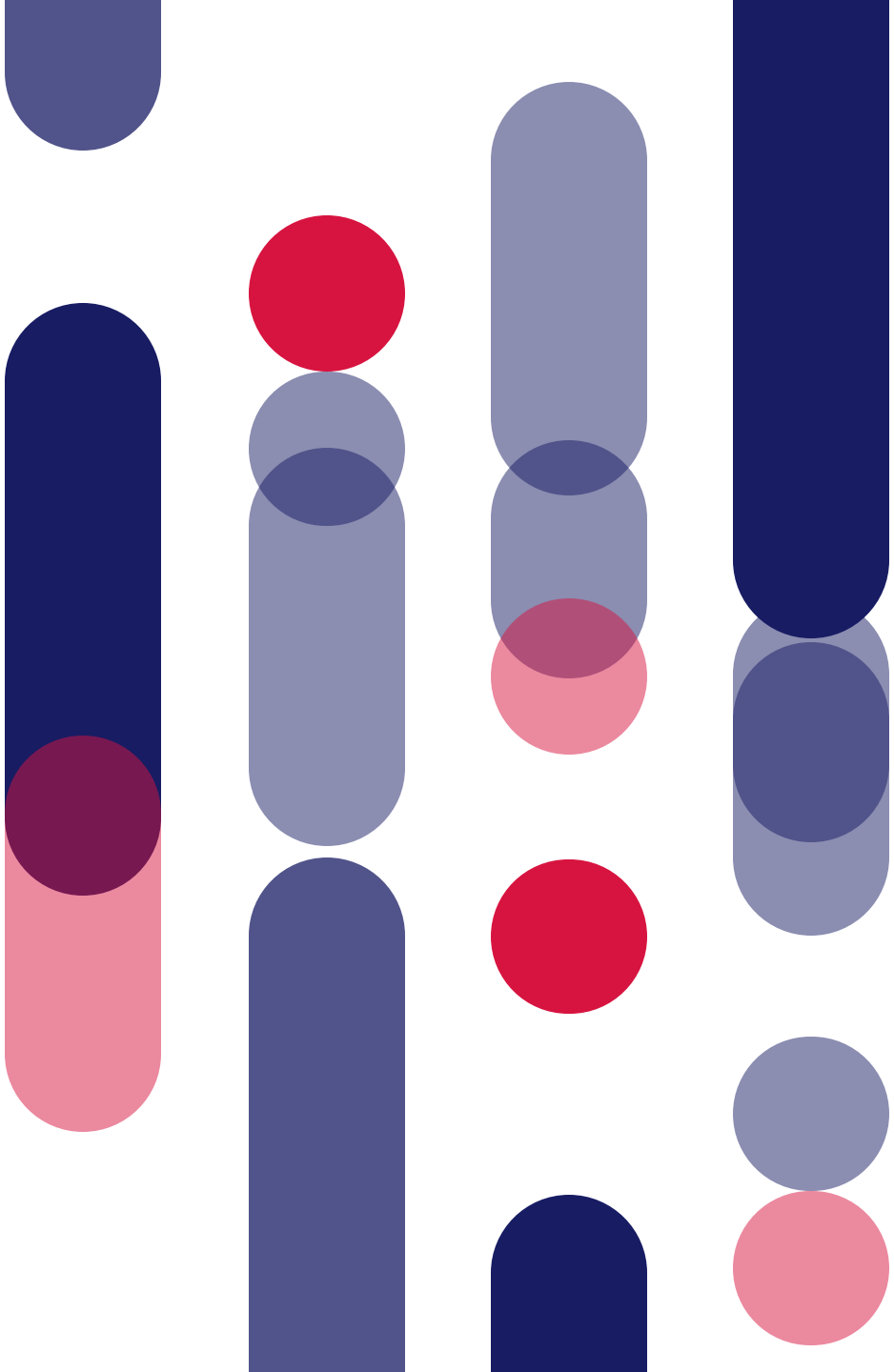
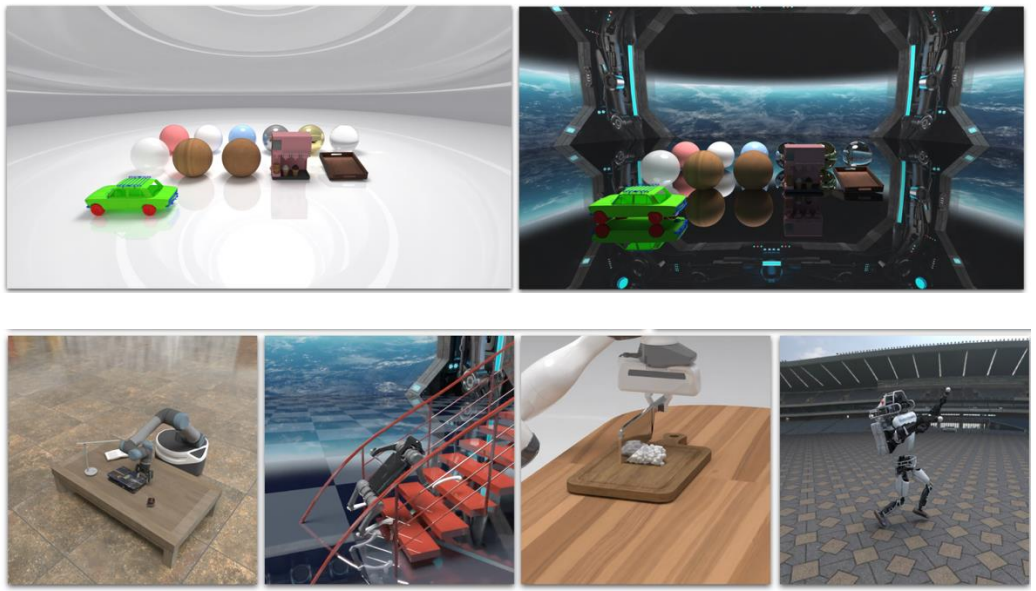
From High-fidelity 3D Generative Models to Dynamic Embodied Learning

Ziwei Liu

S-Lab, Nanyang Technological University

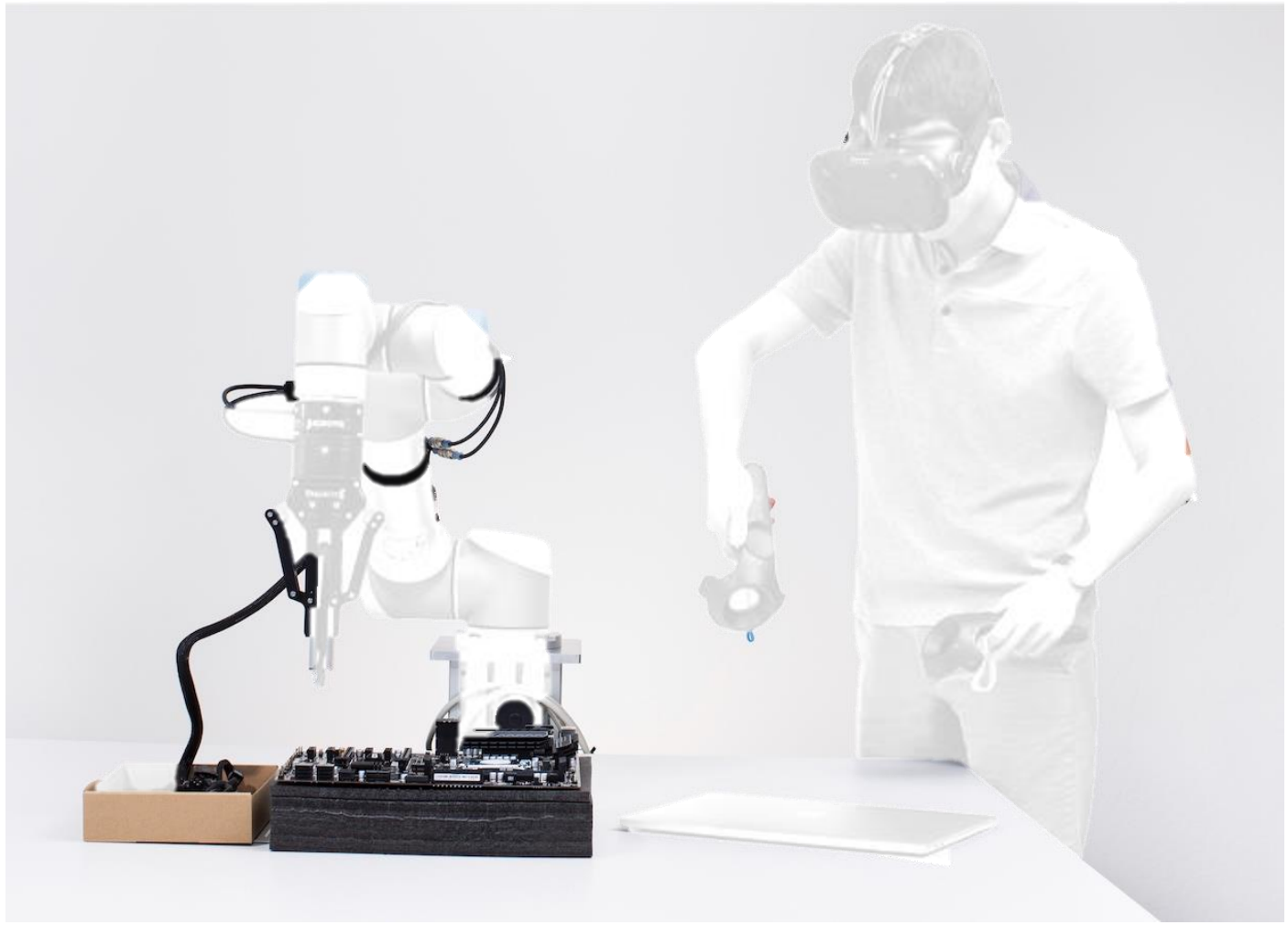


Generative Simulator for Embodied Learning



Generative Simulator

Towards the Generative Simulator



Scene

Towards the Generative Simulator

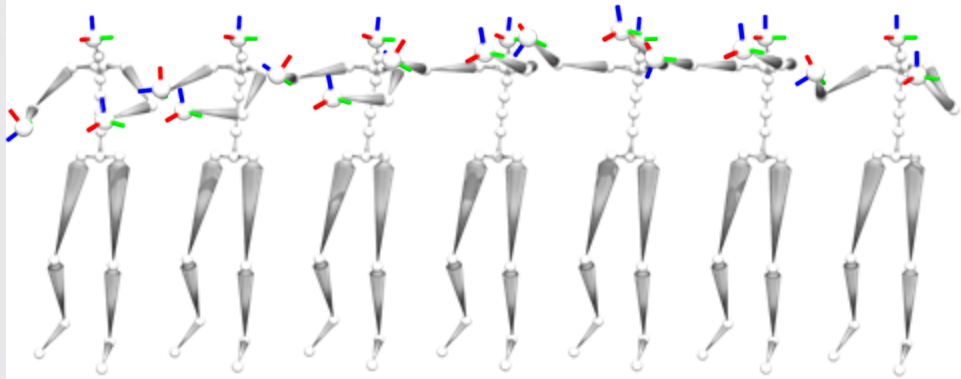
Object

Towards the Generative Simulator

Human

Towards the Generative Simulator

Ego Motion



Overview

Scene

SceneDreamer
[Chen et al. 2023]
CityDreamer
[Xie et al. 2024]

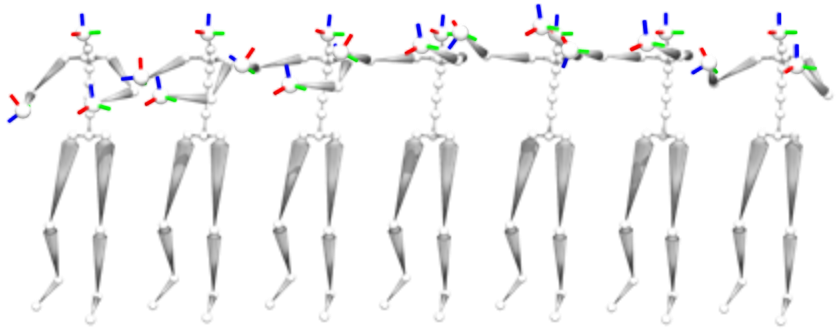
Object

3DTopia-XL
[Chen et al. 2024]

Human

StructLDM
[Hu et al. 2024]

Ego Motion



EgoLM
[Hong et al. 2024]

Overview

Scene

SceneDreamer
[Chen et al. 2023]
CityDreamer
[Xie et al. 2024]

Object

3DTopia-XL
[Chen et al. WIP]

Human

StructLDM
[Hu et al. 2024]

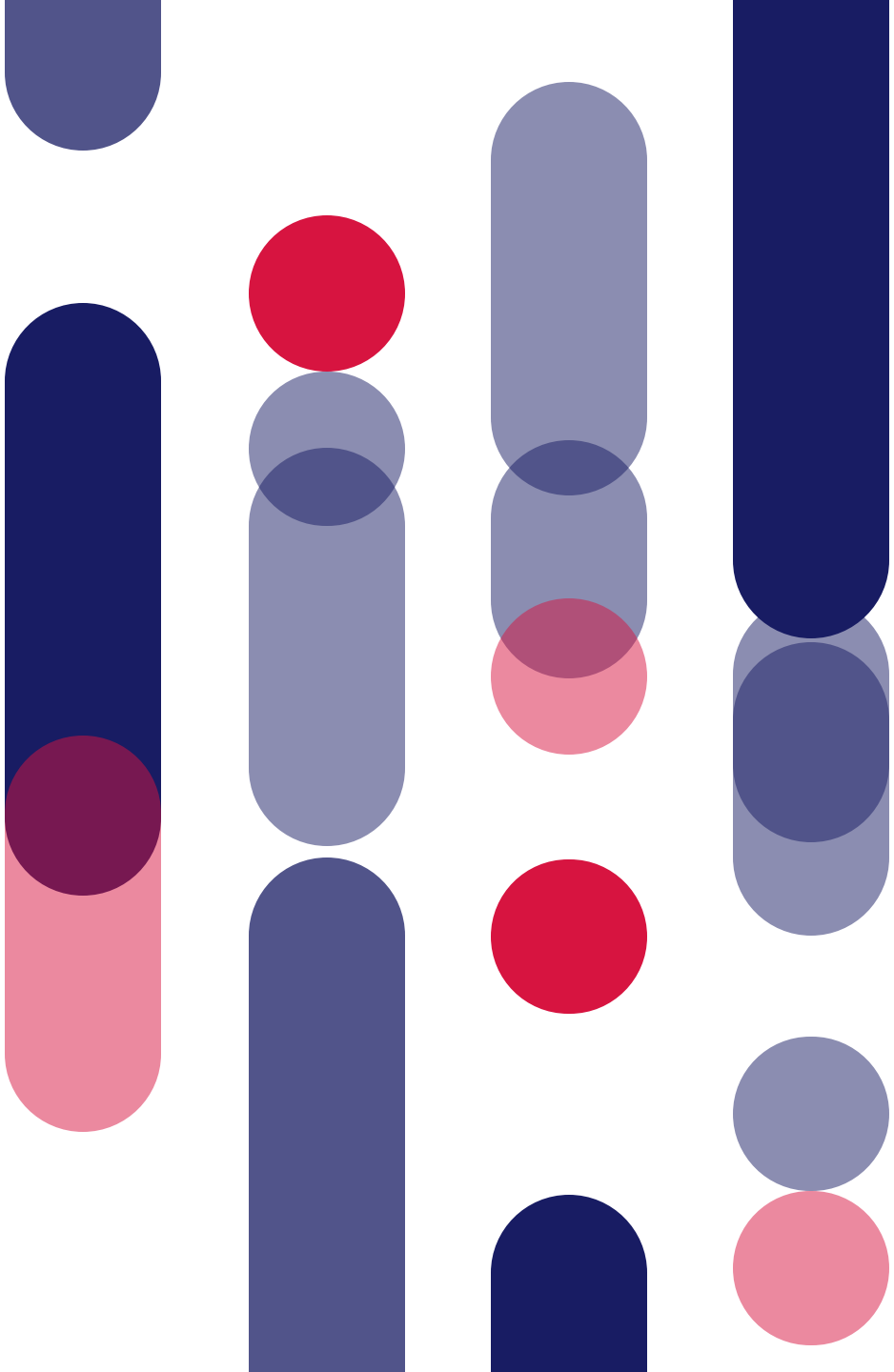
Ego Motion

EgoLM
[Hong et al. 2024]

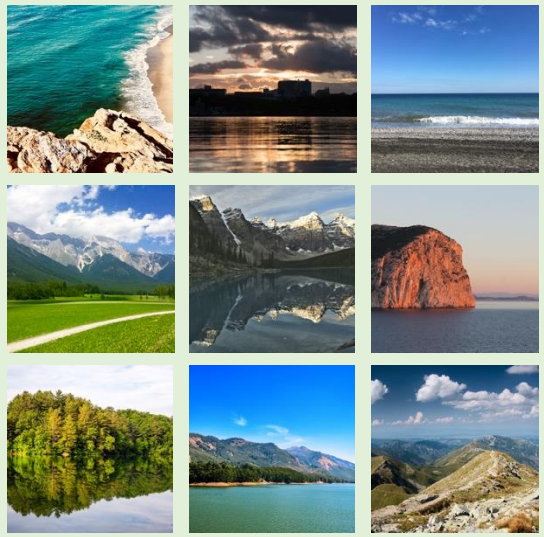
Generate Unbounded 3D Scenes from 2D Images

