

# From Multimodal Generative Models to Dynamic World Modeling

Ziwei Liu 刘子纬

Nanyang Technological University

<https://liuziwei7.github.io>

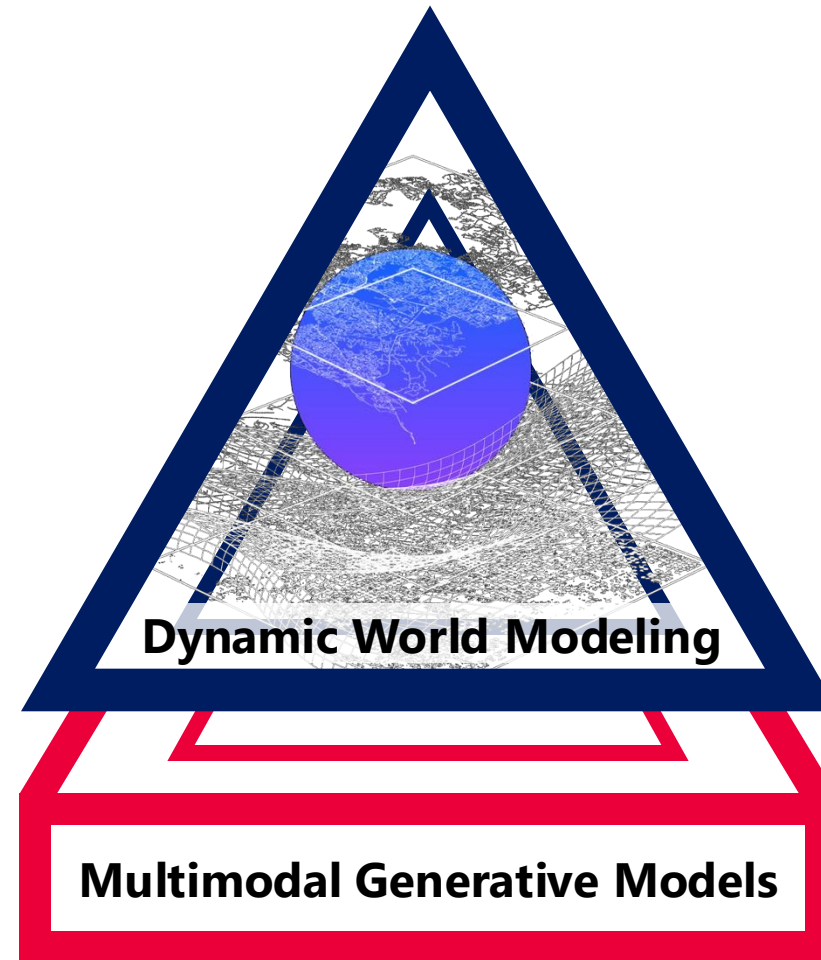


# Be Physical

How to Model Material and Illumination

## Be Dynamic

How to Model  
Dynamic Scenes



Dynamic World Modeling

Multimodal Generative Models

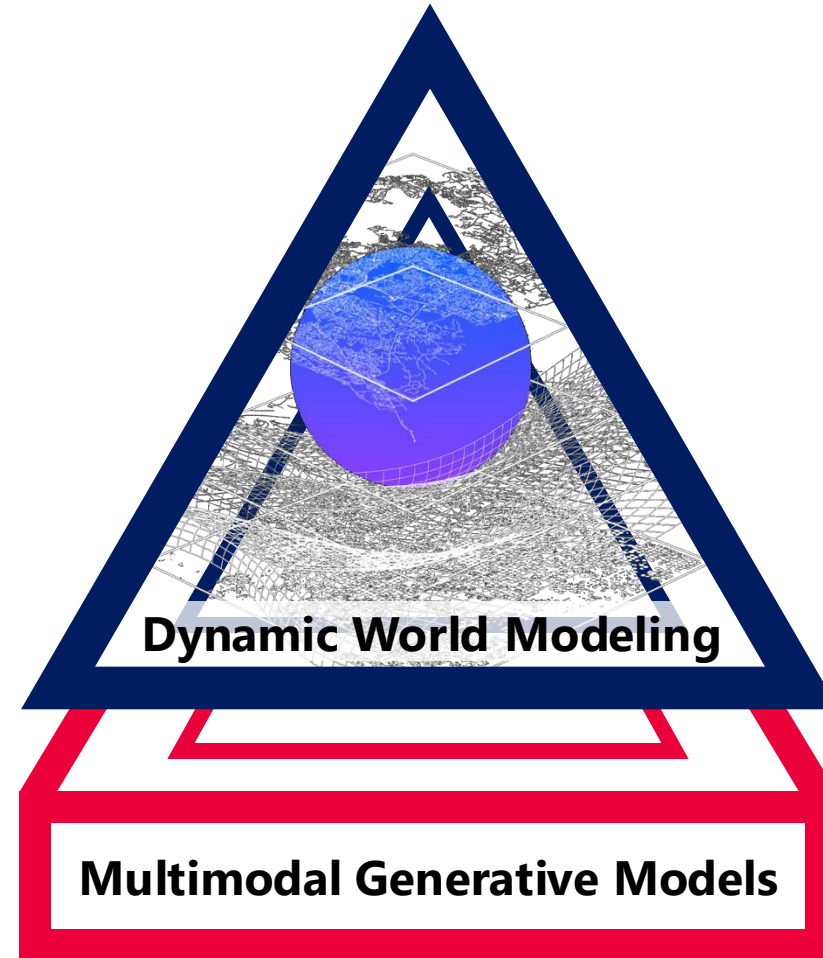
## Be Social

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/ 3DTopia-XL](https://github.com/3DTopia/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

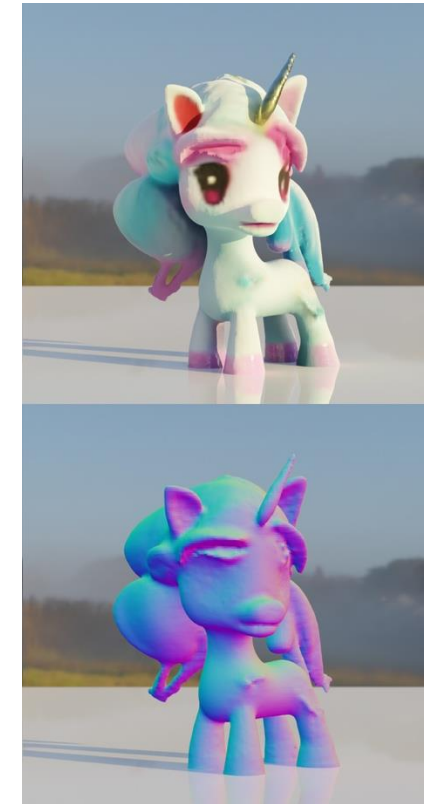
CVPR 2025 Highlight

# 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



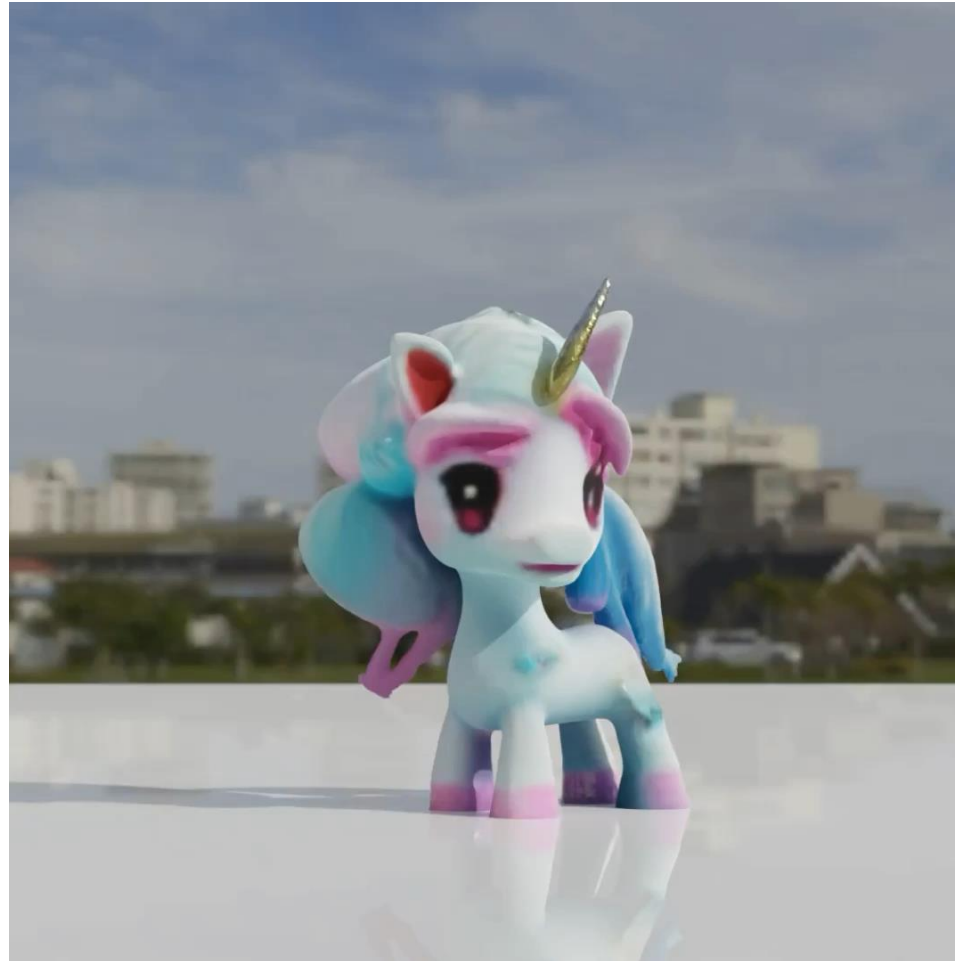
Previous SOTA



Our Goal

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

"A cute unicorn"

