

From Multimodal Generative Models to Dynamic World Modeling

Ziwei Liu 刘子纬
Nanyang Technological University

Presented by Haozhe Xie 谢浩哲

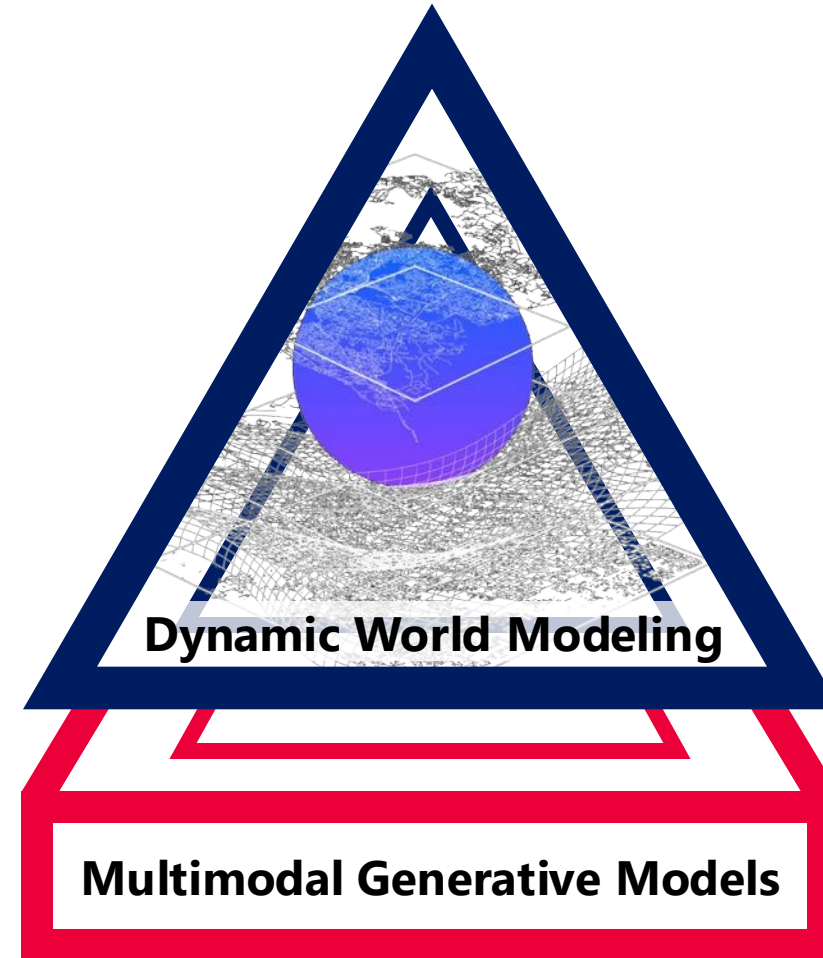


Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model
Dynamic Scenes



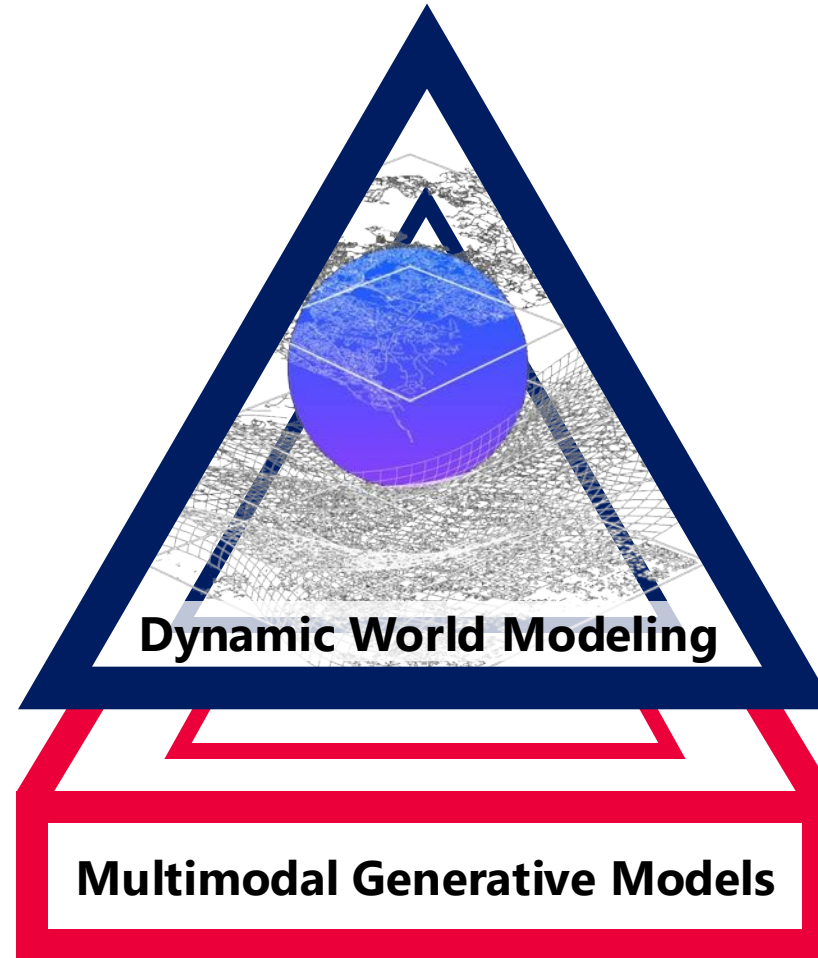
Be Social

How to Model Social
Interactions

Be Physical

How to Model Material and Illumination

Be Dynamic
How to Model
Dynamic Scenes



Be Social
How to Model Social
Interactions

Be Physical: 3DTopia-XL



3DTopia-XL: High-Quality 3D PBR Asset Generation via Primitive Diffusion

Zhaoxi Chen, Jiayang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, Liang Pan, Dahua Lin, Ziwei Liu

CVPR 2025 Highlight

Challenges

- High-resolution Generative 3D Representation
 - **Parameter-efficient**
 - Surface-only
 - As compact as possible
 - **Scalable Tokenization**
 - Rapid tensorization from input
 - Reversible conversion to GLB mesh
 - **Differentiable Rendering**
- Modelling of Physical Light Transport
 - Well-defined Geometry
 - PBR (Physically Based Rendering) Materials



Previous SOTA



Our Goal

3DTopia-XL: A Native 3D Diffusion Model for PBR Asset

"A cute unicorn"

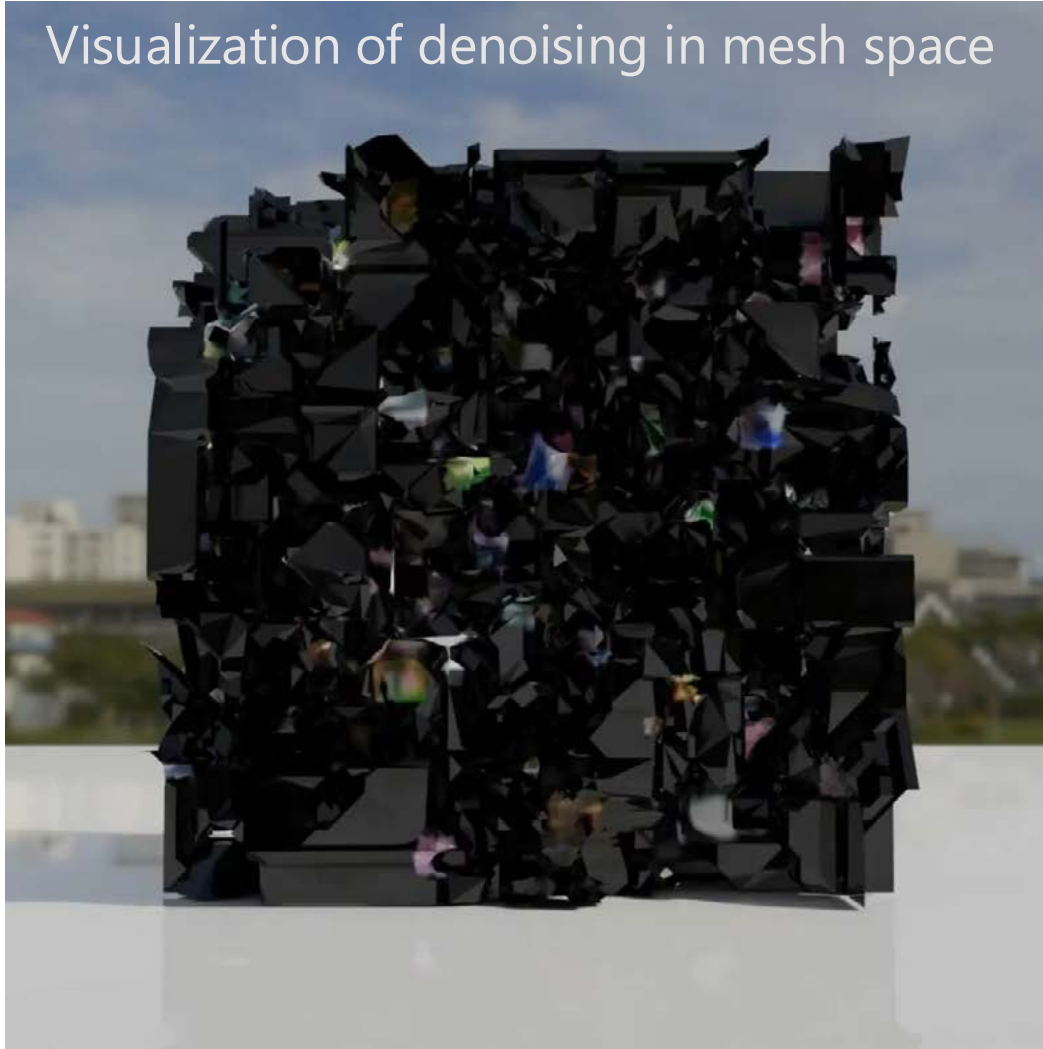


A Single Image / Texts

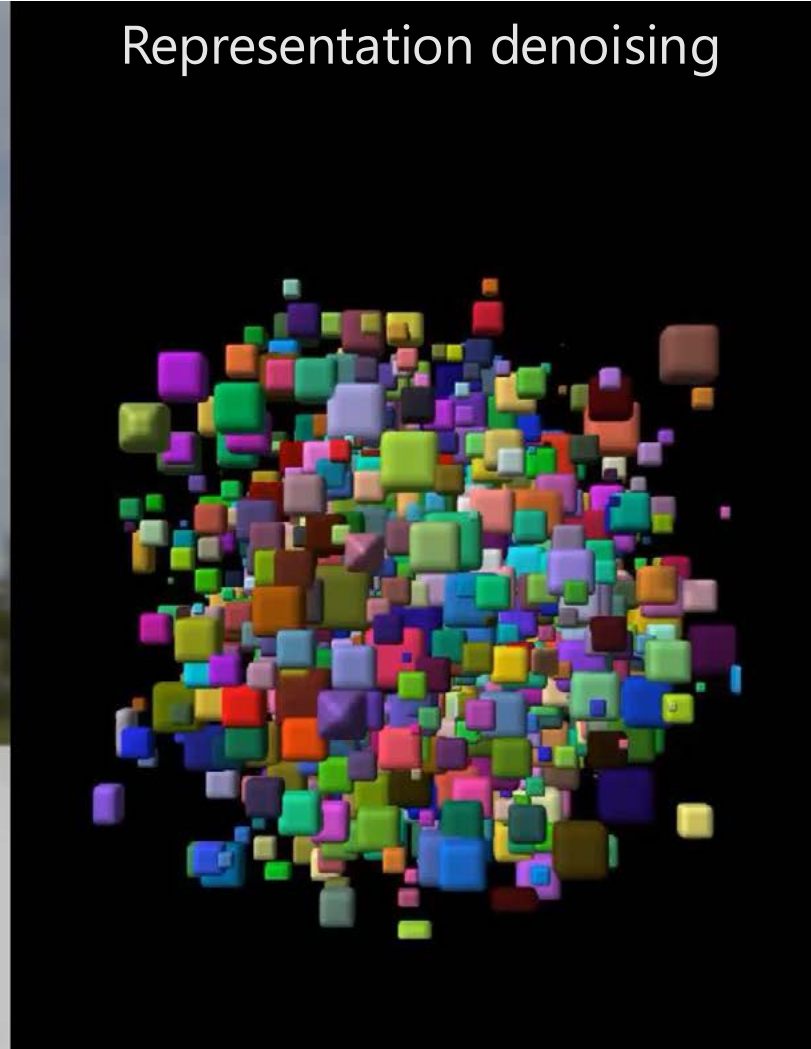
High-quality 3D Asset Ready for Blender 

Key Idea: Primitive Diffusion

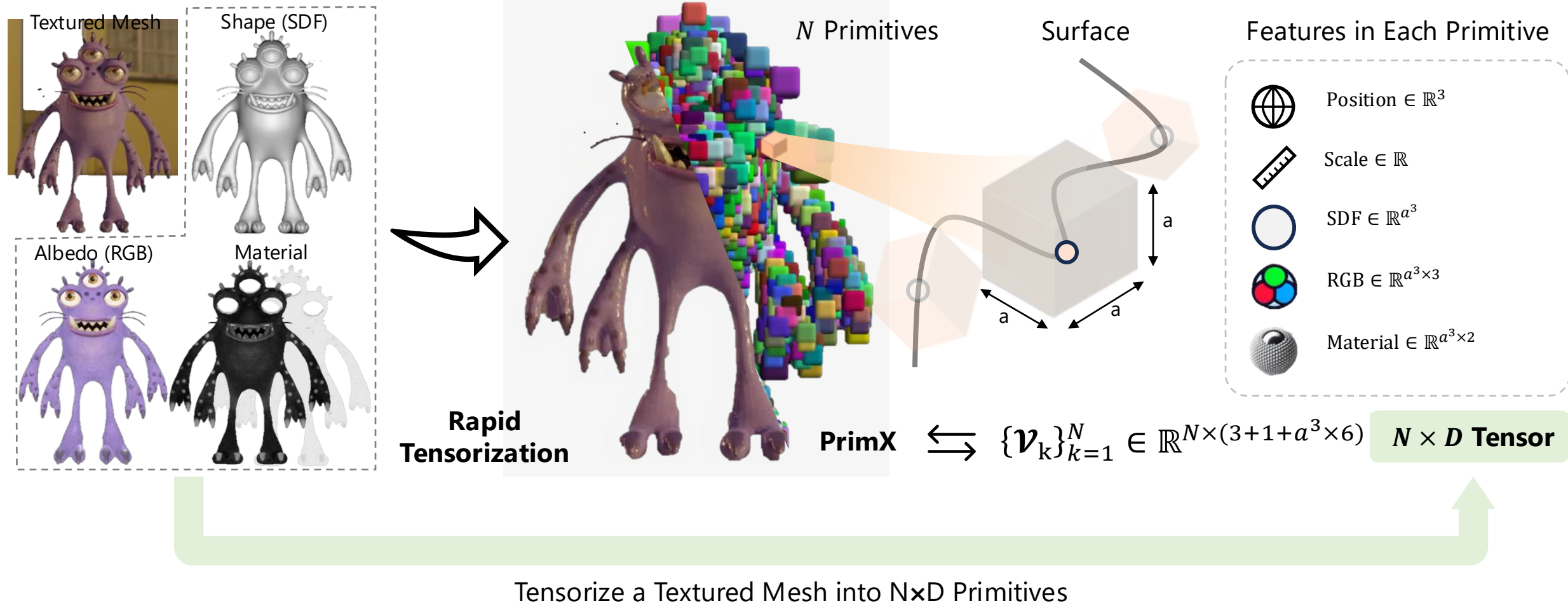
Visualization of denoising in mesh space



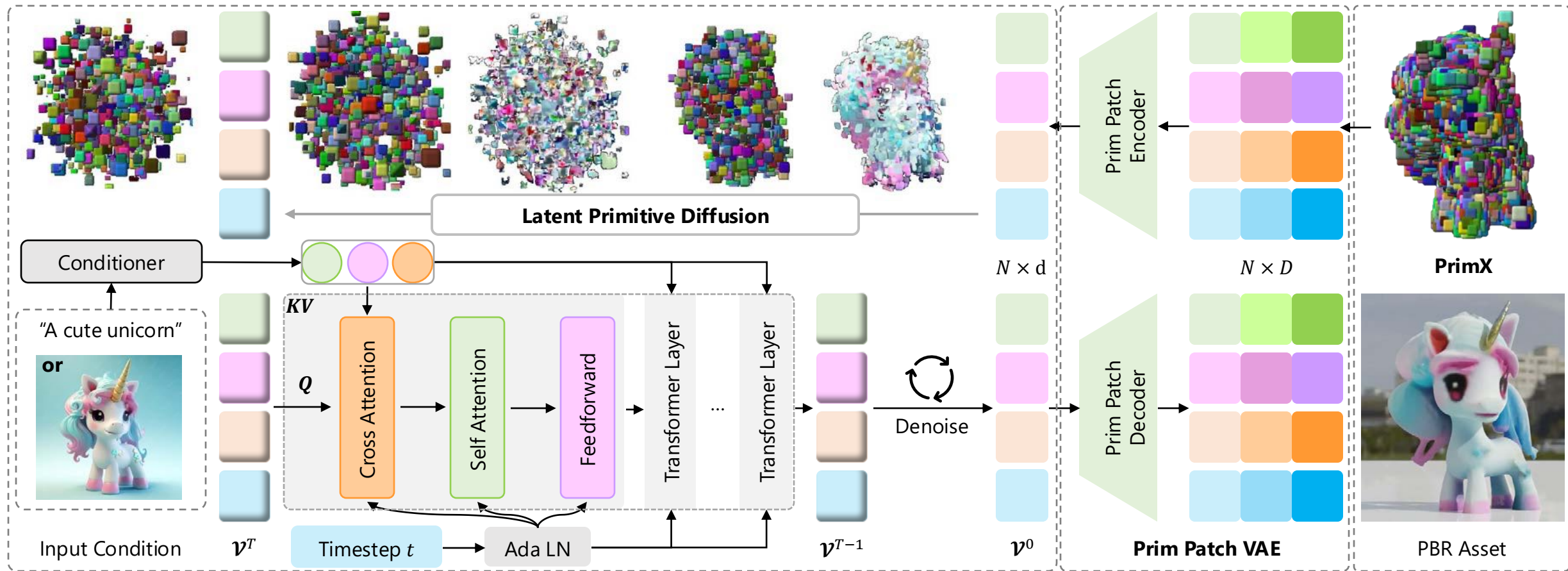
Representation denoising



Stage I: Geometry, Texture, Materials into NxD Primitives



Stage II: Latent Primitive Diffusion



Gallery: Denoising in 5 Seconds



Gallery: Ready for Graphics Engines



Be Physical: Neural LightRig

 [ZexinHe/Neural-LightRig](https://github.com/ZexinHe/Neural-LightRig)