Zhaoxi Chen, Guangcong Wang, Ziwei Liu

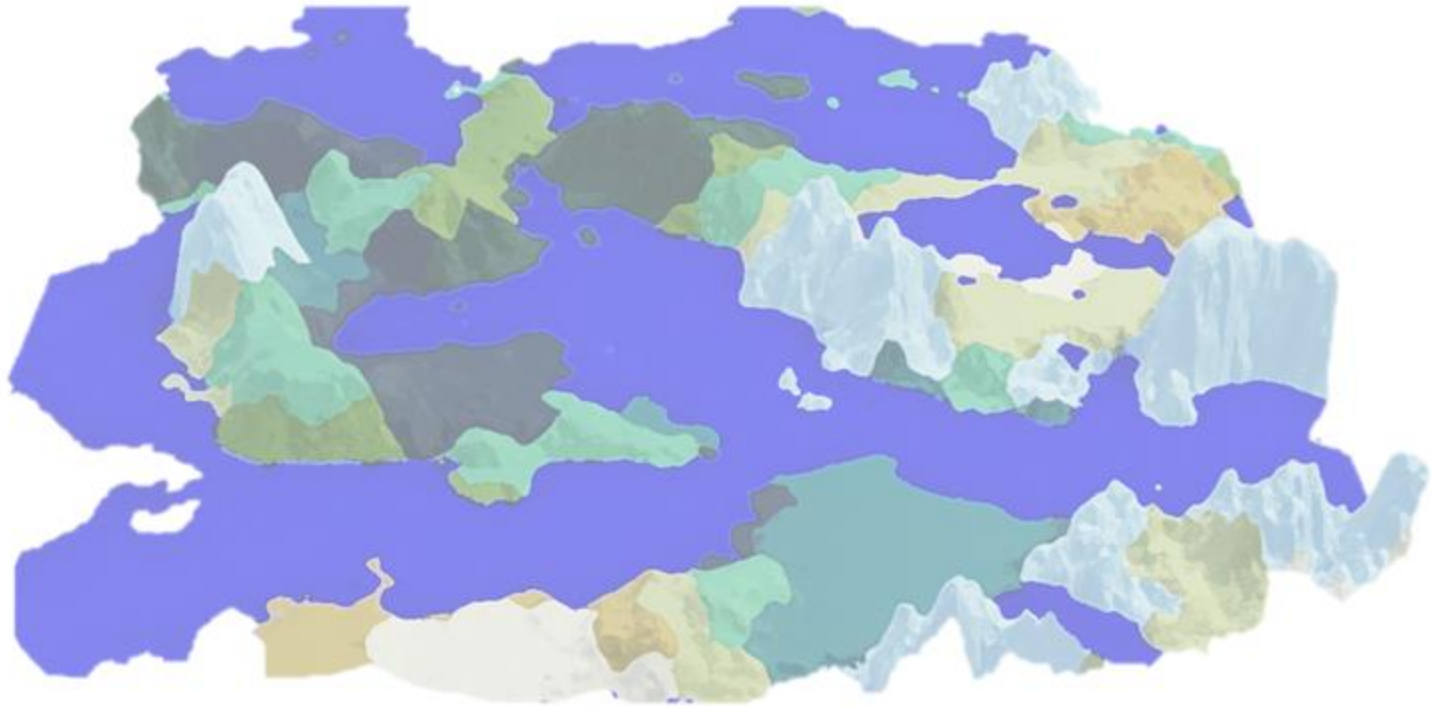
S-Lab, Nanyang Technological University



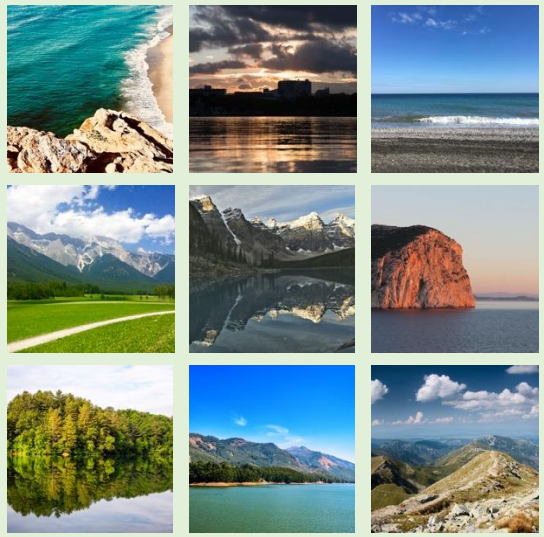
Motivation



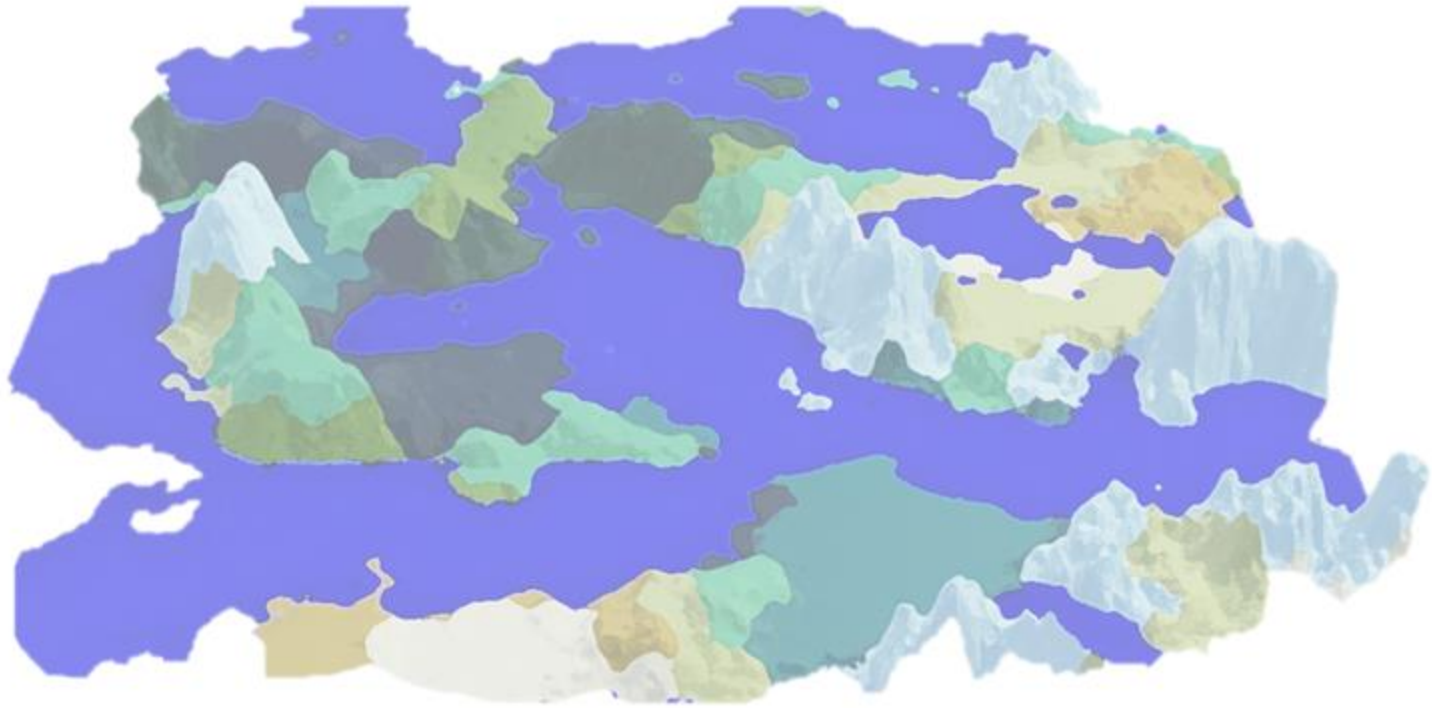
2D Image Collections



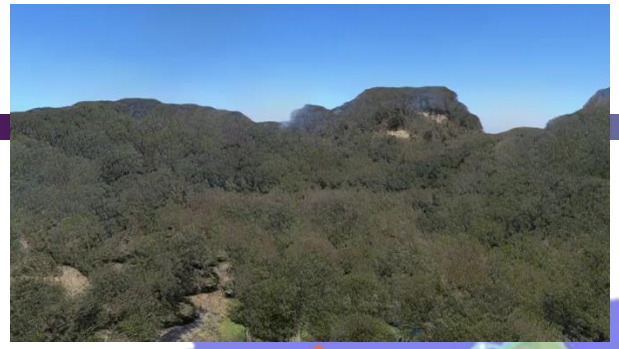
Unbounded 3D Scenes

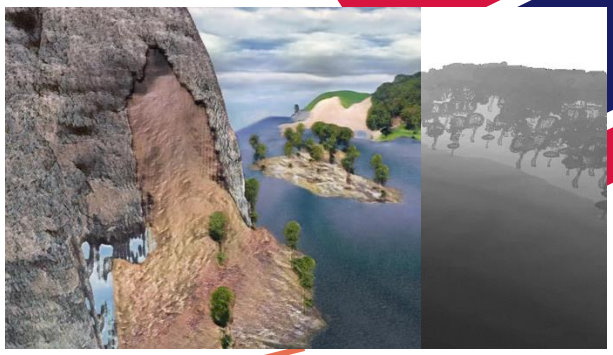


In-the-wild
Image Collections



Unbounded 3D Scenes

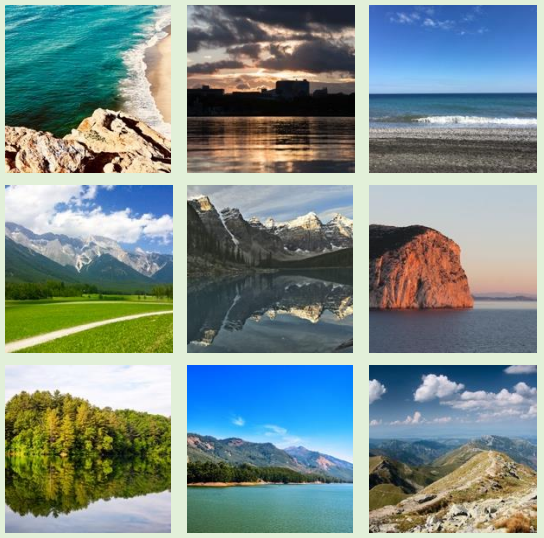





Multi-view consistent

Well-defined geometry

Diverse scenes and styles

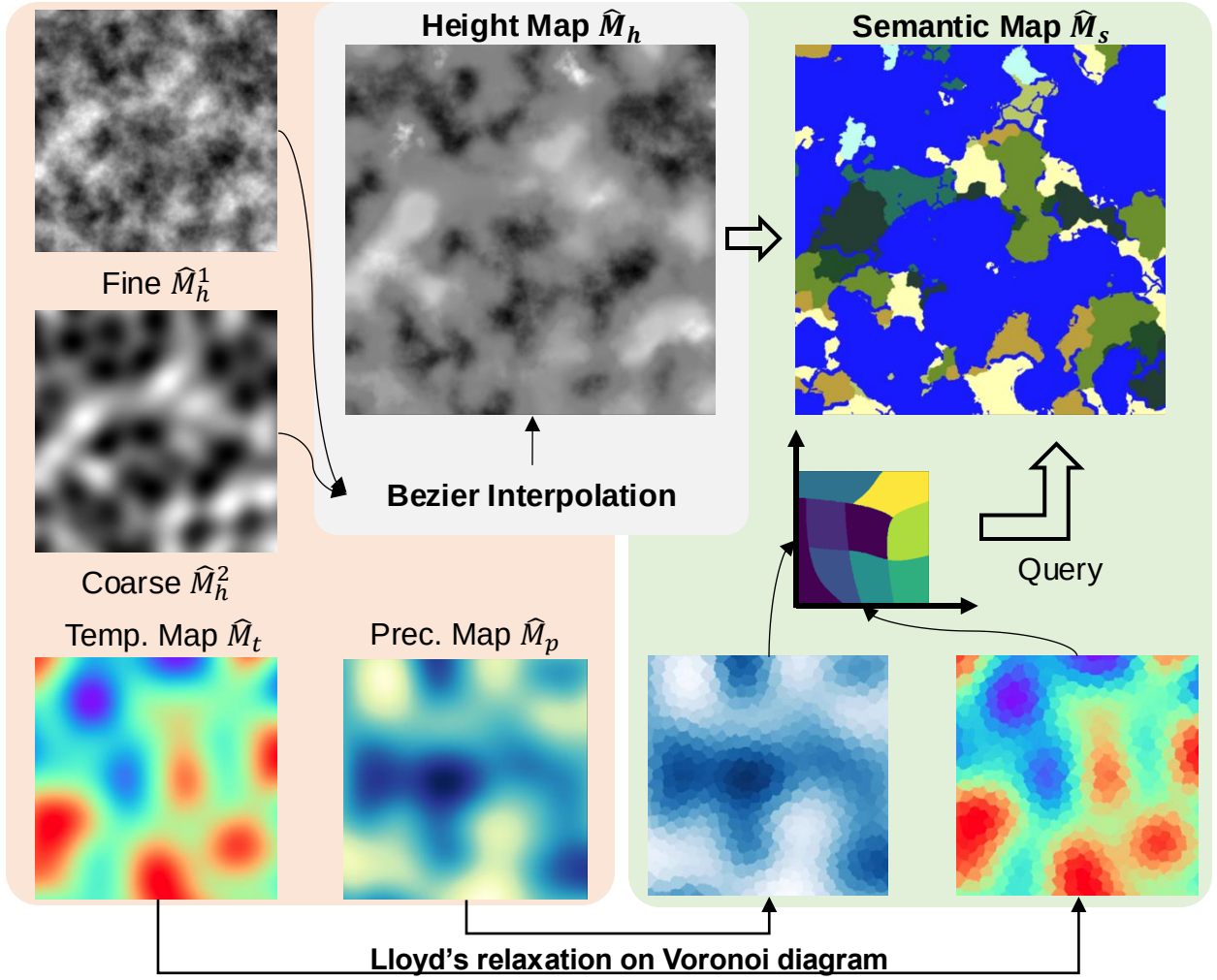


In-the-wild
Image Collections

Unbounded 3D Scenes

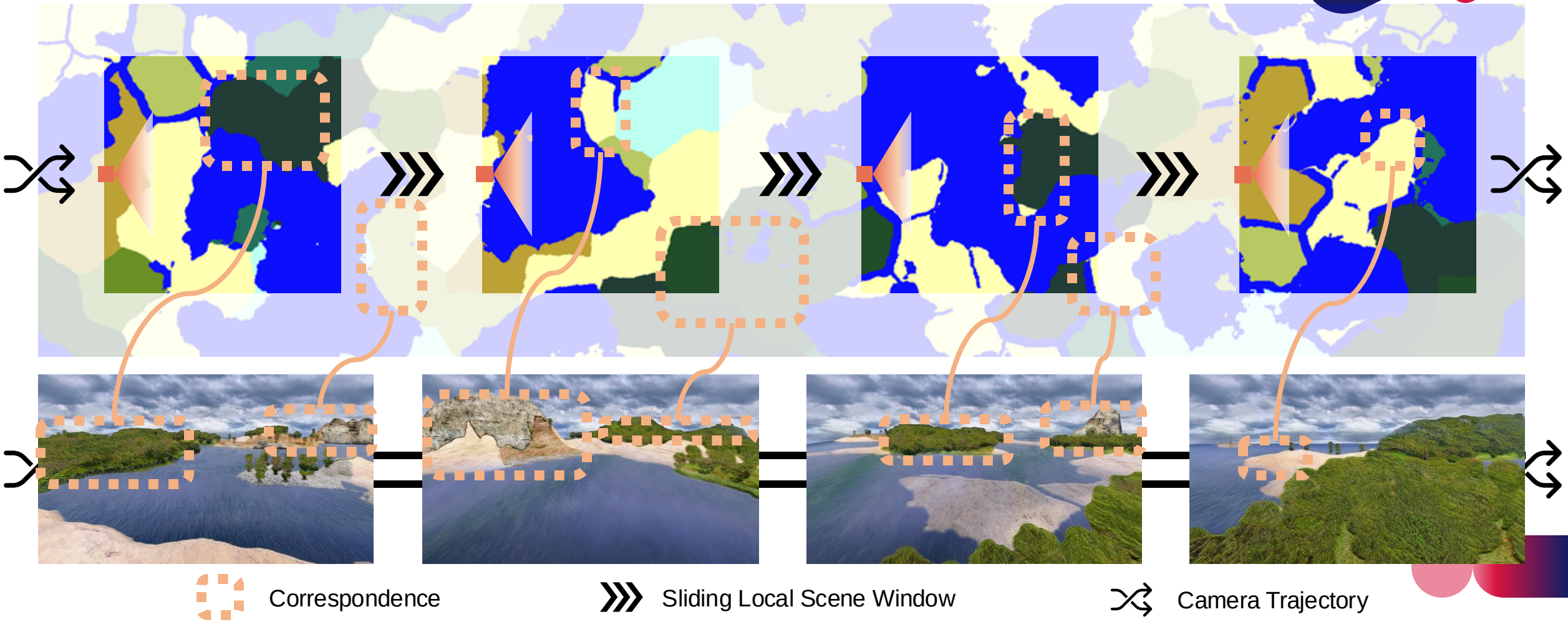
Overview

Procedural Generation of BEV

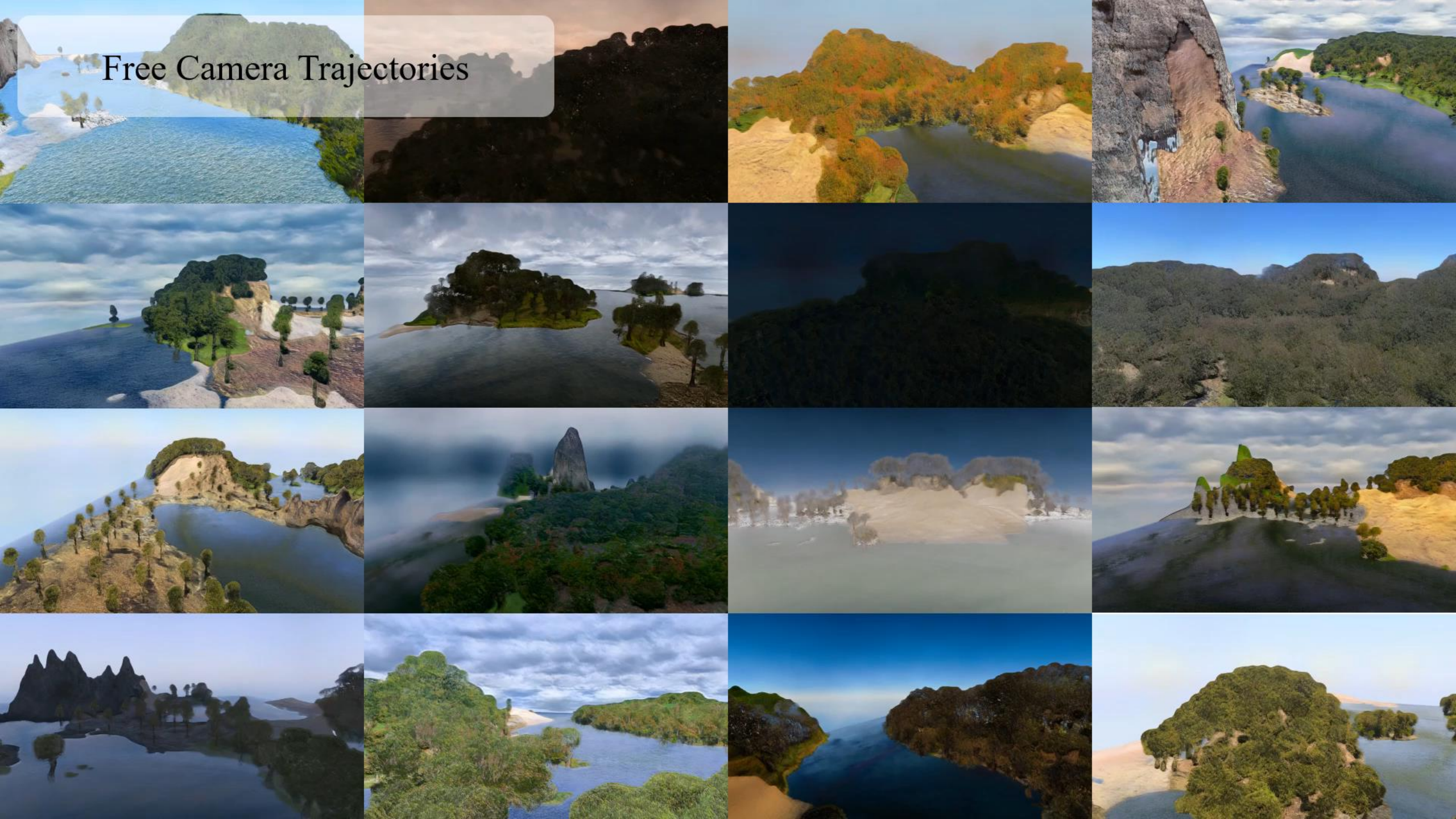



Infinite 3D World!