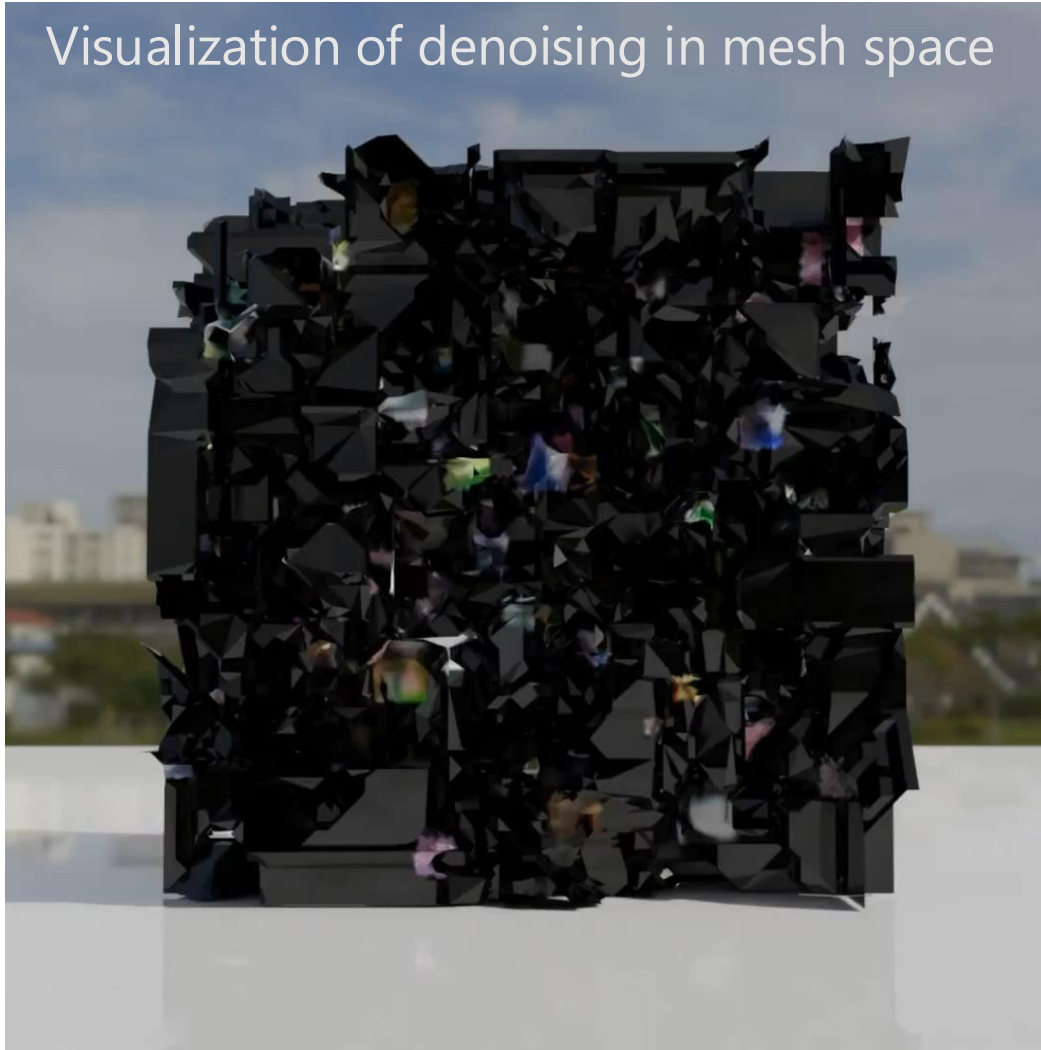


A Single Image / Texts

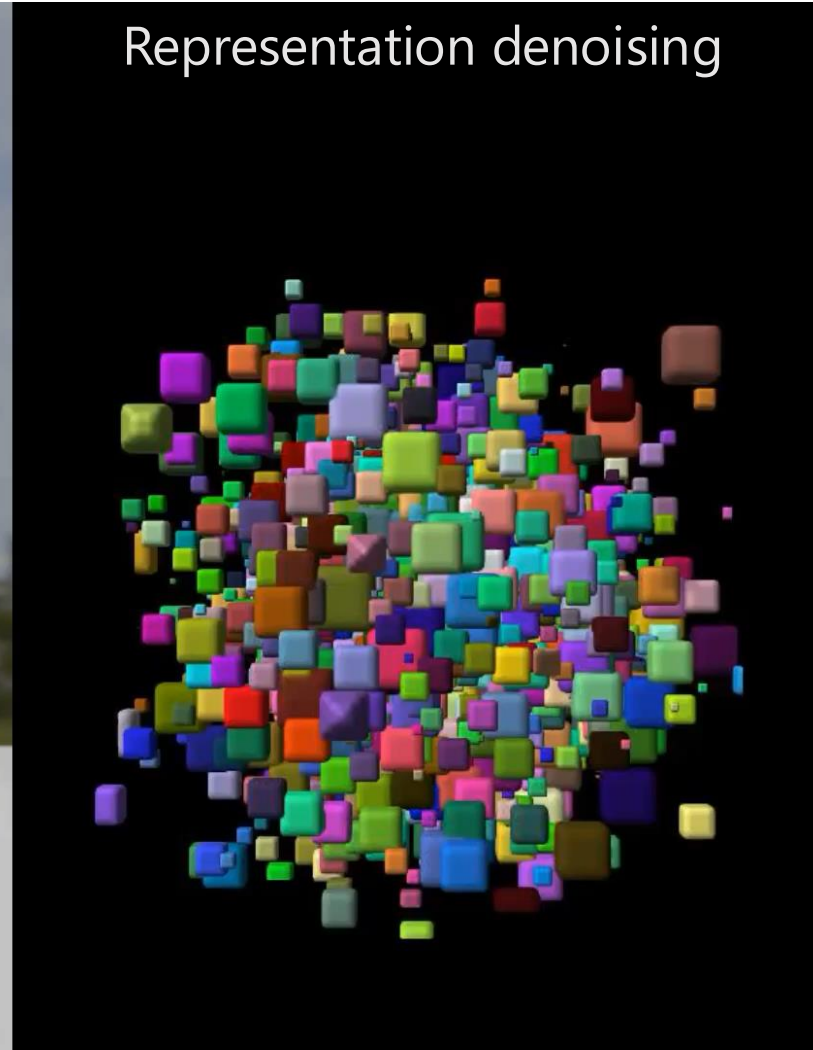
High-quality 3D Asset Ready for Blender 

# Key Idea: Primitive Diffusion

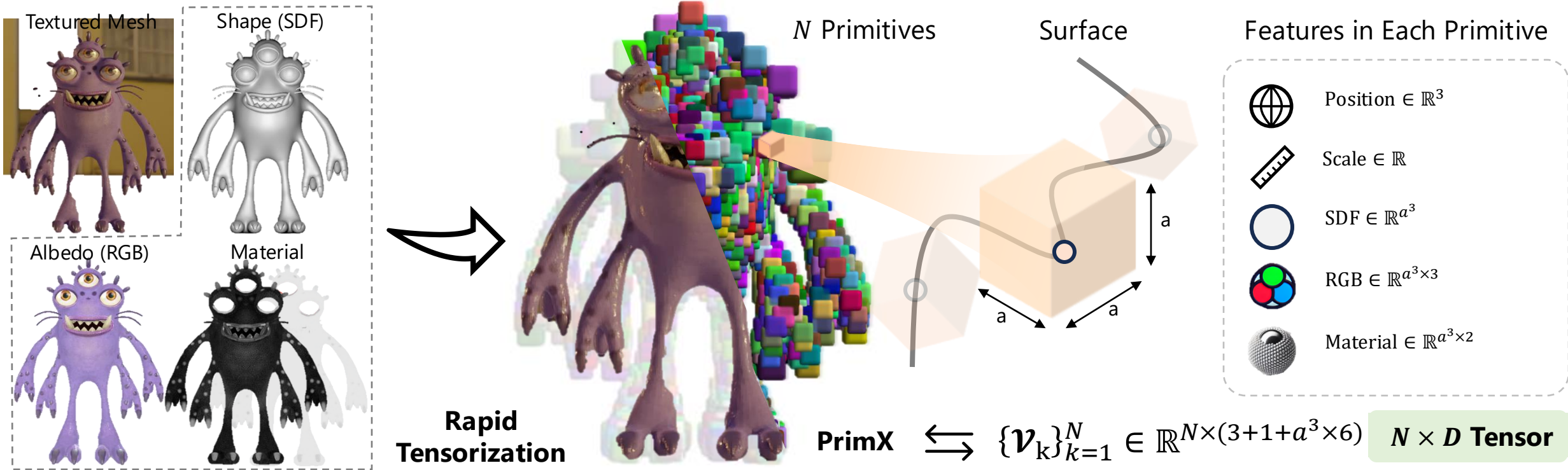
Visualization of denoising in mesh space



Representation denoising



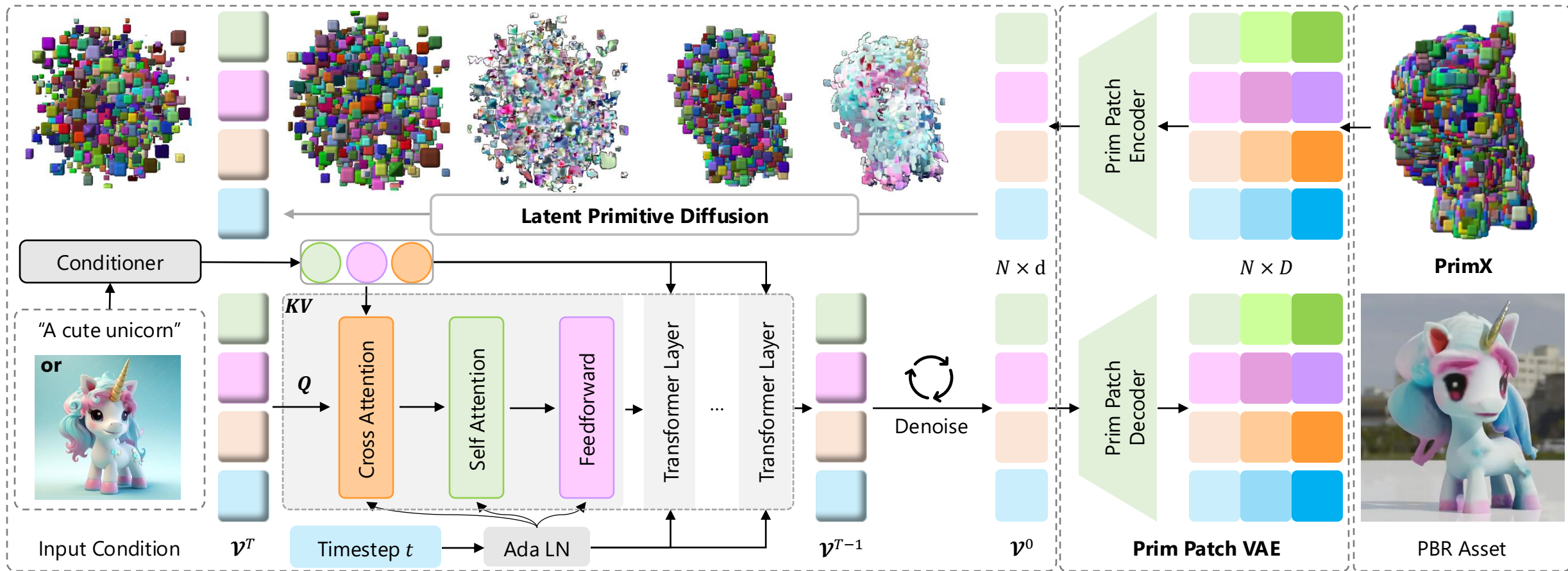
# Stage I: Geometry, Texture, Materials into $N \times D$ Primitives



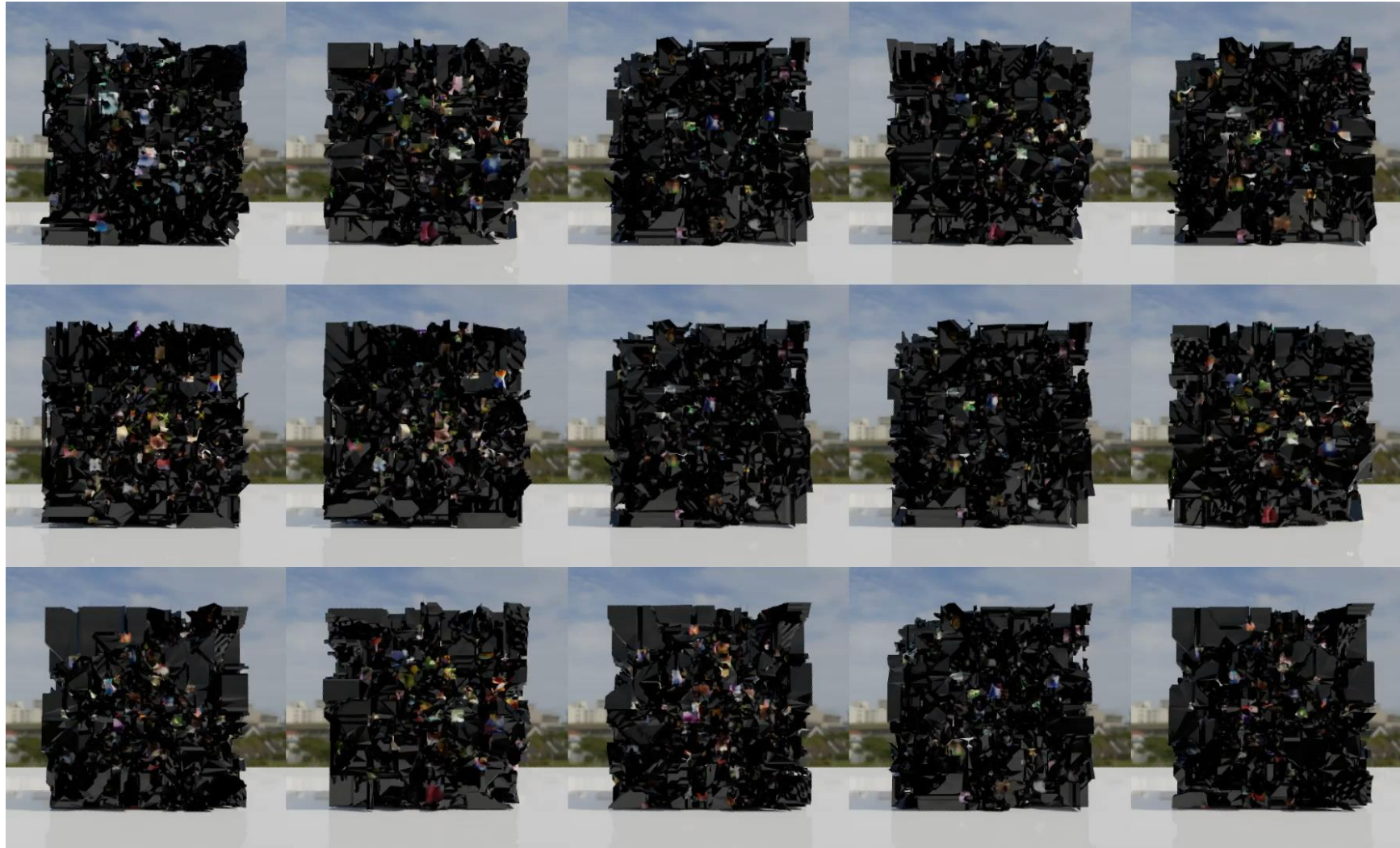
Tensorize a Textured Mesh into  $N \times D$  Primitives



# 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)

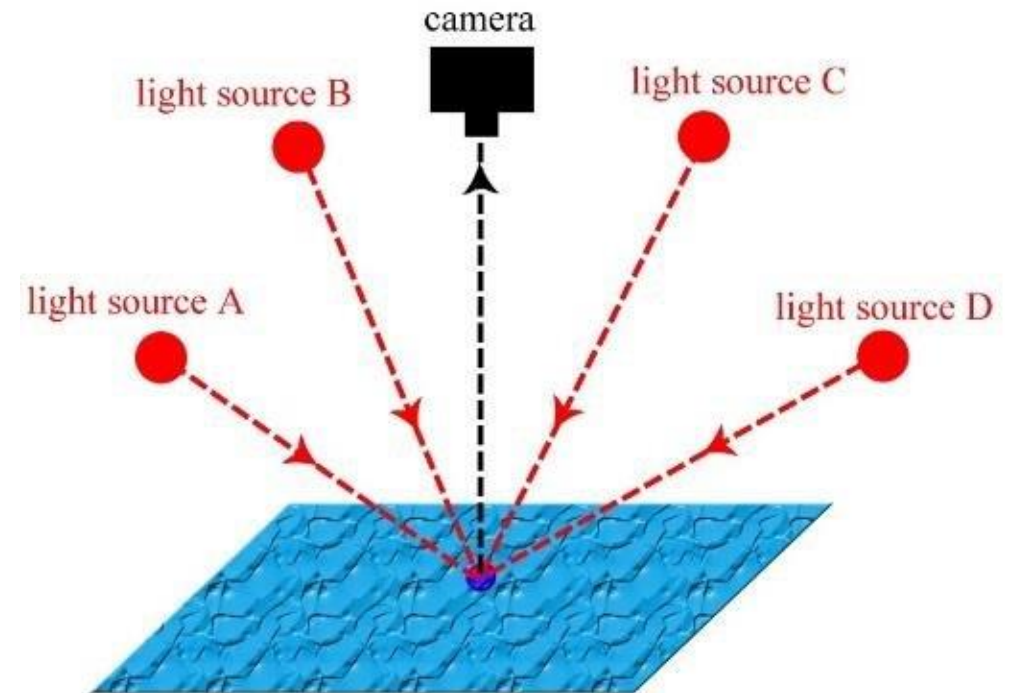
Neural LightRig: Unlocking Accurate Object Normal and Material Estimation with Multi-Light Diffusion

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-constrained**
- Complex interaction among geometry, materials, and environmental lighting
- Traditional methods need photometric stereo setups<sup>[1]</sup> – **impractical** for in-the-wild images

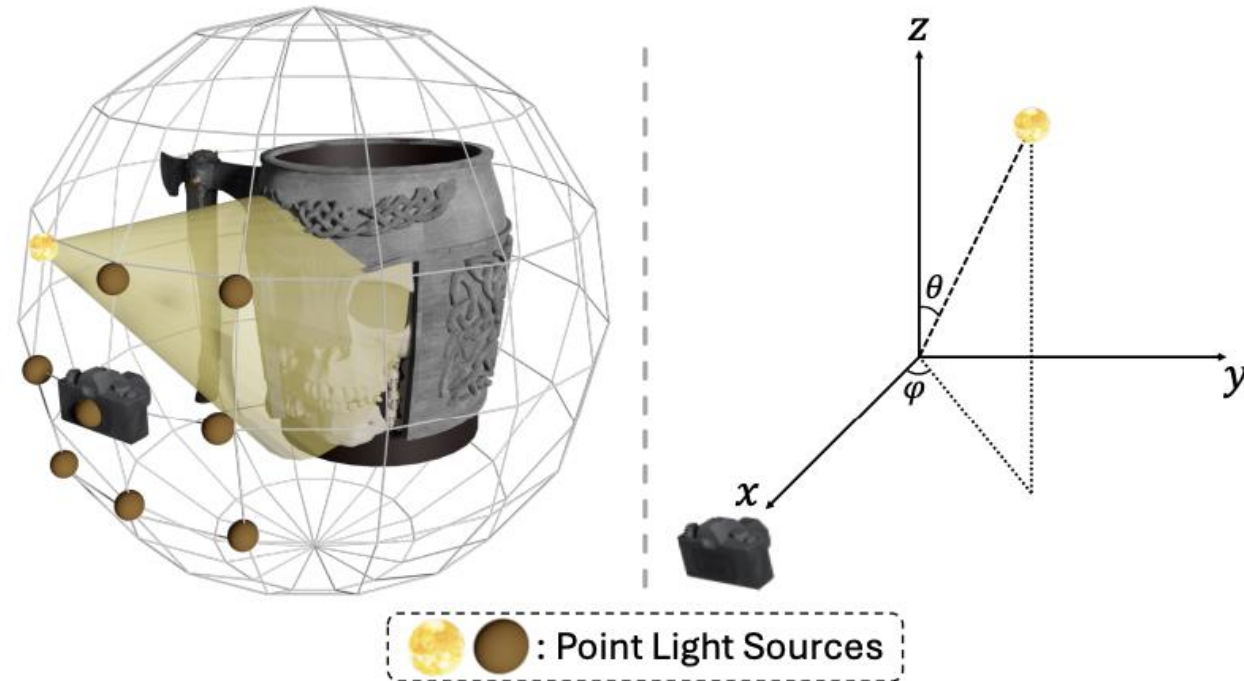


[1] Robert J. Woodham. *Photometric method for determining surface orientation from multiple images*. 1989.