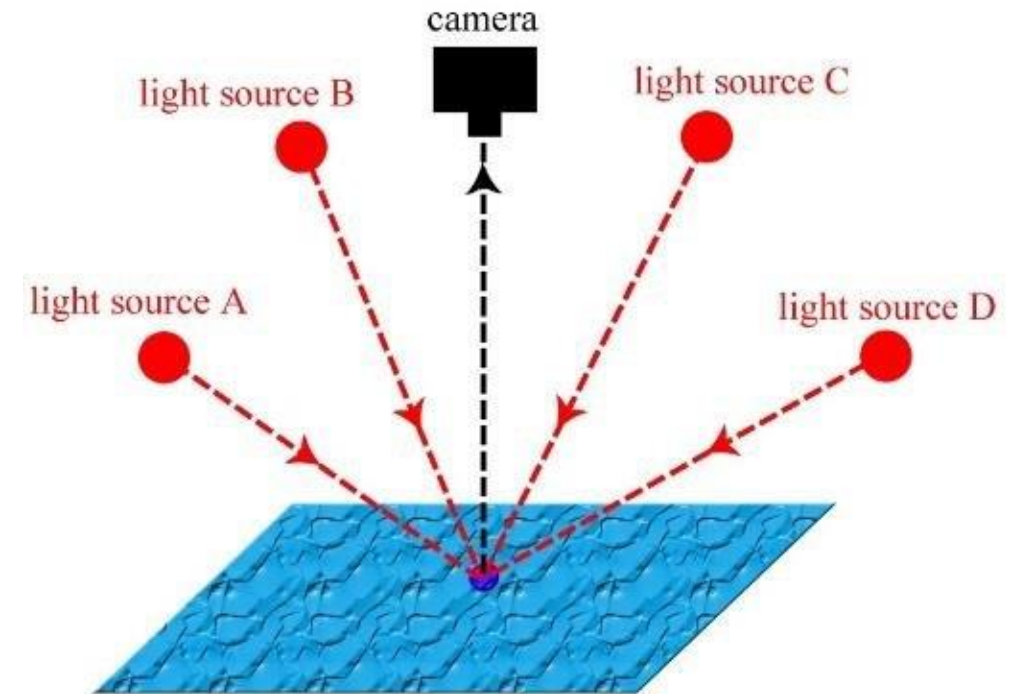
Neural LightRig: Unlocking Accurate Object Normal and Material Estimation with Multi-Light Diffusion

Zexin He, Tengfei Wang, Xin Huang, Xingang Pan, Ziwei Liu

CVPR 2025

A Long-Standing Challenge – Inverse Rendering

- Estimating geometry & materials from a single image is **ill-posed** and **under-constraint**
- Complex interaction among geometry, materials, and environmental lighting
- Traditional methods need photometric stereo setups^[1] – **impractical** for in-the-wild images

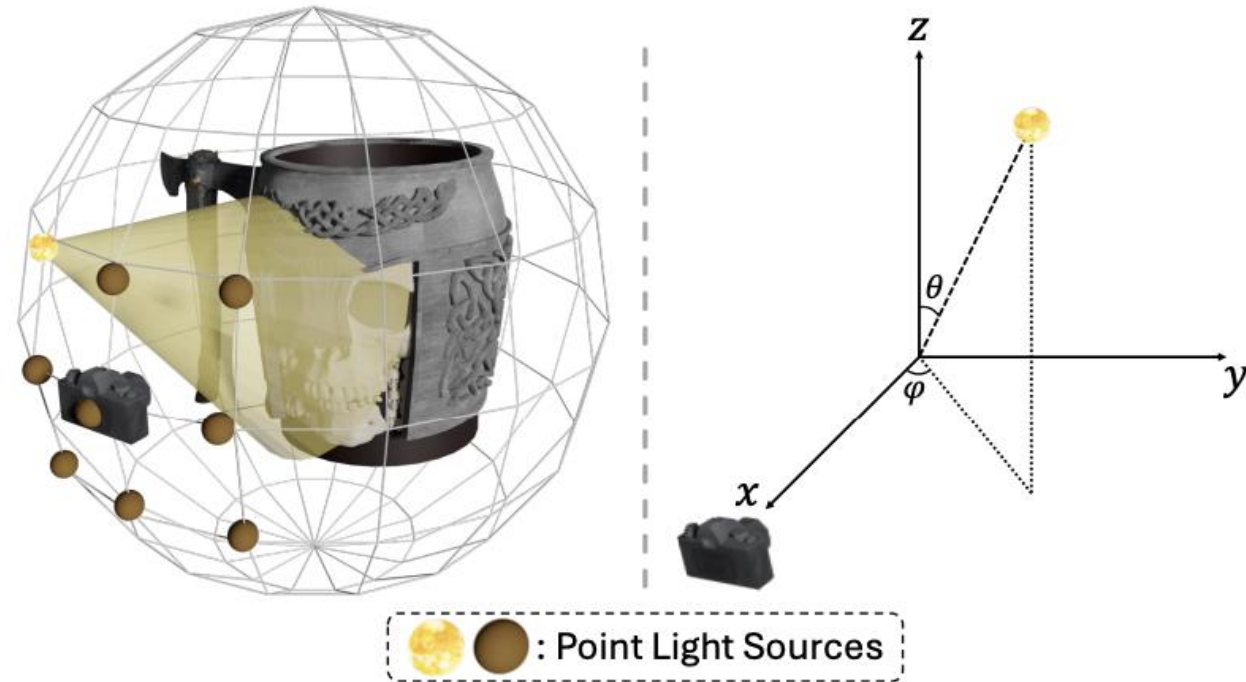


[1] Robert J. Woodham. *Photometric method for determining surface orientation from multiple images*. 1989.

[2] Image source: https://www.researchgate.net/profile/Lyndon-Smith-4/publication/325473321/figure/fig1/AS:666789923020804@1535986514936/The-principle-of-photometric-stereo-which-employs-a-single-camera-to-capture-multiple_W640.jpg.

Insights

- Diffusion models can generate consistent multi-view images^[1]
- Relighting diffusion models can synthesize images under various lighting conditions^[2]
- Relit images reveal different aspects of geometry & material – **reducing ambiguity**

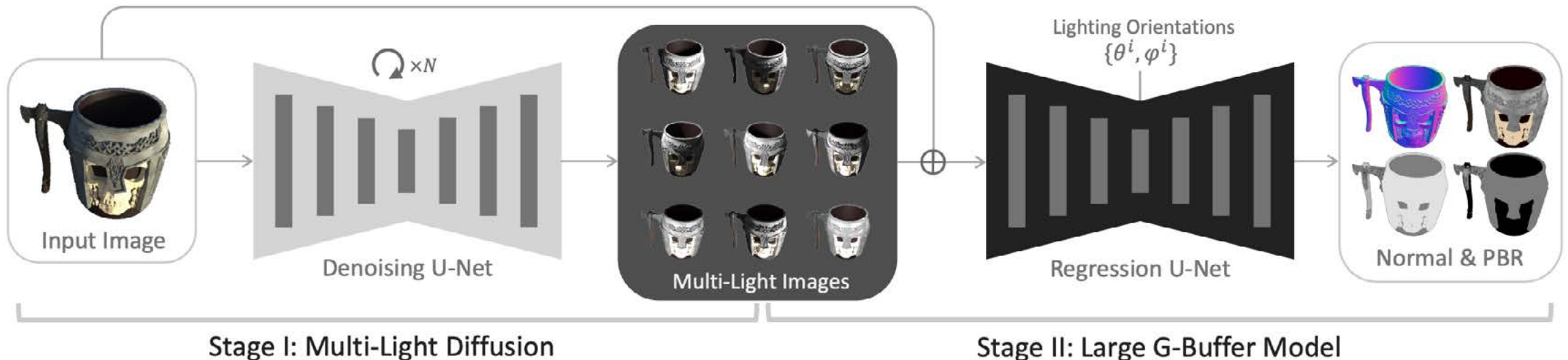


[1] Ruoxi Shi, et al. *Zero123++: A single image to consistent multi-view diffusion base model*. 2023.

[2] Lvmin Zhang, et al. *Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport*. 2025.

Methodology

- Multi-Light Diffusion
 - Fine-tuning a pre-trained image diffusion model to generate consistent relit images
 - These multi-light images enrich information and reduce the inherent uncertainty
- Large G-Buffer Reconstruction
 - Feed-forward regression U-Net to estimate geometry and PBR materials



Quantitative Evaluations

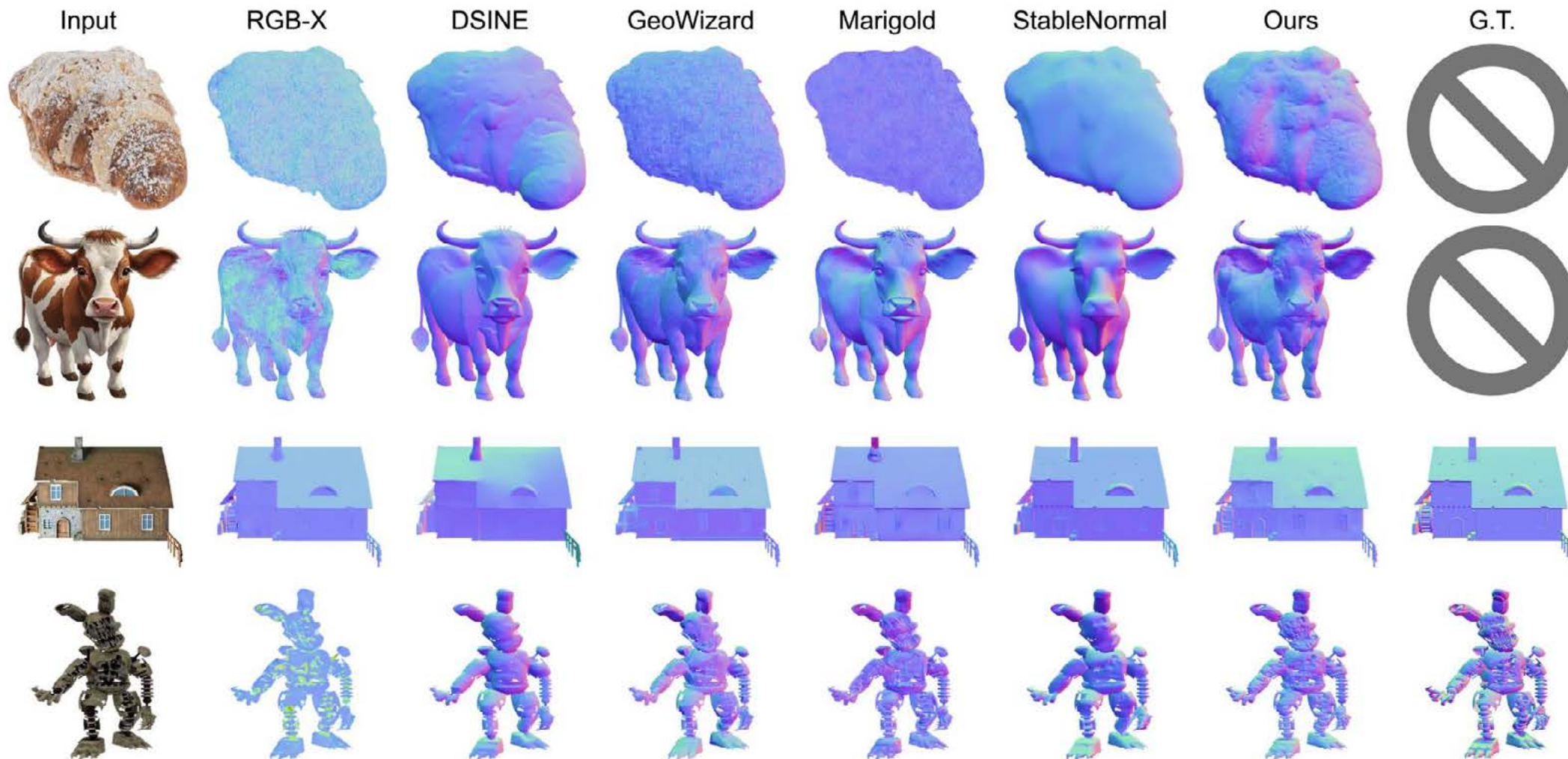
■ Surface Normal Estimation

Method	Mean ↓	Median ↓	3° ↑	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑
RGB↔X [57]	14.847	13.704	11.676	23.073	35.196	49.829	75.777	86.348
DSINE [2]	9.161	7.457	23.565	41.751	57.596	72.003	90.294	95.297
GeoWizard [16]	8.455	6.926	22.245	40.993	58.457	74.916	93.315	97.162
Marigold [25]	8.652	7.078	25.219	42.289	58.062	72.873	92.326	96.742
StableNormal [53]	8.034	6.568	21.393	43.917	63.740	78.568	93.671	96.785
Ours	6.413	4.897	38.656	56.780	70.938	82.853	95.412	98.063

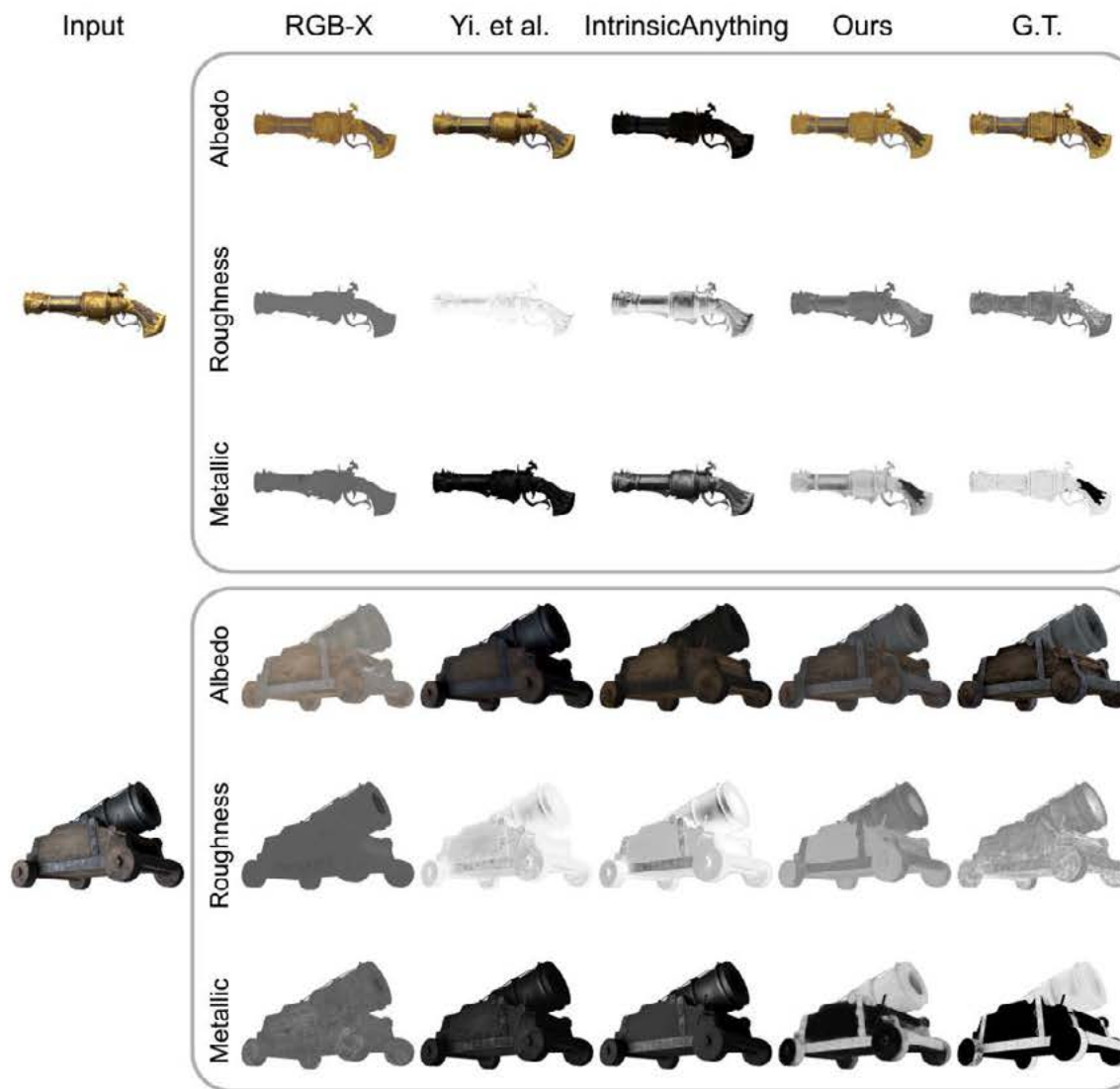
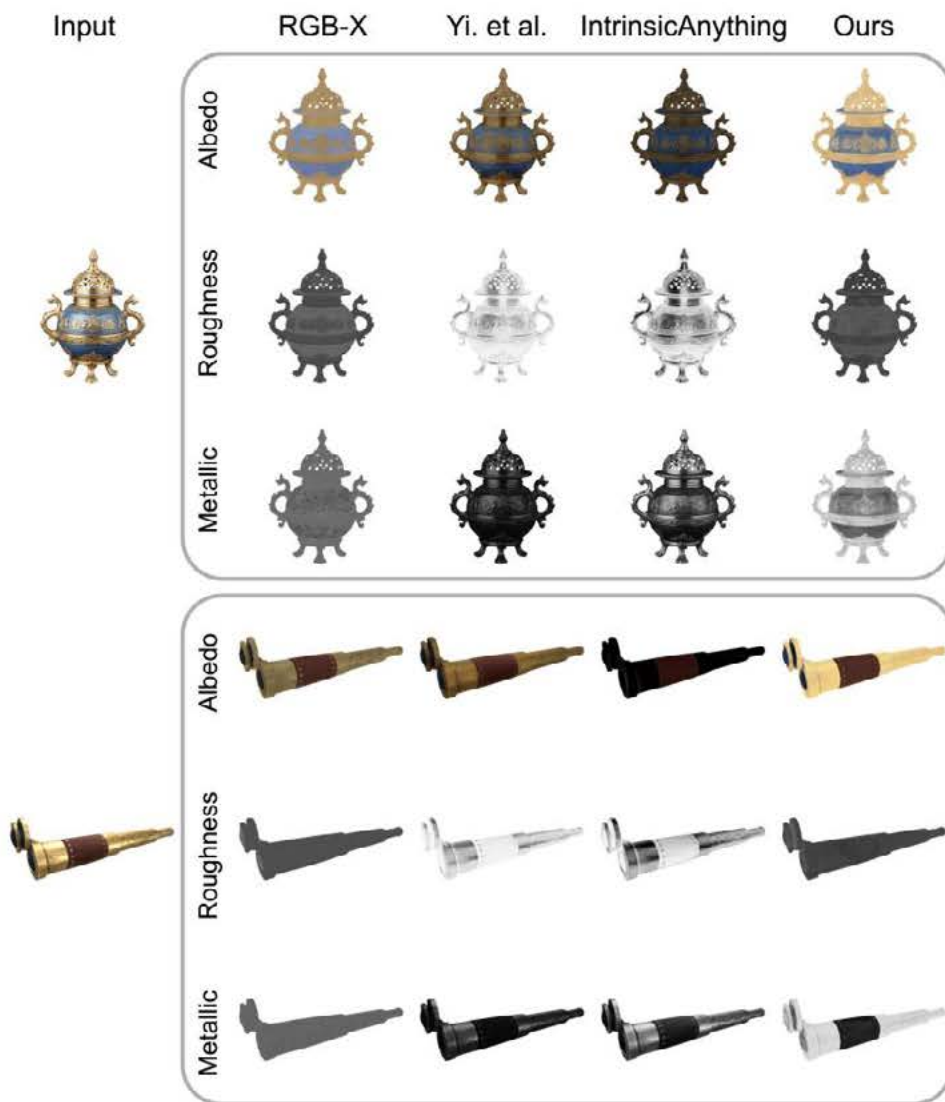
■ PBR Estimation and Single-Image Relighting