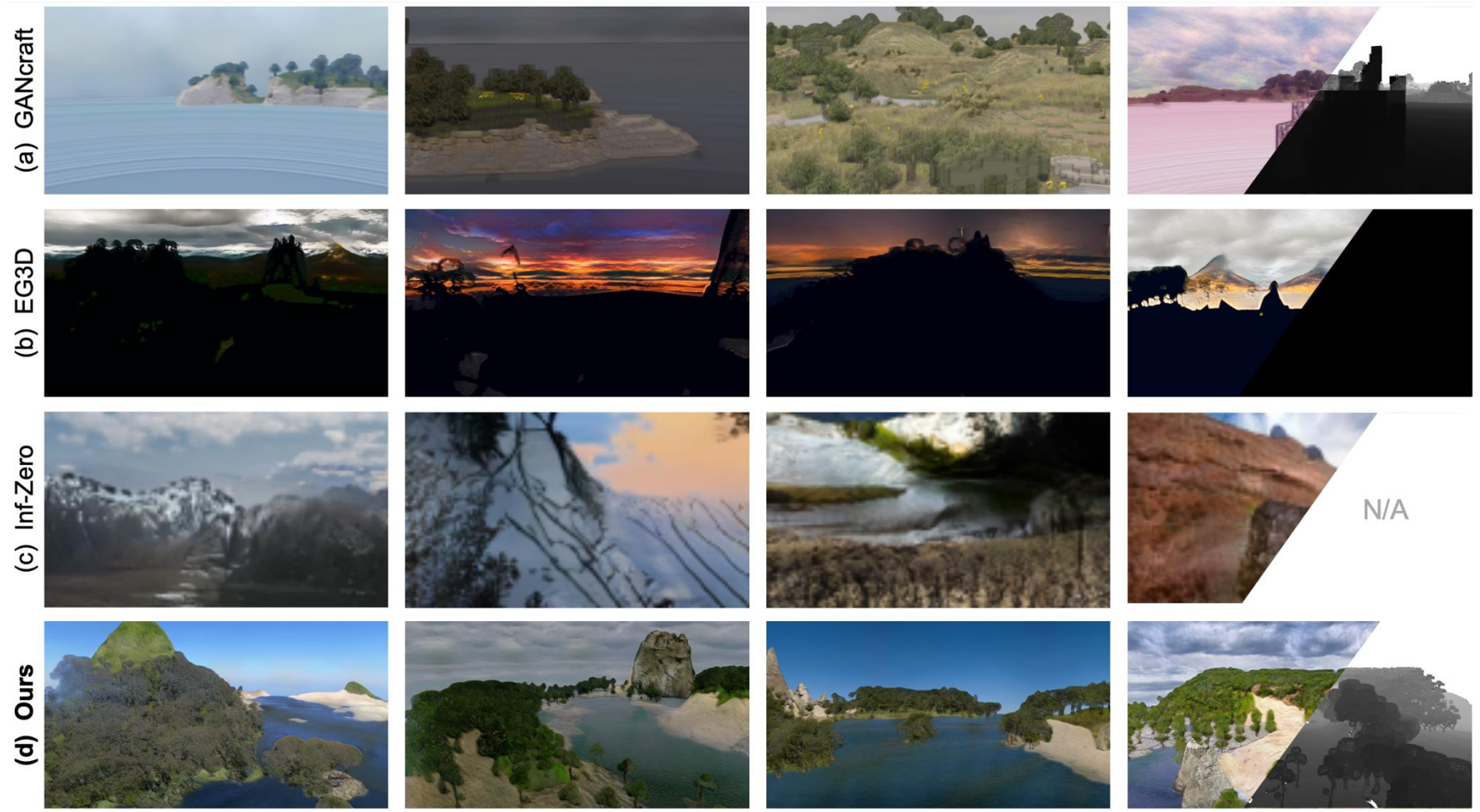
Inference: Sliding Window

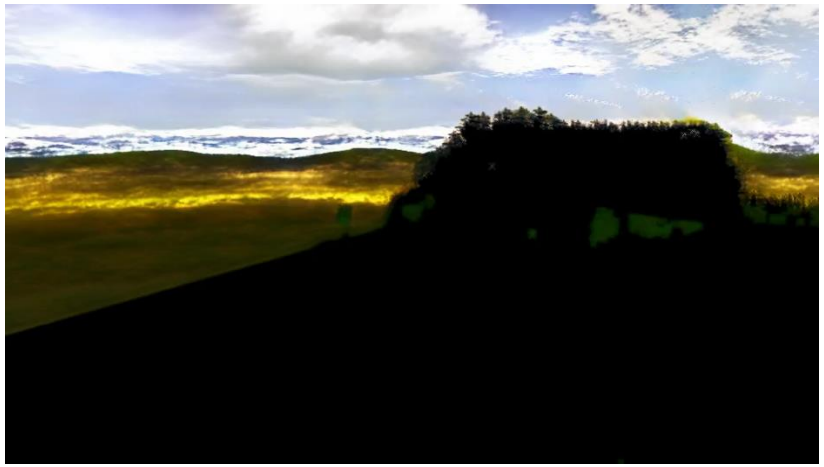



Generate with Different Styles


Free Camera Trajectories

Experiments





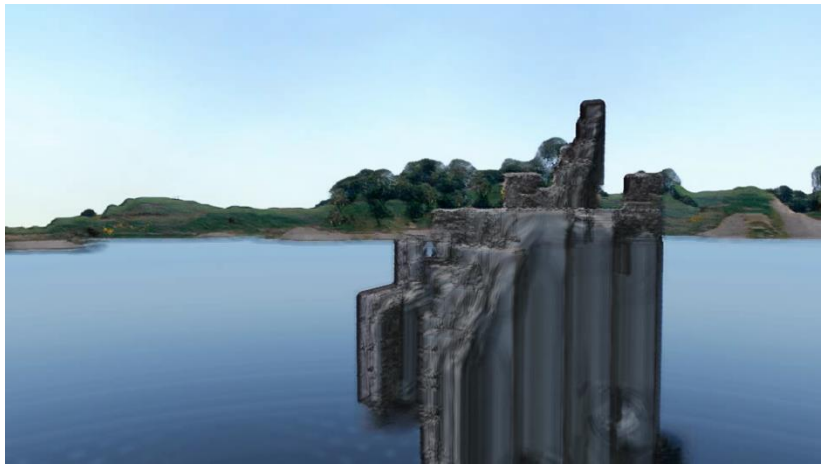
EG3D[1]

GANcraft[2]

Ours

[1] Eric R. Chan, et. al. Efficient Geometry-aware 3D Generative Adversarial Networks. In CVPR 2022.
[2] Zekun Hao, et. al. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In ICCV 2021.

EG3D[1]

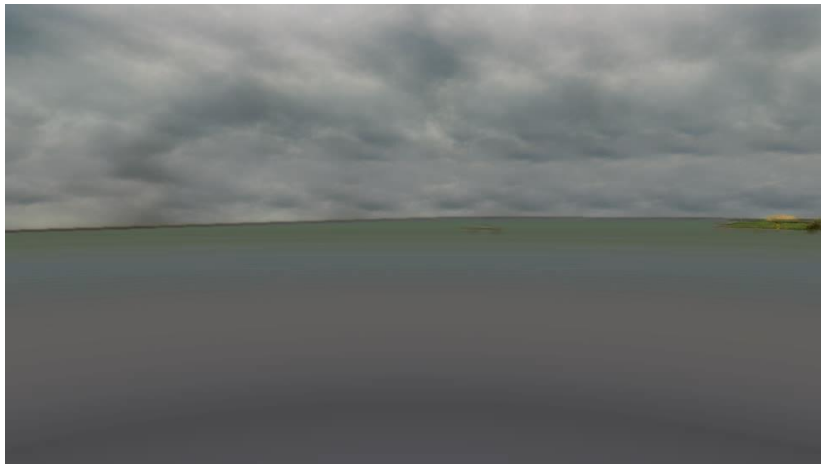


GANcraft[2]

Ours

[1] Eric R. Chan, et. al. Efficient Geometry-aware 3D Generative Adversarial Networks. In CVPR 2022.
[2] Zekun Hao, et. al. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In ICCV 2021.

EG3D[1]



GANcraft[2]

Ours

[1] Eric R. Chan, et. al. Efficient Geometry-aware 3D Generative Adversarial Networks. In CVPR 2022.
[2] Zekun Hao, et. al. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In ICCV 2021.

InfiniteNature-Zero[1]

Ours

[1] Zhengqi Li, et. al. InfiniteNature-Zero: Learning Perpetual View Generation of Natural Scenes from Single Images. In ECCV 2022.

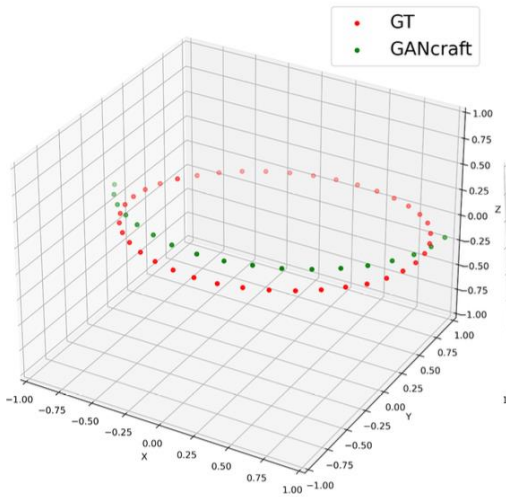
InfiniteNature-Zero[1]

Ours

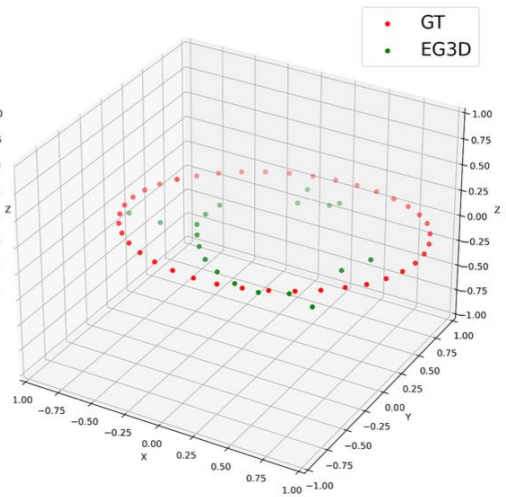
[1] Zhengqi Li, et. al. InfiniteNature-Zero: Learning Perpetual View Generation of Natural Scenes from Single Images. In ECCV 2022.

3D Consistency - SfM

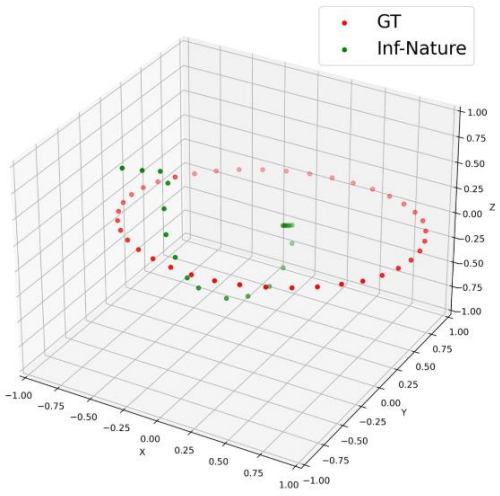
- Reconstruct camera pose using SfM from rendered trajectories