[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](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<sup>[1]</sup>
- Relighting diffusion models can synthesize images under various lighting conditions<sup>[2]</sup>
- 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

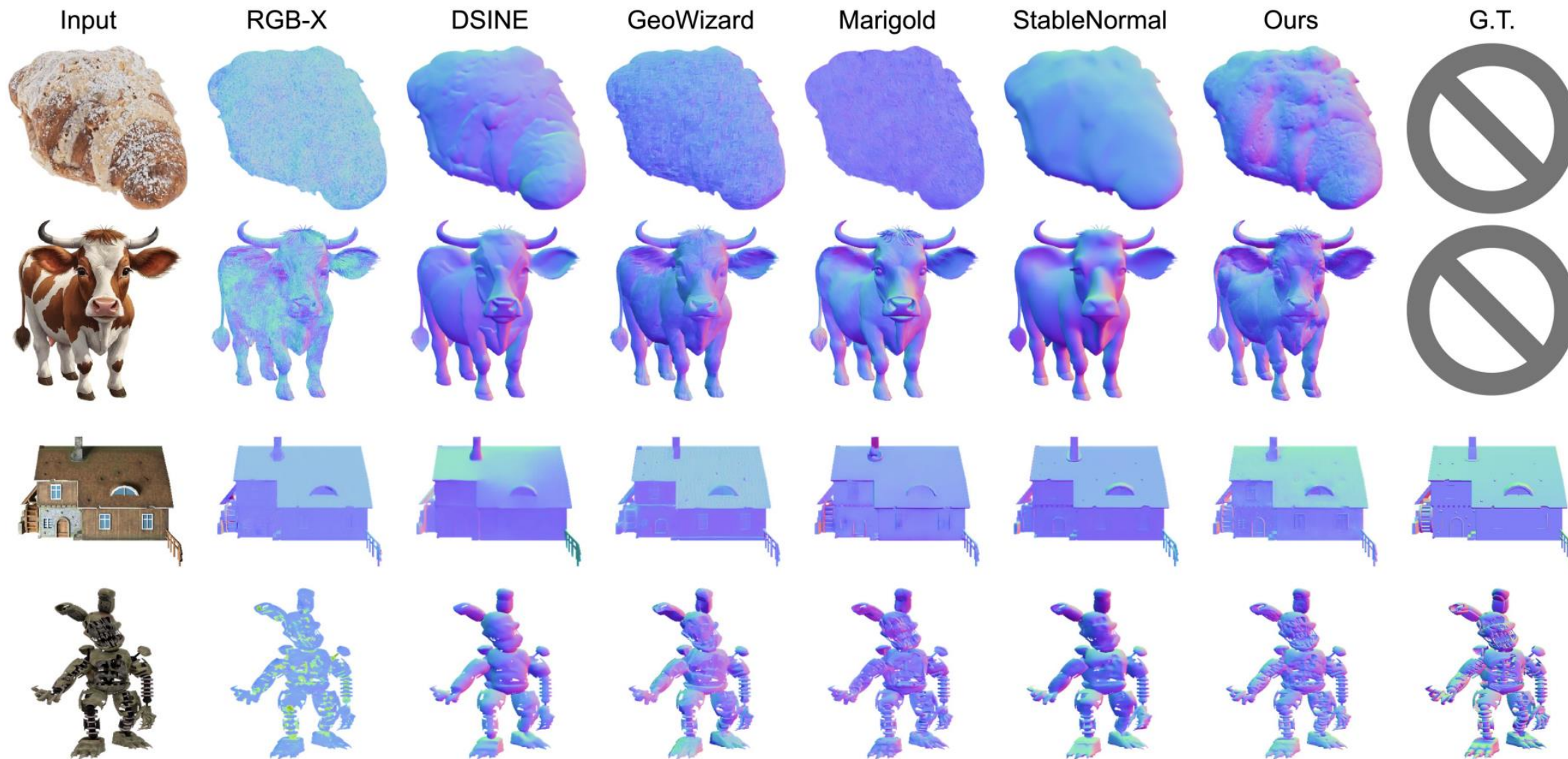
Method	Mean ↓	Median ↓	3° ↑	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑
RGB↔X [57]	14.847	13.704	11.676	23.073	35.196	49.829	75.777	86.348
DSINE [2]	9.161	7.457	23.565	41.751	57.596	72.003	90.294	95.297
GeoWizard [16]	8.455	6.926	22.245	40.993	58.457	74.916	93.315	<u>97.162</u>
Marigold [25]	8.652	7.078	<u>25.219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

## ■ PBR Estimation and Single-Image Relighting

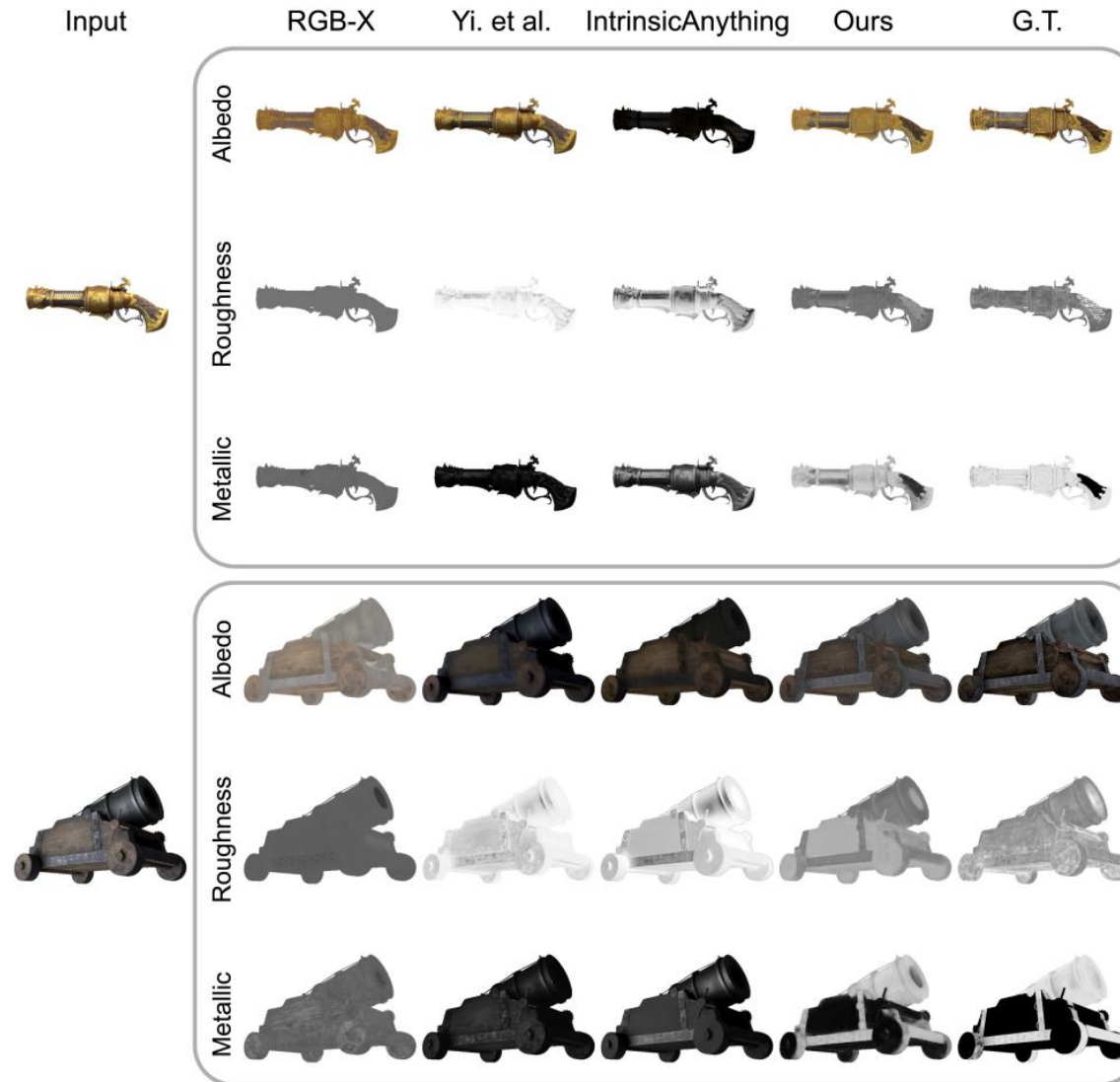
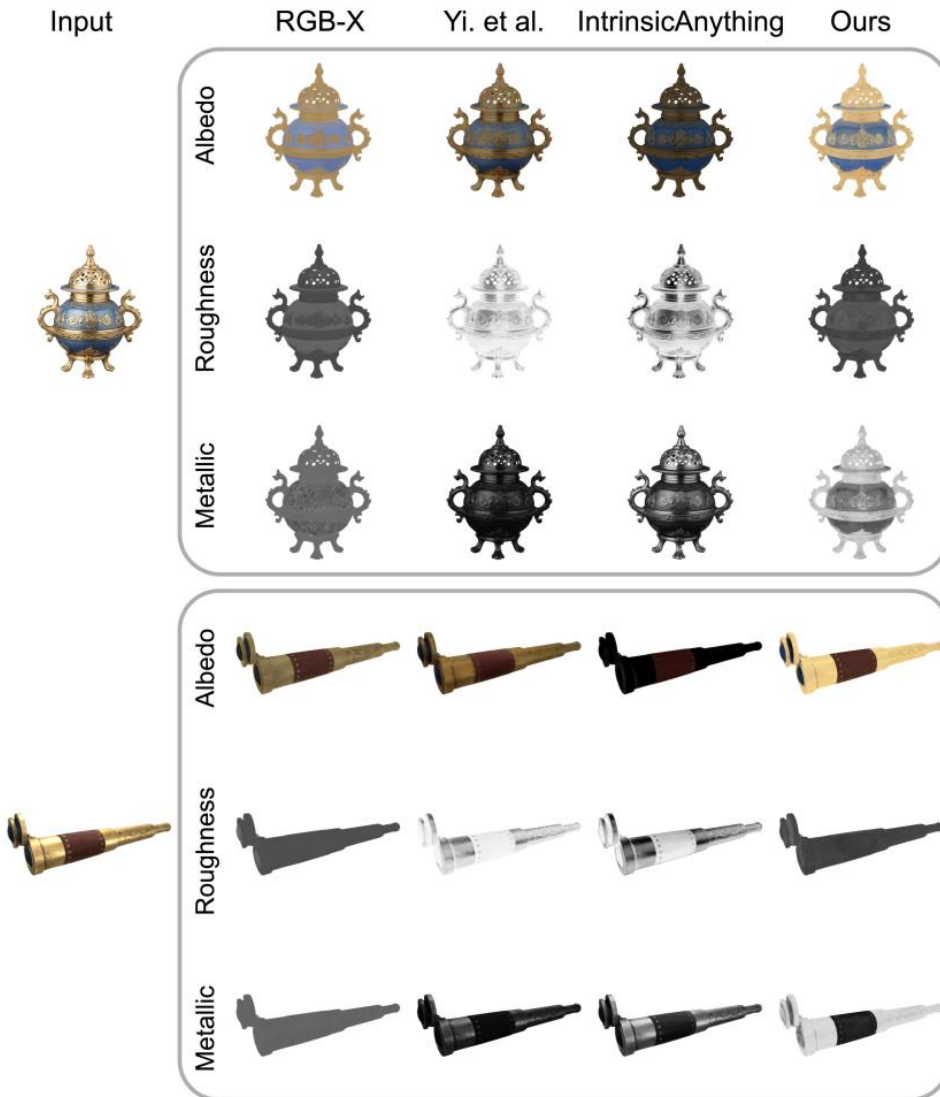
Method	Albedo		Roughness		Metallic		Relighting			Latency
	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Average Time ↓
RGB↔X [57]	16.26	0.176	<u>19.21</u>	<u>0.134</u>	16.65	0.199	20.78	0.8927	0.0781	15s
Yi. et al [54]	21.10	0.106	16.88	0.180	20.30	0.144	26.47	0.9316	0.0691	5s
IntrinsicAnything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