Method	Albedo		Roughness		Metallic		Relighting			Latency Average Time ↓
	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
RGB↔X [57]	16.26	0.176	19.21	0.134	16.65	0.199	20.78	0.8927	0.0781	15s
Yi. et al [54]	21.10	0.106	16.88	0.180	20.30	0.144	26.47	0.9316	0.0691	5s
IntrinsicAnything [8]	23.88	0.078	17.25	0.172	22.00	0.134	27.98	0.9474	0.0490	2min
DiLightNet [56]	-	-	-	-	-	-	22.68	0.8751	0.0981	30s
IC-Light [60]	-	-	-	-	-	-	20.29	0.9027	0.0638	1min
Ours	26.62	0.054	23.44	0.085	26.23	0.109	30.12	0.9601	0.0371	5s

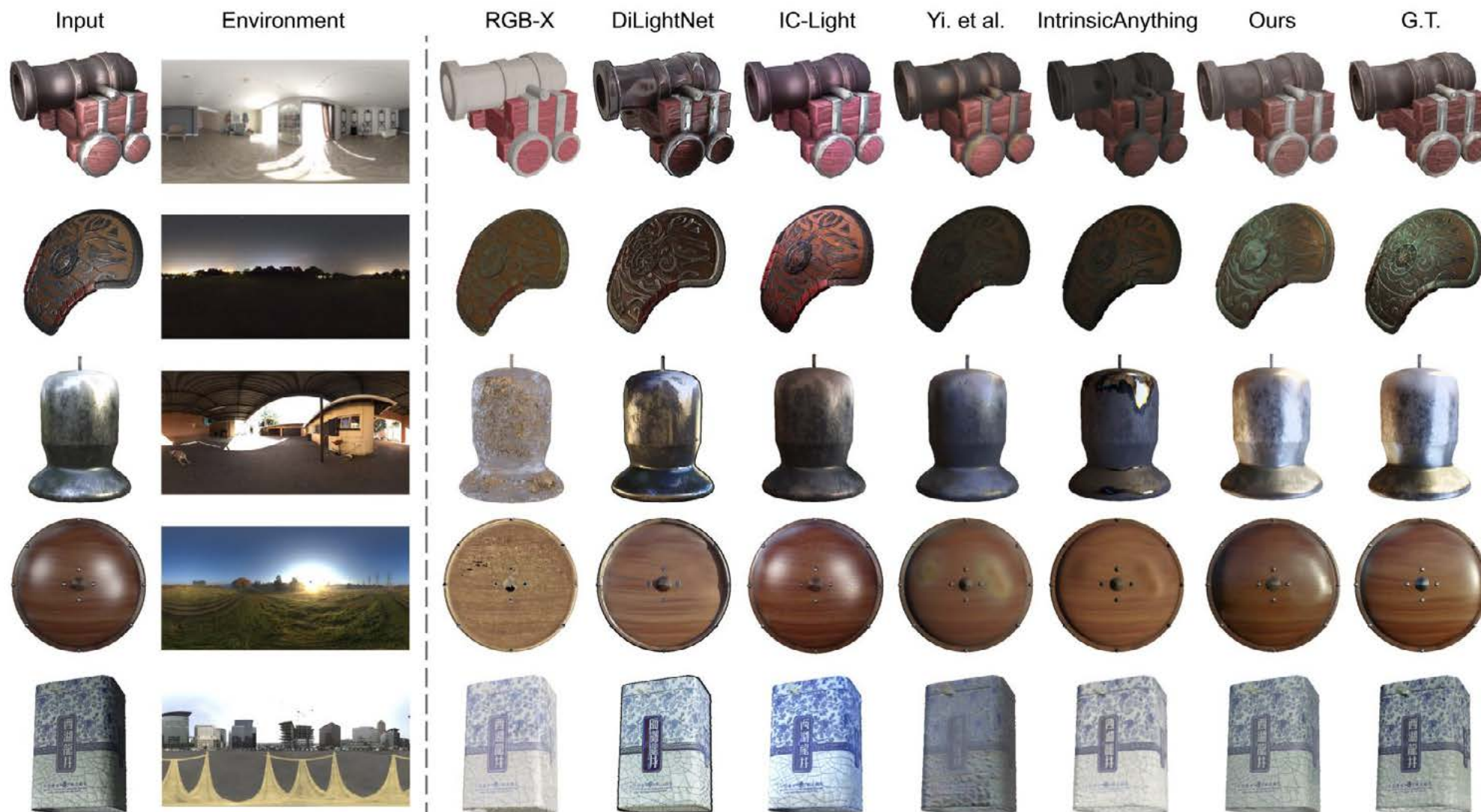
Surface Normal Estimation



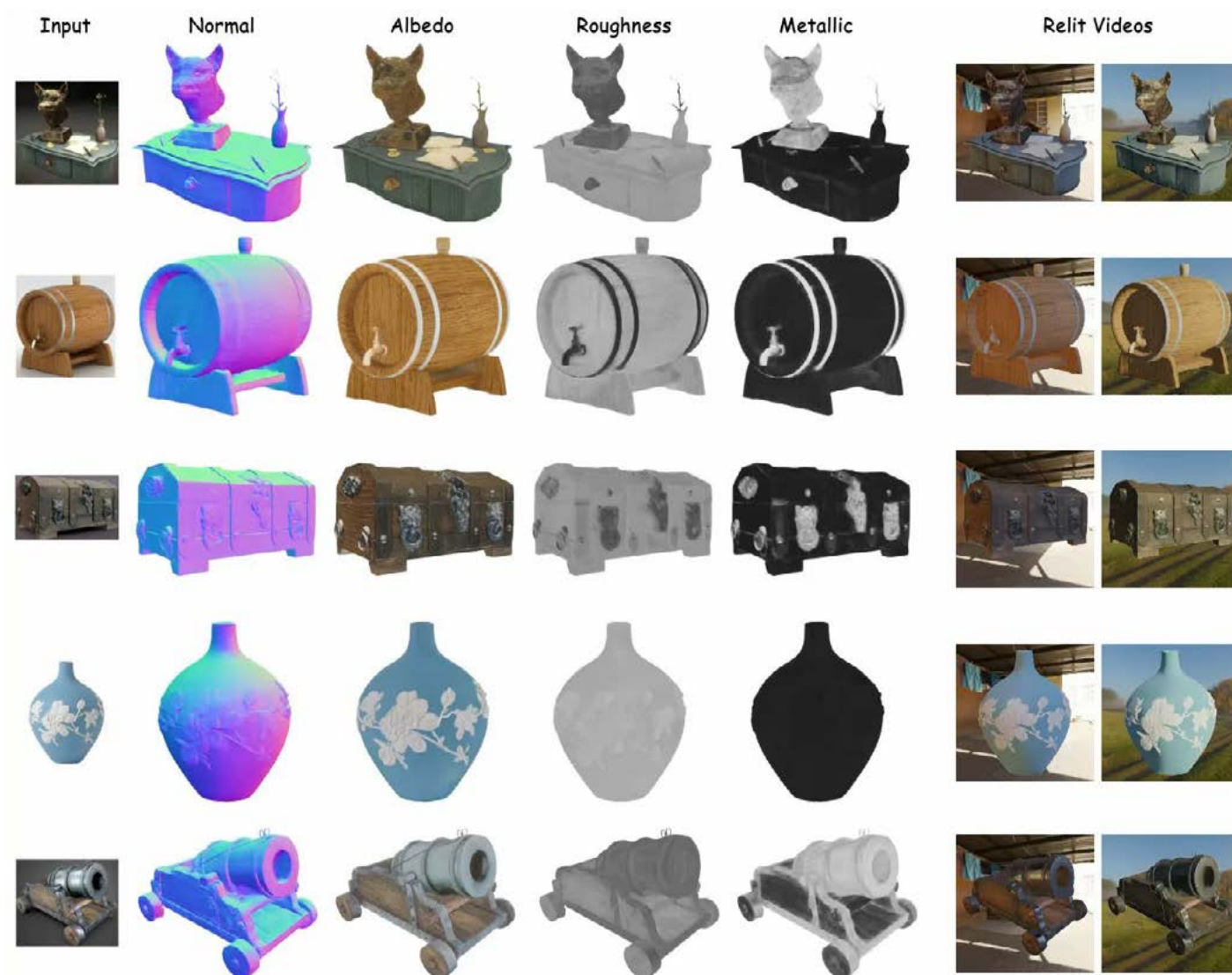
PBR Material Estimation



Single-Image Relighting



Single-Image Relighting

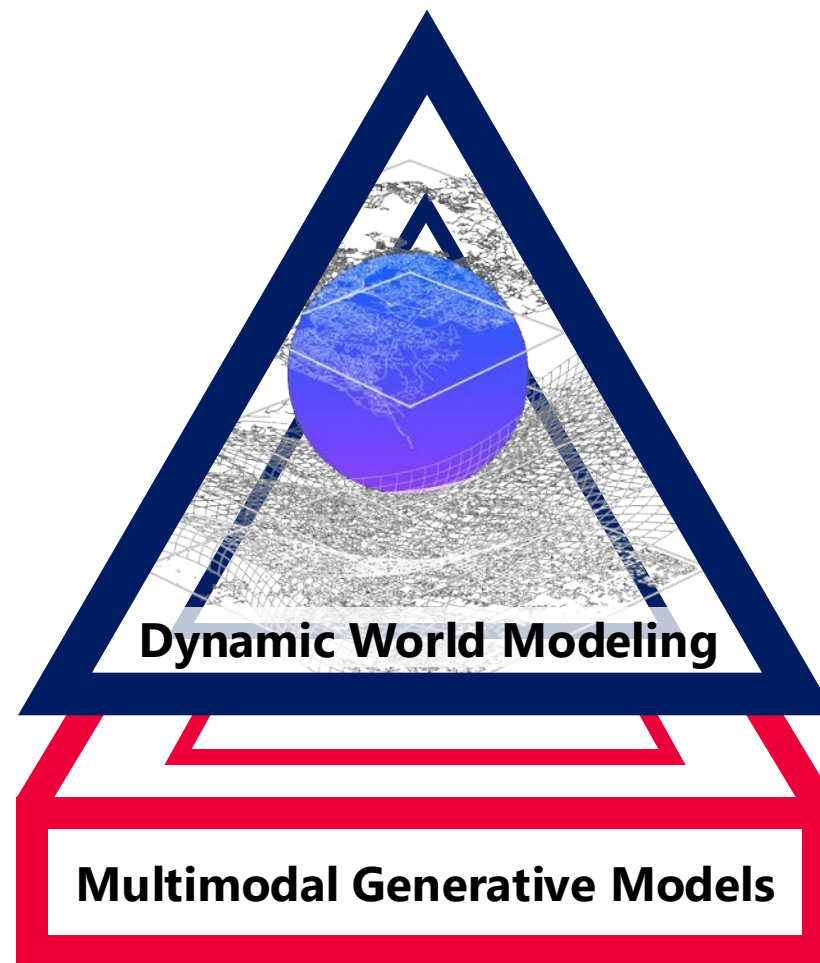


Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model
Dynamic Scenes



Be Social

How to Model Social
Interactions



Be Dynamic: DynamicCity

 [3DTopia/DynamicCity](https://github.com/3DTopia/DynamicCity)

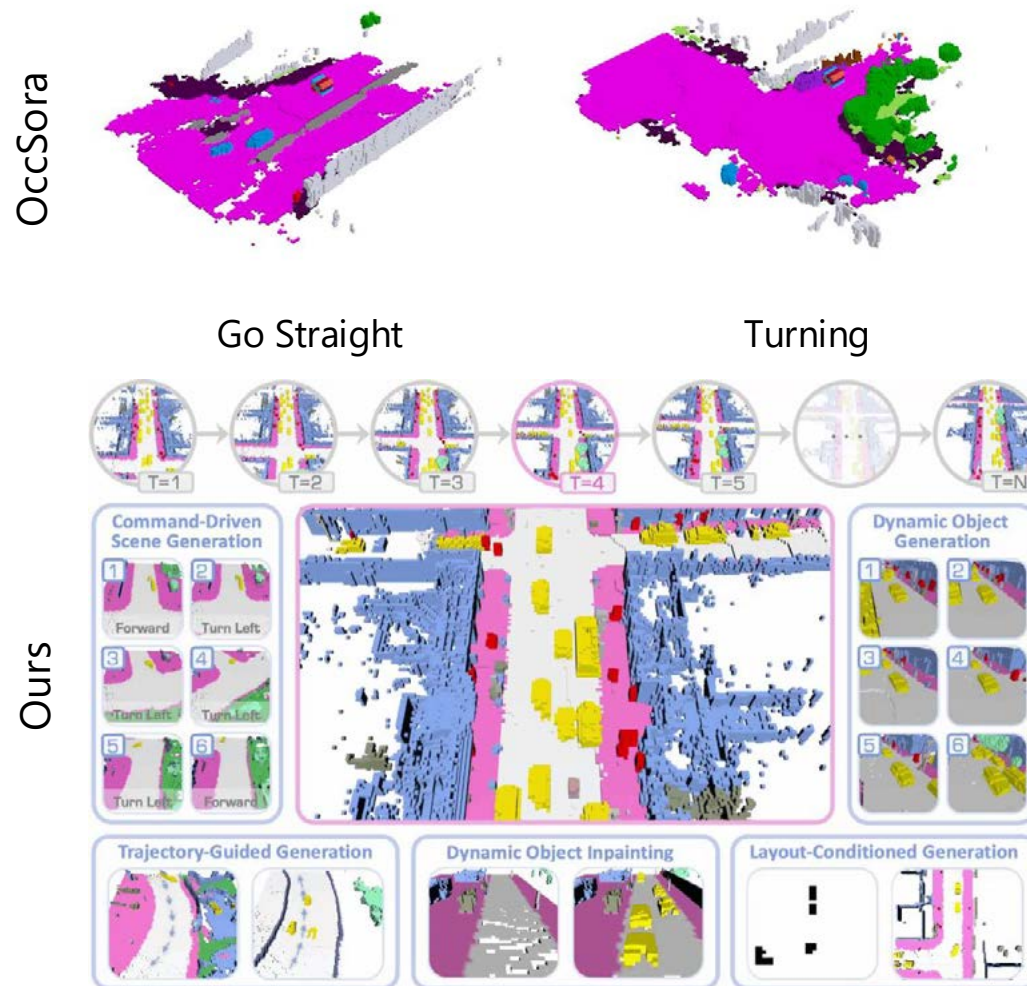
DynamicCity: Large-Scale 4D Occupancy Generation from Dynamic Scenes

Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, Ziwei Liu

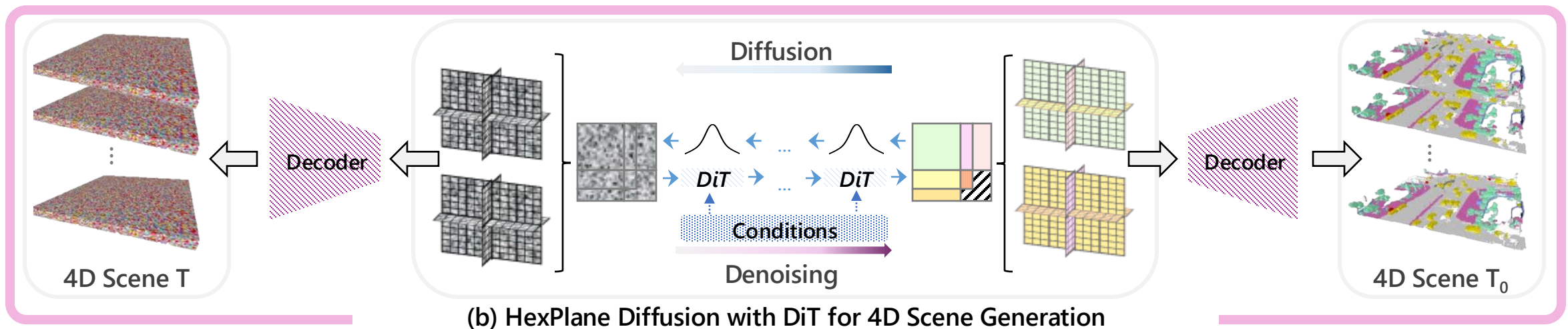
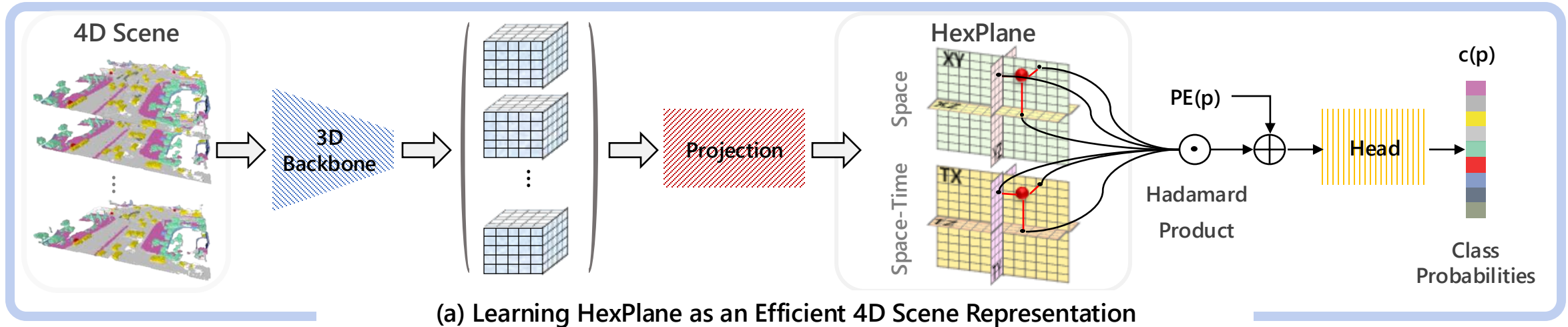
ICLR 2025 Spotlight