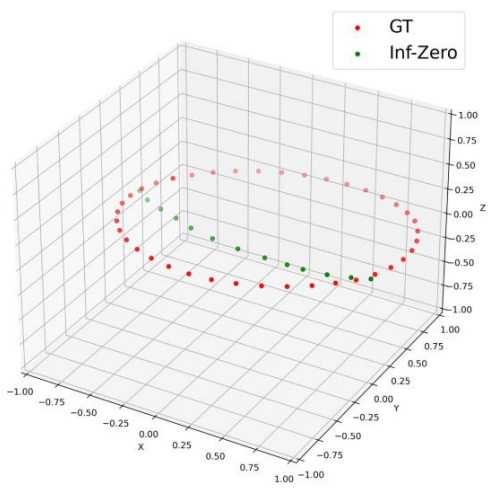
(a) GANcraft



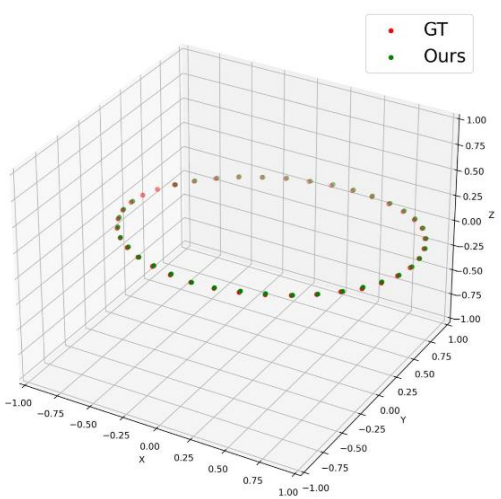
(b) EG3D



(c) Inf-Nature

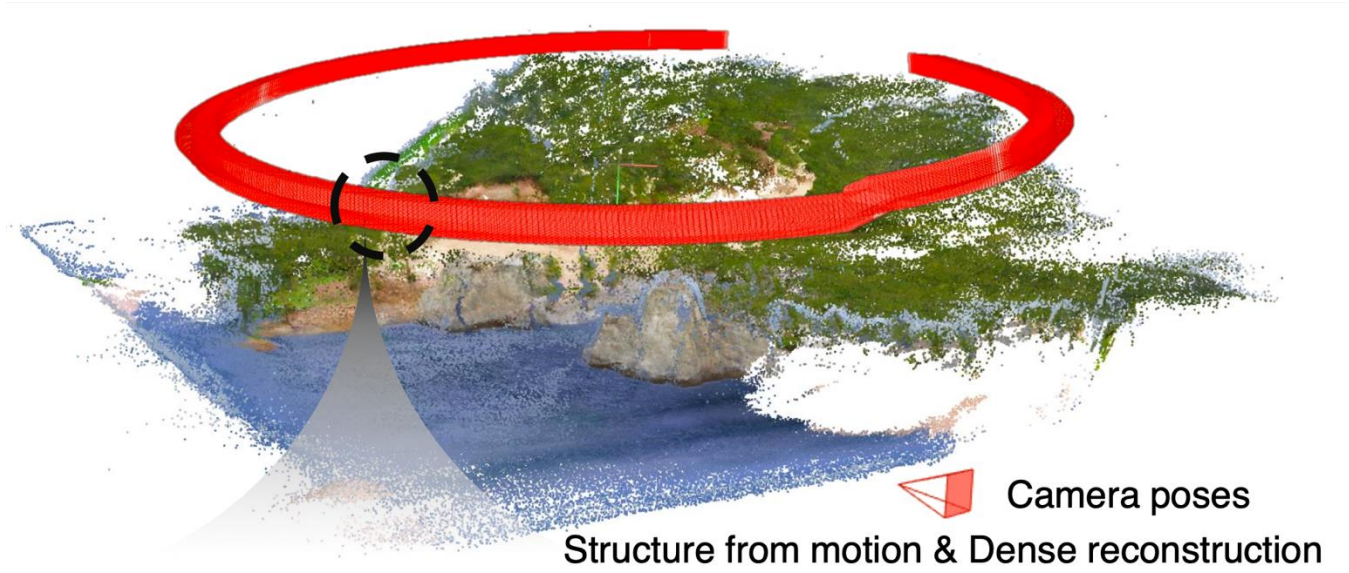


(d) Inf-Zero



(e) Ours

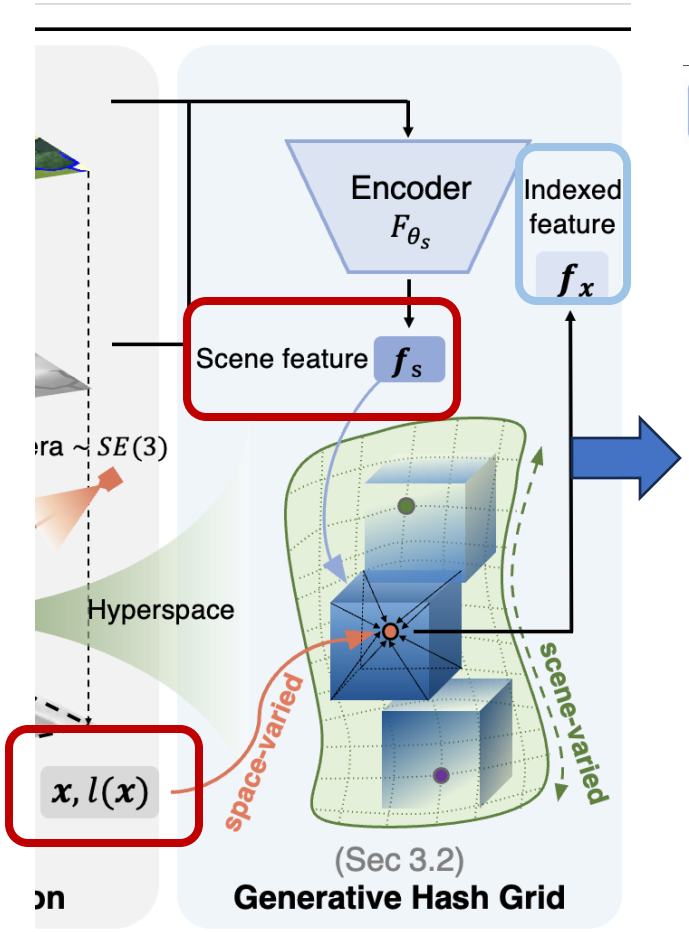
3D Consistency – Dense Reconstruction

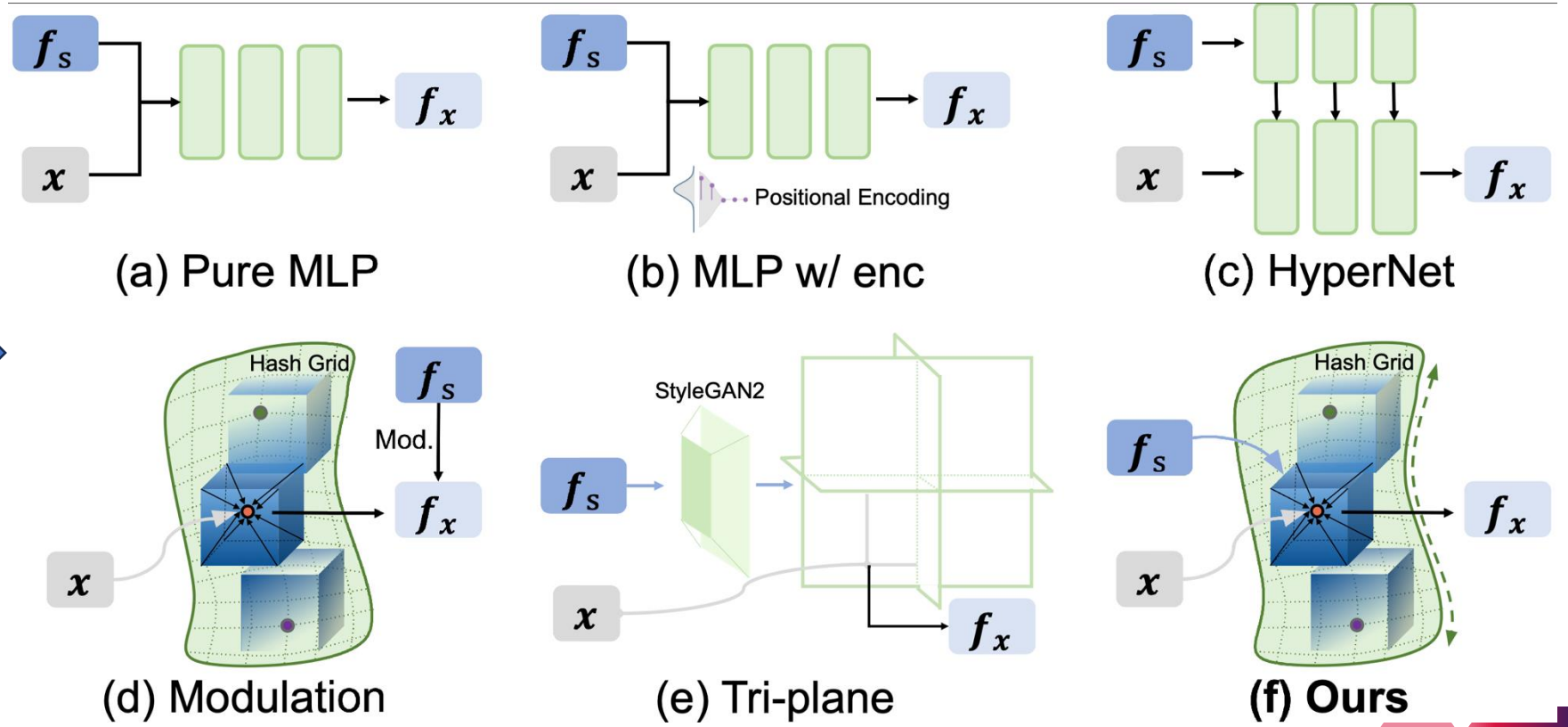


Corresponding rendered view

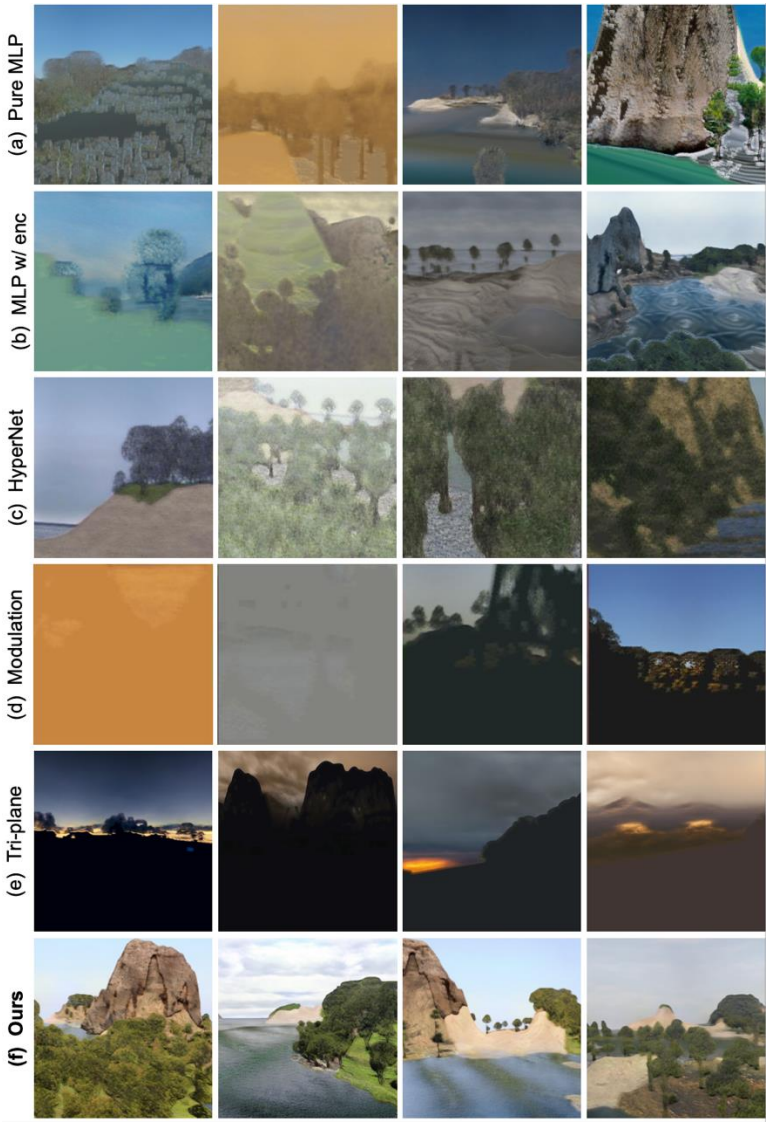
Reconstructed Poisson mesh

Different Generative Parameterizations



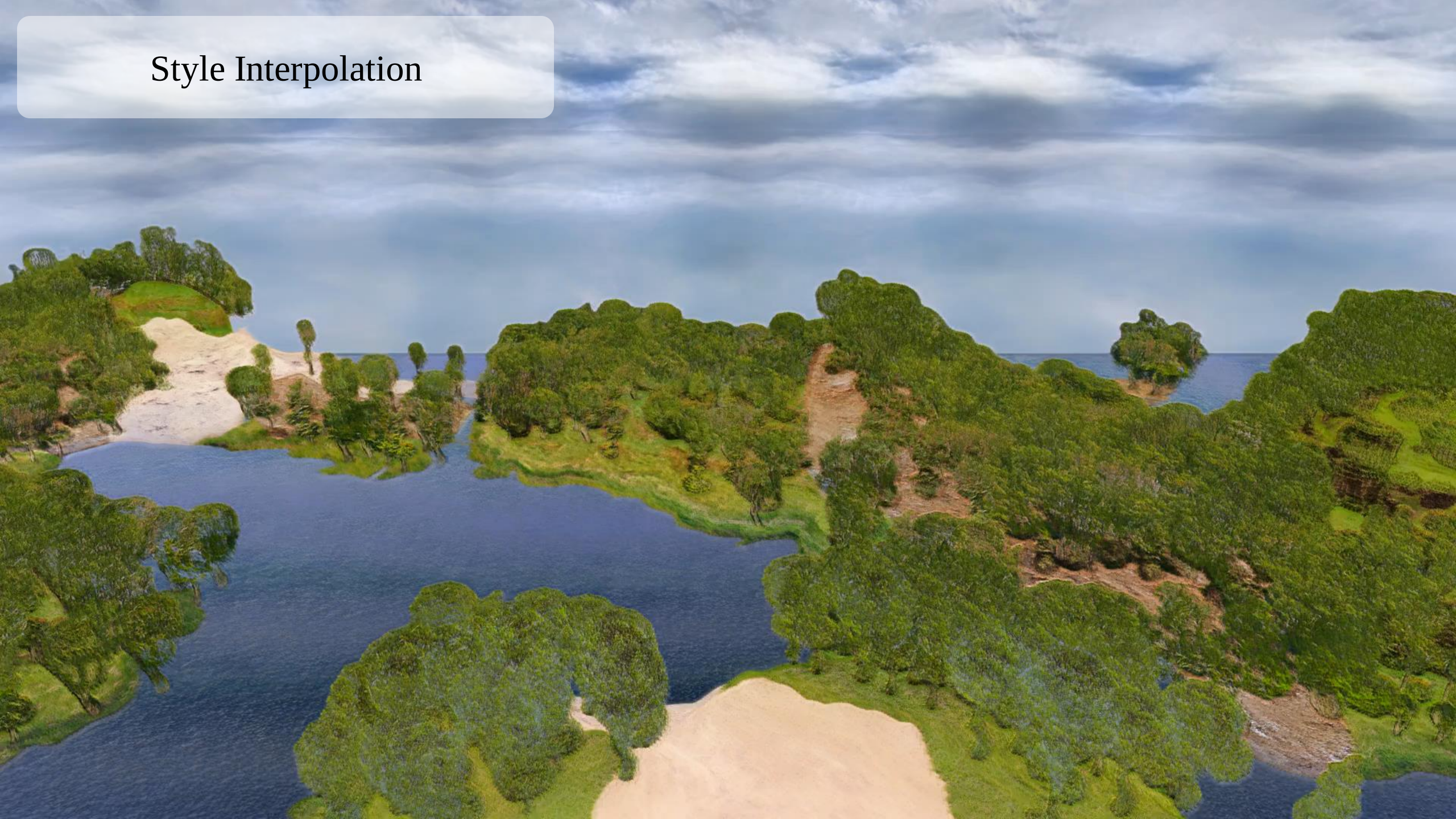


Different Generative Parameterizations



Methods	FID ↓	KID ↓	Depth ↓	CE ↓	SfM rate ↑
Pure MLP	88.76	5.07	0.433	0.194	0.650
MLP w/ enc	83.14	4.70	0.522	0.106	0.815
HyperNet	117.64	6.07	0.781	1.032	0.525
Modulation	120.88	6.94	0.994	1.331	0.325
Tri-plane	103.86	6.20	0.993	1.178	0.475
Ours	76.73	4.52	0.277	0.021	0.935

Style Interpolation




Stylized 3D Scene by ControlNet!

CityDreamer

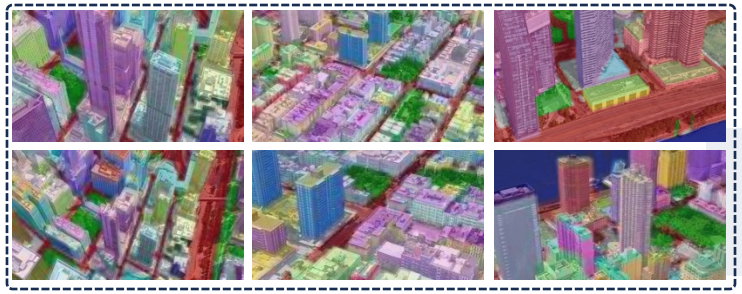
Compositional Generative Model of Unbounded 3D Cities

Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, Ziwei Liu

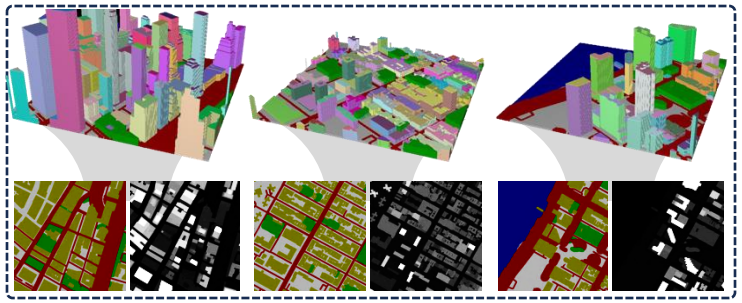
S-Lab, Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

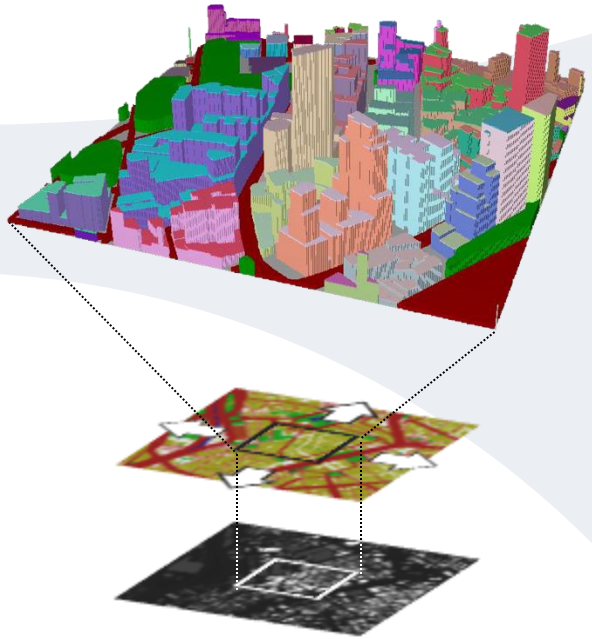
The Proposed Method



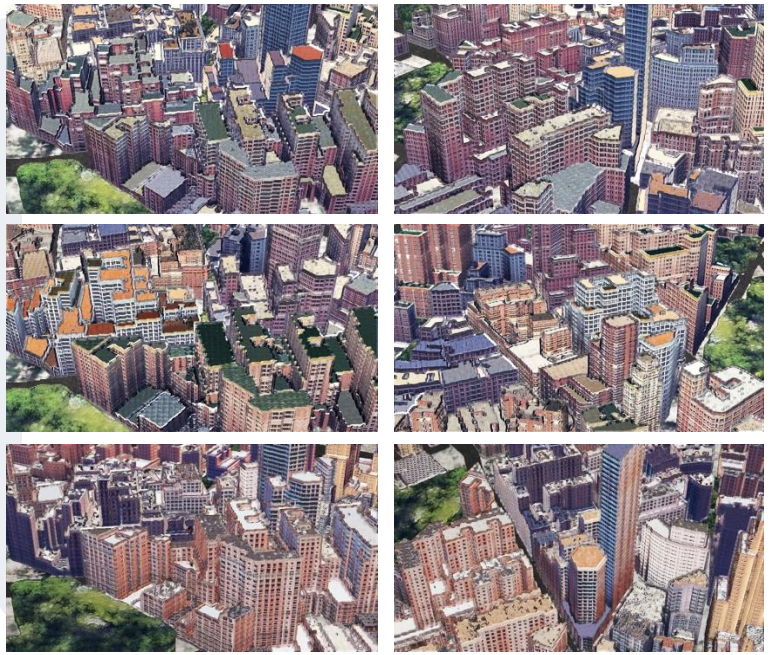
(a) Google Earth Dataset: Real-world City Appearance



(b) OSM Dataset: Real-world City Layout

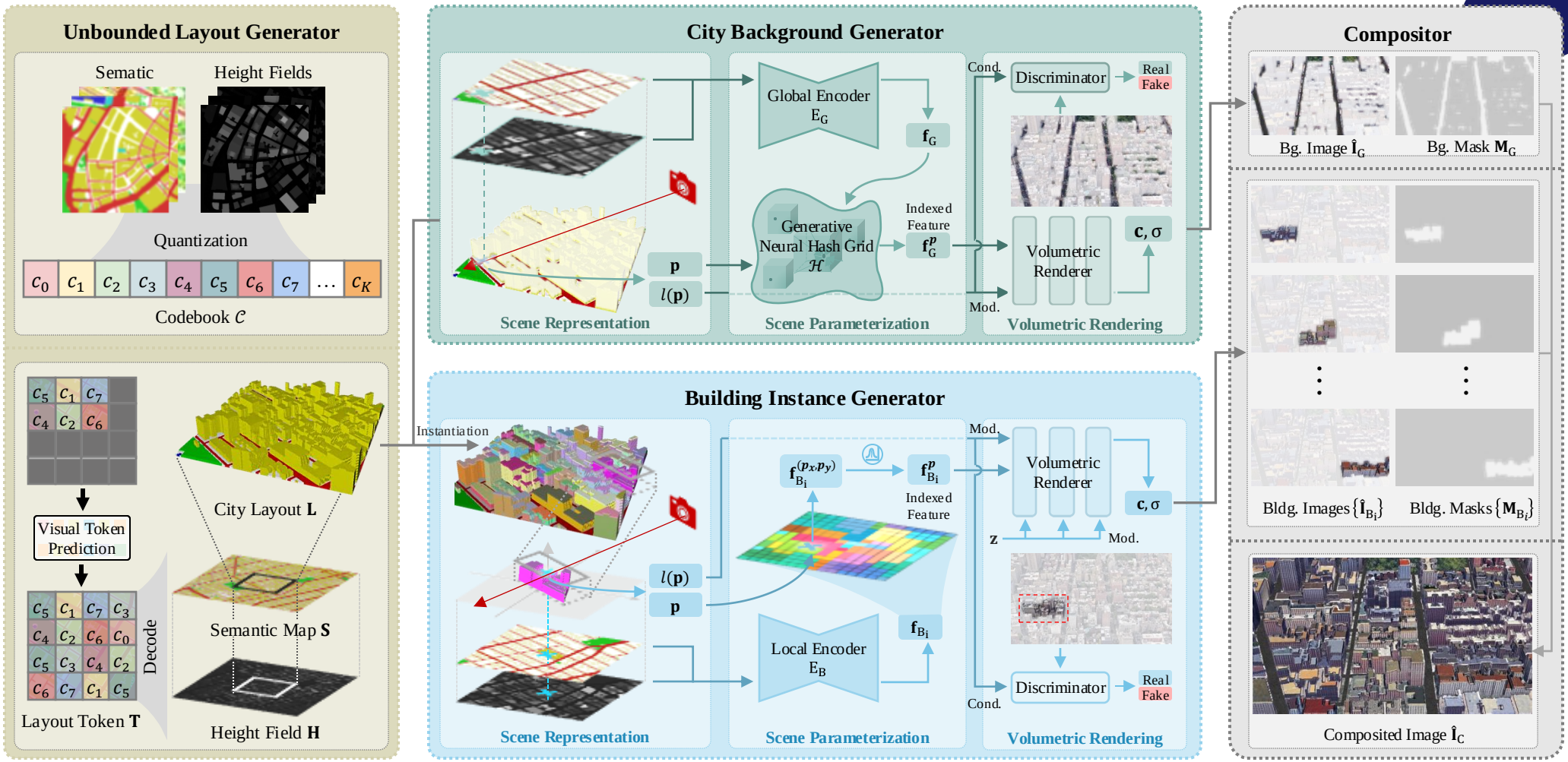


(c) Unbounded City Layout Generation



(d) CityDreamer Generated 3D Cities

The Proposed Method



Comparison to SOTA Methods

Overview

Scene

SceneDreamer
[Chen et al. 2023]
CityDreamer
[Xie et al. 2024]

Object

3DTopia-XL
[Chen et al. 2024]

Human

StructLDM
[Hu et al. 2024]

Ego Motion

EgoLM
[Hong et al. 2024]