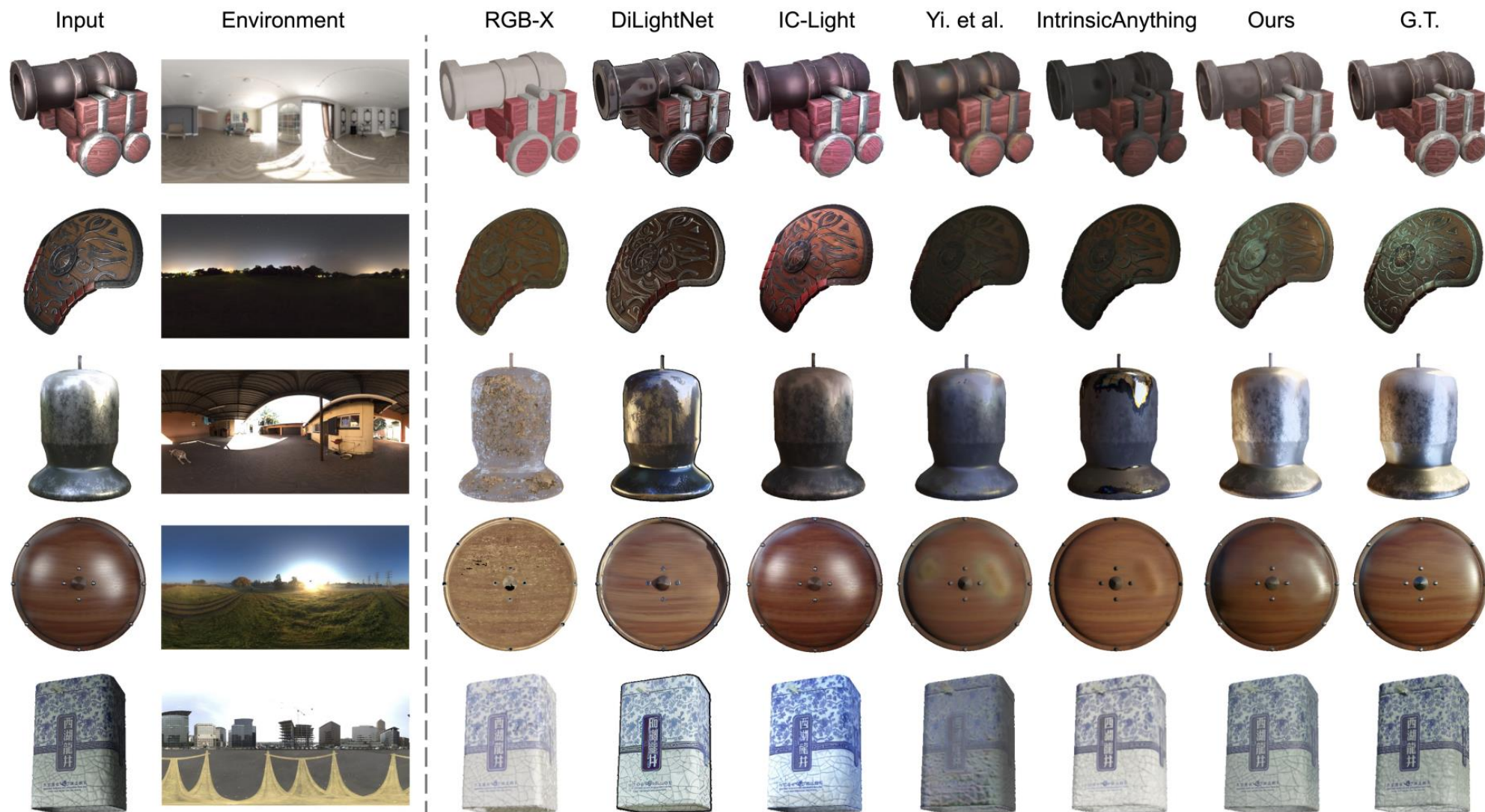
# Surface Normal Estimation



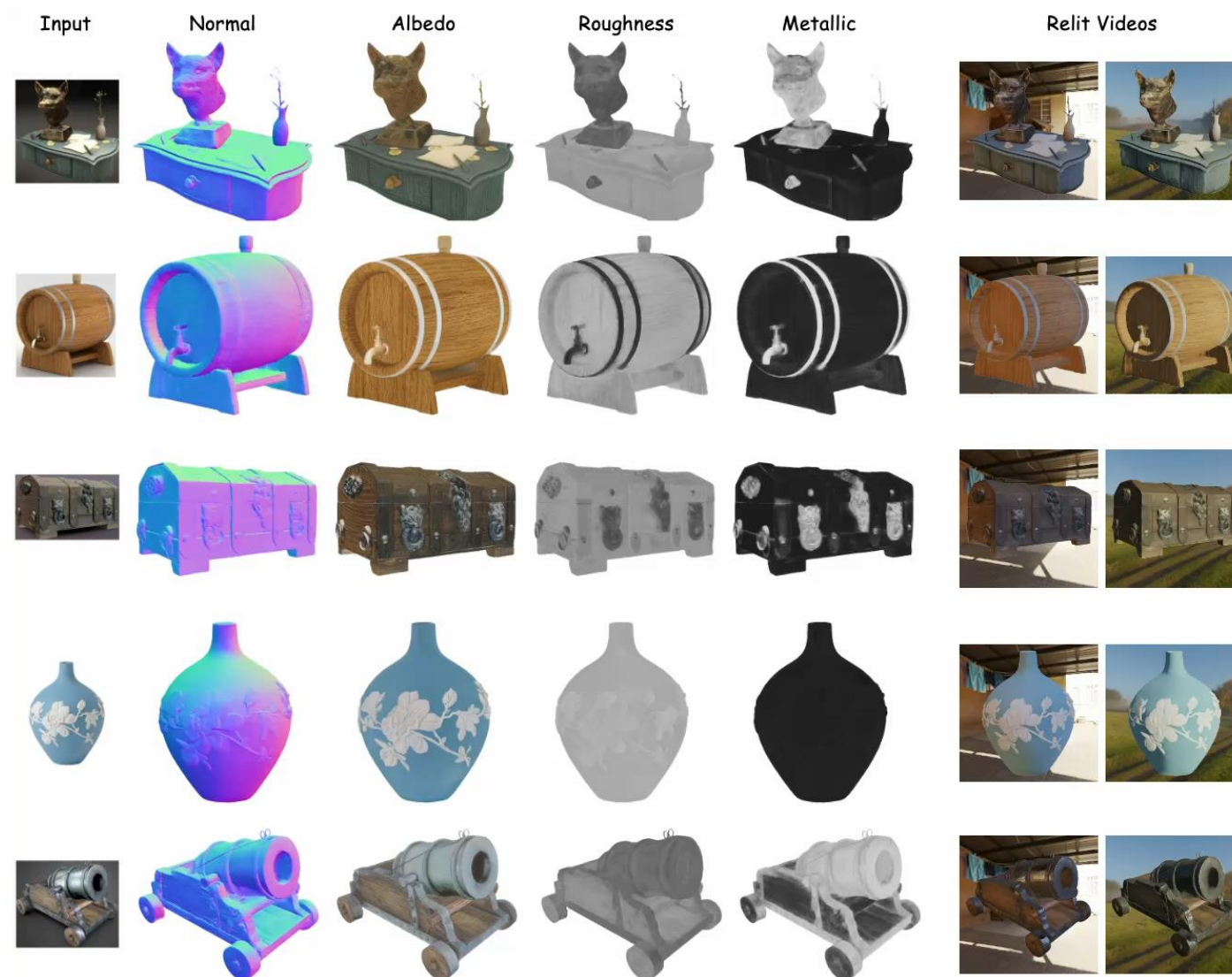
# PBR Material Estimation



# Single-Image Relighting



# Single-Image Relighting

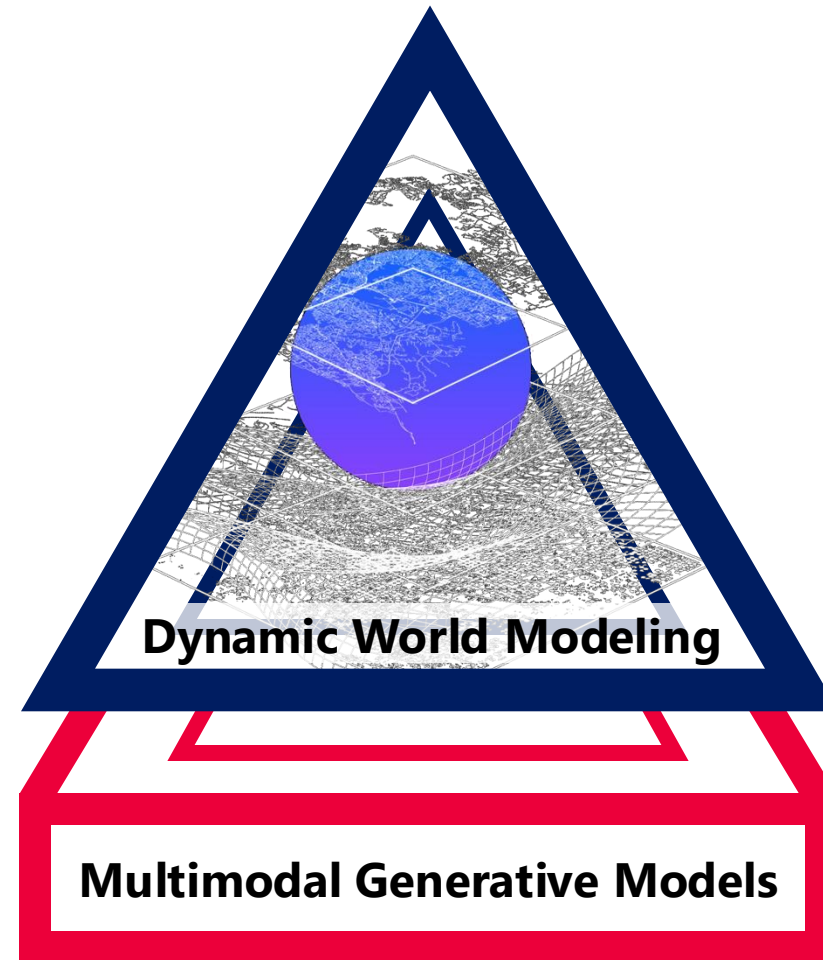


# Be Physical

How to Model Material and Illumination

Be **Dynamic**

How to Model  
Dynamic Scenes



Dynamic World Modeling

Multimodal Generative Models

Be Social

How to Model Social  
Interactions



# Be Dynamic: DynamicCity



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

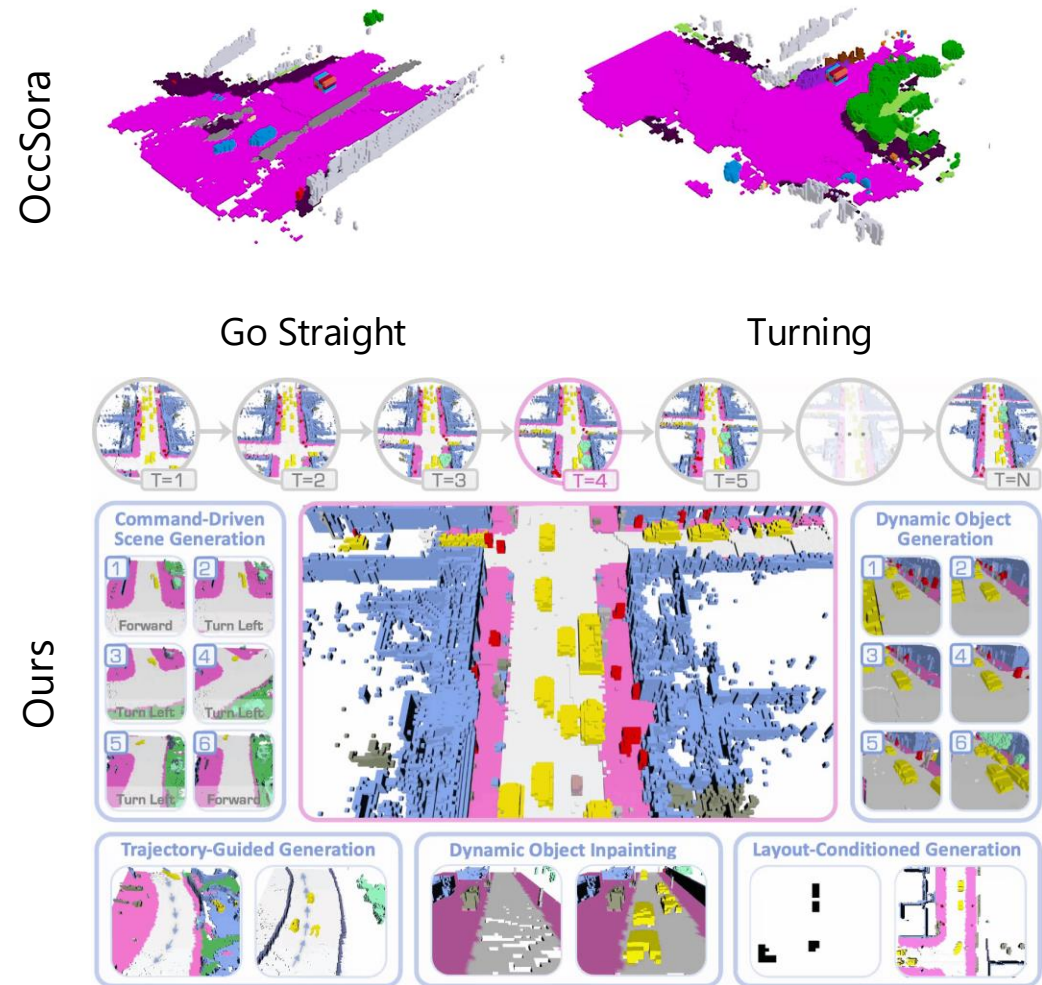
DynamicCity: Large-Scale 4D Occupancy Generation from Dynamic Scenes

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

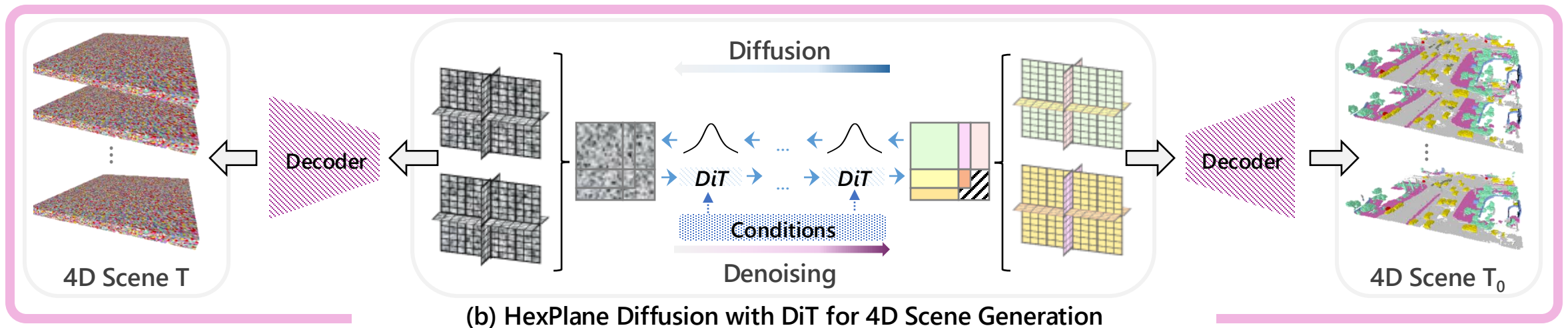
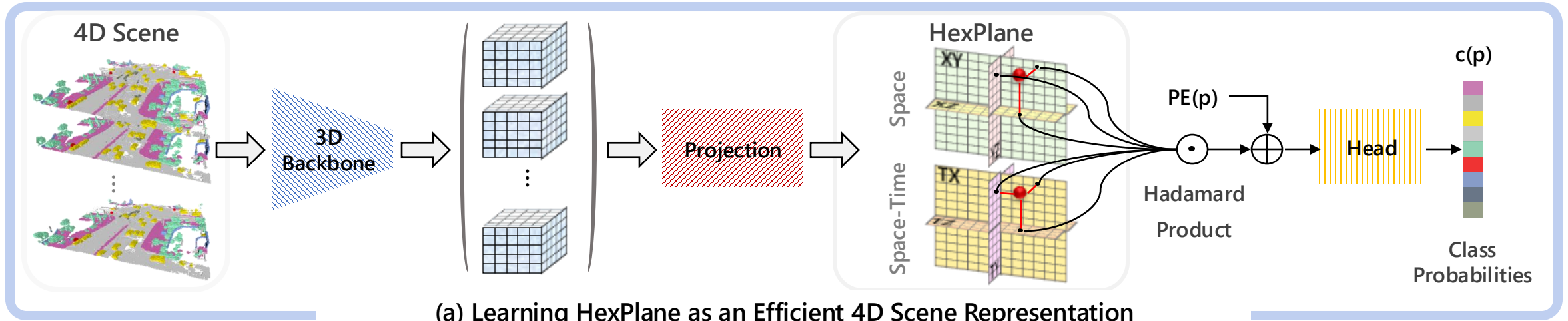
ICLR 2025 Spotlight