Challenges

- Inefficient VAEs for 4D data
 - low compression
 - poor reconstruction
- Suboptimal generation quality
- Limited control over the generation process

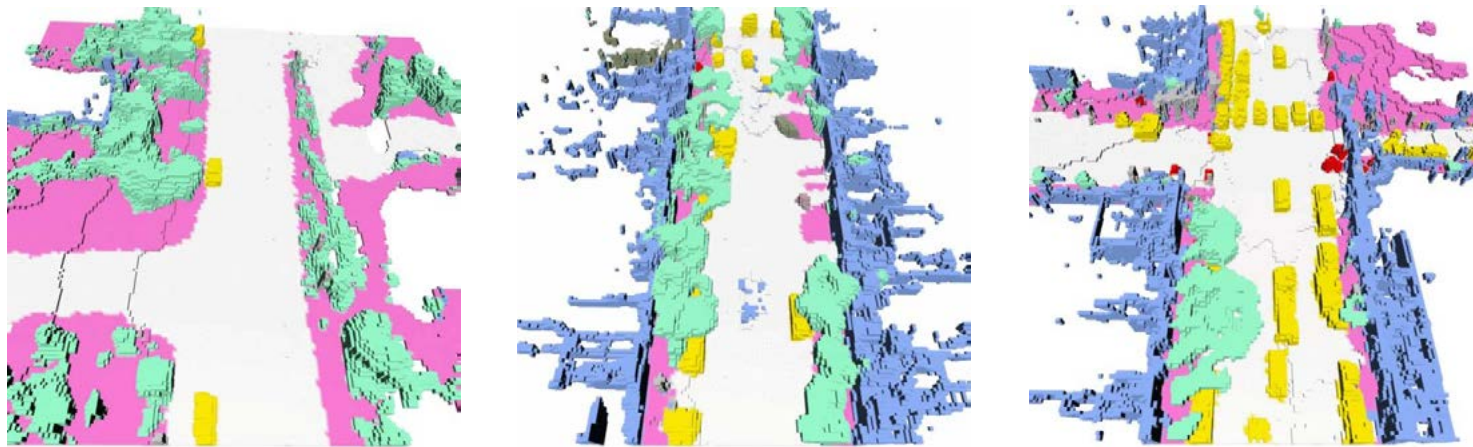


DynamicCity: 4D Occupancy Generation

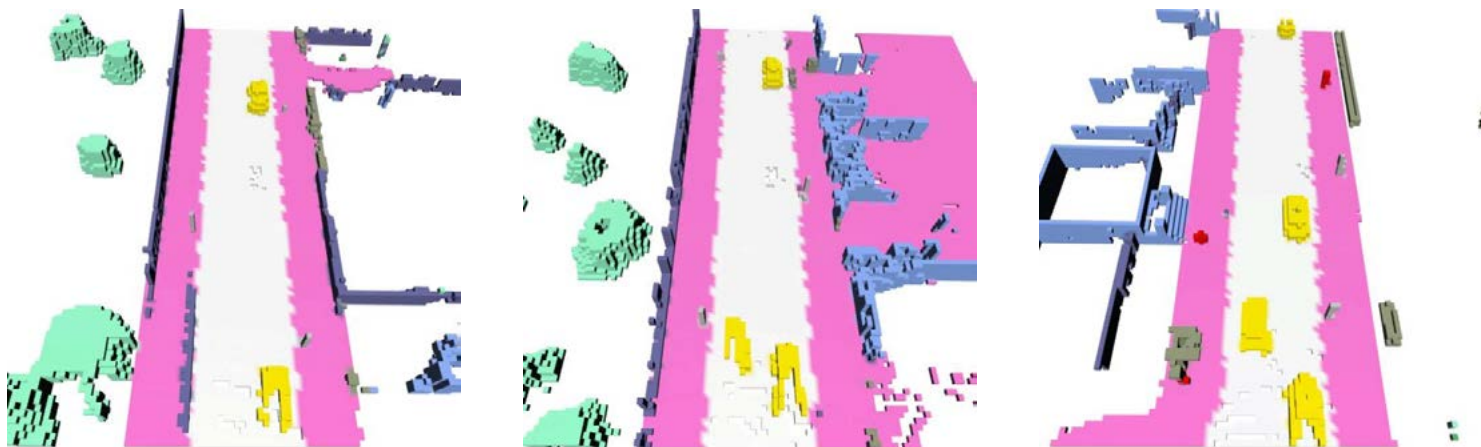


Unconditional 4D Generation

Occ3D-Waymo

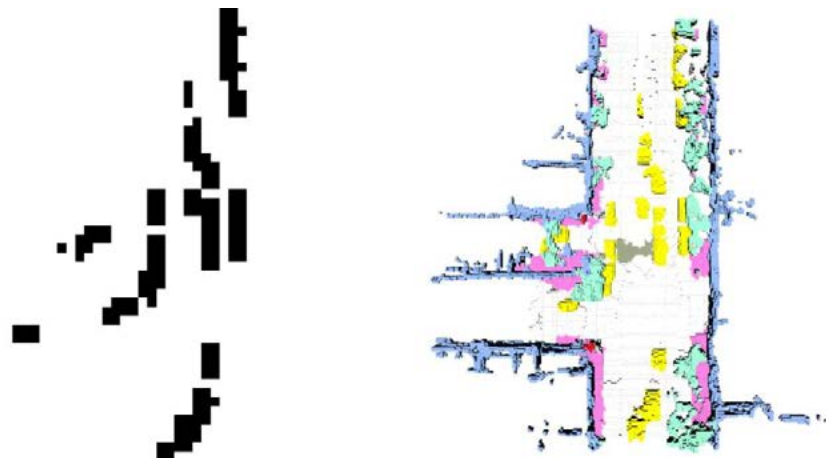


CarlaSC

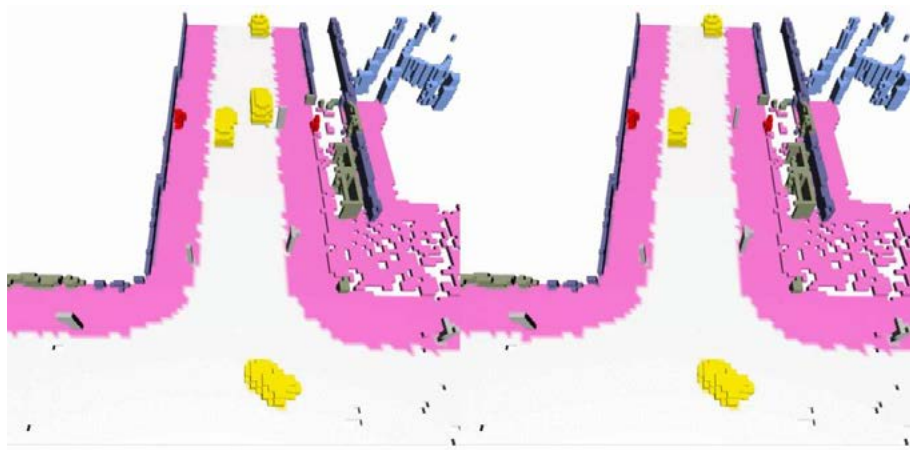


Conditional 4D Generation

Layout-conditioned



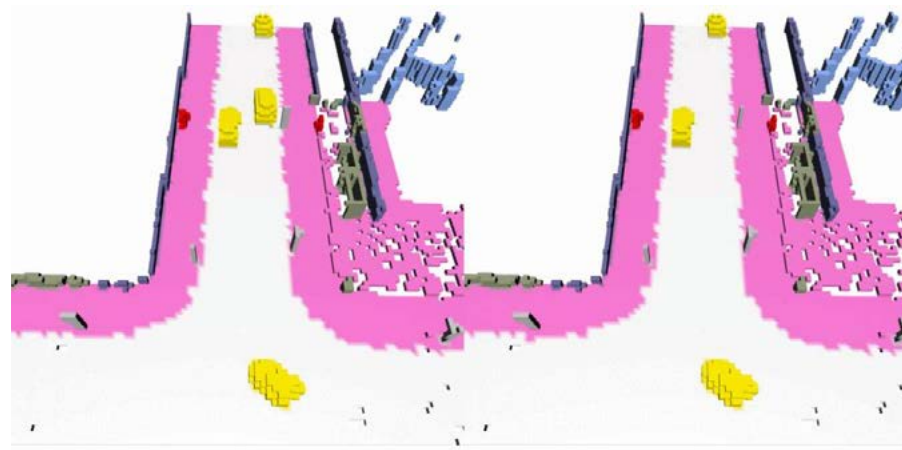
Inpainting



Trajectory-conditioned



Outpainting





Be Dynamic: CityDreamer4D



[hzxie/CityDreamer4D](https://github.com/hzxie/CityDreamer4D)

CityDreamer4D: Compositional Generative Model of Unbounded 4D Cities

Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, Ziwei Liu

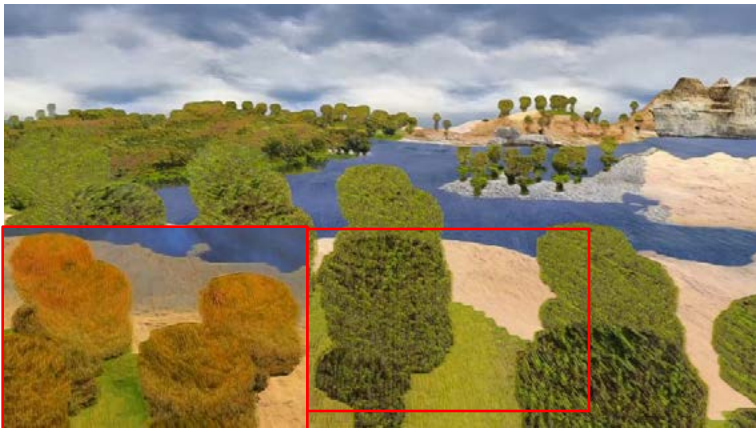
arXiv 2501.08983

How to Generate Unbounded 3D Cities?

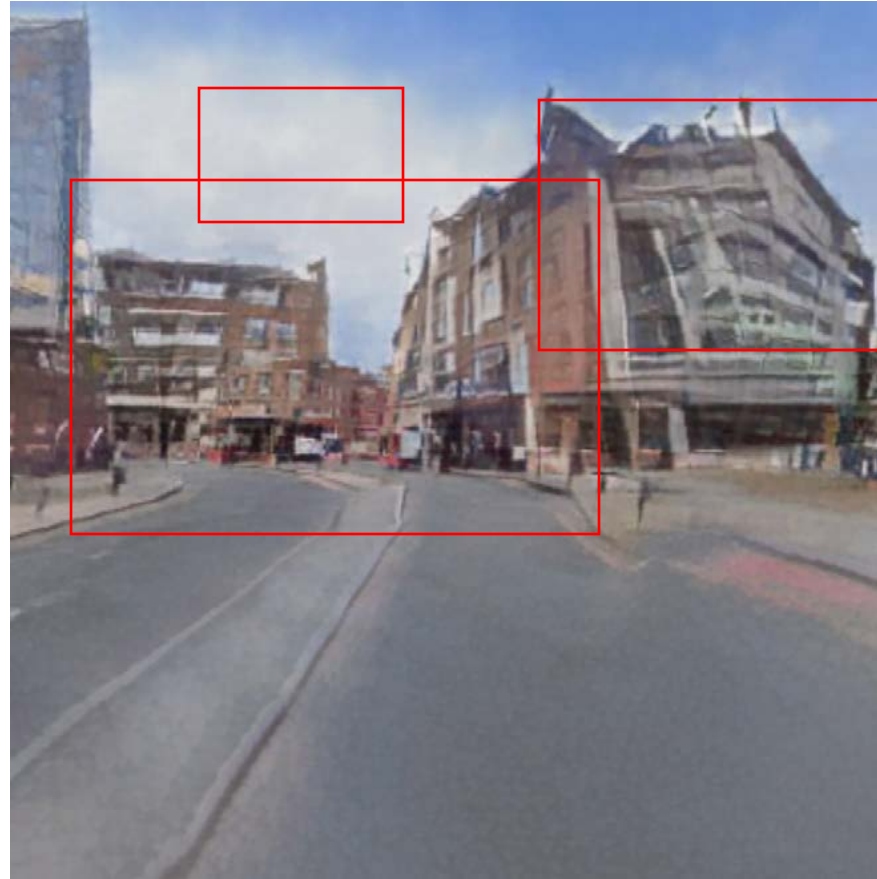
- Creating cities are more challenging than natural scenes



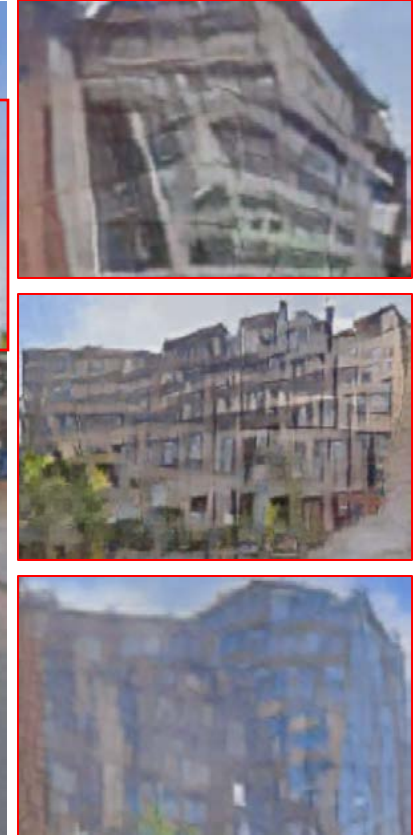
GANCraft [CVPR'21]



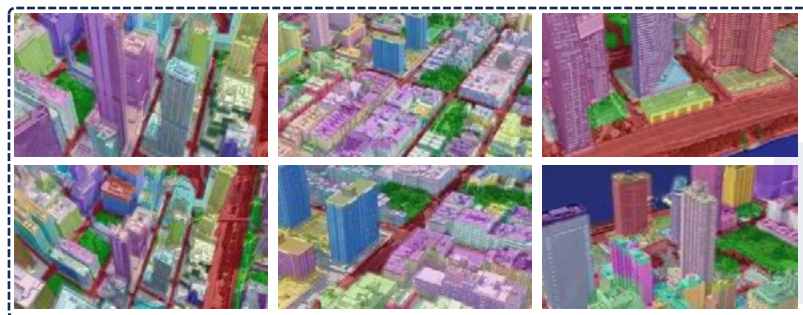
SceneDreamer [TPAMI'23]



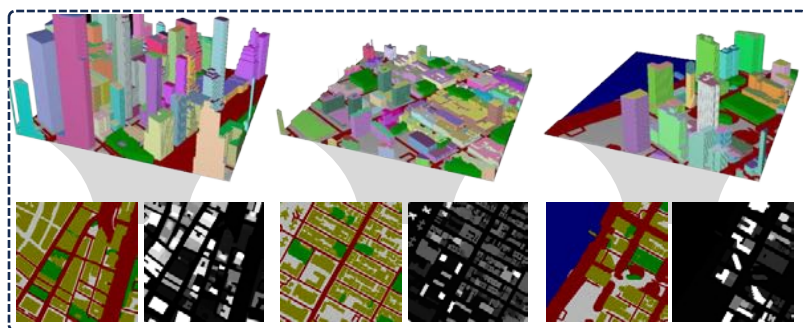
InfiniCity [ICCV'23]



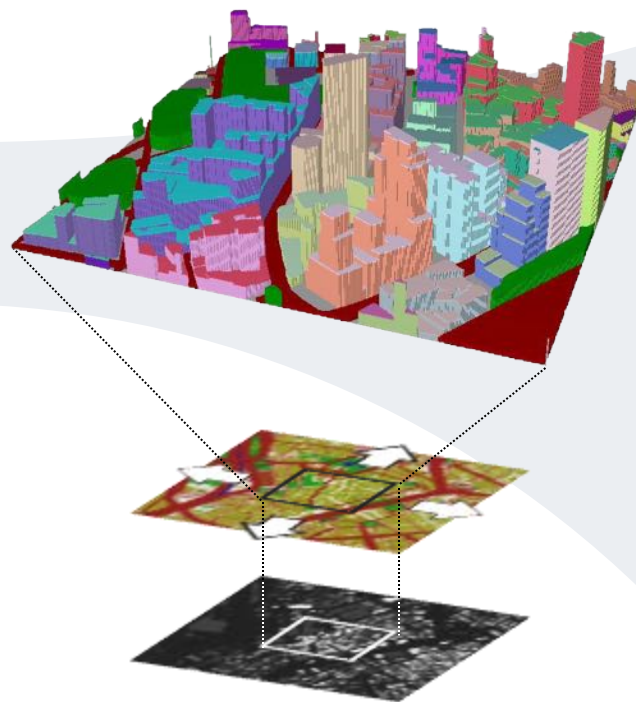
Learning 3D City from Unannotated 2D Images