3D PBR Asset Generation via Primitive Diffusion

S-LAB
FOR ADVANCED
INTELLIGENCE

Motivation

Previous SOTA

Low quality

Deterministic

No PBR -> No Physics

Goal

High quality

Generative

PBR Asset -> Correct Physics

PBR: Physically Based Rendering

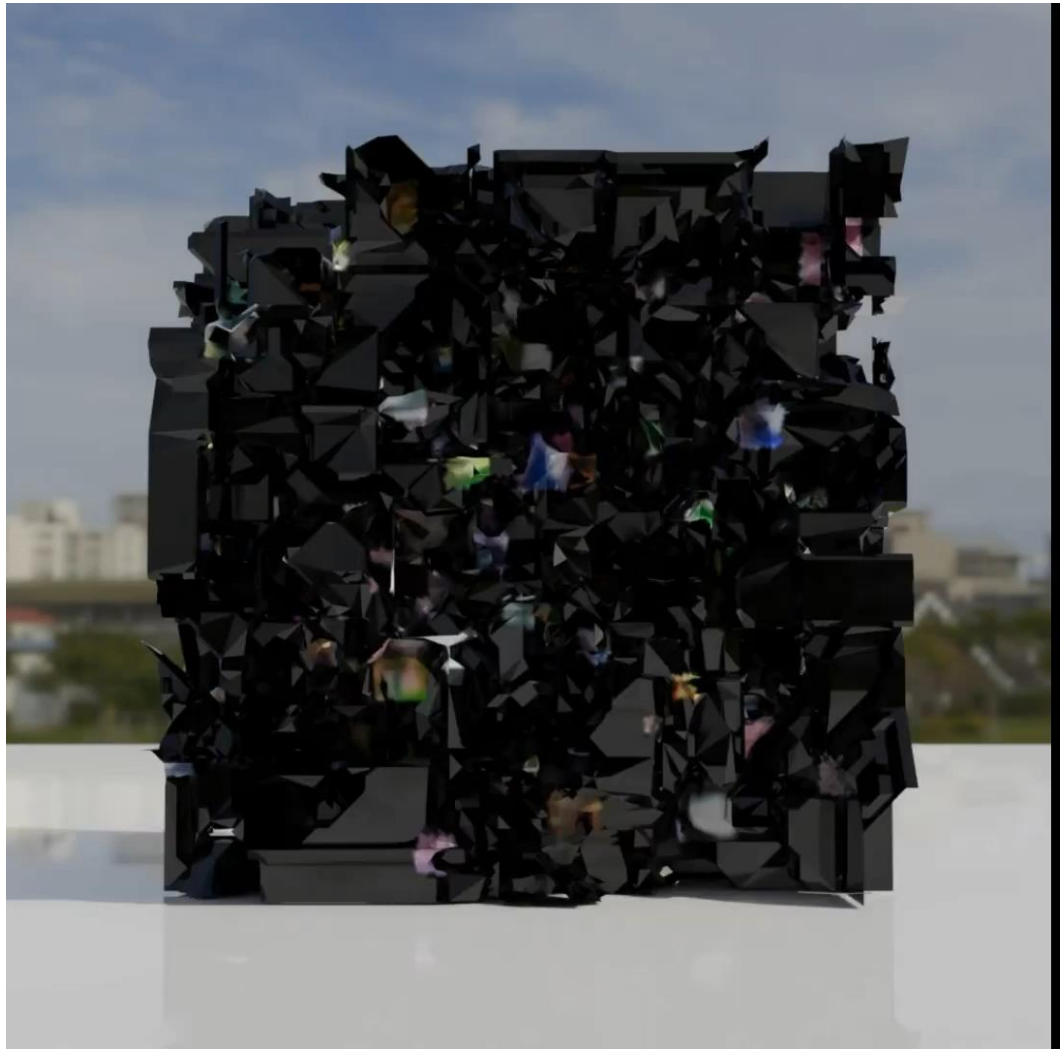
A Native 3D Diffusion Model for PBR Asset Generation

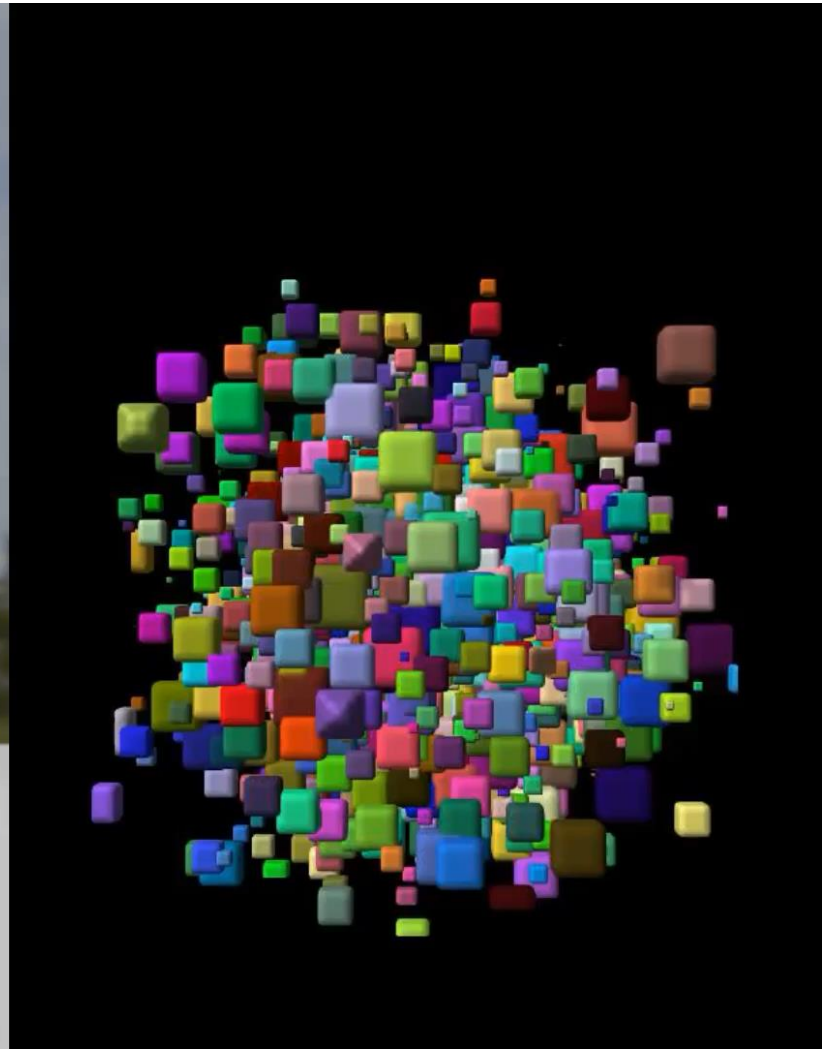
A cute unicorn

A Single Image / Texts

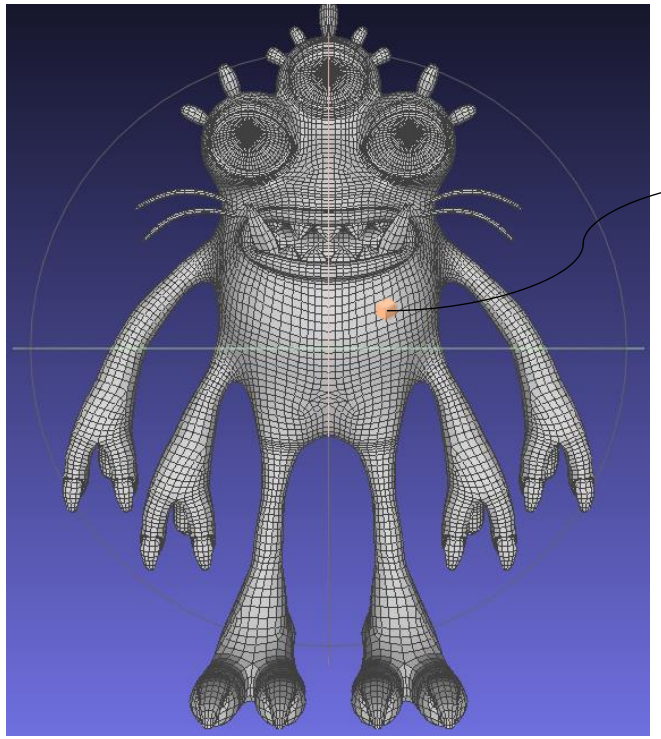
High-quality 3D Asset Ready for Blender

Key Idea: Primitive Diffusion





Stage I: Geometry, Texture, Materials into NxD Primitives



Input: GLB Mesh
Geometry, Texture and Materials

SDF Primitive

3D Anchors $T \in \mathbb{R}^3$

SDF $\alpha \in \mathbb{R}^{8\times8\times8}$

RGB $c \in \mathbb{R}^{8\times8\times8\times3}$

Mat $\rho \in \mathbb{R}^{8\times8\times8\times2}$

Scale $s \in \mathbb{R}^1$

$$V = \{T, \alpha, c, \rho, s\} \in \mathbb{R}^{3+8^3\times6+1}$$

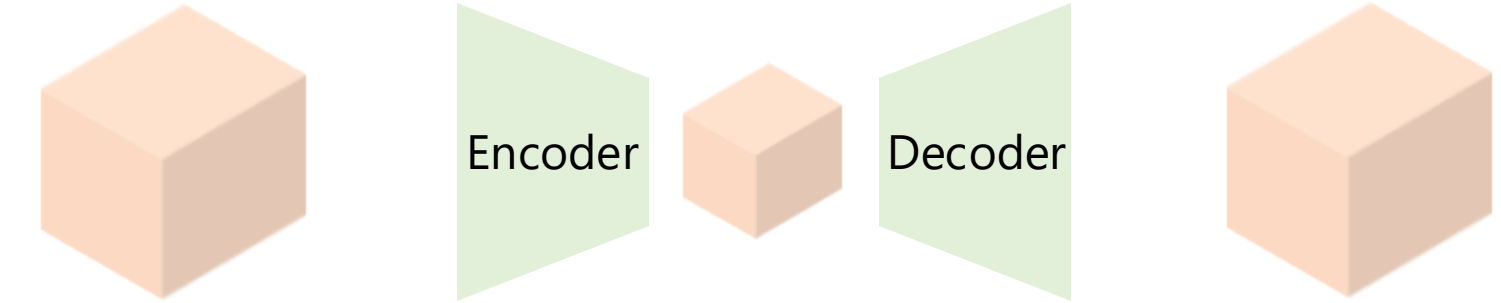
N x D Tensor to represent a textured mesh

$$\{V_i\} \in \mathbb{R}^{n\times(3+8^3\times6+1)}$$

1min Fitted Result

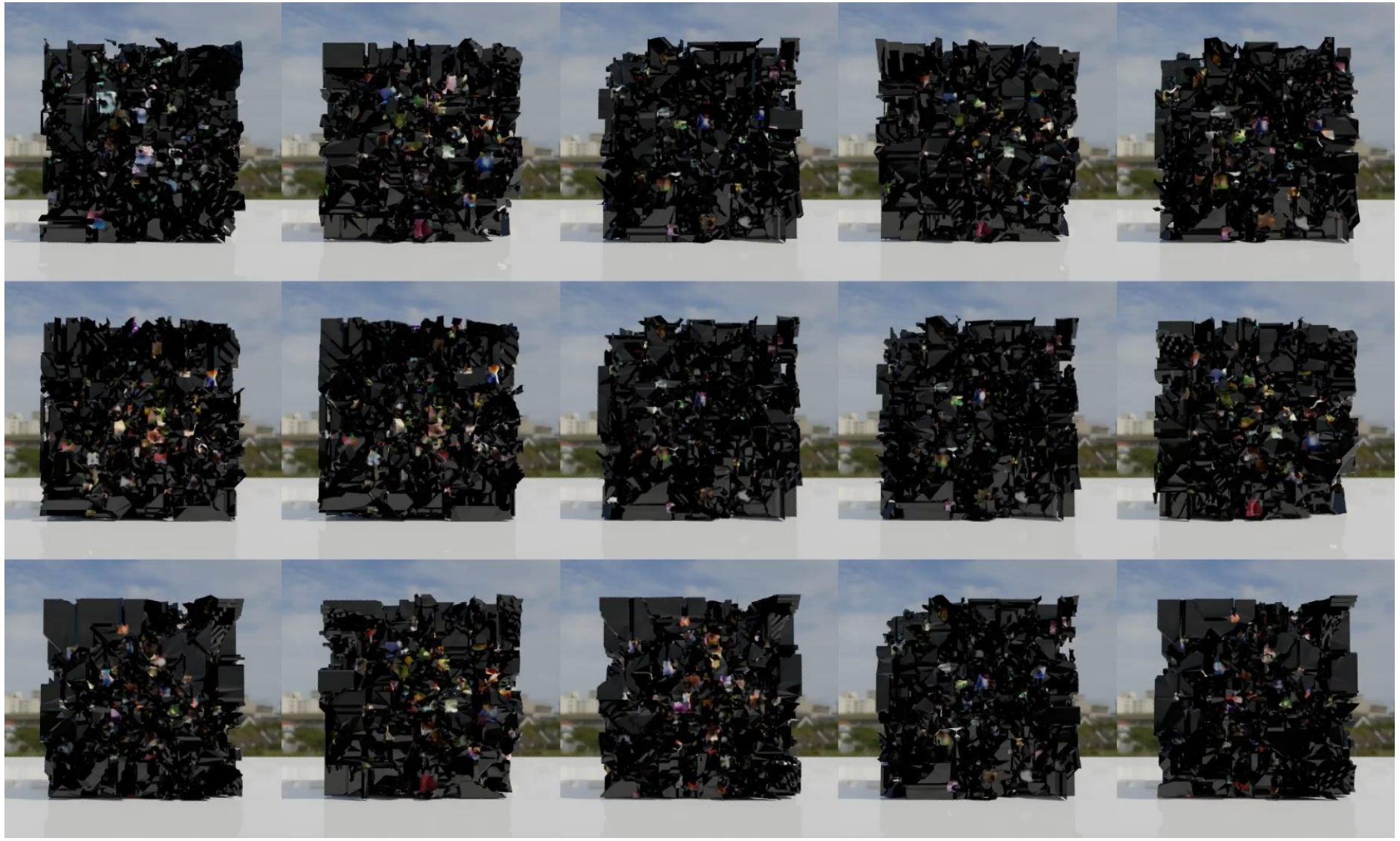
Stage II: Generative Modeling via Diffusion Transformer

- Patch-wise Compression using VAE



- Diffusion Transformer for Generation
  - **Target:** N patches of 3D latent with their 3D coordinates and scales
  - **Scale Up**: 1B #param model and 256K data

Gallery: Denoising in 5 seconds



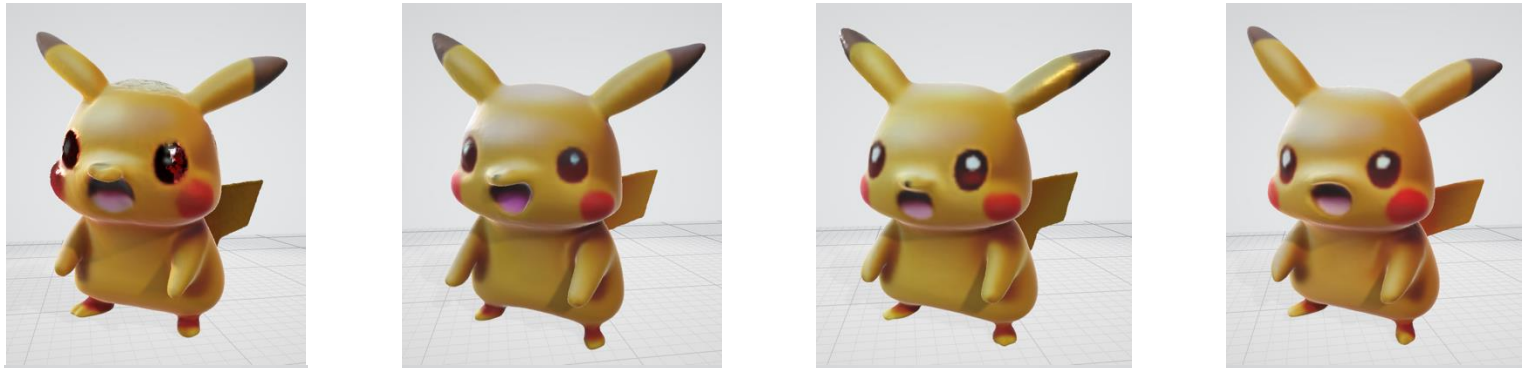
Gallery: Ready for Rendering Engine

InstantMesh | Real3D | CRM | CraftsMan | ShapE | **Ours** | Metallic / Roughness

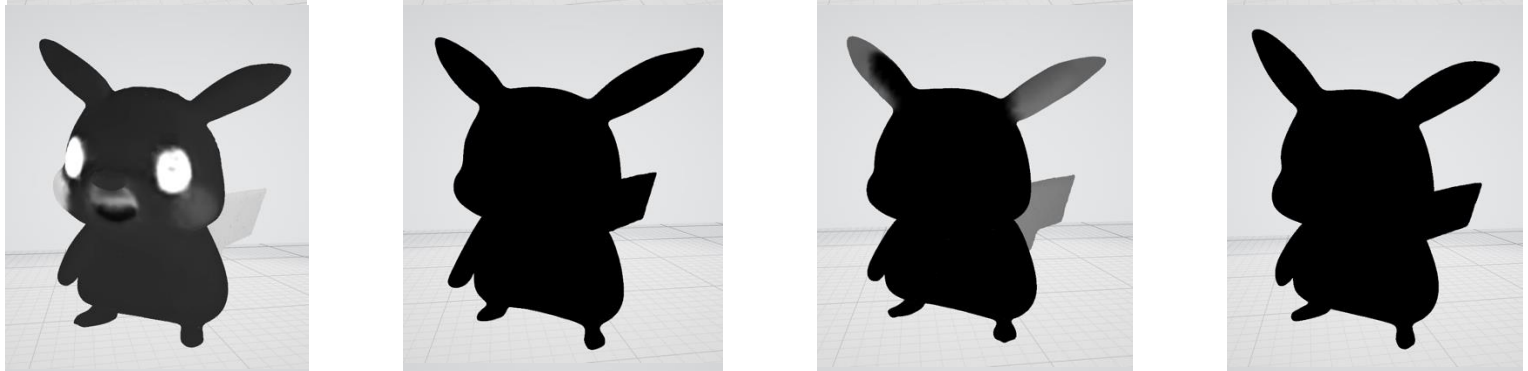
InstantMesh | Real3D | CRM | CraftsMan | ShapE | **Ours** | Metallic / Roughness

Diversity

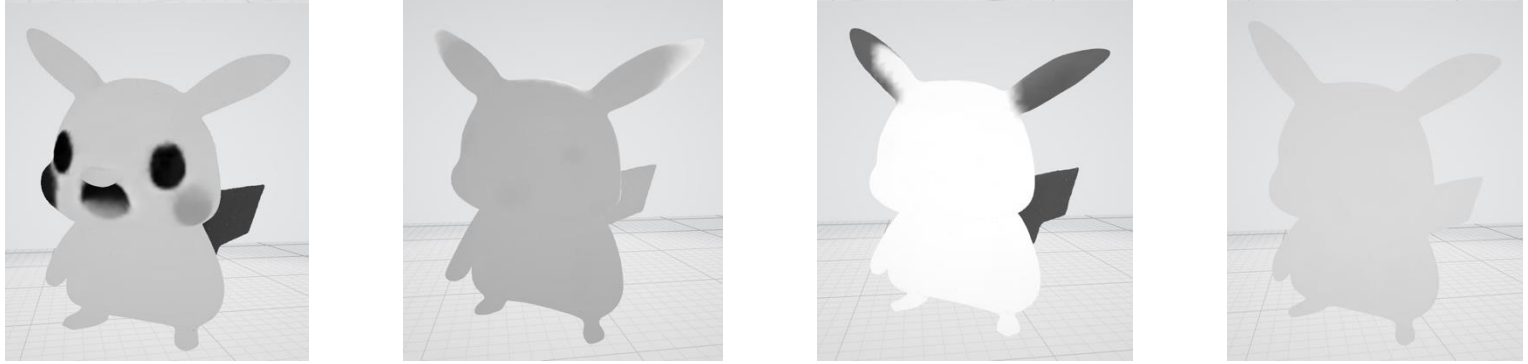
Renderings



Metallic



Roughness



Overview

Scene

SceneDreamer
[Chen et al. 2023]
CityDreamer
[Xie et al. 2024]

Object

3DTopia-XL
[Chen et al. 2024]

Human

StructLDM
[Hu et al. 2024]

Ego Motion

EgoLM
[Hong et al. 2024]

StructLDM: Structured Latent Diffusion for 3D Human Generation

Tao Hu

Fangzhou Hong

Ziwei Liu

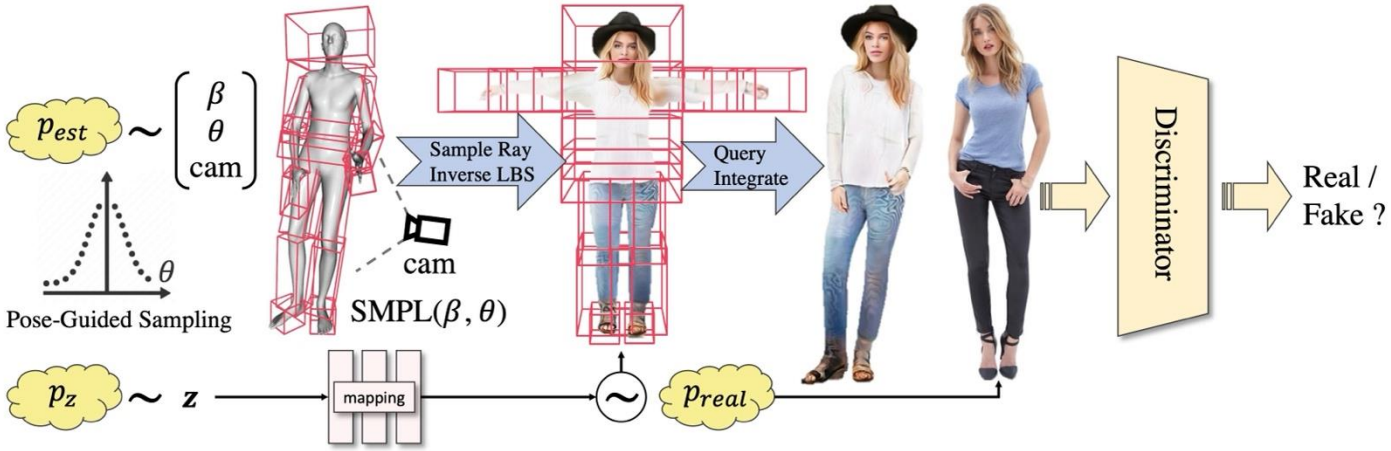
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Background

EVA3D (Hong et al. 2023)

Gaussian Shell (Abdal et al. 2024)

AG3D (Dong et al. 2023)



From 3D-Aware GAN to Diffusion

ICLR 2023

ICCV 2023

CVPR 2024

Photo-realistic 3D human GAN is too hard!