# Challenges

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



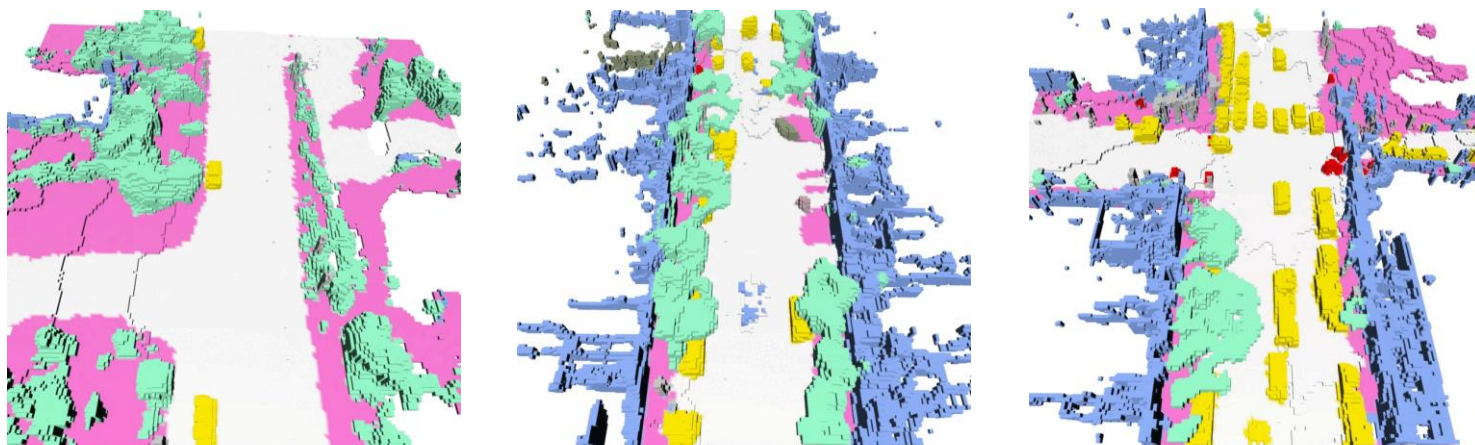
# DynamicCity: 4D Occupancy Generation



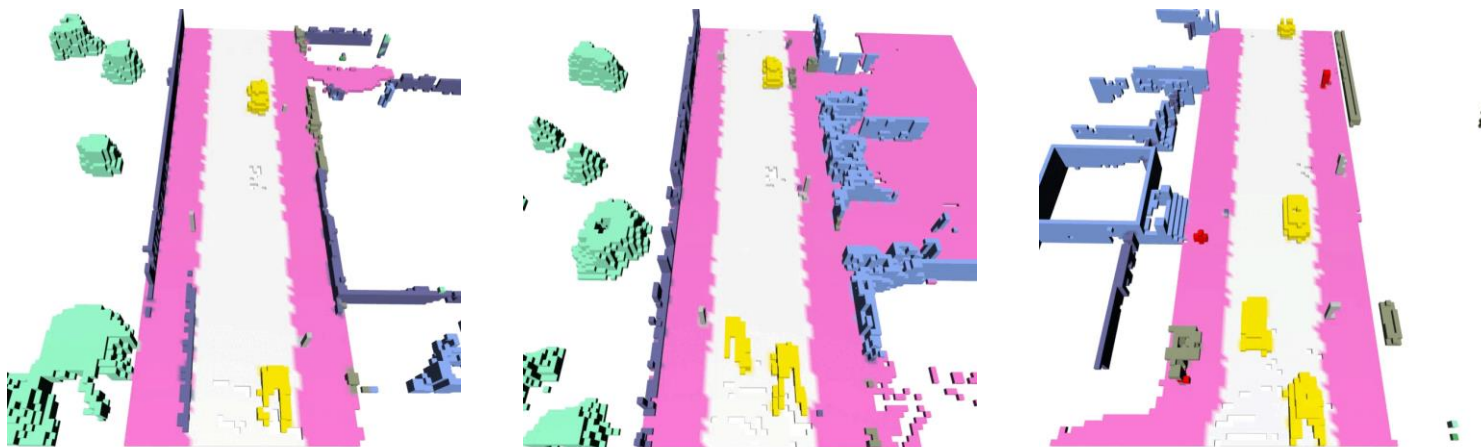


# Unconditional 4D Generation

Occ3D-Waymo

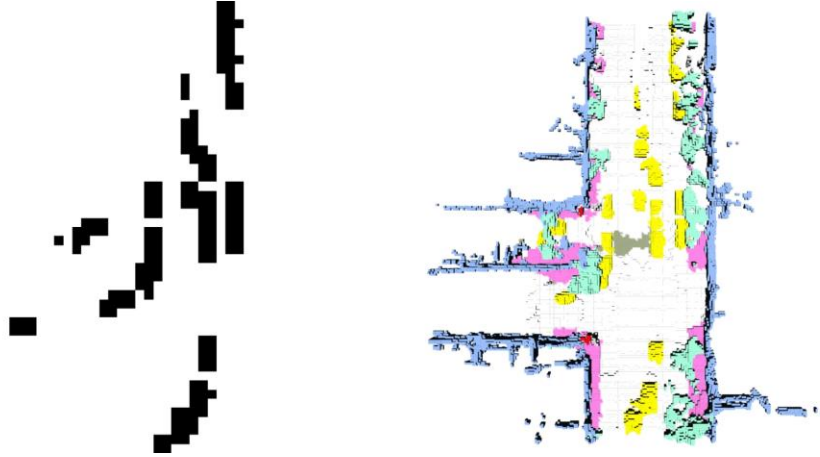


CarlaSC

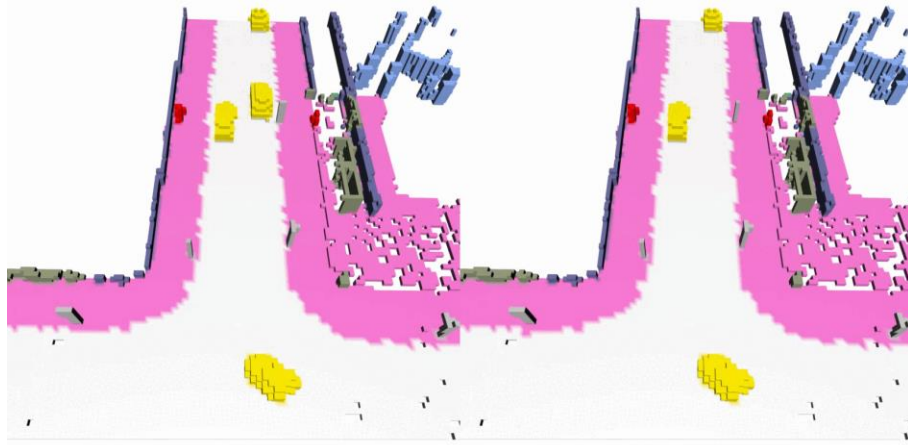


# Conditional 4D Generation

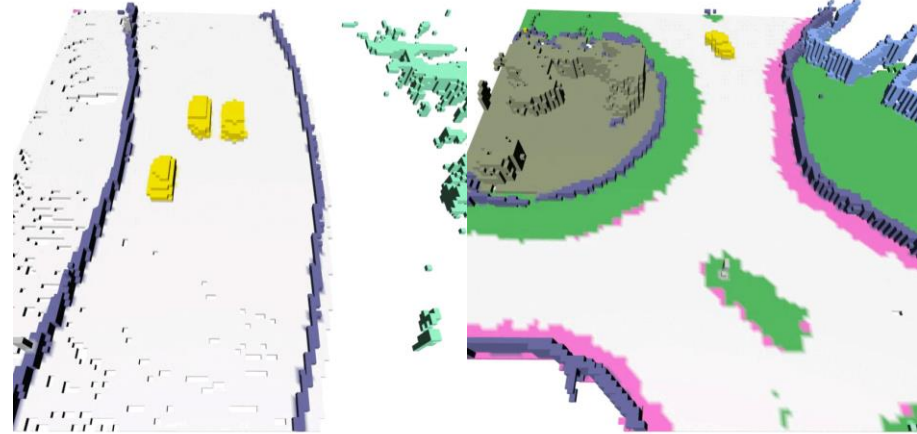
Layout-conditioned



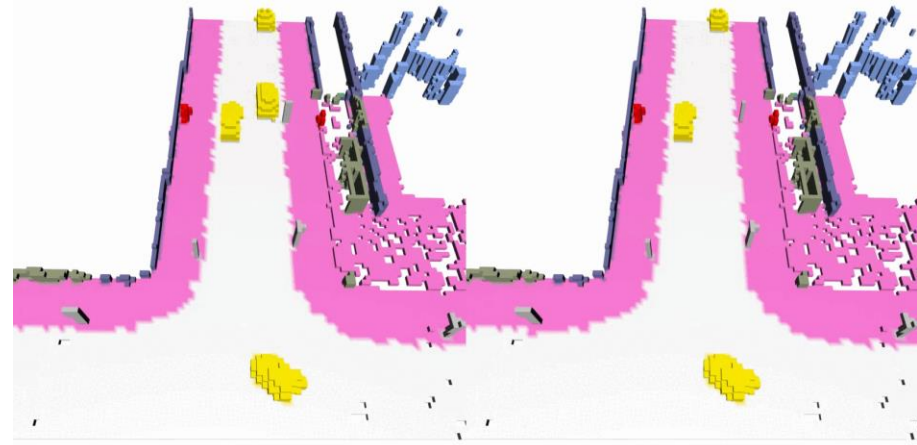
Inpainting



Trajectory-conditioned



Outpainting





# Be Dynamic: CityDreamer4D



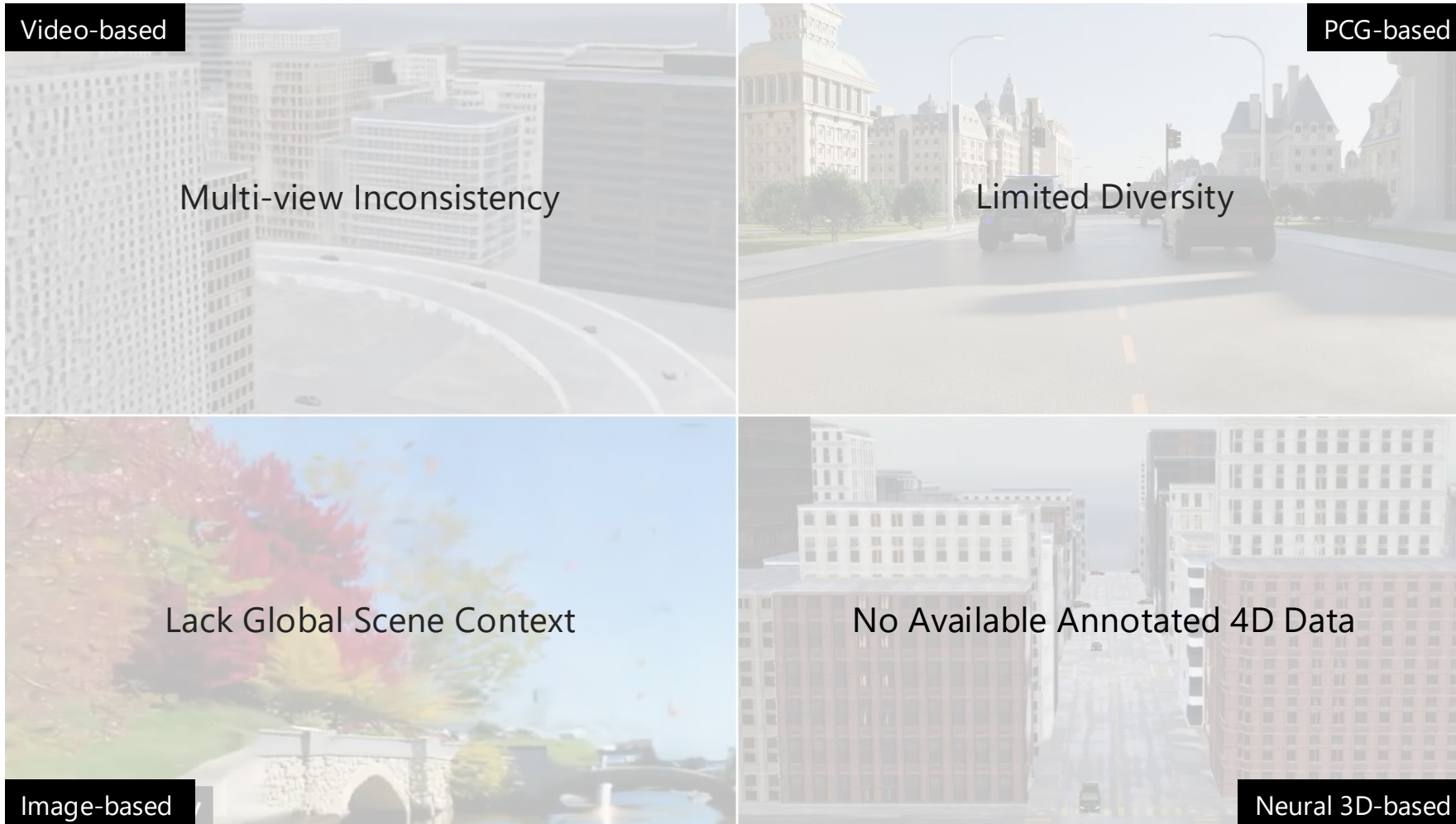
[hzxie/CityDreamer4D](https://github.com/hzxie/CityDreamer4D)

CityDreamer4D: Compositional Generative Model of Unbounded 4D Cities

Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, Ziwei Liu

arXiv 2501.08983

# Challenges



[1] Wonderjourney: Going from Anywhere to Everywhere. CVPR 2024.

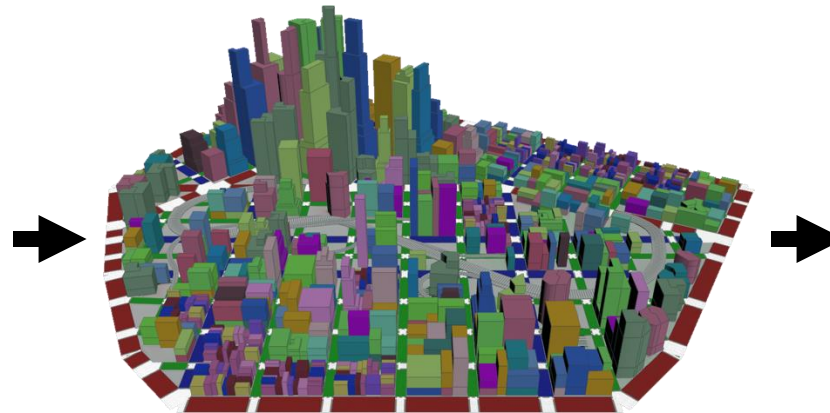
[2] CityX: Controllable Procedural Content Generation for Unbounded 3D Cities. arXiv 2407.17572.

[3] DimensionX: Create Any 3D and 4D Scenes from a Single Image with Controllable Video Diffusion. arXiv 2411.04928.