(a) Google Earth Dataset: Real-world City Appearance



(b) OSM Dataset: Real-world City Layout

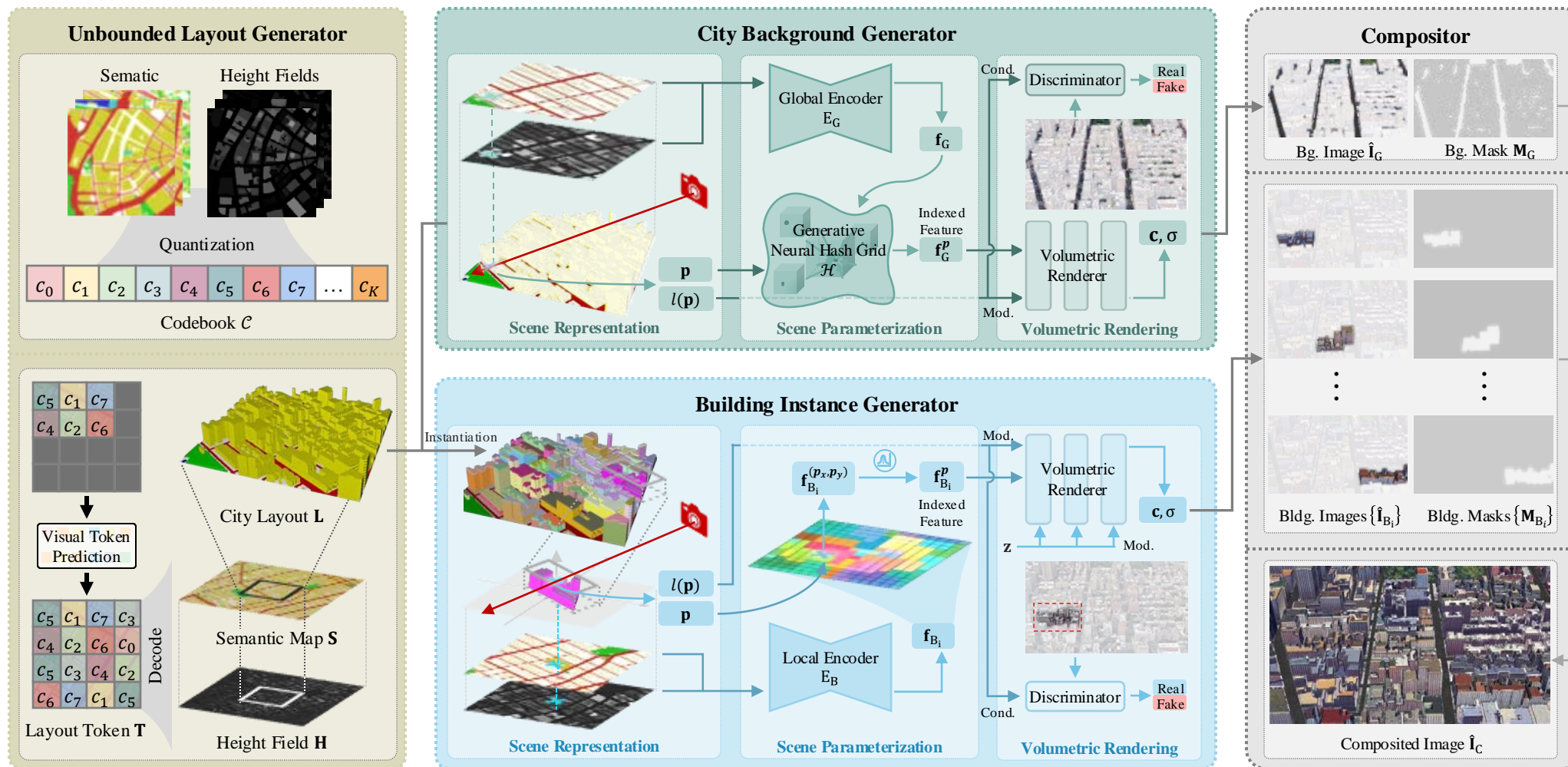


(c) Unbounded City Layout Generation



(d) CityDreamer Generated 3D Cities

CityDreamer Framework



Rendering in Unreal Engine 5



CityDreamer: Compositional Generative Model of Unbounded 3D Cities

The official demo to generate your own city in New York style.

[Source Code](#) [Project Page](#)



Layout
Generate the city layout

Topography
Set up camera trajectories

Render
Generate your own city

Data Source: Layout Generator

Layout Size: 4096x4096

Generate

Segmentation Map

Height Field

Press and hold the Ctrl Command key to enter the edit mode.

CityDreamer: Compositional Generative Model of Unbounded 3D Cities

The official demo to generate your own city in New York style.

[Source Code](#) [Project Page](#)



Layout
Generate the city layout

Topography
Set up camera trajectories

Render
Generate your own city

Data Source: Layout Generator

Layout Size: 4096x4096

Generate

Segmentation Map

Height Field

Press and hold the Ctrl Command key to enter the edit mode.

CityDreamer: Compositional Generative Model of Unbounded 3D Cities. CVPR 2024.

How to Make the City Generation Faster?

■ Challenge 1

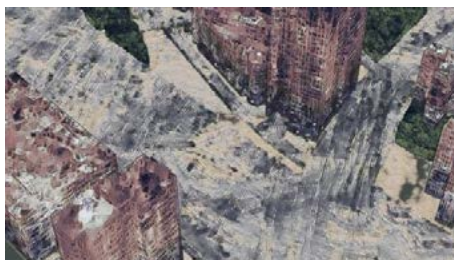
NeRF-based Methods are Not Efficient



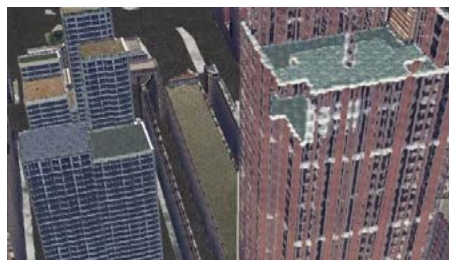
Pers.Nature (5.99 FPS)



InfiniCity (Unknown)



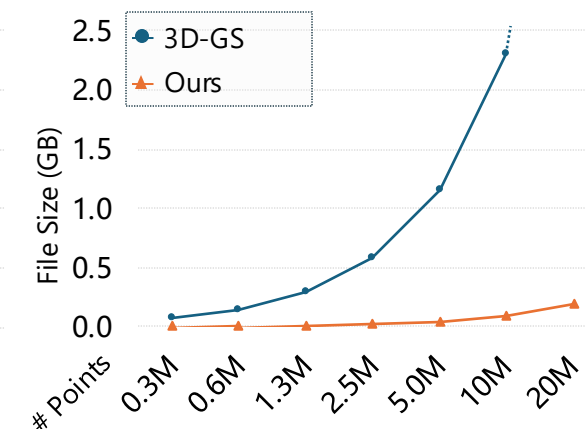
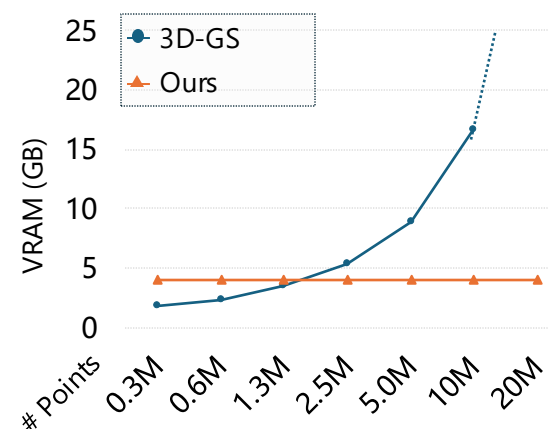
SceneDreamer (1.61 FPS)



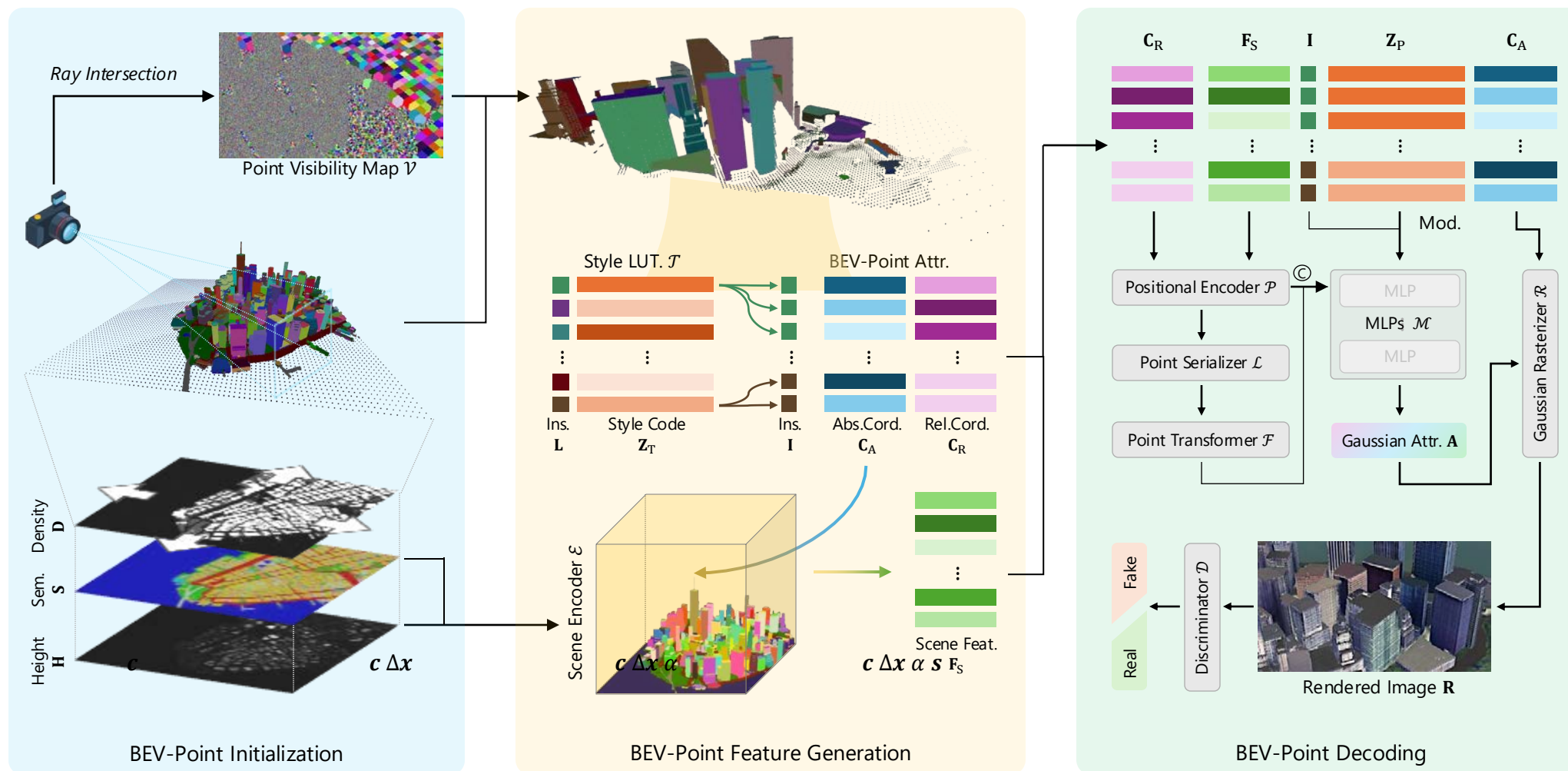
CityDreamer (0.18 FPS)

■ Challenge 2

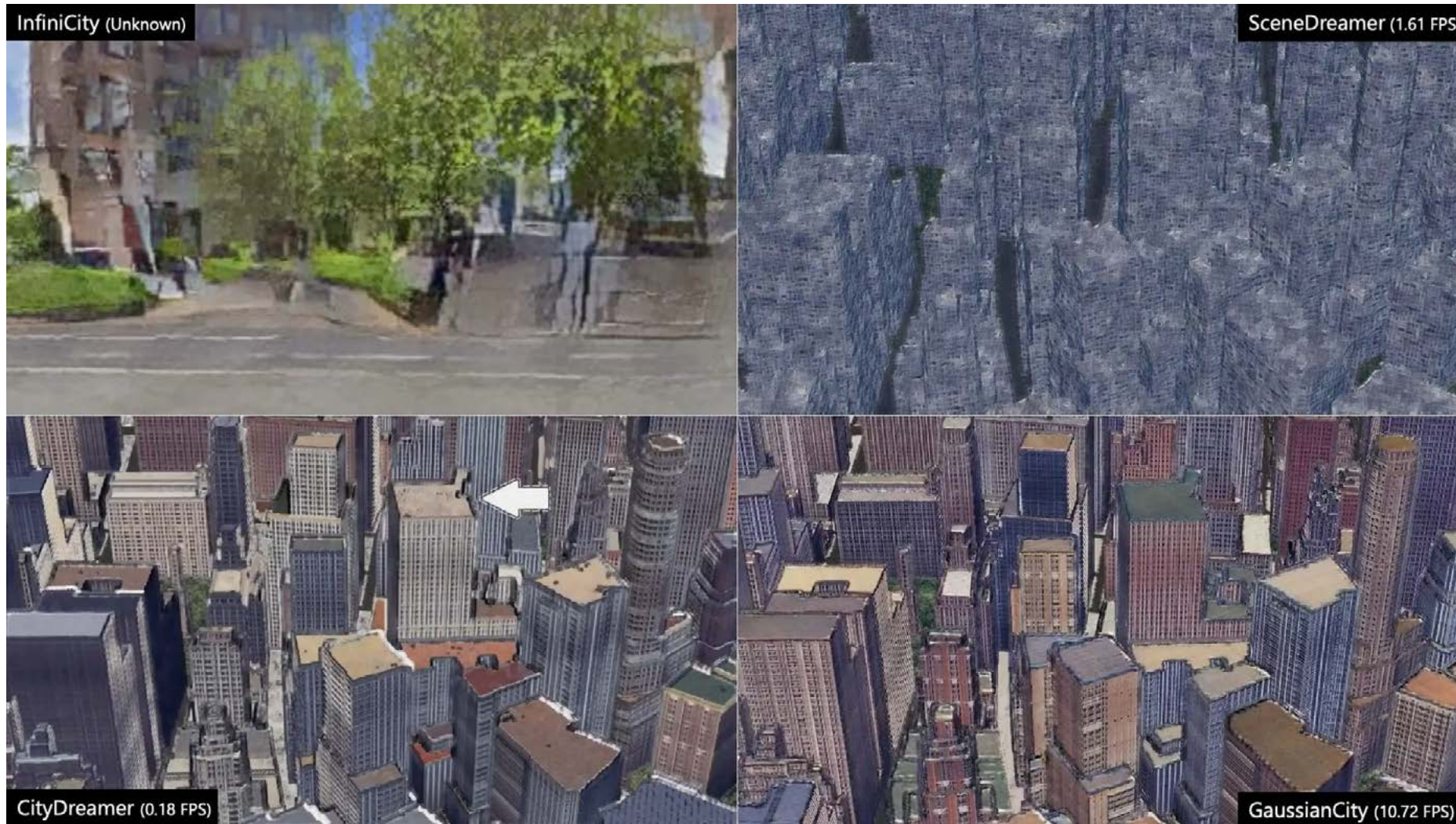
3D-GS are not Storage Efficient



GaussianCity Framework



60x Faster City Generation with GaussianCity



GaussianCity: Generative Gaussian Splatting for Unbounded 3D City Generation. CVPR 2025.

How to Generate 4D Cities?

Video-based

Multi-view Inconsistency

PCG-based

Limited Diversity

Lack Global Scene Context

No Available Annotated 4D Data

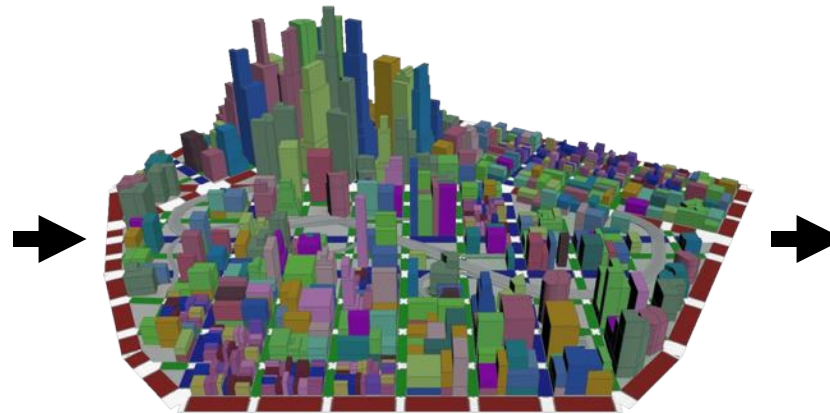
Image-based

Neural 3D-based

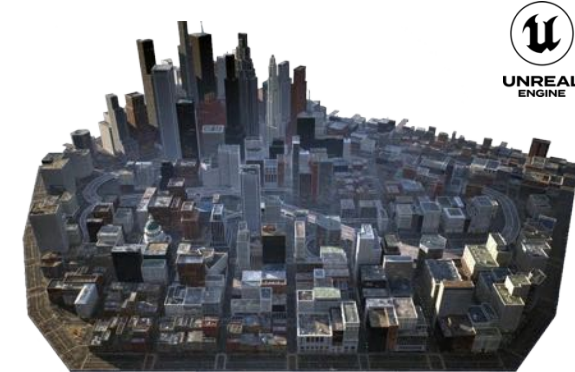
Learning 4D City from 3D Data Annotations



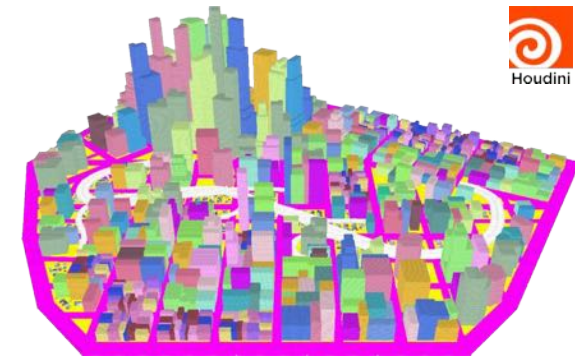
3D Assets
(Small set for Visualization)



