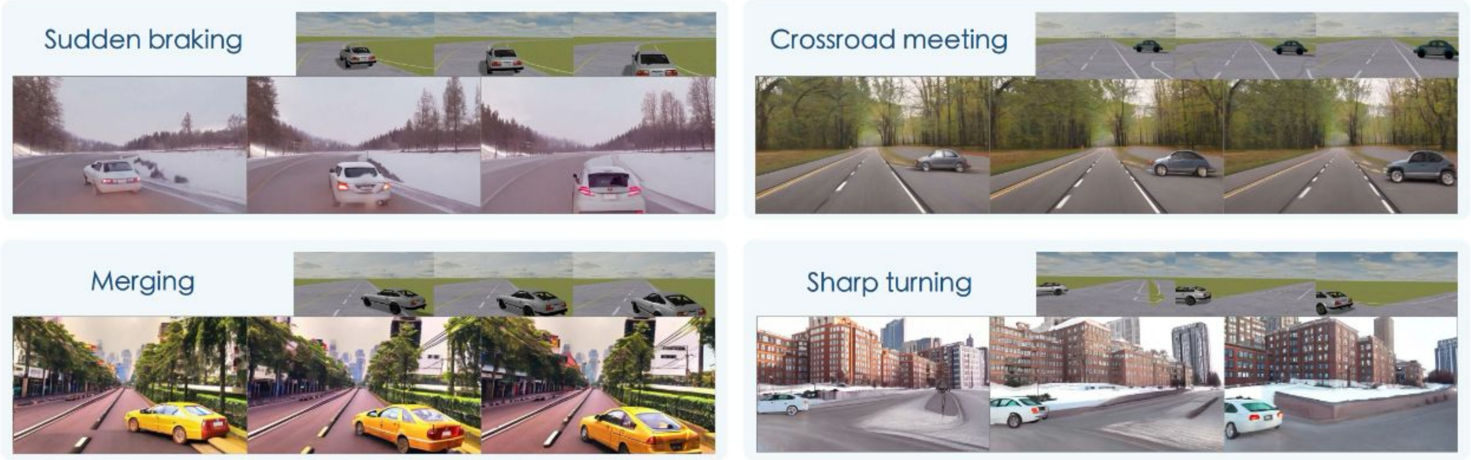
Experiments

Diverse Appearances

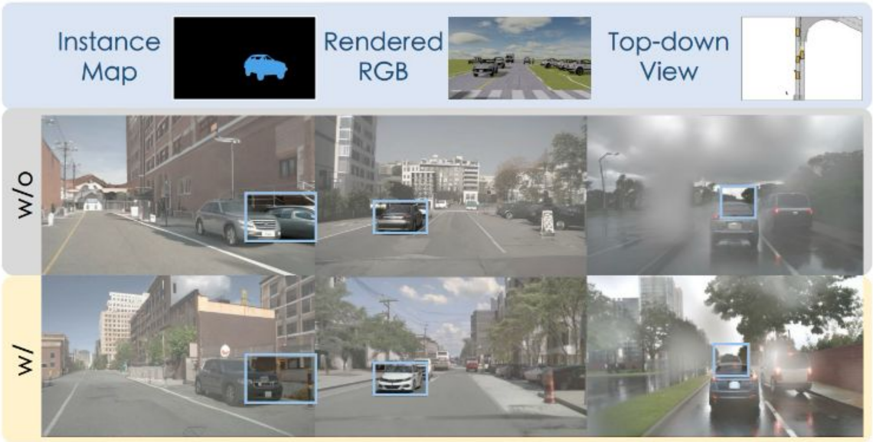
LEGO | Ukiyo-e | Minecraft | Super Mario

Experiments

Safety-critical Layouts



Efficiency of Simu-conditions



Applications on Closed-loop Evaluation

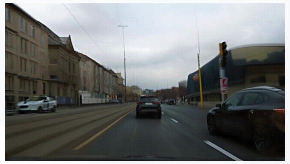
Conclusions

Grab-and-go

- A simulator-conditioned diffusion model, SimGen, that learns to generate diverse driving scenarios by mixing data from the simulator and the web.
- A novel dataset containing massive web and simulated driving videos that ensure diverse scene generation and advanced simulation-to-reality research is collected.

Limitations

- SimGen is not designed for video generation.



- SimGen does not cope with common settings such as multi-view generation.

- Inheriting the drawbacks of diffusion models, SimGen suffers from long inference time, which may impact the applications like closed-loop training.

END