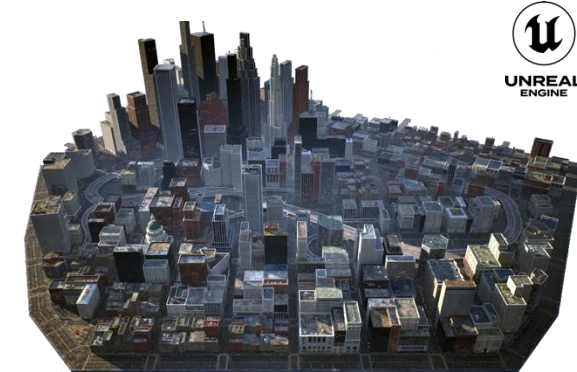
# Learning 4D City from 3D Data Annotations



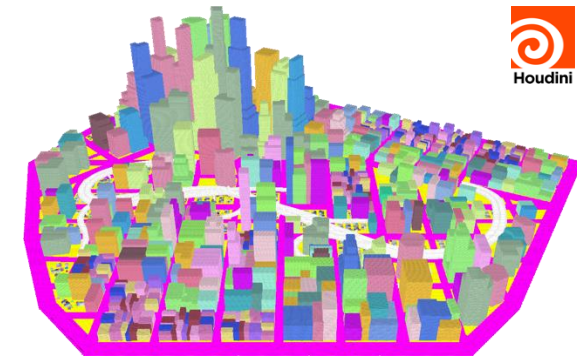
3D Assets  
(Small set for Visualization)



City Prototype



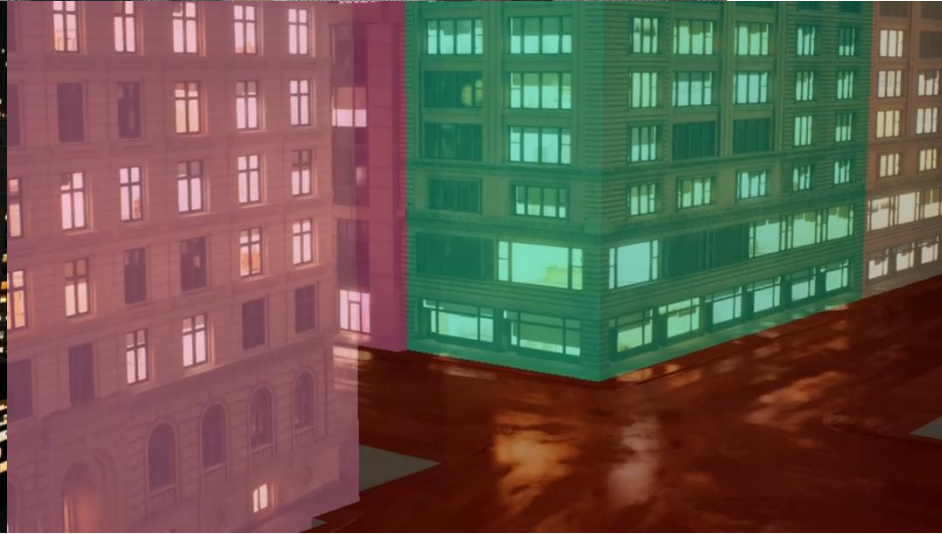
Generated 3D City



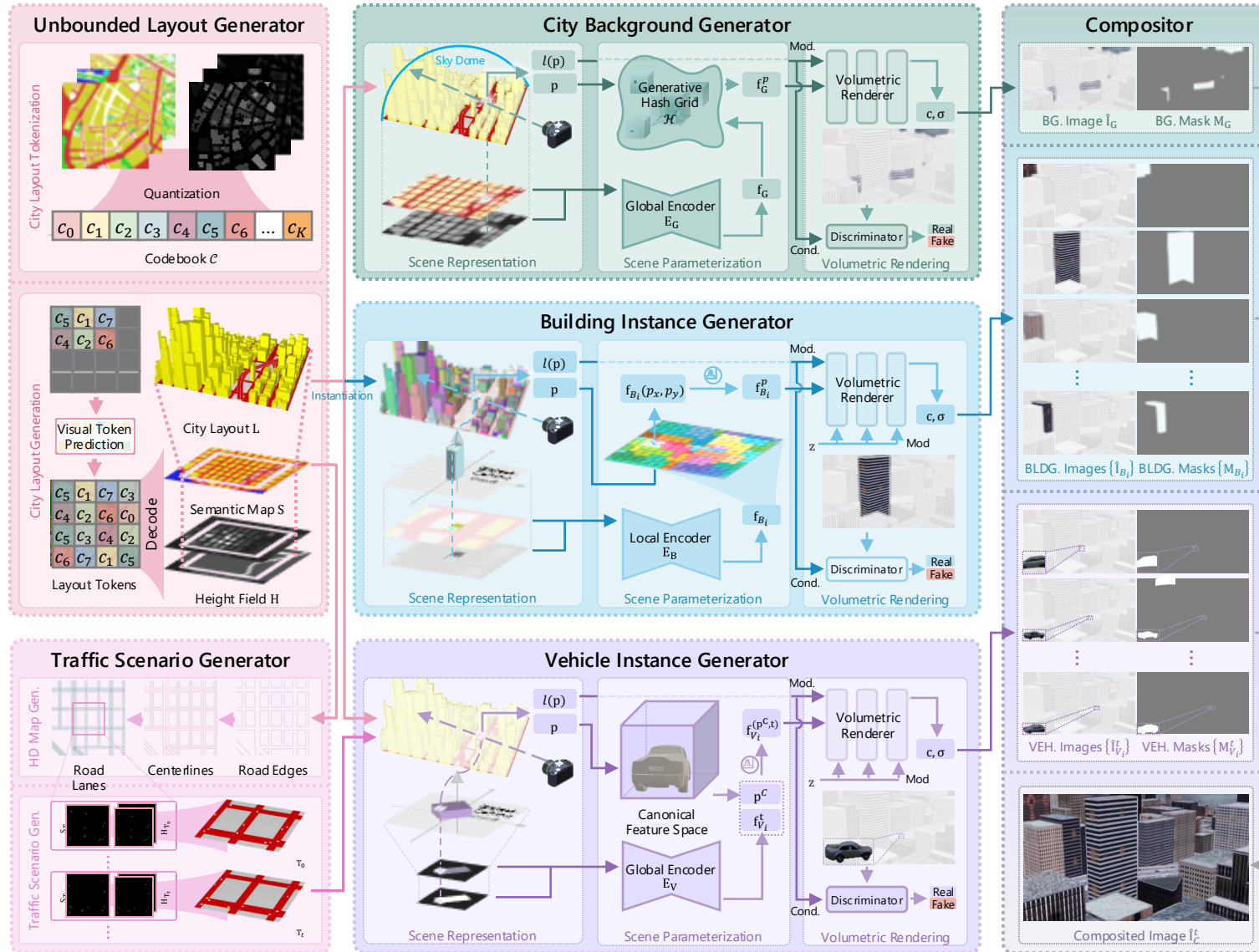
3D Instance Annotation



# The CityTopia Dataset



# CityDreamer4D Framework



# Comparison to SOTA Methods

InfiniCity



SceneDreamer



PersistentNature



CityDreamer4D





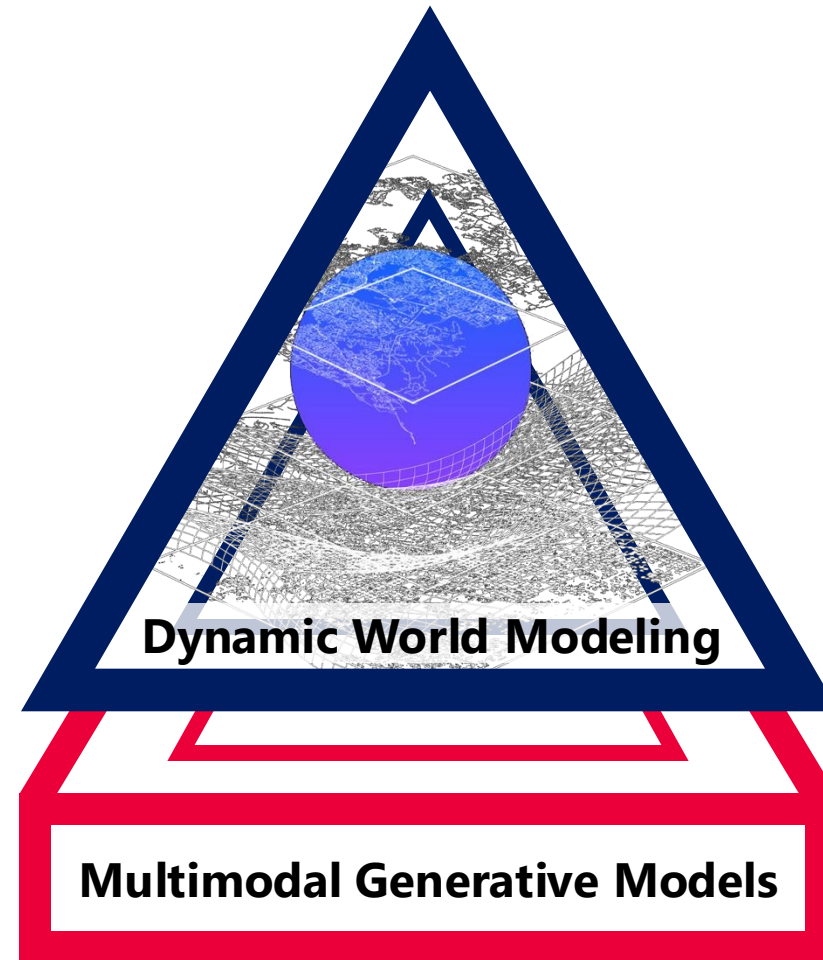
# Arbitrary View Rendering



# 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

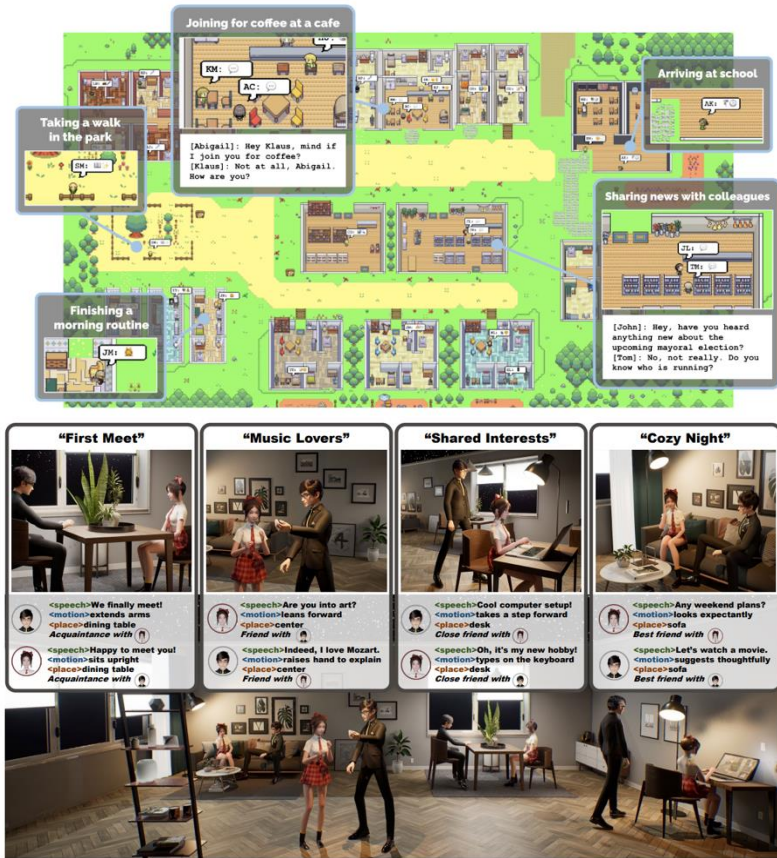
# 3D Characters with Social Intelligence

## Modeling with LLM-Agent Framework

## Limitations

- Scalable Formulation
- Multimodal Coherence
- Latency

Generative Agents [1]  
Life Project [2]  
Digital



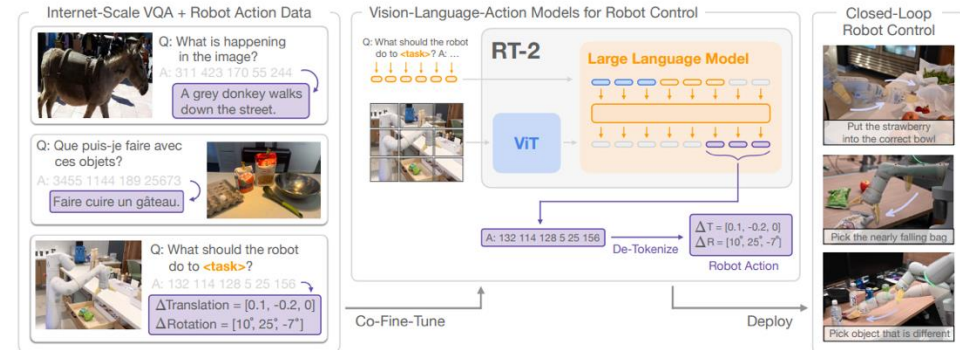
[1] Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023.

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

# Motivation: Avatar as Virtual Robot



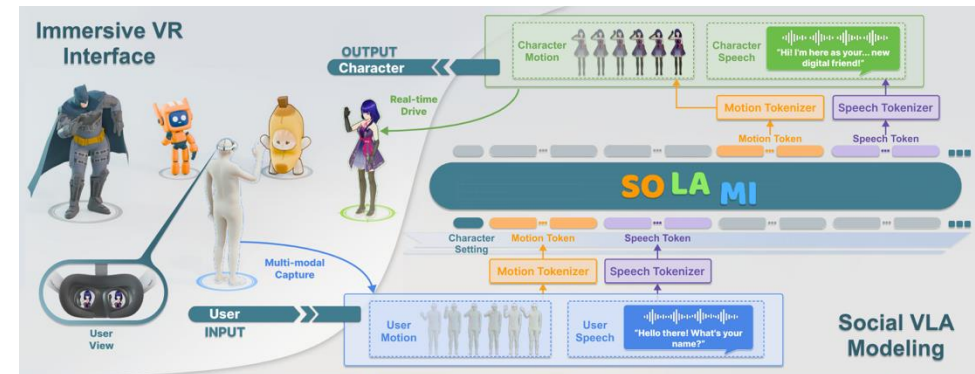
Robot  
3D Agent with Real Embodiment  
(Real-world Task & Interaction)



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



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



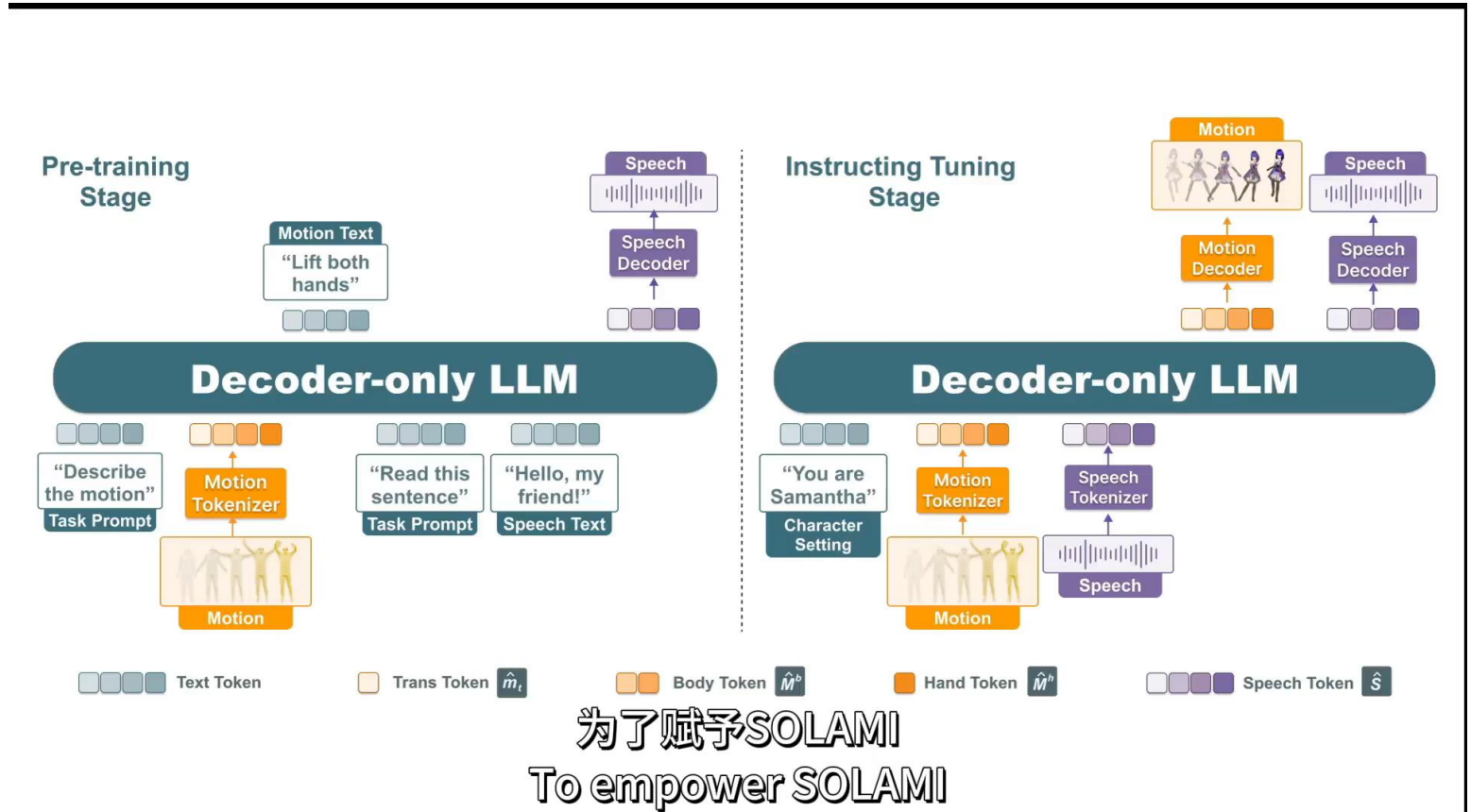
Social VLA for Immersive Interaction with 3D Characters