From 3D-Aware GAN to Diffusion

Human Dataset

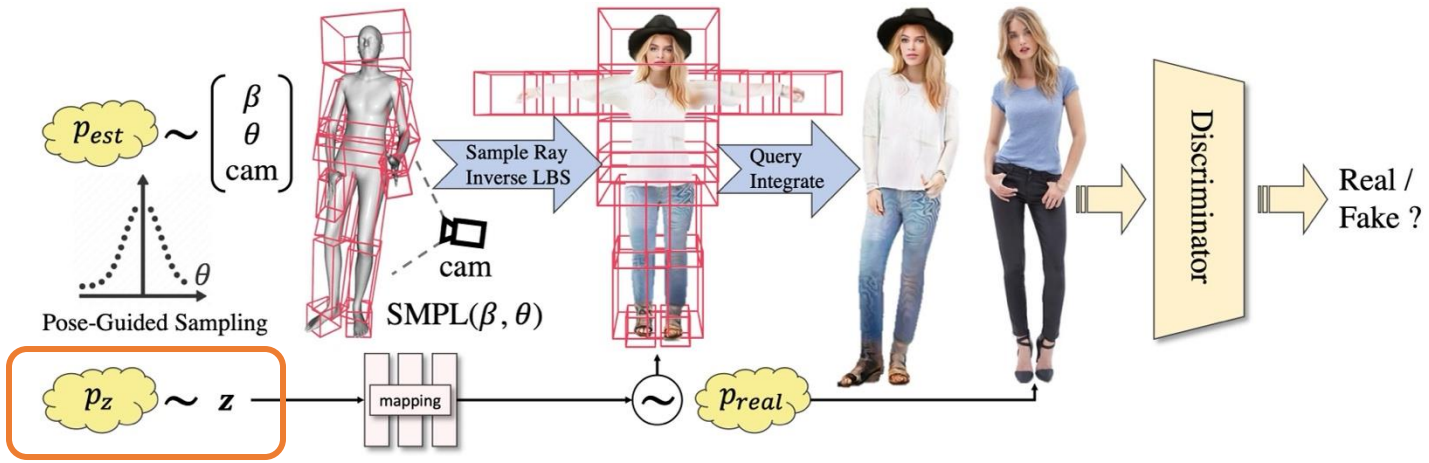
Some Human Representation

Latent Diffusion Model

✓ Expressive enough for high-fidelity 3D human reconstruction
✓ Compact enough for efficient diffusion model training

2D Structured Latent Representation

From 1D Latent Space to 2D Structured Latent Space

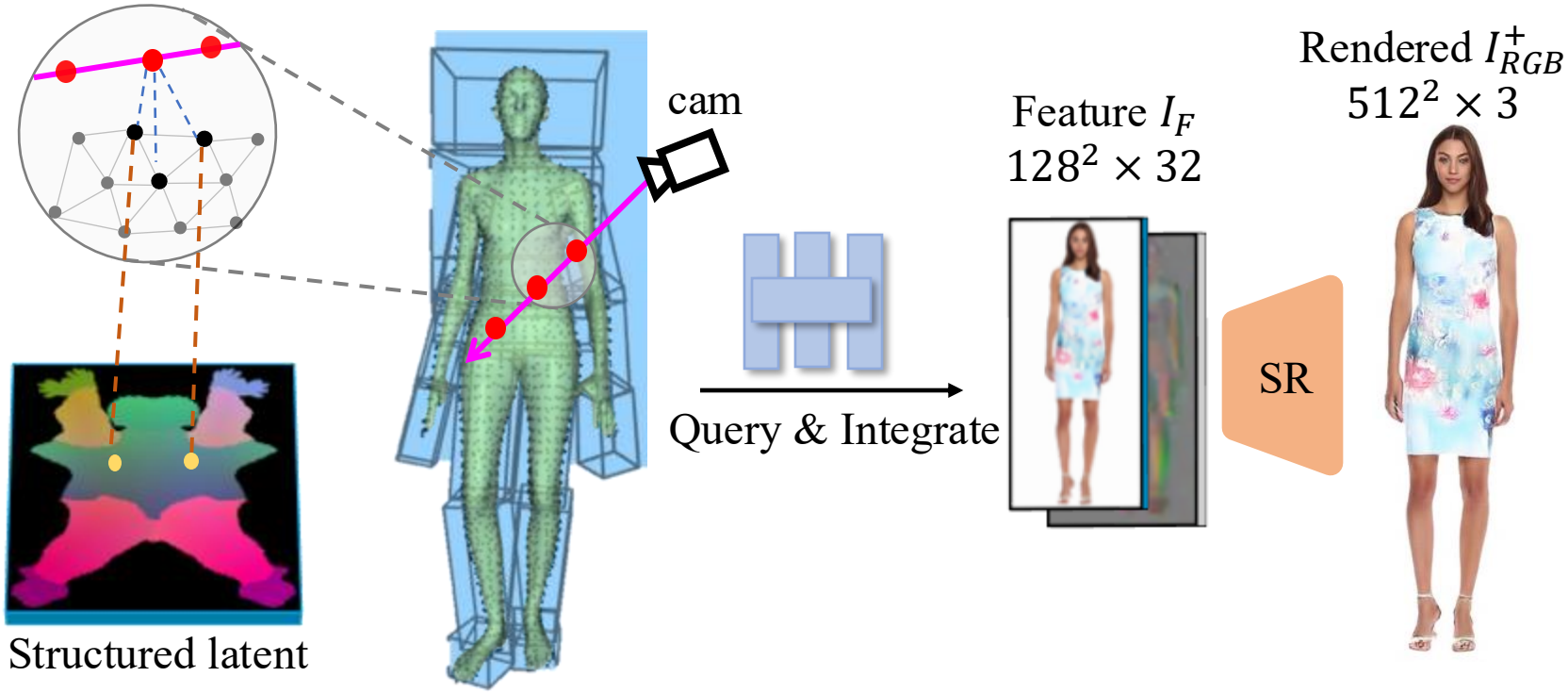


1D Latent Space

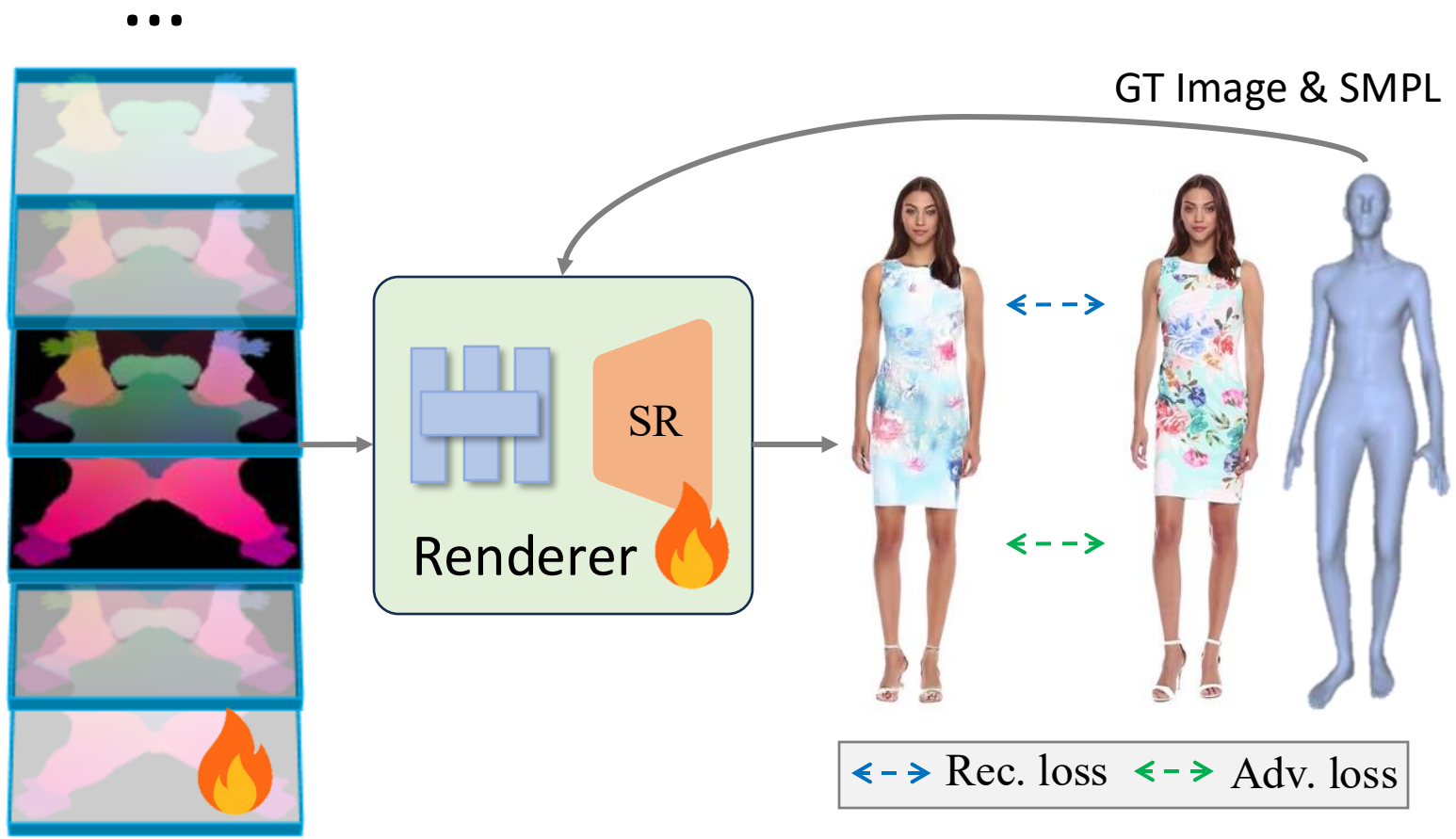
2D Structured Latent Space

✓ Easy Local Editing

Structured 3D Human Representation



Auto-Decoder for Latent Space Learning



2D Human Image Collections

Monocular Human Videos

3D Human Dataset

Experiments

FID@50k

Method	UBCFashion	RenderPeople	THUman2.0	DeepFashion
StyleSDF	18.52	51.27	-	92.40
EG3D	23.95	24.32	-	26.38
EVA3D	12.61	44.37	124.54	15.92
AG3D	11.04	-	-	**10.93**
PrimDiff	-	17.95	-	-
Ours	**9.56**	**13.98**	**25.22**	20.82

Qualitative Results

UBCFashion

DeepFashion

Controllable 3D Human Generation

Pose Control

Shape Control

Interpolation in **2D** latent space

Local Editing via Structured Latent Space

Discussion

- Auto-decoder is the bottleneck
- Loose clothes
- Scale-up to larger human dataset

DNA-Rendering
[Cheng et al. 2023]

WildAvatar
[Huang et al. 2024]

Overview

Scene

SceneDreamer
[Chen et al. 2023]
CityDreamer
[Xie et al. 2024]

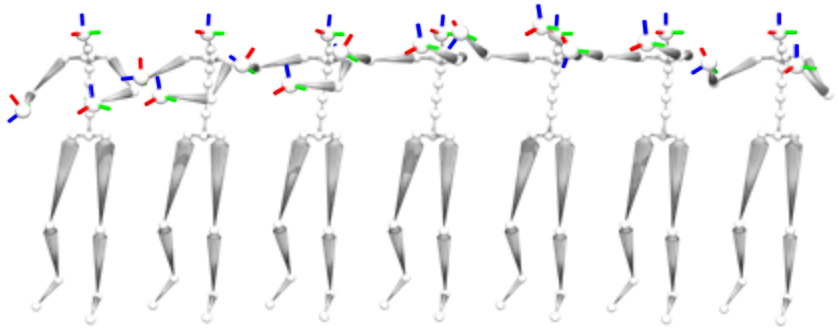
Object

3DTopia-XL
[Chen et al. 2024]

Human

StructLDM
[Hu et al. 2024]

Ego Motion



EgoLM
[Hong et al. 2024]

EgoLM: Multi-Modal Language Model of Egocentric Motions

Fangzhou Hong

Vladimir Guzov

Hyo Jin Kim

Yuting Ye

Richard Newcombe

Ziwei Liu

Lingni Ma

Meta

REALITY LABS

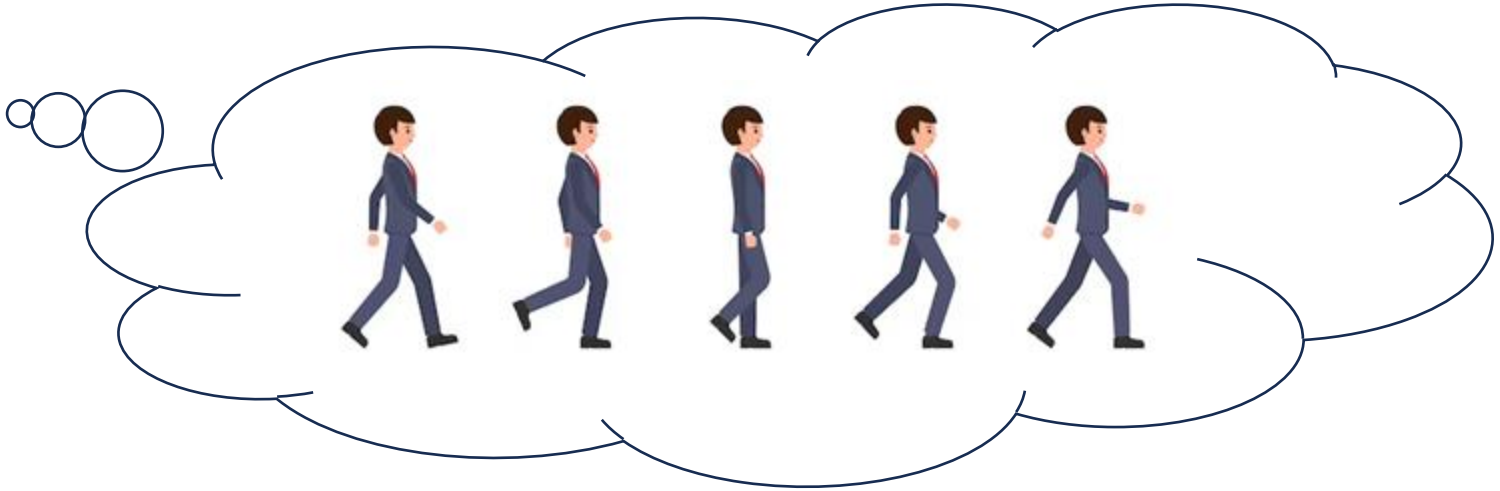
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Background

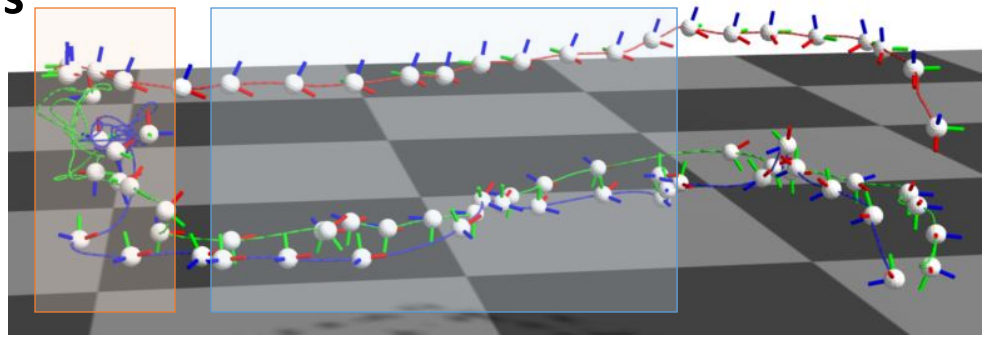
Multi-Modal LLM, e.g., GPT-4o, are great.



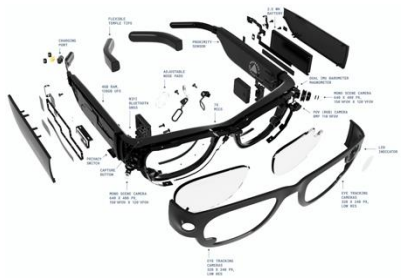
Need the understanding of egocentric motions.