City Prototype



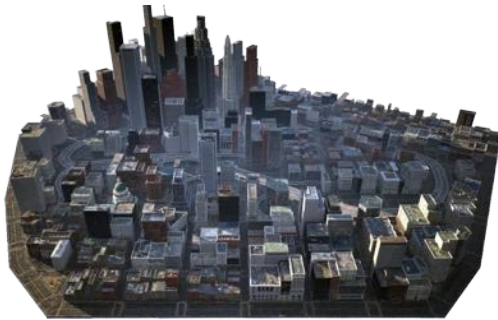
Generated 3D City



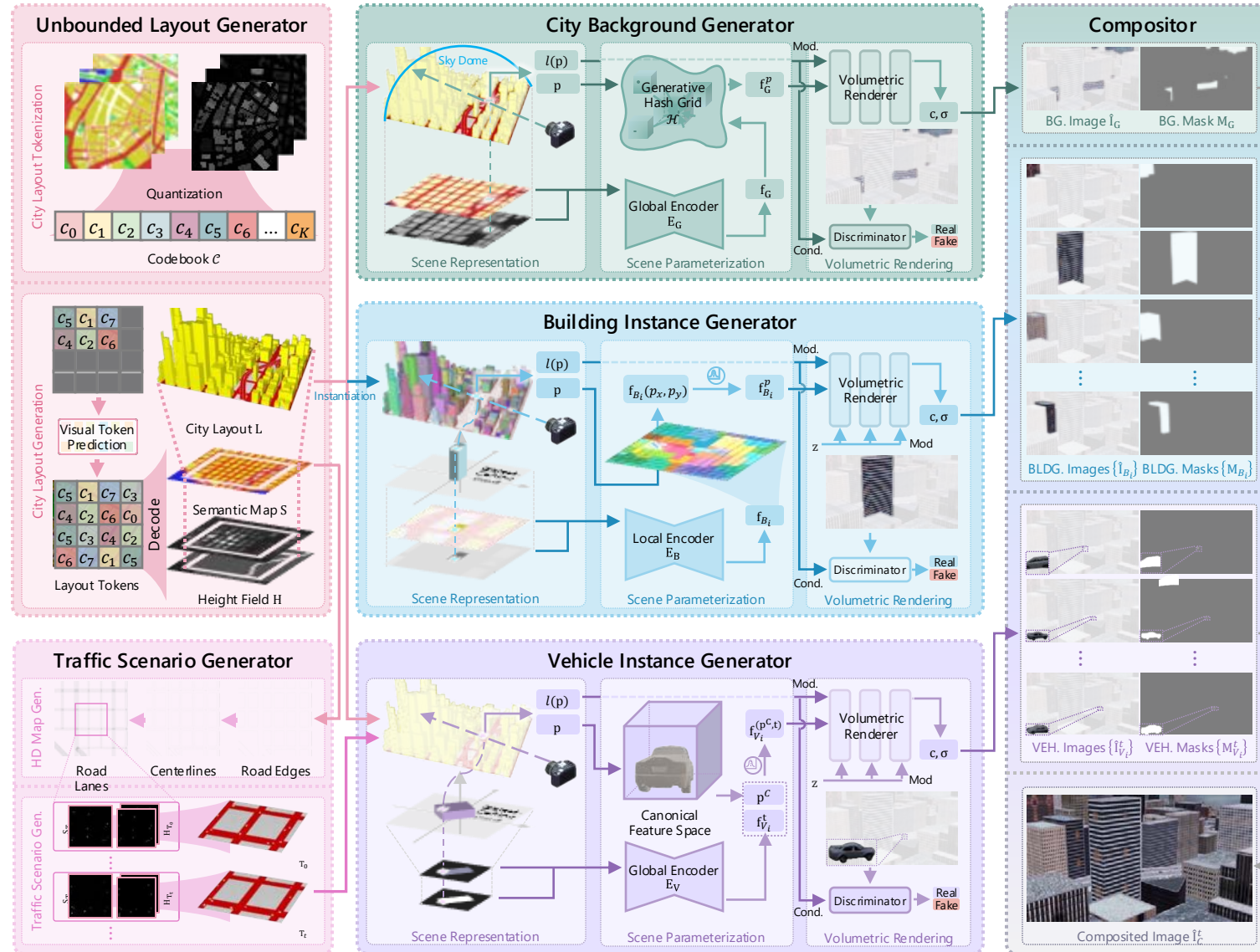
3D Instance Annotation



The CityTopia Dataset



CityDreamer4D Framework



Comparison to SOTA Methods

InfiniCity



SceneDreamer



PersistentNature



CityDreamer4D



Arbitrary View Rendering

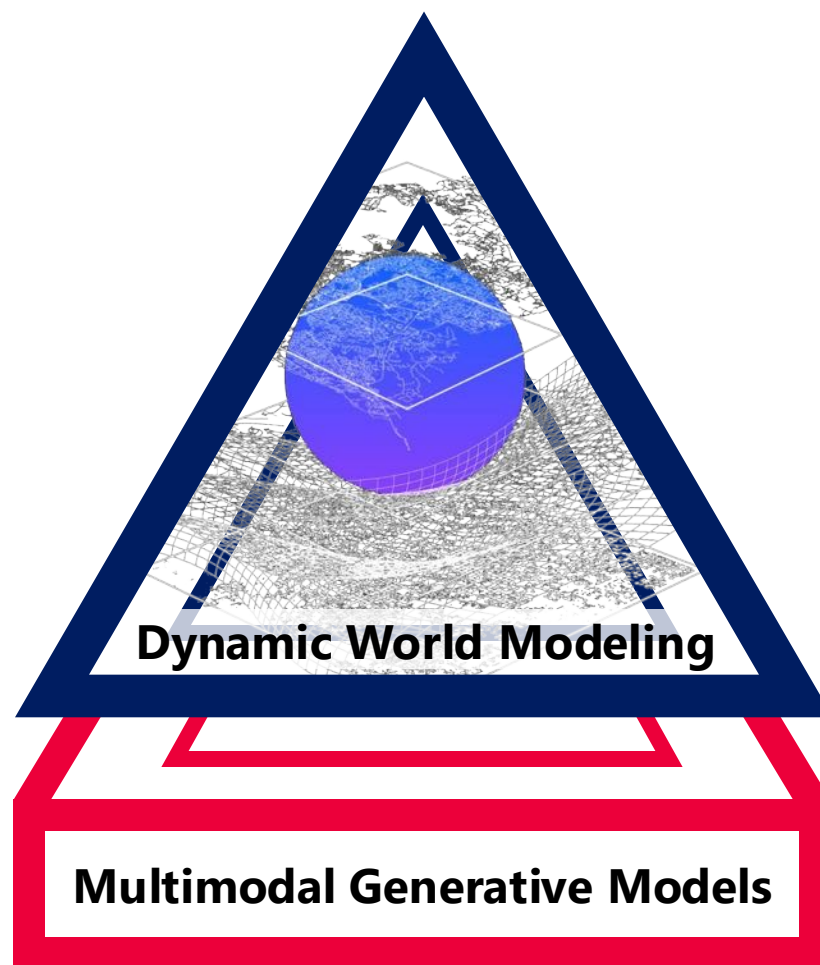


Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model
Dynamic Scenes



Be Social

How to Model Social
Interactions

Be Social: SOLAMI

SOLAMI: Social Vision-Language-Action Modeling for Immersive Interaction with 3D Autonomous Characters

Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, Ziwei Liu
CVPR 2025

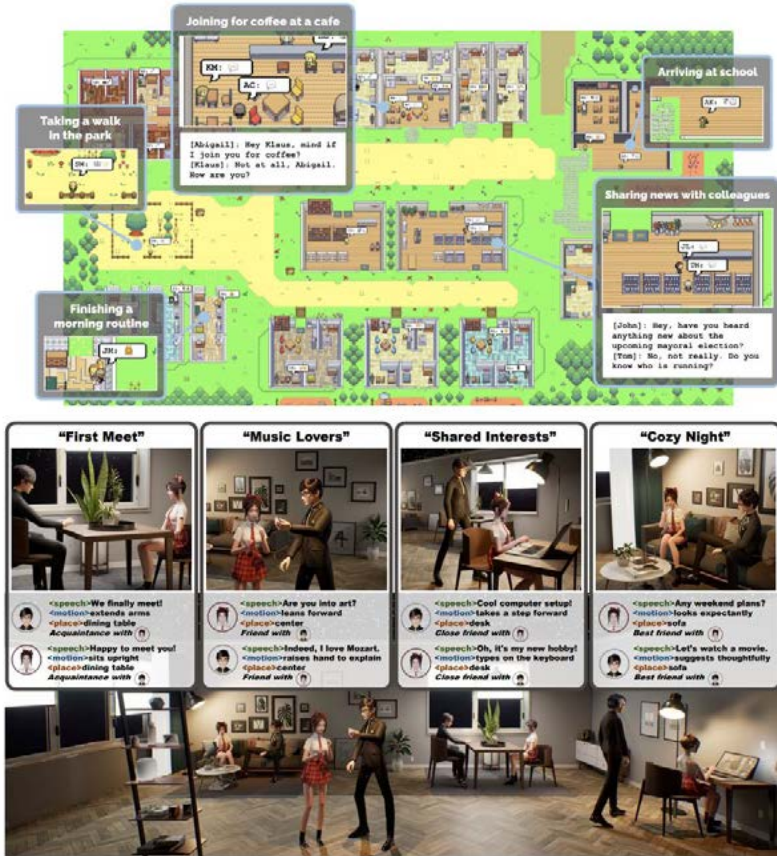
3D Characters with Social Intelligence

■ Modeling with LLM-Agent Framework

■ Limitations

- Scalable Formulation
- Multimodal Coherence
- Latency

Generative Agents [1]
Life Project [2]
Digital



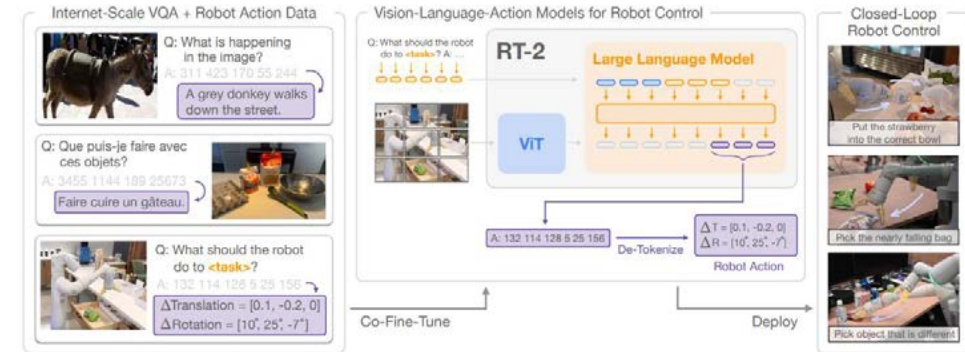
[1] Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023.

[2] Digital Life Project: Autonomous 3D Characters with Social Intelligence. CVPR 2024.

Motivation: Avatar as Virtual Robot



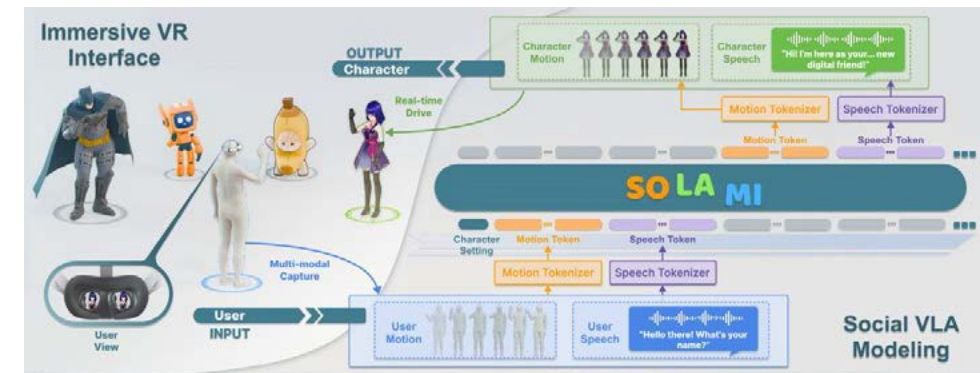
Robot
3D Agent with Real Embodiment
(Real-world Task & Interaction)



RT-2 ^[1]: Vision-Language-Action Models



3D Avatar
3D Agent with Virtual Embodiment
(Natural Appearance & Behavior)



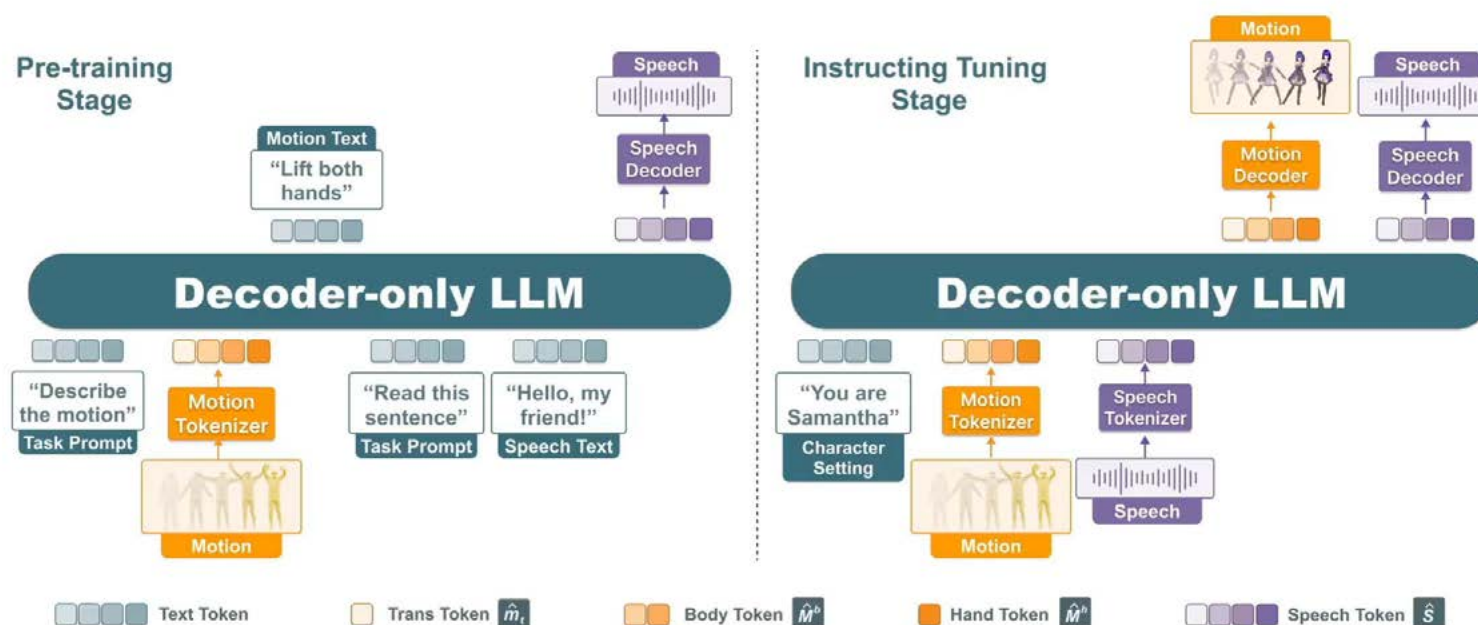
Social VLA for Immersive Interaction with 3D Characters

[1] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. CoRL 2023.

Training Recipe

■ Training Stages

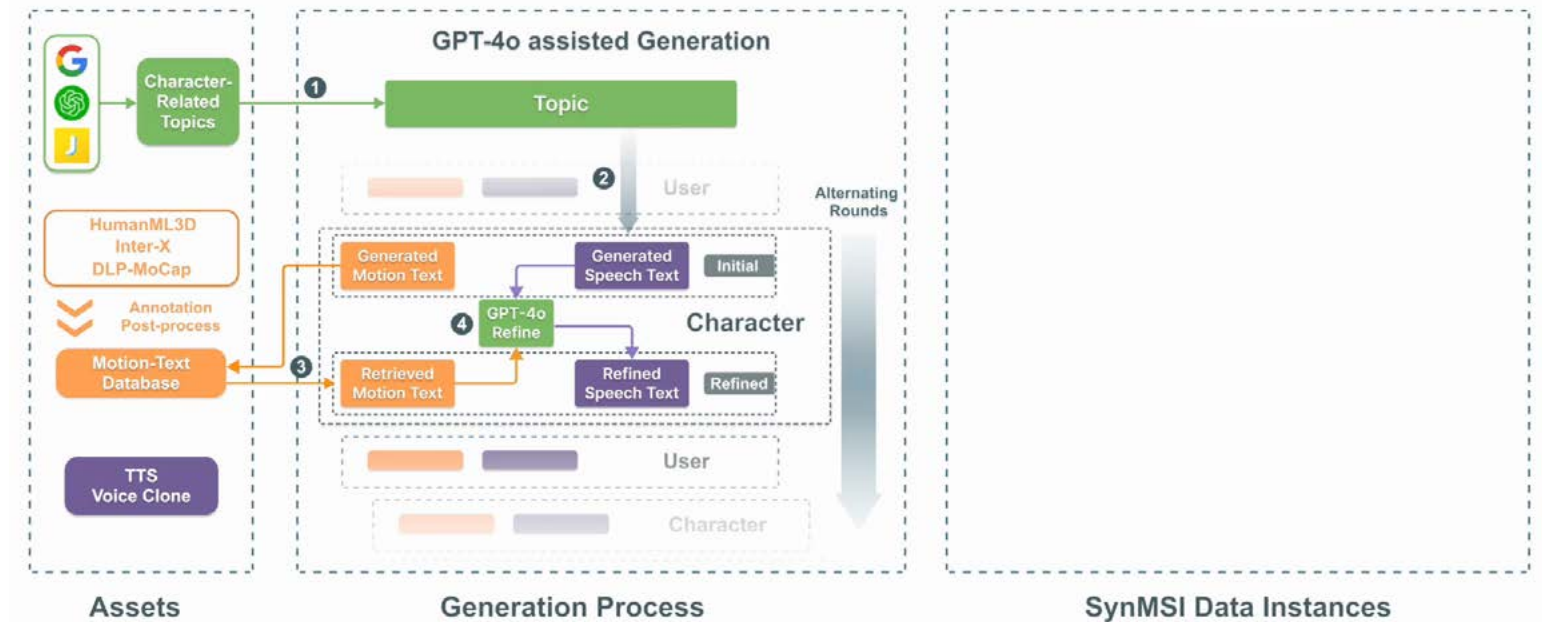
- Stage1: Motion & Speech Tokenizer Training
- Stage2: Motion-Text-Speech Alignment with Multi-Task Pretraining
- Stage3: Instruction Tuning for Multimodal Chat