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

# Training Recipe

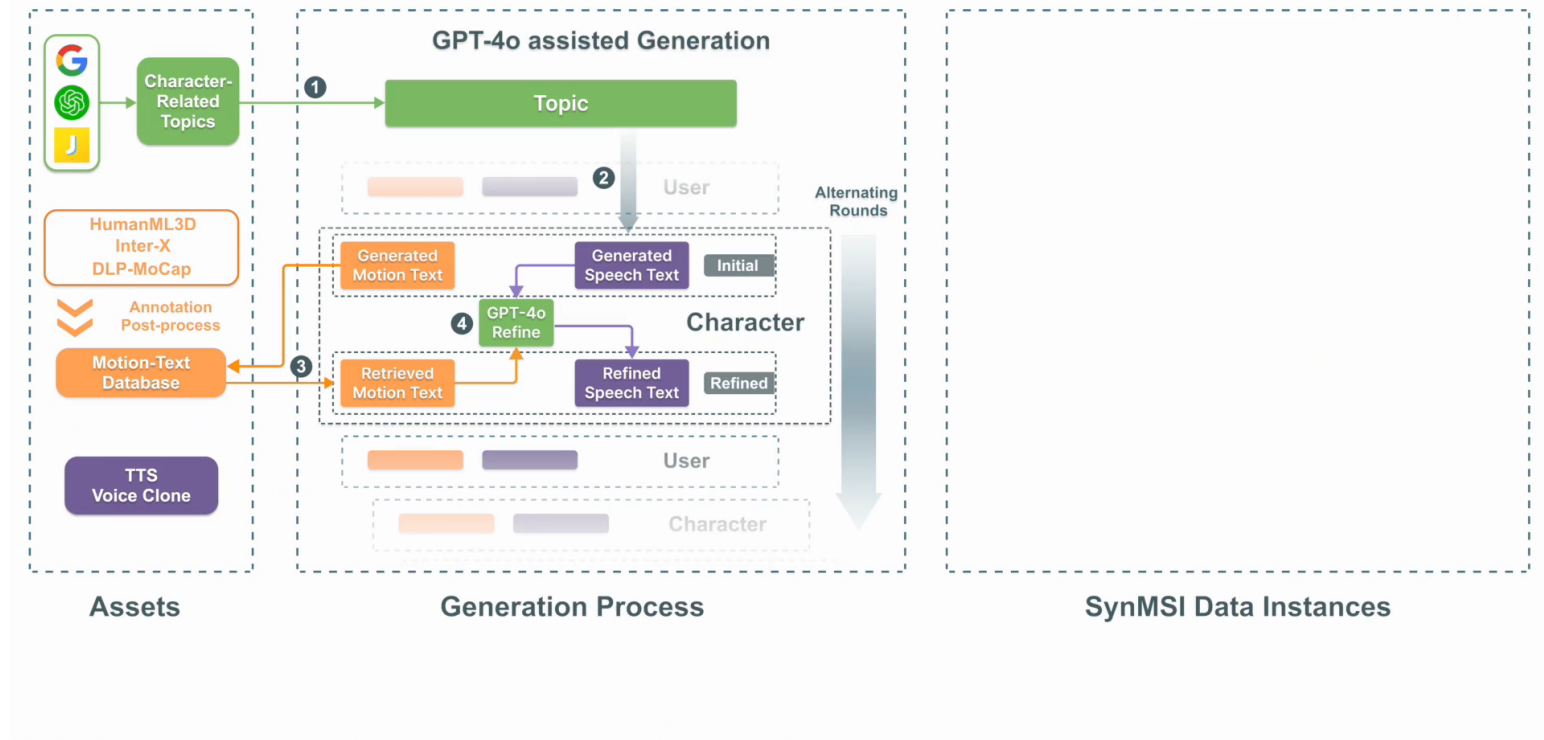
## Training Stages

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



# 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

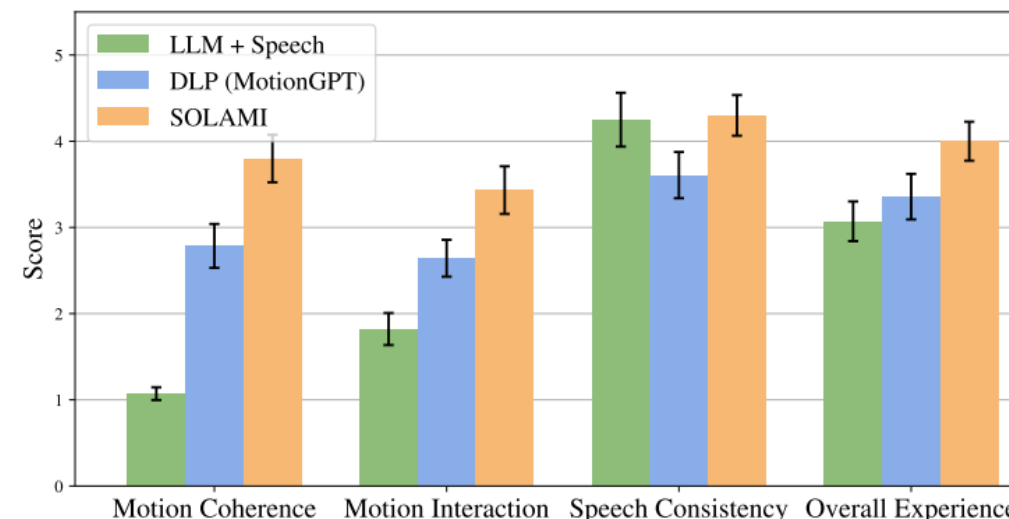


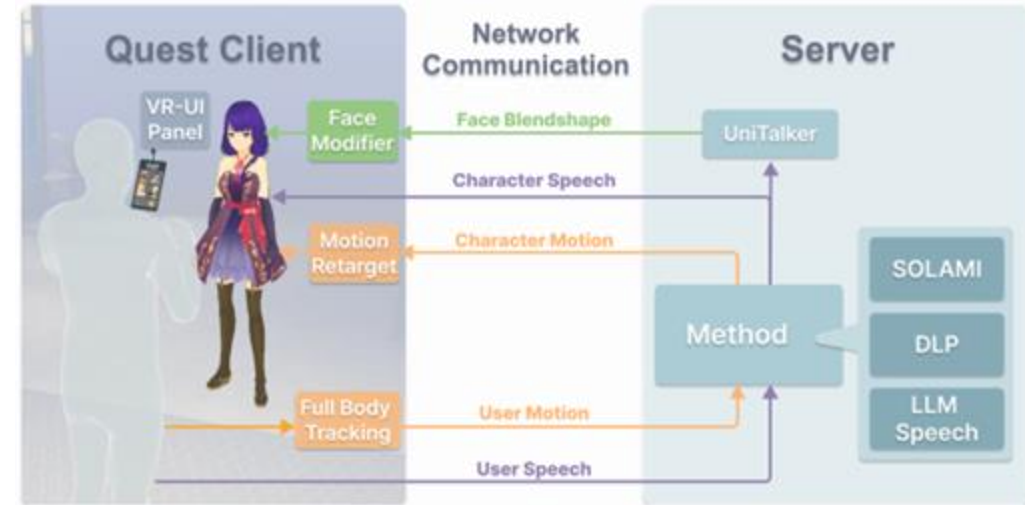
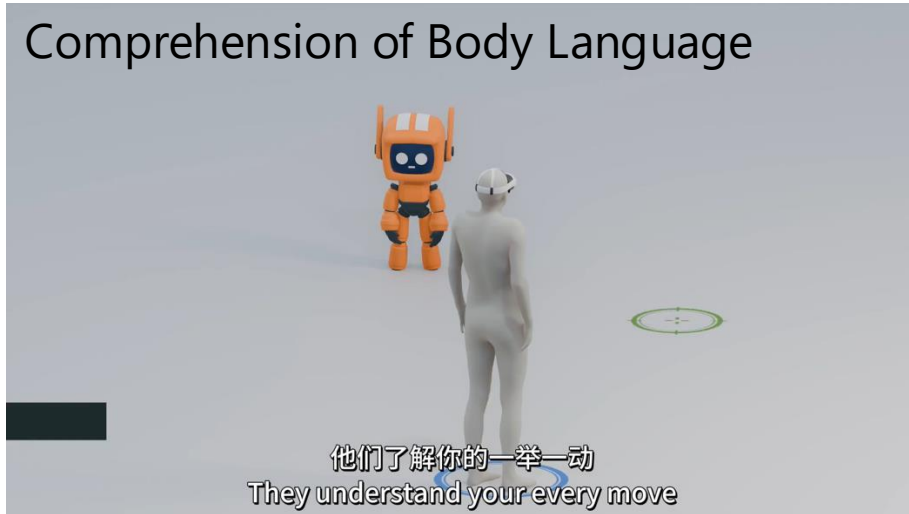
Table 1. **Quantitative results of baselines and SOLAMI.** ‘↑’(‘↓’) indicates that the values are better if the metrics are larger (smaller). We run all the evaluations 5 times and report the average metric. The best results are in bold and the second best results are underlined.

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	<b>3.859</b>	3.157
AnyGPT (fine-tune) [81]	-	-	-	-	0.819	3.502	3.803	<b>2.588</b>
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)	<b>3.443</b>	<b>8.853</b>	<b>151.500</b>	<b>0.360</b>	<b>0.824</b>	<b>3.634</b>	<u>3.824</u>	<u>2.639</u>