Multi-Modal Egocentric Capture

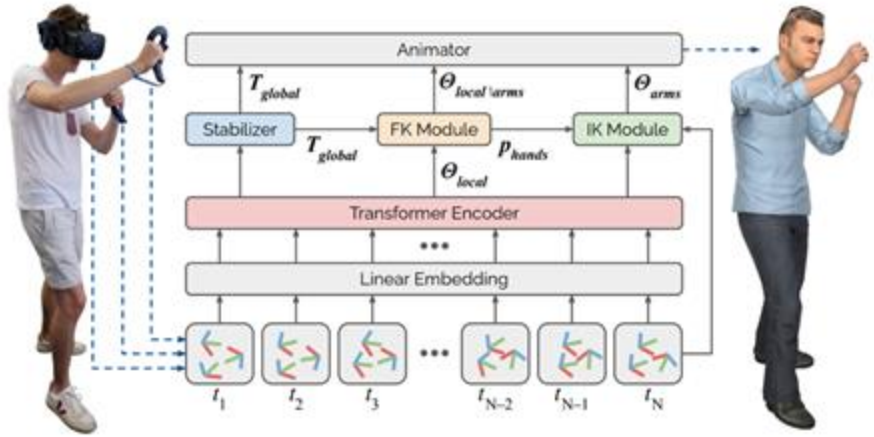
Sparse Motion Sensors

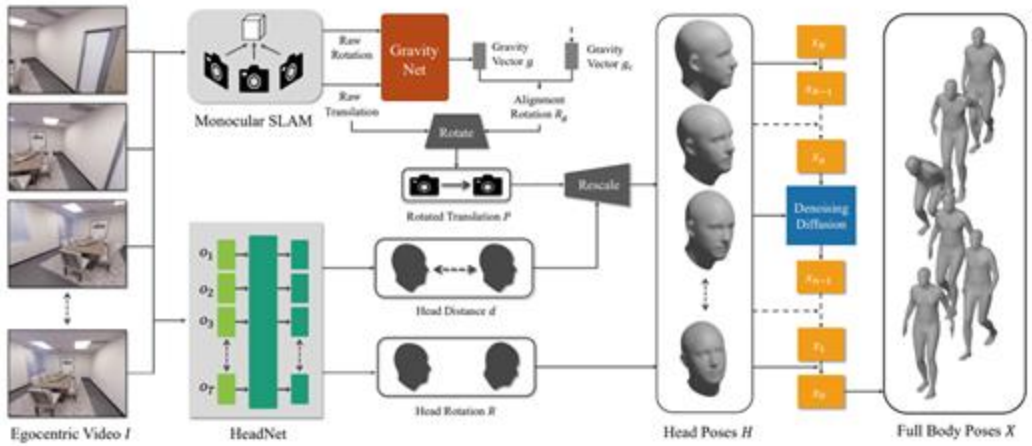


Egocentric Videos



Related Works – Egocentric Motion Tracking





Jiaxi Jiang[1] Paul Streli[2] Huajian Qiu[3] Andreas Fender[1] Larissa Laich[1] Patrick Snape[2] Christian Holz[1]

AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing

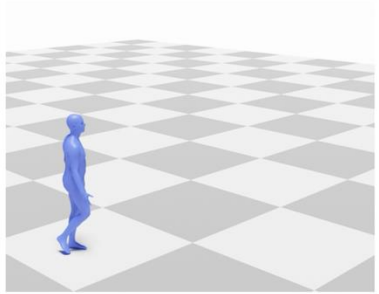
ECCV 2022 siplab.org/projects/AvatarPoser

siplab ETH zürich Meta

Three-Points Body Tracking

EgoEgo on Kinpoly-MoCap

Input



Output

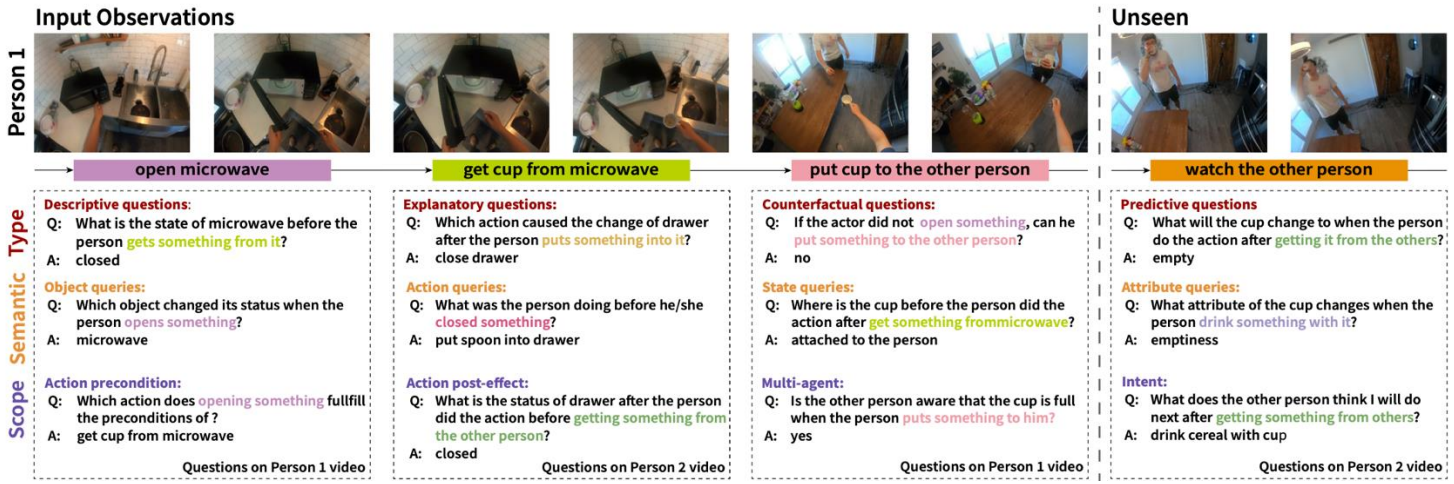
One-Point Body Tracking

AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing, Jiang et al., ECCV 2022

Ego-Body Pose Estimation via Ego-Head Pose Estimation, Li et al., CVPR 2023

Related Works – Egocentric Video Understanding

Ego4D

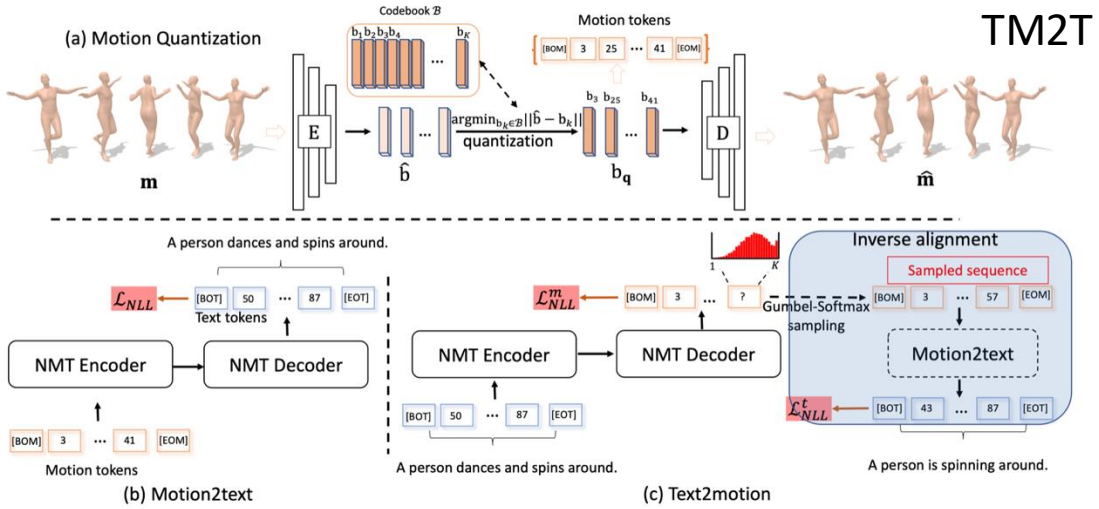


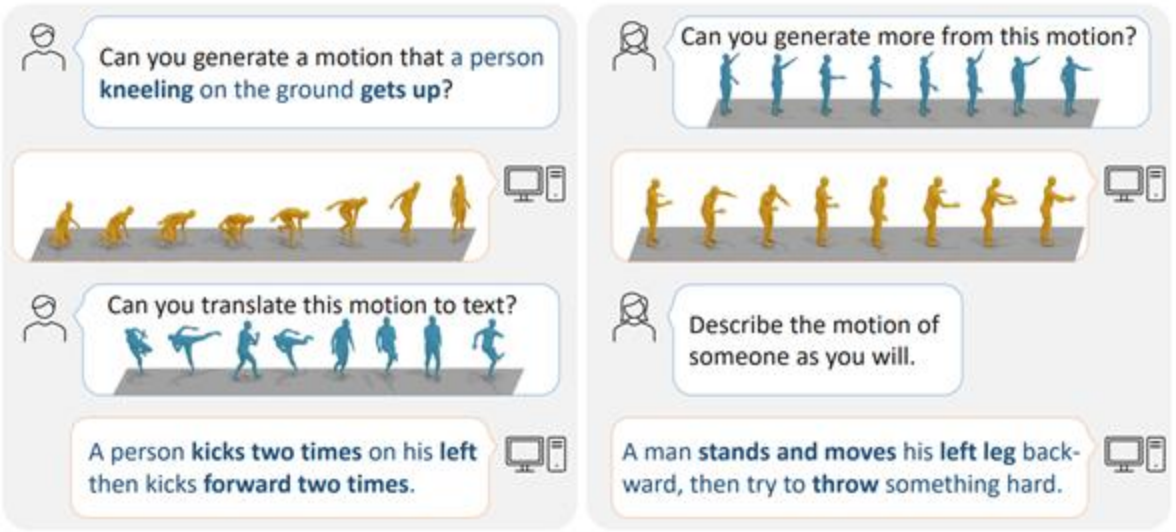
EgoTaskQA

Ego4D: Around the World in 3,000 Hours of Egocentric Video, Grauman et al., CVPR 2022
EgoTaskQA: Understanding Human Tasks in Egocentric Videos, Jia et al., NeurIPS 2022

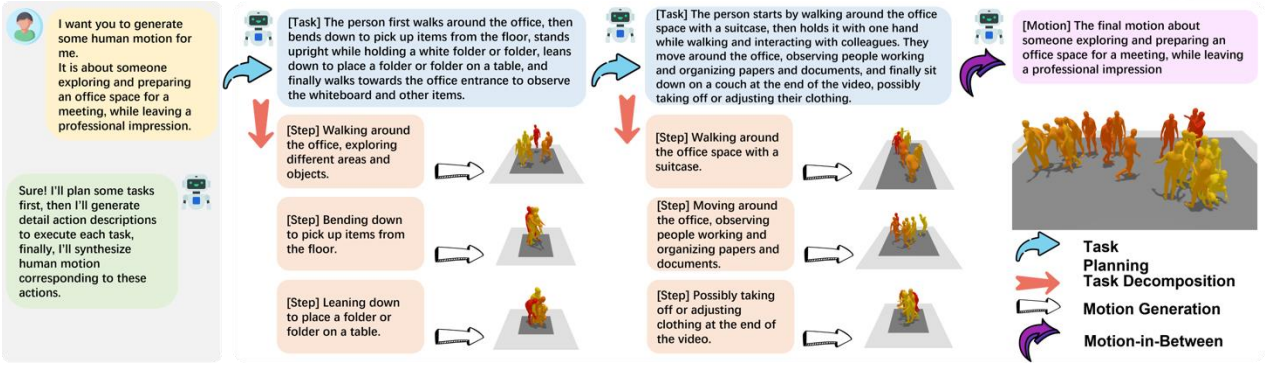
Related Works – Motion and Language

TM2T





MotionGPT



AvatarGPT

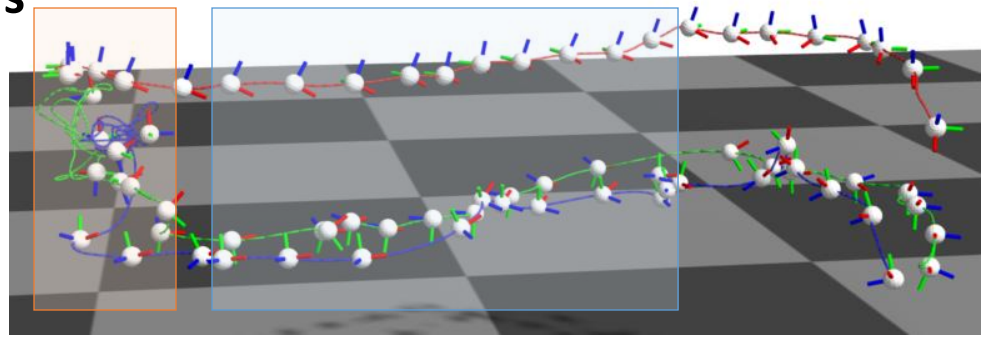
TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts, Guo et al., ECCV 2022

MotionGPT: Human Motion as a Foreign Language, Jiang et al., NeurIPS 2023

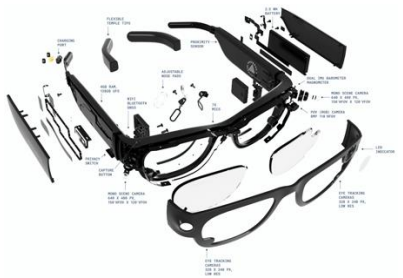
AvatarGPT: All-in-One Framework for Motion Understanding, Planning, Generation and Beyond, Zhou et al., arXiv 2024

Egocentric Motion Tracking and Understanding

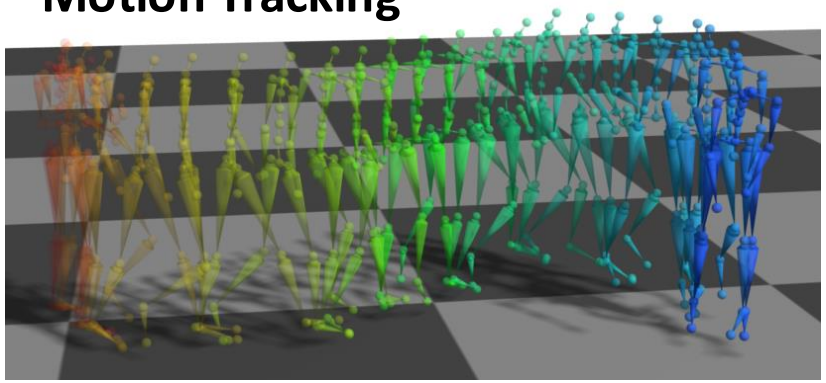
Sparse Motion Sensors



Egocentric Videos



Motion Tracking

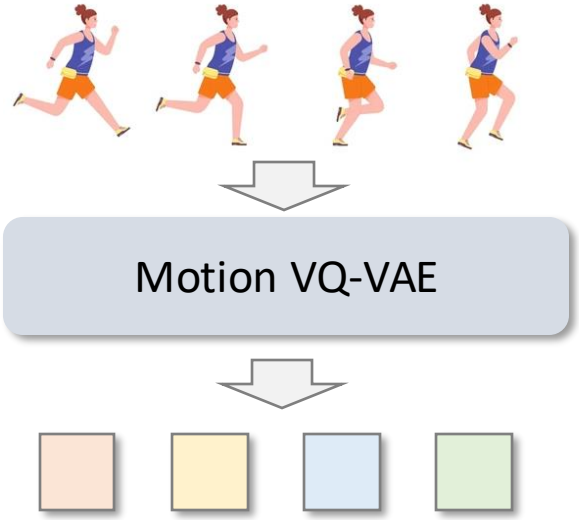


Motion Understanding

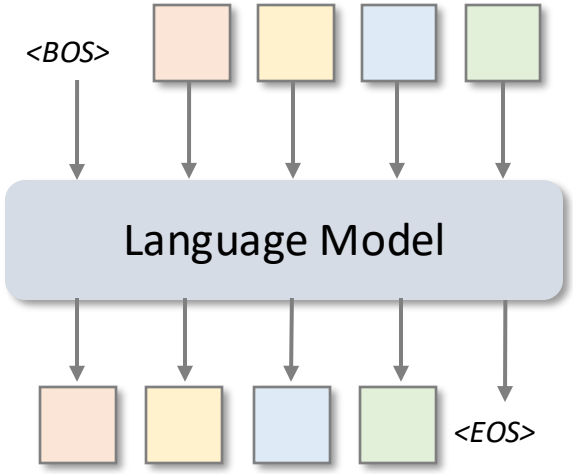
"The person is standing straight as she puts the piece of clothing on the hanger."

"The person turns around then walks out of the bedroom."

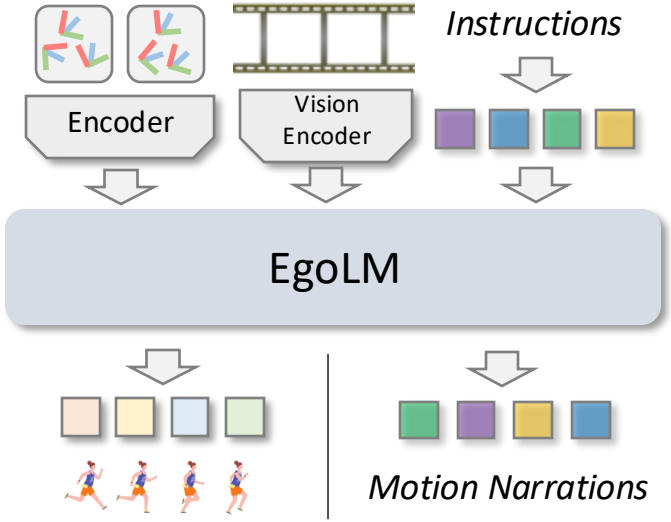
Multi-Modal Multi-Tasking LM for Ego Motion