为了赋予SOLAMI
To empower SOLAMI

Data Generation

- Multimodal Chat Data Synthesize
 - LLM-Generated Scripts
 - Diverse Topics
 - Refined Process
 - Motion-Text Dataset
 - Large-Scale



Evaluation: Quantitative & Qualitative

- Compared to Speech-Only Method
 - Better User Experience
- Compared to LLM-Agent Framework
 - Low Latency & Multimodal Coherence
 - Alignment Tax on Text

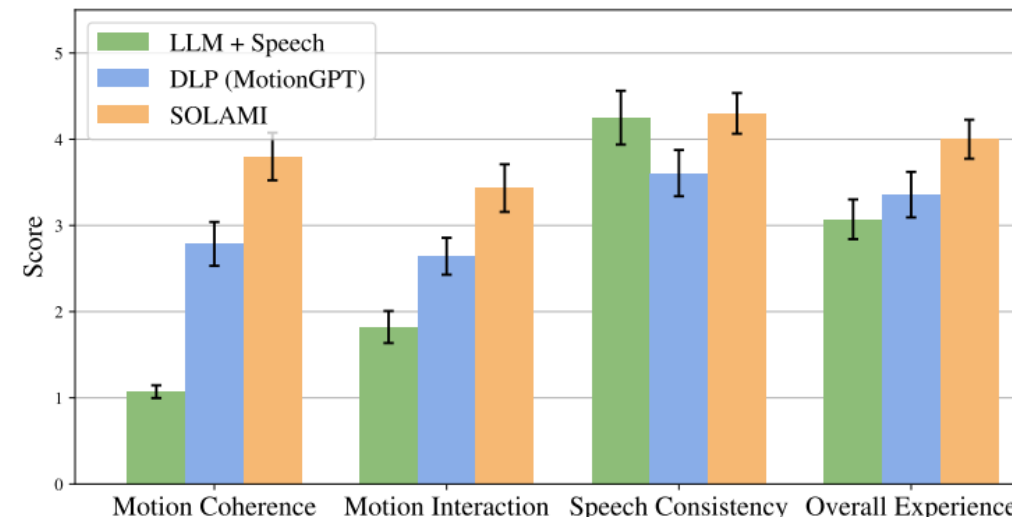
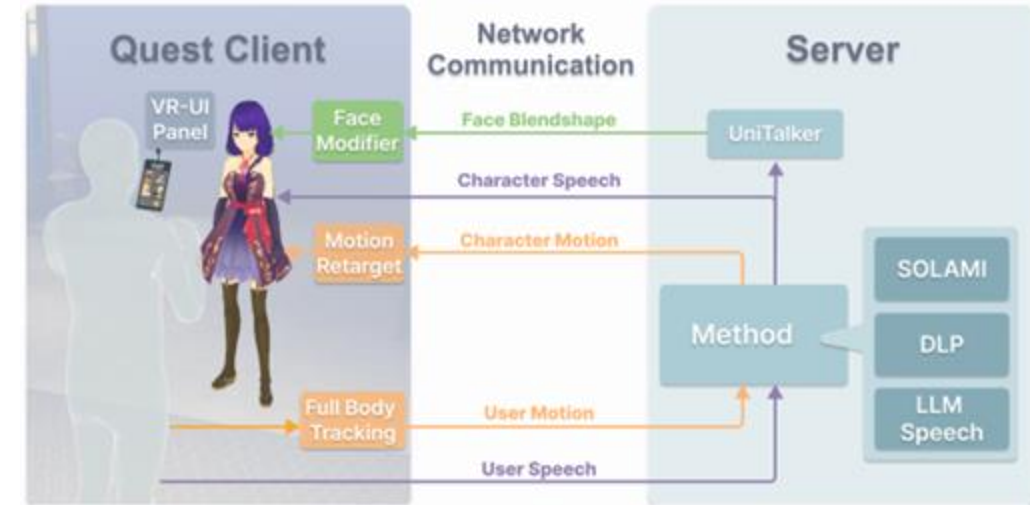


Table 1. **Quantitative results of baselines and SOLAMI.** ‘↑’(‘↓’) indicates that the values are better if the metrics are larger (smaller). We run all the evaluations 5 times and report the average metric. The best results are in bold and the second best results are underlined.

Methods	Motion Metrics					Speech Metrics		Inference Latency ↓
	FID↓	Diversity↑	PA-MPJPE↓	Angle Error↓	VC Similarity↑	Context Relevance↑	Character Consistency↑	
SynMSI Dataset	-	9.136	-	-	-	4.888	4.893	-
LLM+Speech (Llama2) [69]	-	-	-	-	0.818	3.527	3.859	3.157
AnyGPT (fine-tune) [81]	-	-	-	-	0.819	3.502	3.803	2.588
DLP (MotionGPT) [17]	<u>4.254</u>	8.259	165.053	0.495	0.812	<u>3.577</u>	3.785	5.518
SOLAMI (w/o pretrain)	5.052	<u>8.558</u>	<u>159.709</u>	<u>0.387</u>	<u>0.820</u>	3.541	3.461	2.657
SOLAMI (LoRA)	15.729	8.145	167.149	0.400	0.770	3.251	3.423	2.710
SOLAMI (full params)	3.443	8.853	151.500	0.360	0.824	3.634	<u>3.824</u>	<u>2.639</u>

Demo: VR Interface

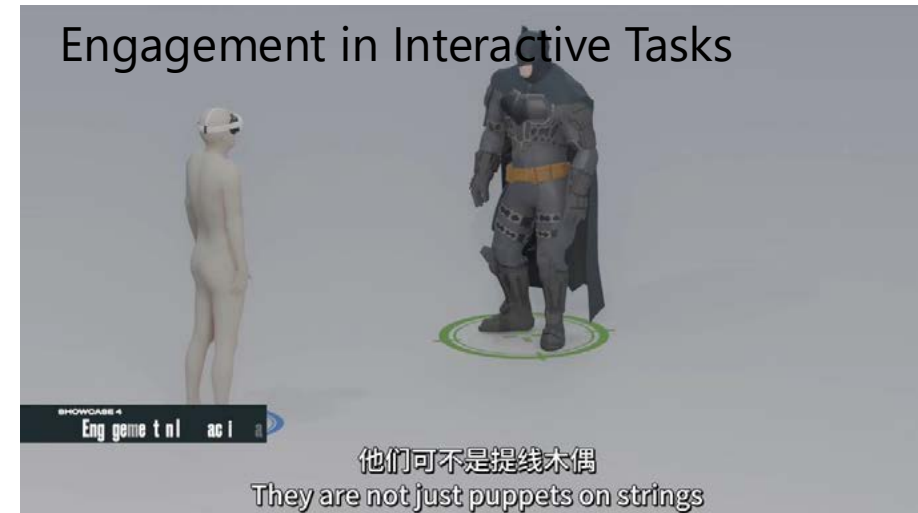
Comprehension of Body Language



Execution of Motion Commands



Engagement in Interactive Tasks





Be Social: EgoLife



[EvolvingLMs-Lab/EgoLife](https://github.com/EvolvingLMs-Lab/EgoLife)

EgoLife: Towards Egocentric Life Assistant

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, Ziwei Liu

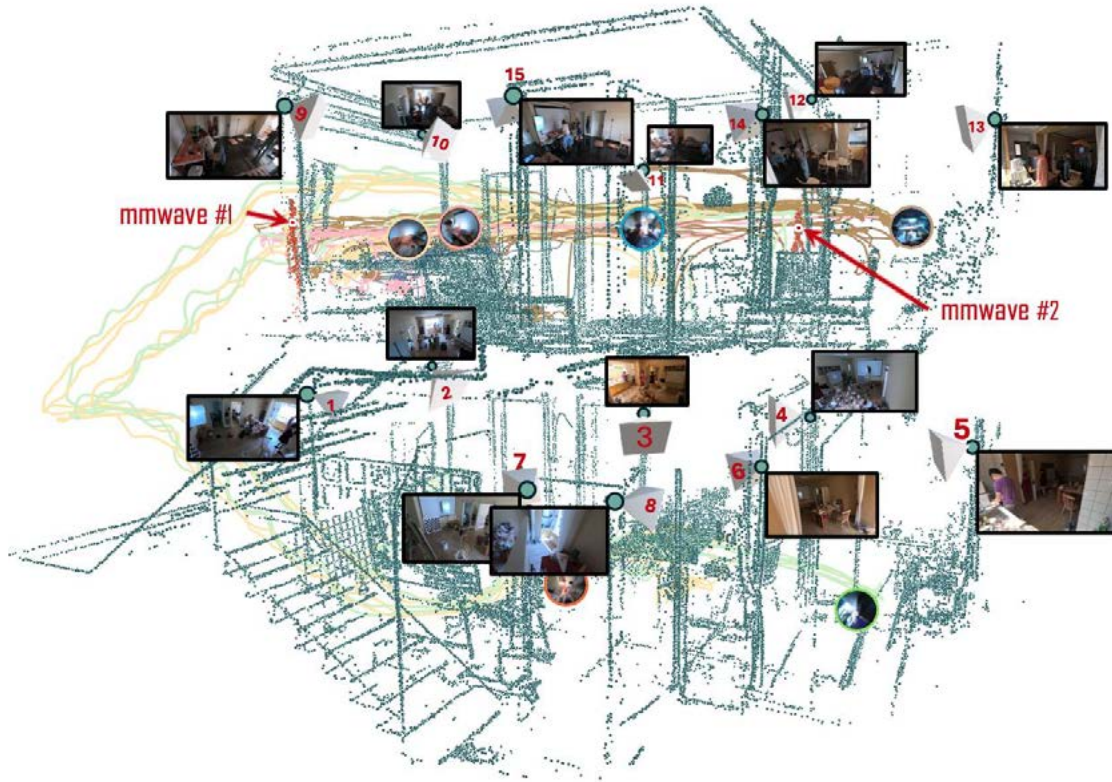
CVPR 2025

We invited **6** people living together
for **7** days in **egolife**



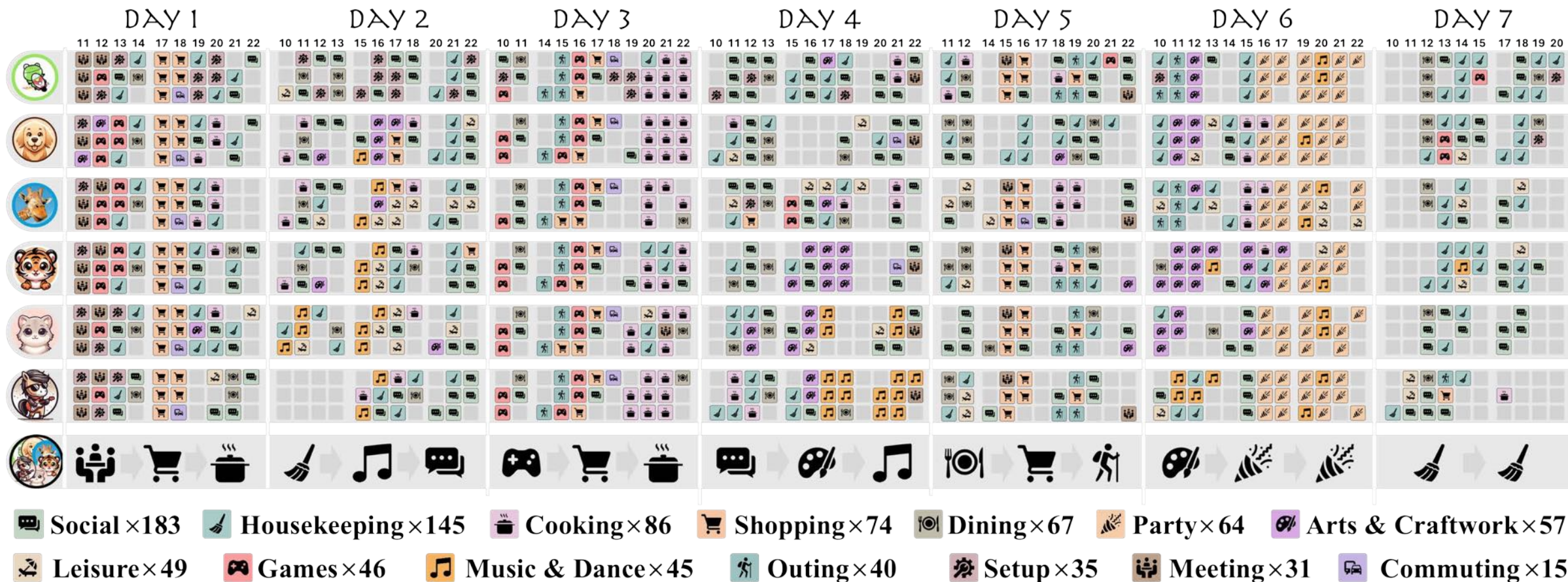
Each one wearing Meta Aria glasses
(almost) all day long.

The EgoLife Collected Data



Ego video, audio, mmwave, wifi, Ego/Exo signals synchronization.

The EgoLife Timeline



The EgoLifeQA Benchmark

6 x 500 = 3000 QAs



Event Recall

Past Events of Interest

Day 1: 21:48:21.200



What was the first song mentioned after planning to dance?

- A. Why Not Dance
- B. Mushroom
- C. I Wanna Dance with Somebody
- D. Never Gonna Give You Up

Answer: A. Evidence:

Shure sang after Jake asked us to dance.



@ Day 1
11:46:59.050



Entity Log

Past Objects of Interest

Day 4: 11:34:05.400



Which price is closest to what we paid for one yogurt?

- A. RMB 2
- B. RMB 3
- C. RMB 4
- D. RMB 5

Answer: B. Evidence:

The yogurt is on sale, RMB19.9 for 6 cups @ Day 3: 17:00:04.450



TaskMaster

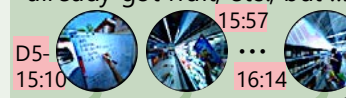
Tasks Assignment and Review

Many things are in my cart already. What items that we previously discussed have I not bought yet?



- A. Milk
- B. Chicken wings
- C. Strawberries
- D. Bananas

Answer: A. Evidence:
I made a shopping list, and already got fruit, etc., but ...



Day 5: 16:20:46.350



Day 1

Day 2

Day 3

Day 4

Day 5

Day 6

Day 7



What activity do I usually do while drinking coffee?

- A. Scrolling through TikTok
- B. Texting on the phone
- C. Tidying up the room
- D. Doing Craftwork

Day 4: 12:08:50.600

Answer: D. Evidence:



D1-16:14



D2-10:40



D2-10:52

...



D4-11:39

I had coffee a total of five times, three of which were while doing crafts...



Shure is playing the guitar now. Who else usually joins us playing guitar together?

- A. Choizst
- B. Jake
- C. Nicous
- D. Lucia

Day 6: 19:50:19.750

Answer: C. Evidence:

D4-17:19

D4-17:22

D4-22:00

D5-22:52

Nicous played the guitar with Shure and me twice, more frequently than anyone else.