# Demo: VR Interface

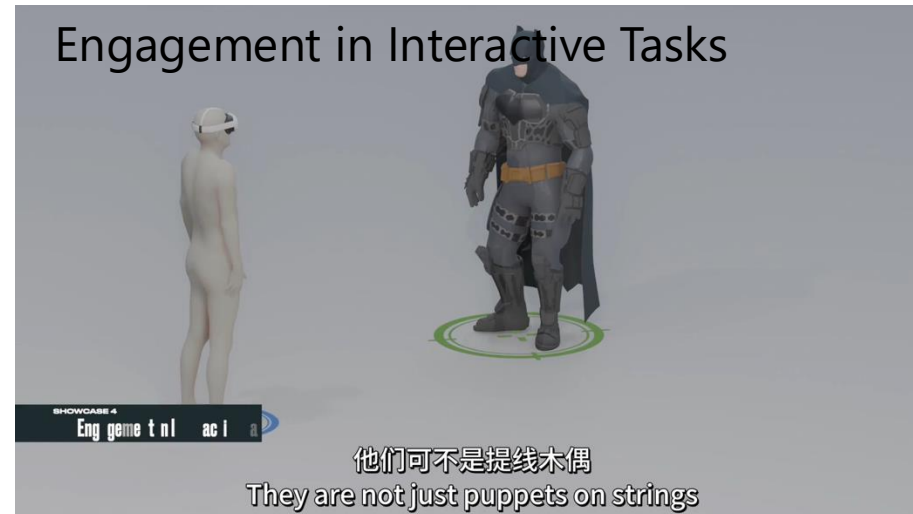
## Comprehension of Body Language



## Execution of Motion Commands



## Engagement in Interactive Tasks





## Be Social: EgoLife



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

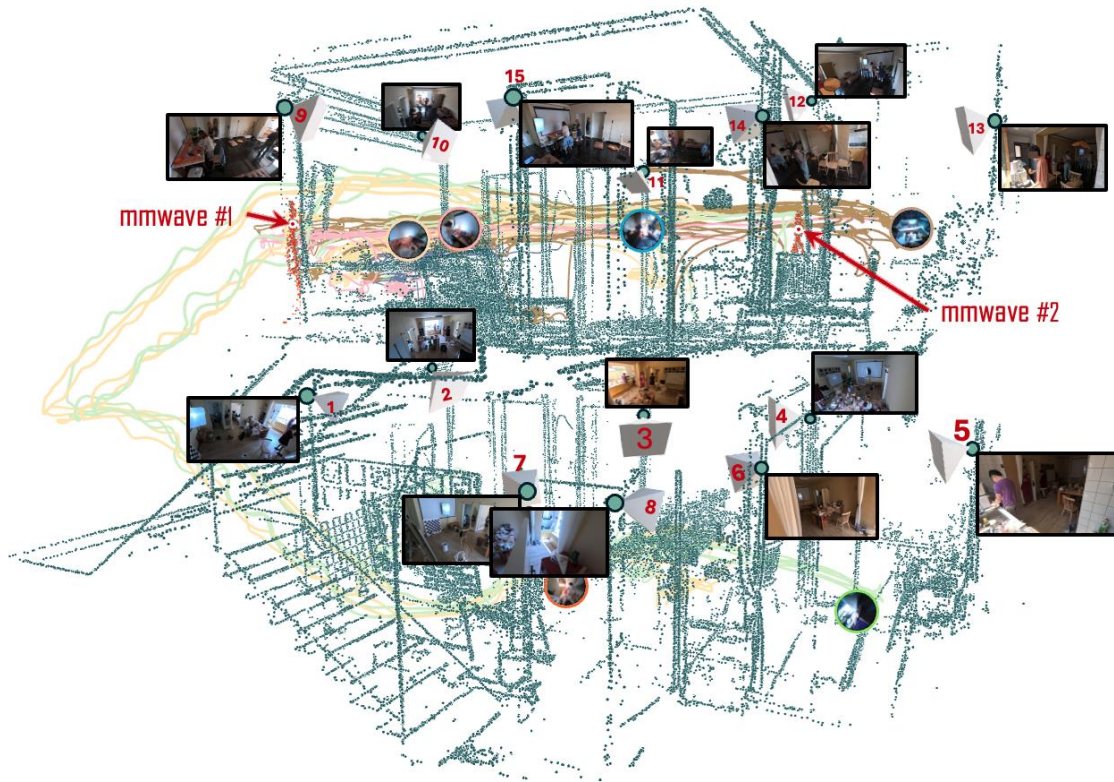
CVPR 2025

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



missing signal

DAY1, L1  
egolife  
11:22:30:00

DAY1, L2  
egolife  
11:22:30:00

15 fixed third-person-view videos

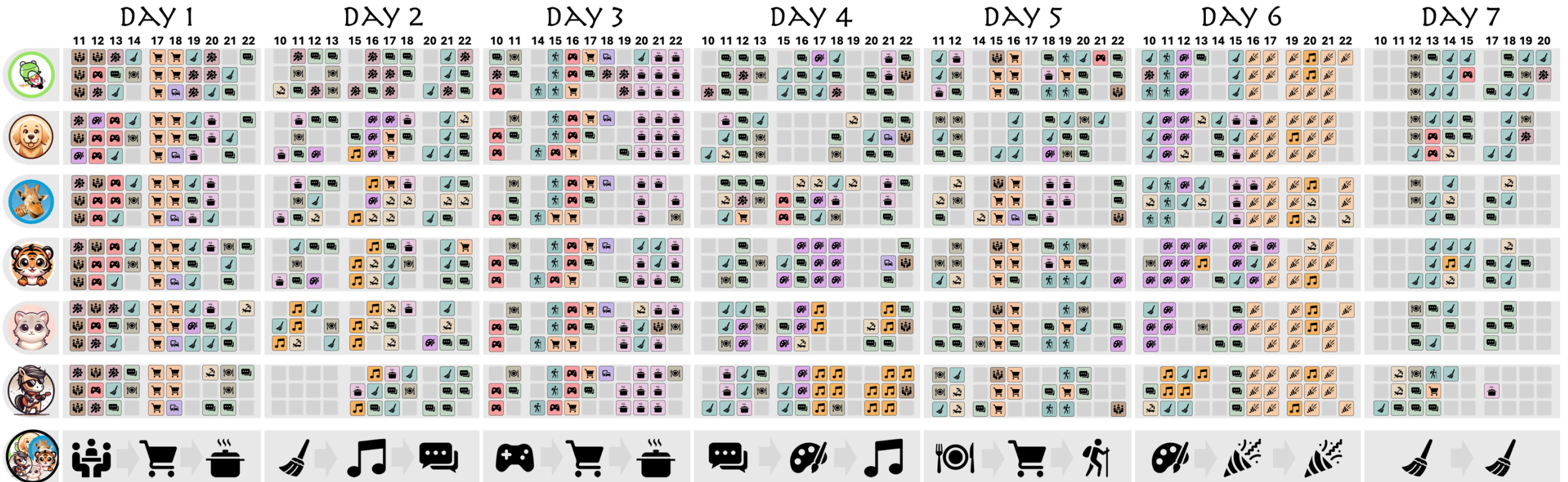
11:22:30:00 DAY1 A1\_JAKE 11:22:30:00 DAY1 A2\_LUICE 11:22:30:00 DAY1 A3\_TASHA 11:22:30:00 DAY1 A4\_LUCIA 11:22:30:00 DAY1 A5\_KATRINA 11:22:30:00 DAY1 A6\_SHURE

6 egocentric videos

ALL THE EGO AND EXO VIDEOS ARE SYNCHRONIZED WITH THE REAL TIME AND DATE.

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

# The EgoLife Timeline



🗣️ Social ×183
🧹 Housekeeping ×145
🍳 Cooking ×86
🛒 Shopping ×74
🍽️ Dining ×67
🎉 Party ×64
🎨 Arts & Craftwork ×57
  
🎮 Leisure ×49
🎮 Games ×46
🎵 Music & Dance ×45
🚶 Outing ×40
🔧 Setup ×35
👥 Meeting ×31
🚗 Commuting ×15

# The EgoLifeQA Benchmark


6 x 500 = 3000 QAs

## EventRecall 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:  @ Day 1 11:46:59.050

## EntityLog 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**



**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.

## HabitInsight Personal Habit Patterns

## RelationMap Interpersonal Interaction Patterns

# The EgoLifeQA Benchmark

**LMMS-Lab**

**MMLab @NTU**

**1-Jake**

**REC**

**Day4 9:12pm**  
**Jake cracks eggs into a bowl**

**EntityLog**  
实体日志  
PAST OBJECTS OF INTEREST

**EventRecall**  
事件回忆  
PAST EVENTS OF INTEREST

**HabitInsight**  
习惯洞察  
PERSONAL HABIT PATTERNS

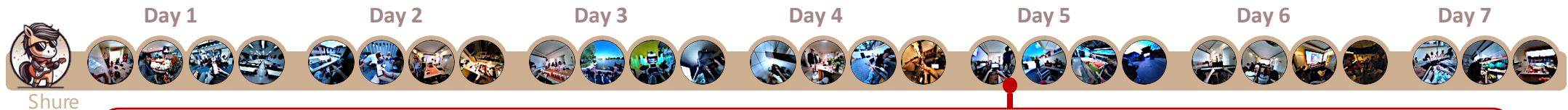
**RelationMap**  
关系网络  
INTERPERSONAL INTERACTION PATTERNS

**TaskMaster**  
任务管理  
TASKS ASSIGNMENT AND REVIEW

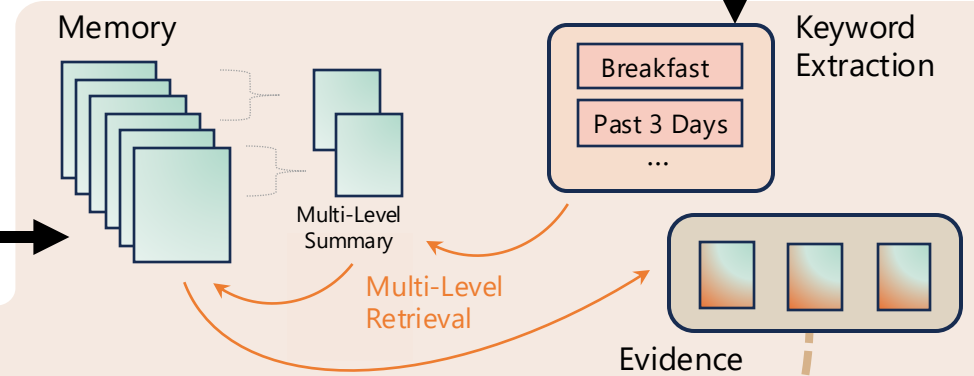
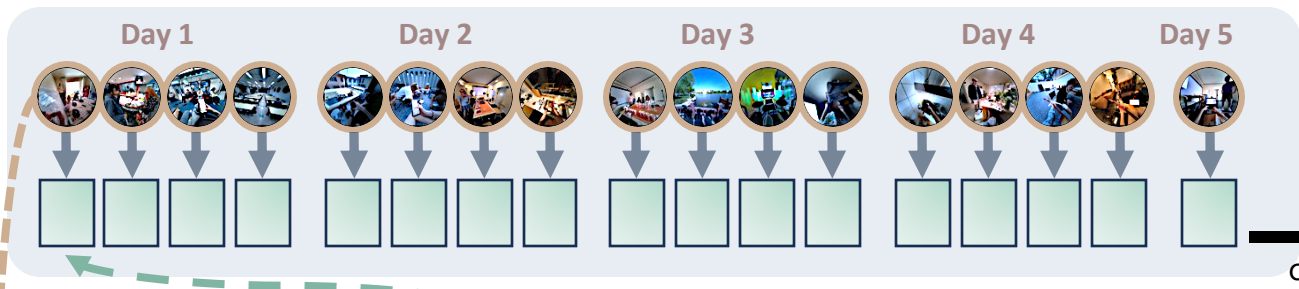
**egollife**

EntityLog, EventRecall, HabitInsight, RelationMap, TaskMaster, egollife

# EgoButler



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



**Caption**

Timestamp: Day 1, 11:12:23 - 11:12:31

I was operating my phone, then looked up to the left at Jake and adjusted my glasses. Jake said, "Here is a stopwatch." I responded, "Ok, stamp the time." We were sitting at a long table with some boxes and stationery on it. Pictures were hanging on the wall, and there was also a projector and a whiteboard in the room.

Egocentric omni-modal model as captioner  
**EgoGPT**

video audio

Timestamp: Day 1, 11:12:23 - 11:12:31

Q: What did I do in the last clip, including both visuals and sounds?

(a) Captioning Stage with EgoGPT

**Answer**

In the past 3 days, breakfast was either skipped or combined with lunch: 3 days ago, we made pan-fried pancakes, and next we had pizza. Yesterday's meal featured Chinese toon pancakes and dumplings for the Grain Rain festival. For a new experience, try dim sum, tamagoyaki, or shakshuka.

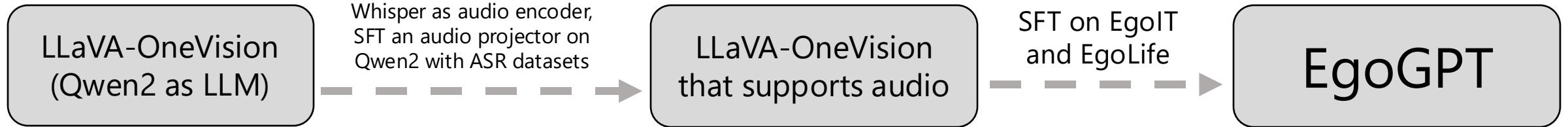
Egocentric omni-modal model as evidence verifier  
**EgoGPT**

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

(b) Question Answering Stage with EgoRAG



# EgoButler – The EgoGPT Component

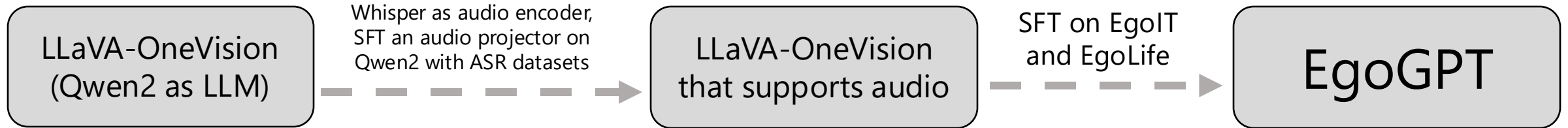


Overview of Classic Ego-centric Dataset

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
<b>Total</b>	<b>43.16h</b>	<b>1529 (686)</b>	<b>99.48K</b>

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:

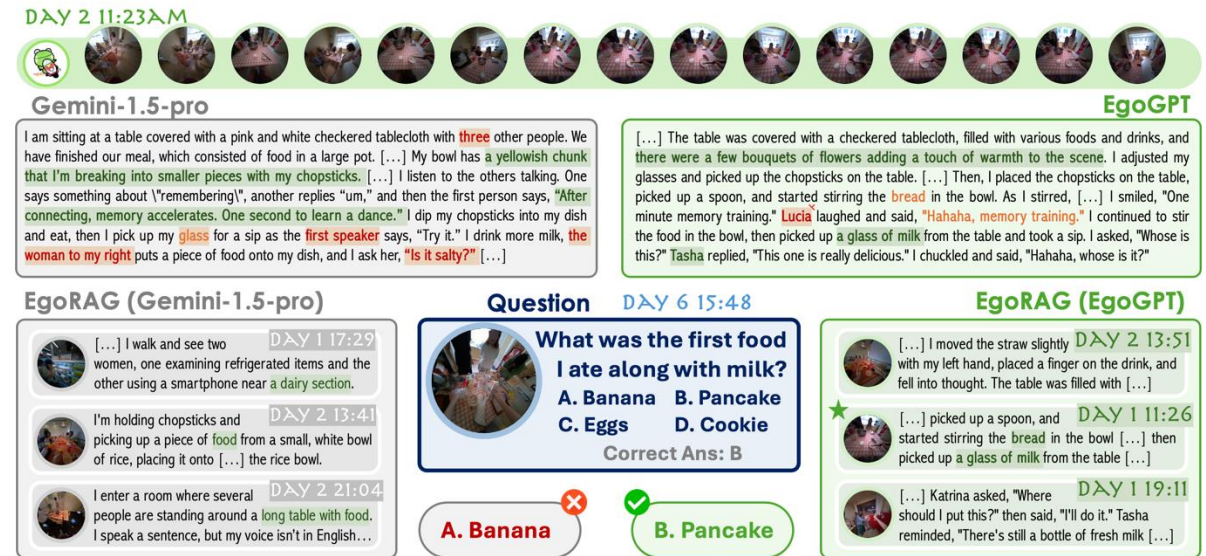
- In-depth egocentric video familiarity
- Omni-modal comprehension — effectively integrating both visual and audio signals

## 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



The screenshot displays a sequence of video thumbnails at the top, labeled 'DAY 2 11:23AM'. Below, it compares the outputs of Gemini-1.5-pro and EgoGPT. Gemini-1.5-pro provides a detailed, somewhat verbose text description of a dining scene. EgoGPT provides a more concise summary. Below this, a 'Question' box asks 'What was the first food I ate along with milk?' with options A. Banana, B. Pancake, C. Eggs, and D. Cookie. The correct answer is B. Pancake. To the right, EgoRAG (EgoGPT) shows a sequence of video thumbnails and text-based outputs for a different question, demonstrating its ability to handle long-horizon queries.

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

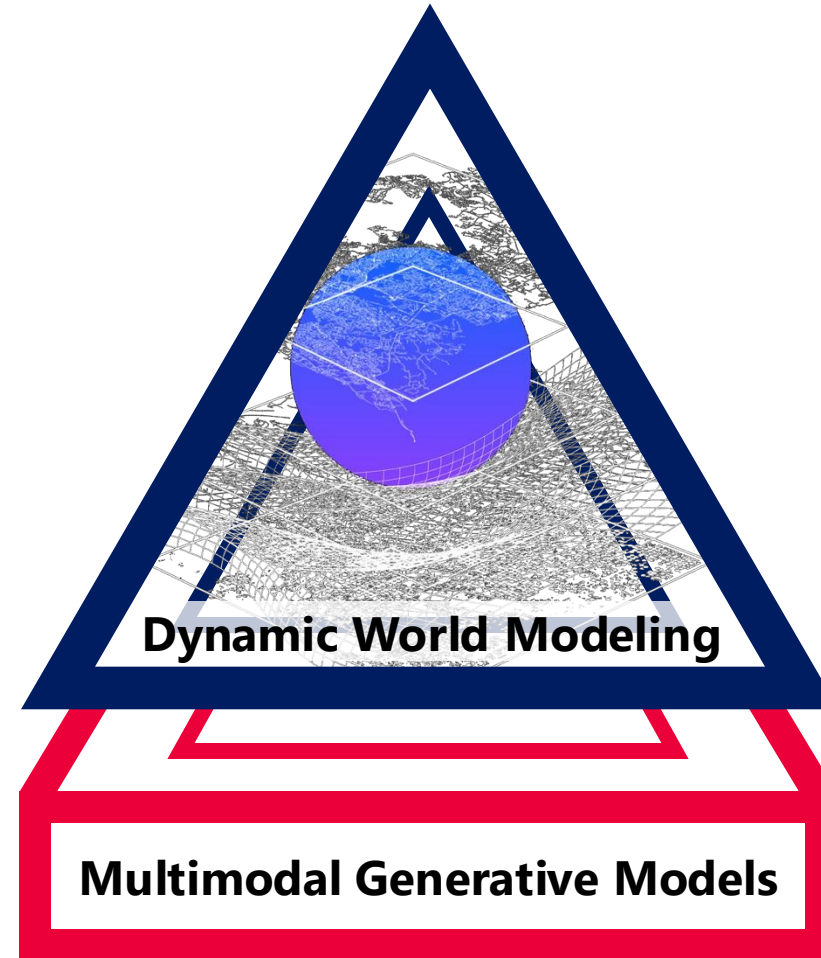
[egolife-ai.github.io](http://egolife-ai.github.io)

# Be Physical

How to Model Material and Illumination

## Be Dynamic

How to Model  
Dynamic Scenes



Dynamic World Modeling

Multimodal Generative Models

## Be Social

How to Model Social  
Interactions

# Thank You

Ziwei Liu 刘子纬

Nanyang Technological University