1) Motions Tokenization

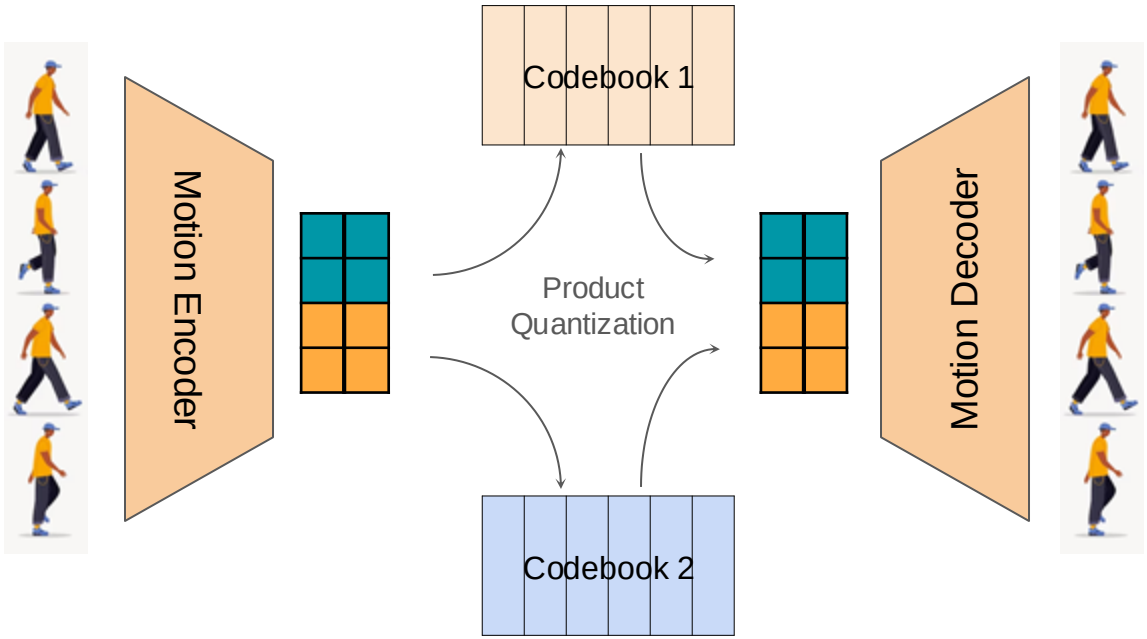


2) Motion Pre-Training



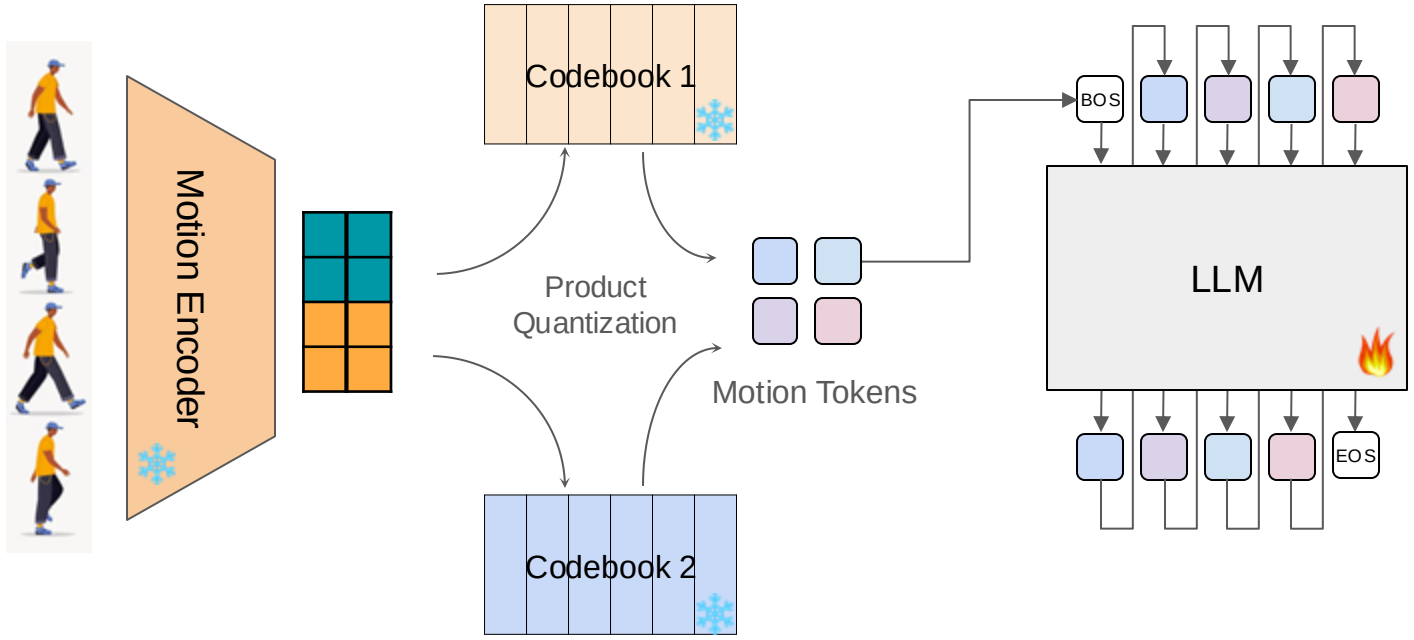
3) Multi-Modal Instruction Tuning

Step 1: Motion VQ-VAE

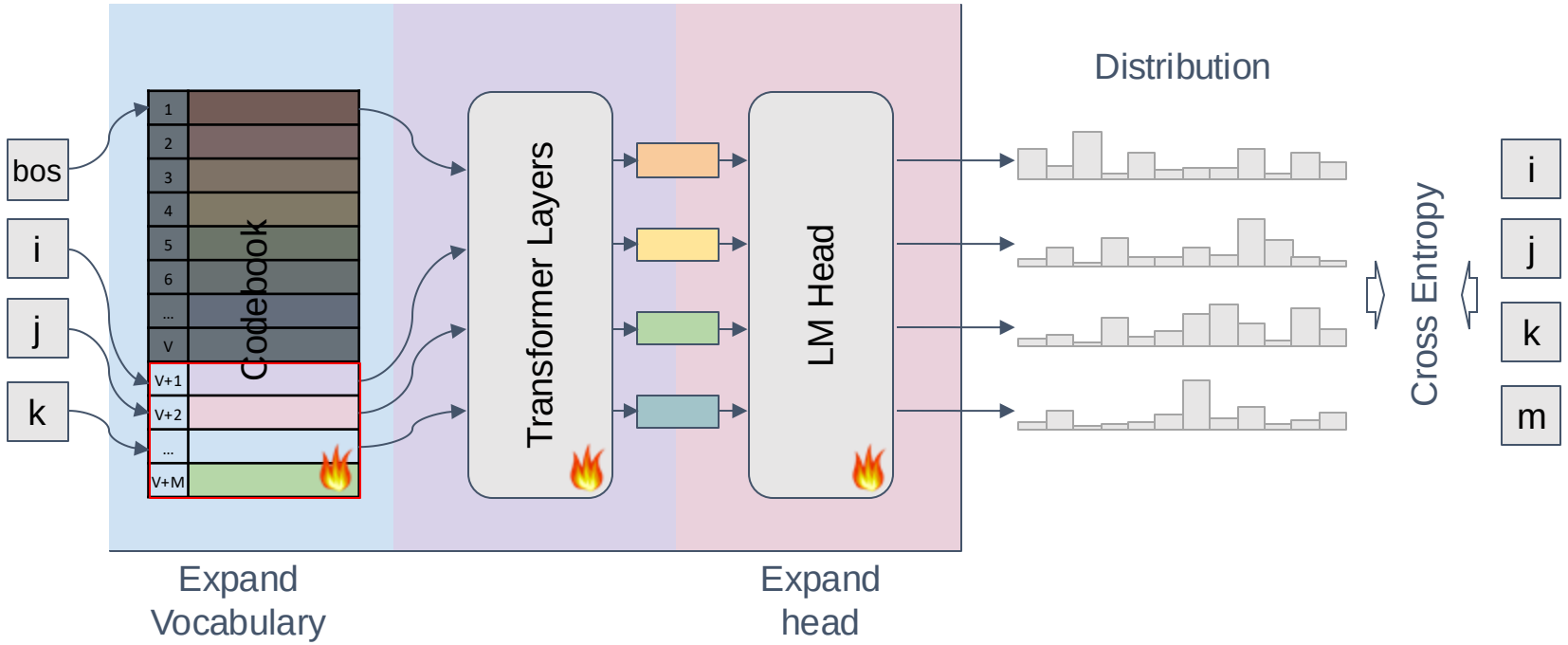


PQ Num	CB Num	CB Dim	MPJPE	PA-MPJPE	ACCEL
1	2048	512	51.60	37.55	1.09
2	2048	512	39.63	29.77	0.71
2	16384	256	39.13	29.78	1.08
2	16384	64	**34.49**	**26.83**	**0.67**

Step 2: Motion Pre-Training

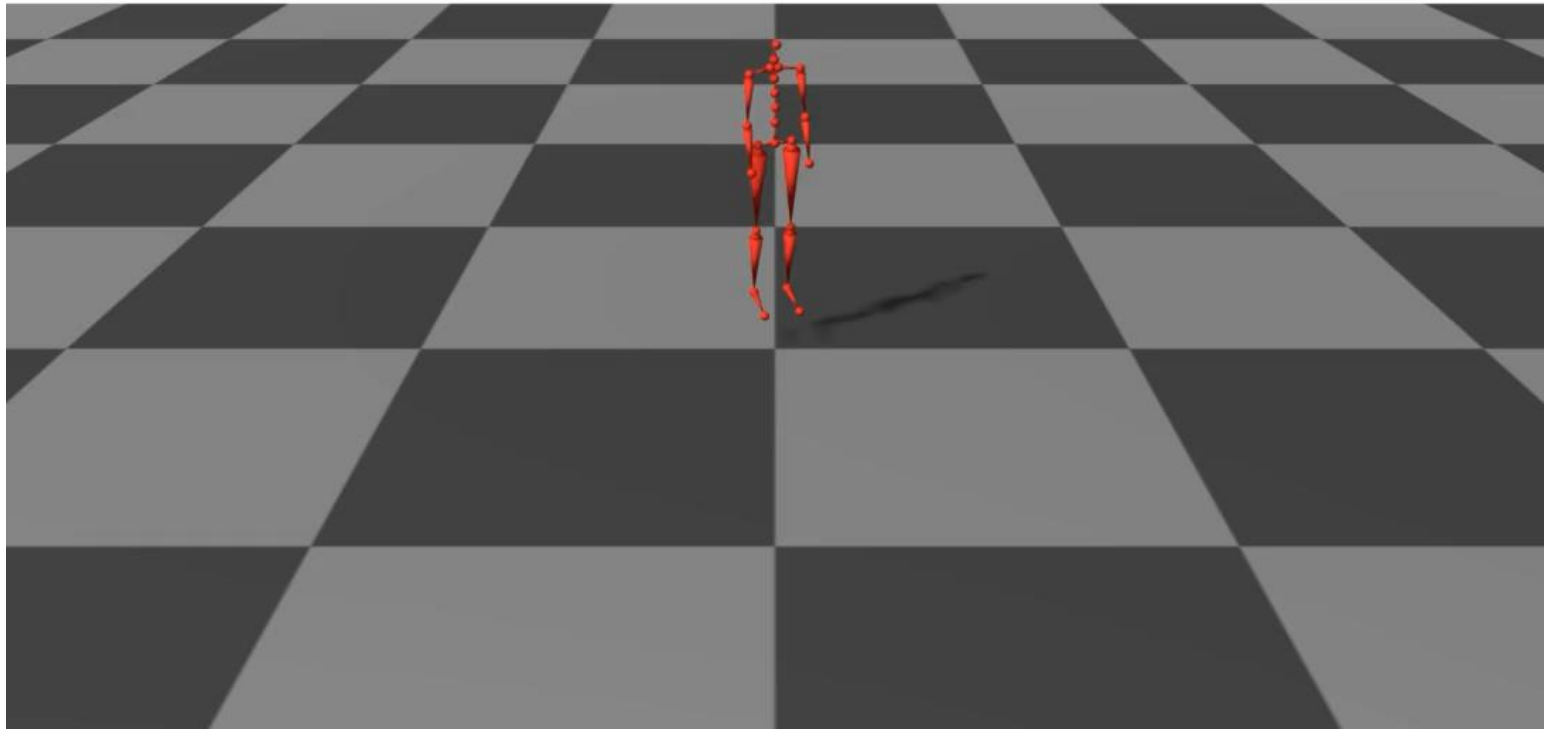


Step 2: Motion Pre-Training



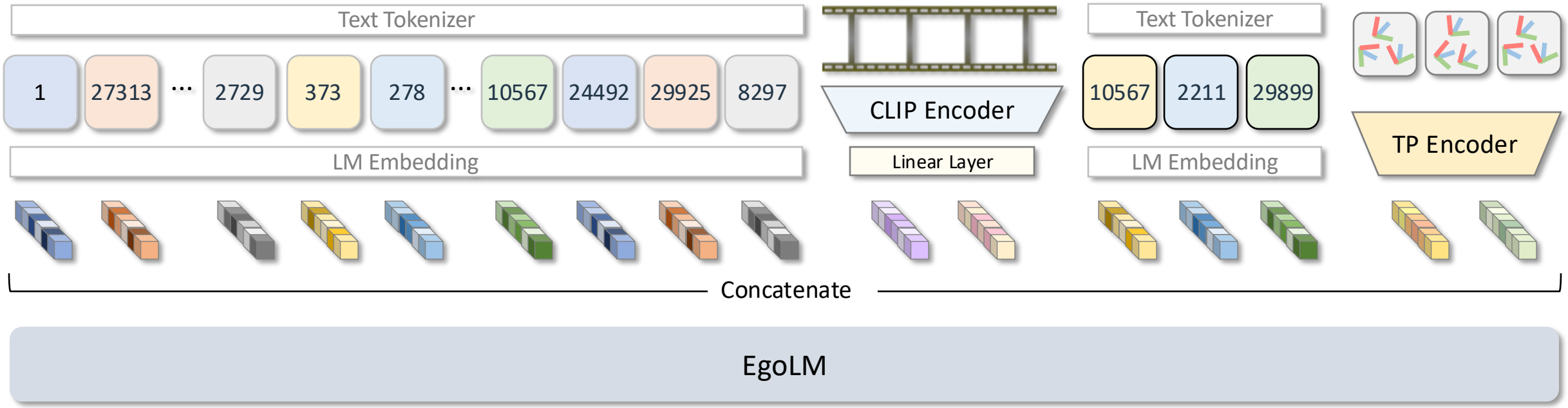
By-Product: Unconditional Motion Generator

2X



Step 3: Instruction Tuning

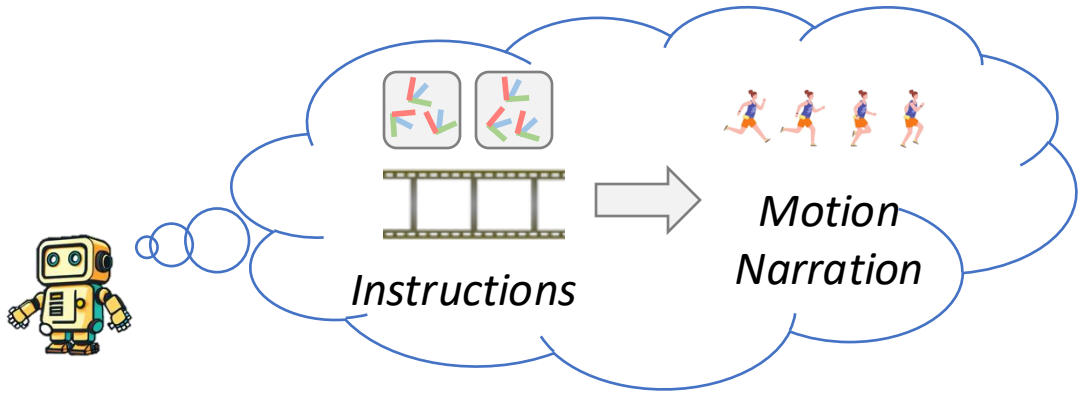
"<s> Perform ... based on the given ... Input CLIP embeddings: <CLIP_Placeholder>*. Input three-points:* <TP_Placeholder>"



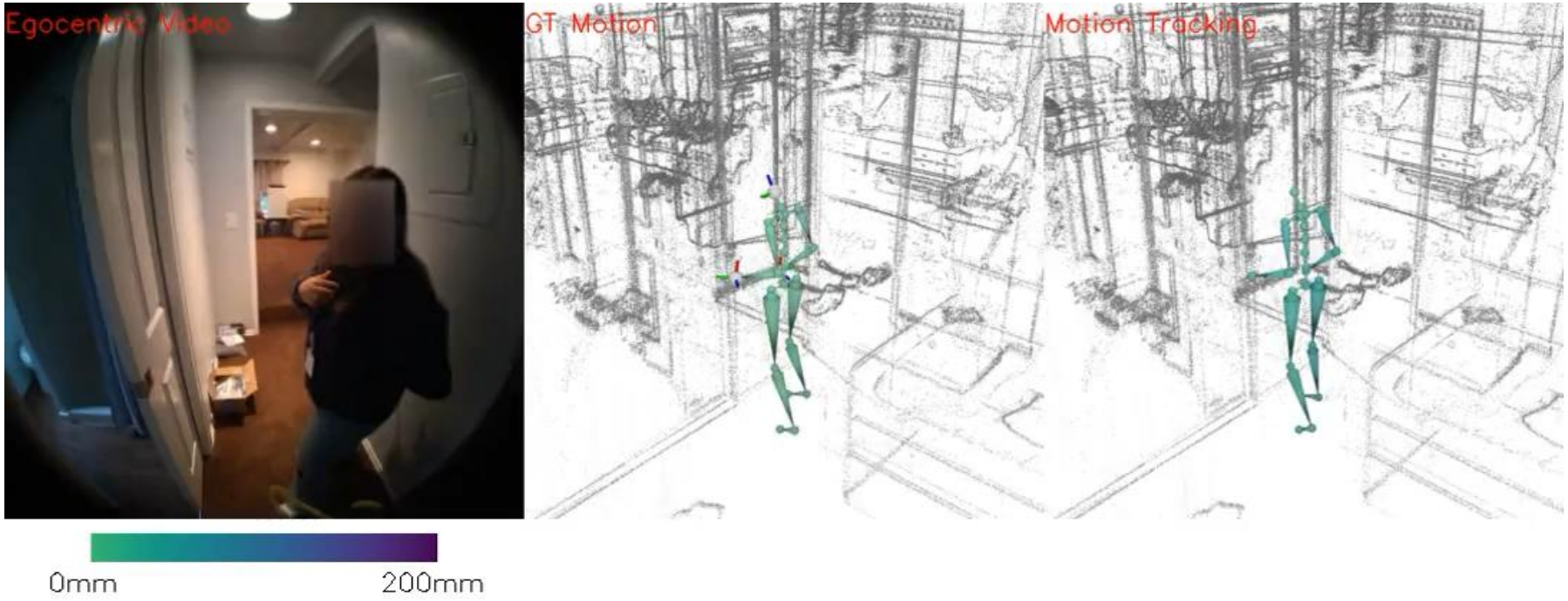
Experiments

Q1: Does it work? **- Yes!**

Q2: Is multi-modality helpful? **- Yes!**



It works!



Three-Points / One-Point Body Tracking

Method	Input Modality			Full	Upper	Lower	J.A.	Root
	3pts	1pt	Vid.					
AvatarPoser	✓			85.89	52.78	165.18	**12.41**	14.78
Bodiffusion	✓			79.80	52.79	152.68	12.74	**13.09**
Ours	✓			83.88	54.06	148.37	13.31	14.13
AvatarPoser		✓		129.23	94.19	192.34	16.55	21.60
EgoEgo		✓		132.16	100.02	190.32	18.90	21.80
Ours		✓		127.45	97.87	174.92	16.97	20.57

Three-Points Motion Understanding

Method	Input Modality			Bert↑	Bleu@1↑	Bleu@4↑	RougeL↑
	3pts	Motion	Vid.				
TM2T		✓		11.08	40.11	8.99	30.70
MotionGPT		✓		14.09	42.22	10.31	32.33
Ours (M2T&T2M)		✓		15.90	42.68	11.06	33.71
Ours (TPV2T)	✓		✓	18.38	44.55	12.12	33.80

Discussion

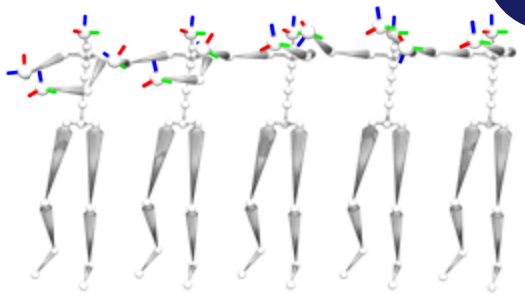
- **Larger** Language Models
- Egocentric Video Encoding
- Other Modalities / Tasks

Summary

Scene

Object

Human



Ego Motion

Current Progress

Relighting

Physical Property

Generative Simulator

Future Work

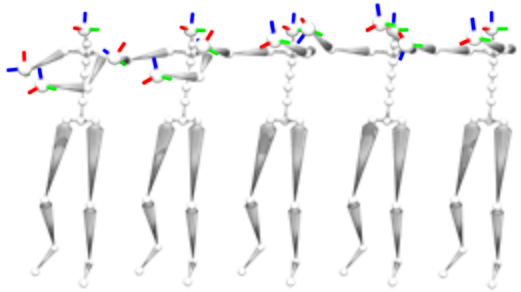
Embodied AI Learning

Thank You! Q&A

Scene

Object

Human



Ego Motion