Habit Insight

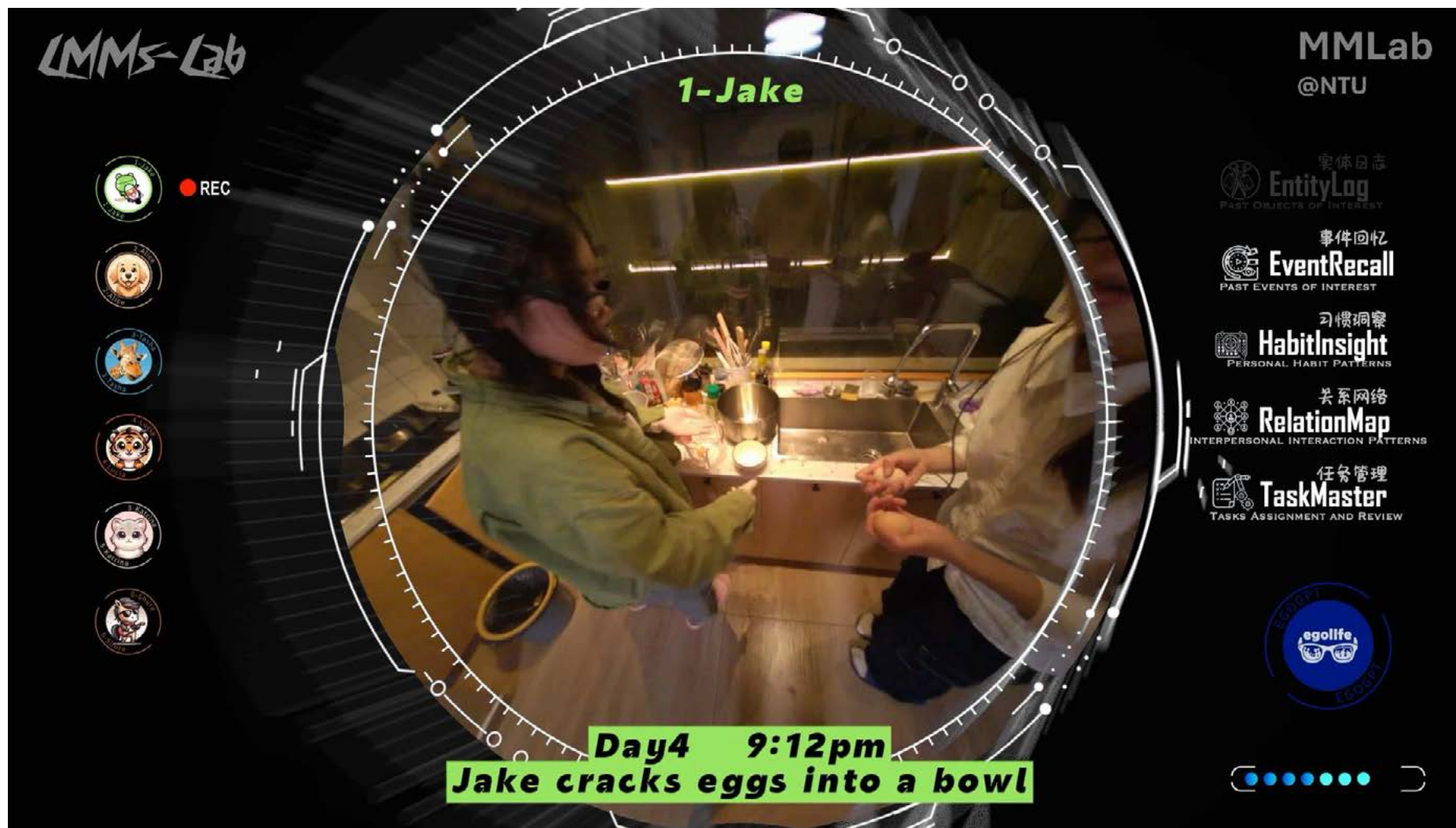
Personal Habit Patterns



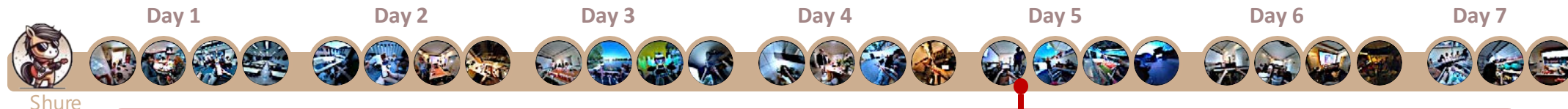
RelationMap

Interpersonal Interaction Patterns

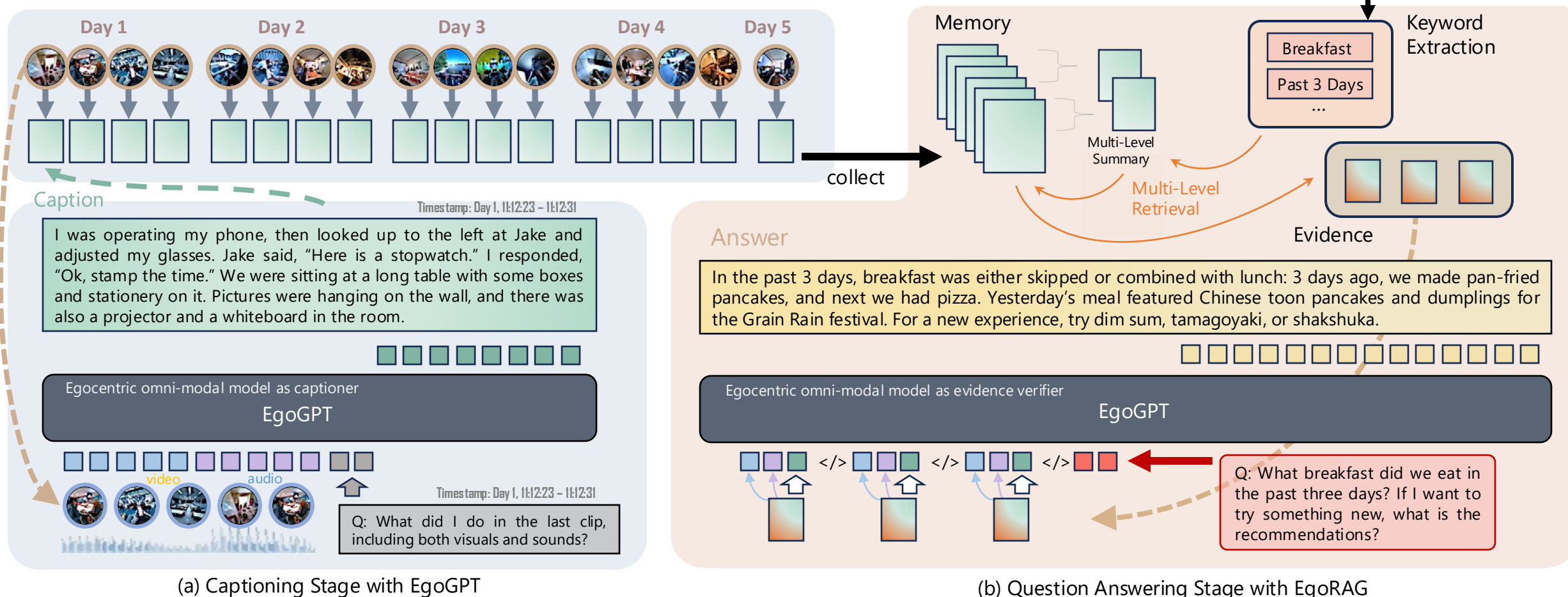
The EgoLifeQA Benchmark



EgoButler



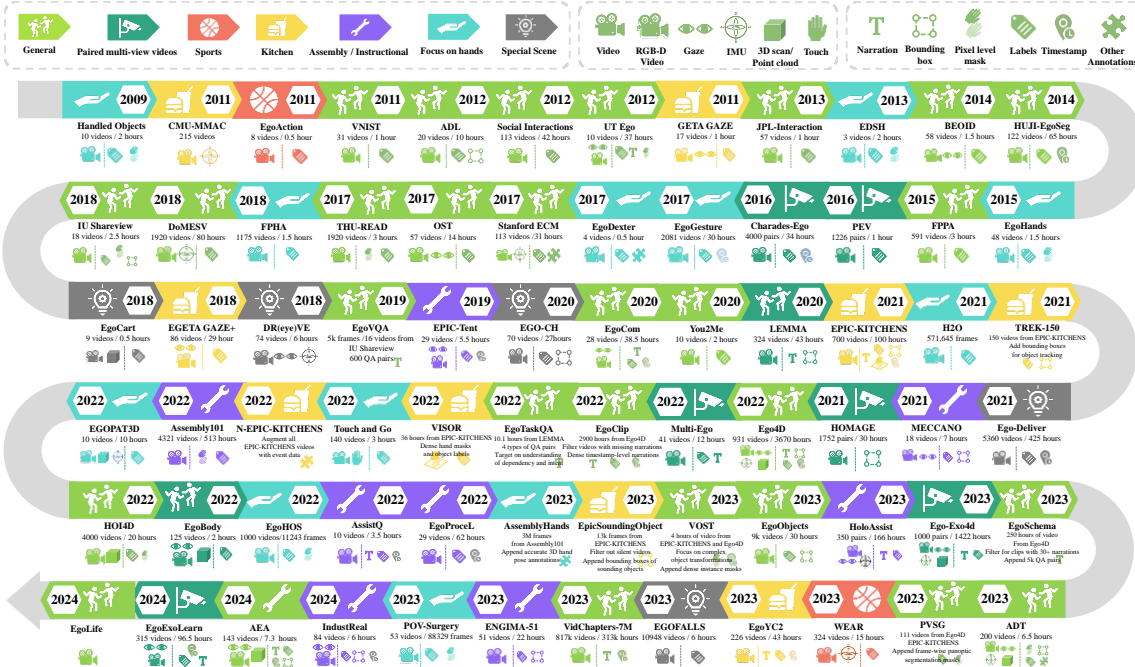
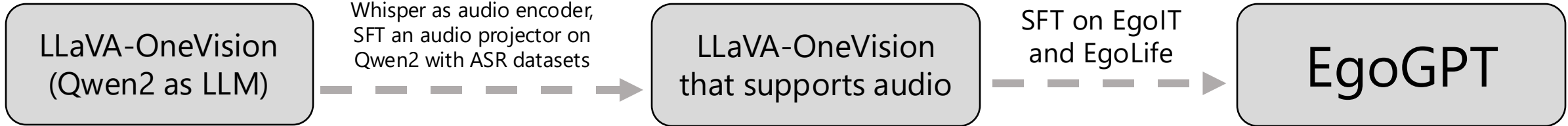
Q: What breakfast did we eat in the past three days? If I want to try something new, what is the recommendations?



(a) Captioning Stage with EgoGPT

(b) Question Answering Stage with EgoRAG

EgoButler – The EgoGPT Component

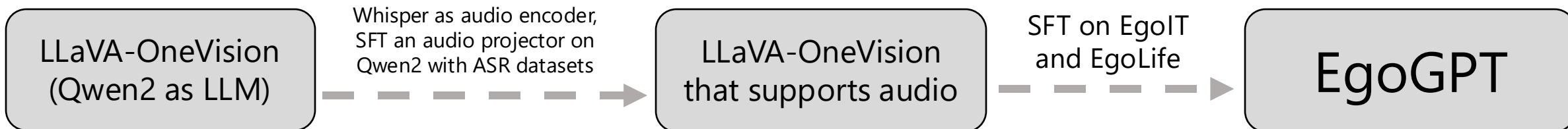


Overview of Classic Egocentric Dataset

Performance of EgoGPT-7B. The table presents a comprehensive comparison of EgoGPT against state-of-the-art commercial and open-source models on existing egocentric benchmarks. With EgoIT and EgoLife Day 1 data, EgoGPT achieve impressive performance on ego setting.

Model	#Param	#Frames	EgoSchema	EgoPlan	EgoThink
GPT-4v [95]	-	32	56.6	38.0	65.5
Gemini-1.5-Pro [96]	-	32	72.2	31.3	62.4
GPT-4o [97]	-	32	72.2	32.8	65.5
LLaVA-Next-Video [98]	7B	32	49.7	29.0	40.6
LongVA [99]	7B	32	44.1	29.9	48.3
IXC-2.5 [100]	7B	32	54.6	29.4	56.0
InternVideo2 [101]	8B	32	55.2	27.5	43.9
Qwen2-VL [94]	7B	32	66.7	34.3	59.3
Oryx [57]	7B	32	56.0	33.2	53.1
LLaVA-OV [55]	7B	32	60.1	30.7	54.2
LLaVA-Videos [102]	7B	32	57.3	33.6	56.4
EgoGPT (EgoIT)	7B	32	73.2	32.4	61.7
EgoGPT (EgoIT+EgoLifeD1)	7B	32	75.4	33.4	61.4

EgoButler – The EgoGPT Component



Dataset Composition of EgoIT-99K. We curated 9 classic egocentric video datasets and utilized their annotations to generate captioning and QA instruction-tuning data for fine-tuning EgoGPT, #AV indicates the number of videos with audio used for training.

Dataset	Duration	#Videos (#AV)	#QA
Ego4D [5]	3.34h	523 (458)	1.41K
Charades-Ego [25]	5.04h	591 (228)	18.46K
HoloAssist [29]	9.17h	121	33.96K
EGTEA Gaze+ [26]	3.01h	16	11.20K
IndustReal [28]	2.96h	44	11.58K
EgoTaskQA [93]	8.72h	172	3.59K
EgoProceL [27]	3.11h	18	5.90K
Epic-Kitchens [4]	4.15h	36	10.15K
ADL [24]	3.66h	8	3.23K
Total	43.16h	1529 (686)	99.48K

Performance of EgoGPT-7B. The table presents a comprehensive comparison of EgoGPT against state-of-the-art commercial and open-source models on existing egocentric benchmarks. With EgoIT and EgoLife Day 1 data, EgoGPT achieve impressive performance on ego setting.

Model	#Param	#Frames	EgoSchema	EgoPlan	EgoThink
GPT-4v [95]	-	32	56.6	38.0	65.5
Gemini-1.5-Pro [96]	-	32	72.2	31.3	62.4
GPT-4o [97]	-	32	72.2	32.8	65.5
LLaVA-Next-Video [98]	7B	32	49.7	29.0	40.6
LongVA [99]	7B	32	44.1	29.9	48.3
IXC-2.5 [100]	7B	32	54.6	29.4	56.0
InternVideo2 [101]	8B	32	55.2	27.5	43.9
Qwen2-VL [94]	7B	32	66.7	34.3	59.3
Oryx [57]	7B	32	56.0	33.2	53.1
LLaVA-OV [55]	7B	32	60.1	30.7	54.2
LLaVA-Videos [102]	7B	32	57.3	33.6	56.4
EgoGPT (EgoIT)	7B	32	73.2	32.4	61.7
EgoGPT (EgoIT+EgoLifeD1)	7B	32	75.4	33.4	61.4

EgoButler – The EgoRAG Component




Boosted by EgoGPT, EgoButler achieves SOTA:

- In-depth egocentric video familiarity
- Omni-modal comprehension — effectively integrating both visual and audio signals

Powered by EgoRAG, EgoGPT enables:

- Week-long memory retrieval, answering complex, long-horizon questions
- Robust grounding and context-aware reasoning, where others often fail

Limitations

-  One-Time Retrieval → Agentic Search
-  Better Person Identification Modeling
-  Pattern Tracker: Building a habit and behavior pattern engine for continuous insight generation

DAY 2 11:23AM

Gemini-1.5-pro

I am sitting at a table covered with a pink and white checkered tablecloth with **three** other people. We have finished our meal, which consisted of food in a large pot. [...] My bowl has a yellowish chunk that I'm breaking into smaller pieces with my chopsticks. [...] I listen to the others talking. One says something about 'remembering!', another replies "um," and then the first person says, "After connecting, memory accelerates. One second to learn a dance." I dip my chopsticks into my dish and eat, then I pick up my **glass** for a sip as the **first speaker** says, "Try it." I drink more milk, **the woman to my right** puts a piece of food onto my dish, and I ask her, "**Is it salty?**" [...]

EgoGPT

[...] The table was covered with a checkered tablecloth, filled with various foods and drinks, and there were a few bouquets of flowers adding a touch of warmth to the scene. I adjusted my glasses and picked up the chopsticks on the table. [...] Then, I placed the chopsticks on the table, picked up a spoon, and started stirring the **bread** in the bowl. As I stirred, [...] I smiled, "One minute memory training." **Lucia** laughed and said, "Hahaha, memory training." I continued to stir the food in the bowl, then picked up a glass of milk from the table and took a sip. I asked, "Whose is this?" **Tasha** replied, "This one is really delicious." I chuckled and said, "Hahaha, whose is it?"

EgoRAG (Gemini-1.5-pro)

[...] I walk and see two women, one examining refrigerated items and the other using a smartphone near a **dairy** section. DAY 1 17:29

I'm holding chopsticks and picking up a piece of food from a small, white bowl of rice, placing it onto [...] the rice bowl. DAY 2 13:41

I enter a room where several people are standing around a **long table with food**. I speak a sentence, but my voice isn't in English... DAY 2 21:04

Question DAY 6 15:48

What was the first food I ate along with milk?
A. Banana B. Pancake
C. Eggs D. Cookie
Correct Ans: B

A. Banana ✗ B. Pancake ✓

EgoRAG (EgoGPT)

[...] I moved the straw slightly DAY 2 13:51 with my left hand, placed a finger on the drink, and fell into thought. The table was filled with [...]

[...] picked up a spoon, and DAY 1 11:26 started stirring the bread in the bowl [...] then picked up a glass of milk from the table [...]

[...] Katrina asked, "Where DAY 1 19:11 should I put this?" then said, "I'll do it." Tasha reminded, "There's still a bottle of fresh milk [...]"

Table 5. Performance comparison of EgoGPT with state-of-the-art models on EgoLifeQA benchmarks. For a fair comparison on EgoLifeQA, EgoGPT was replaced with the corresponding models in the EgoButler pipeline to evaluate their performance under the same conditions. Models that provide captions for EgoLifeQA use 1 FPS for video sampling.

Model	#Frames	Audio	Identity	EgoLifeQA					
				EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Average
Gemini-1.5-Pro [95]	-	✓	✗	36.0	37.3	45.9	30.4	34.9	36.9
GPT-4o [96]	1 FPS	✗	✗	34.4	42.1	29.5	30.4	44.4	36.2
LLaVA-OV [55]	1 FPS	✗	✗	36.8	34.9	31.1	22.4	28.6	30.8
EgoGPT (EgoIT-99K)	1 FPS	✓	✗	35.2	36.5	27.9	29.6	36.5	33.1
EgoGPT (EgoIT-99K+D1)	1 FPS	✓	✓	39.2	36.5	31.1	33.6	39.7	36.0



Towards

Extremely Long, Egocentric, Interpersonal, Multi-view, Multi-modal, Daily Life Video Understanding



More to explore:

Dense Caption, Transcript,
Gaze, Multiple Third-Person
View, SLAM

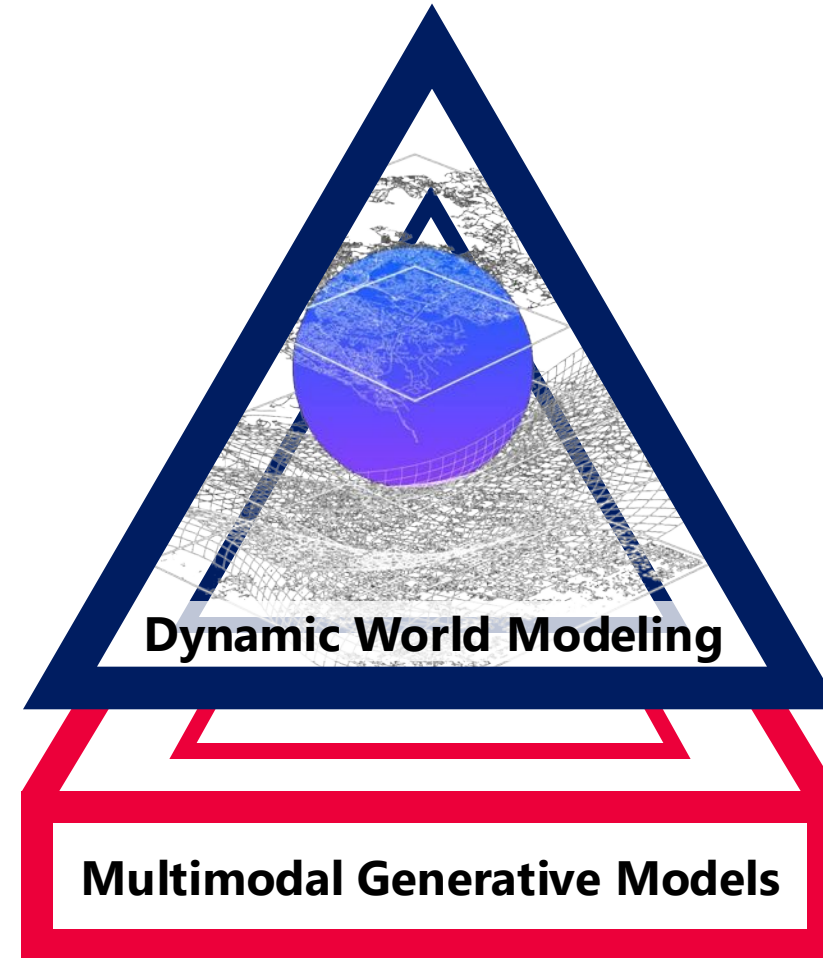
egolife-ai.github.io

Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model
Dynamic Scenes



Be Social

How to Model Social
Interactions



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

S-LAB
FOR ADVANCED
INTELLIGENCE

Thank You

Haozhe Xie 谢浩哲

Nanyang Technological University

