NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

From Multimodal Generative Models to Unified World Modeling

Ziwei Liu 刘子纬

Nanyang Technological University

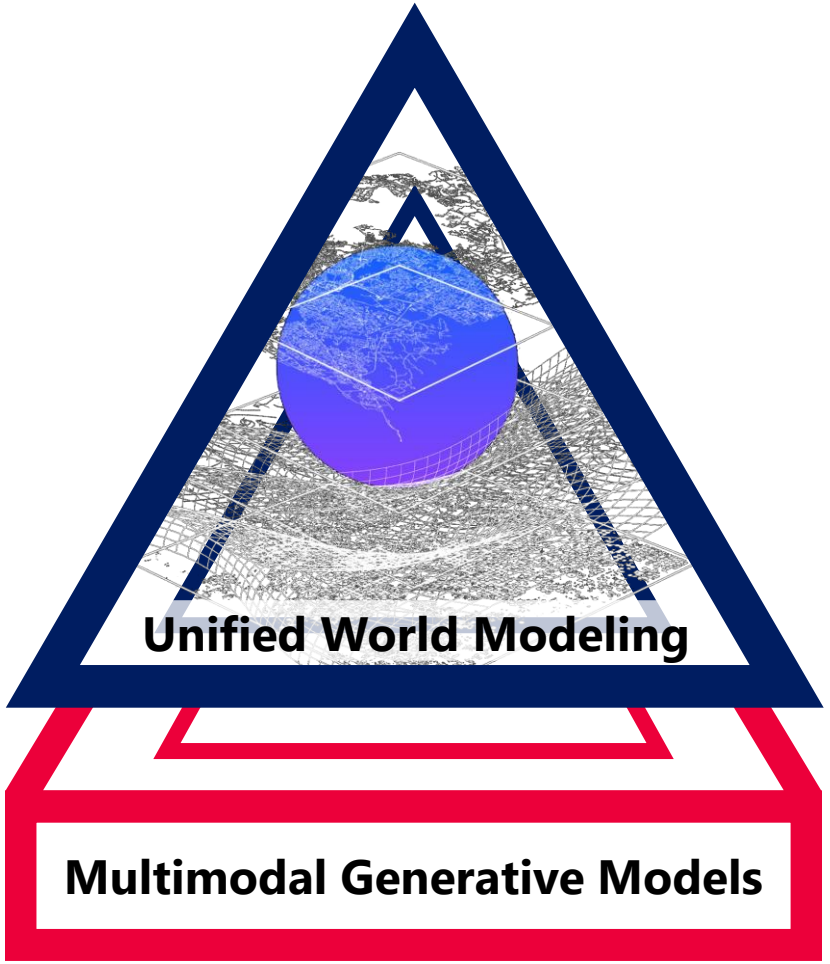
https://liuziwei7.github.io

Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model Dynamic Scenes

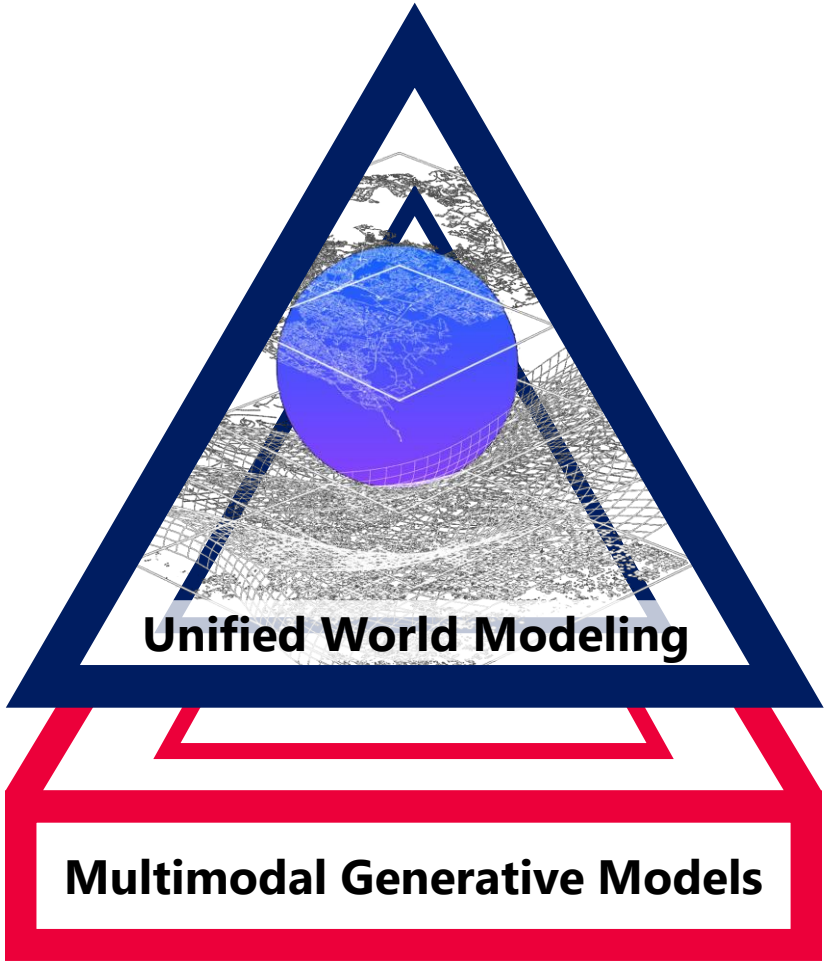


Be Actionable

How to Interact with the Physical World

Be Physical

How to Model Material and Illumination



Be Dynamic

How to Model Dynamic Scenes

Be Actionable

How to Interact with the Physical World

Be Physical: PhysX-Anything

Ziangcao0312/PhysX-Anything

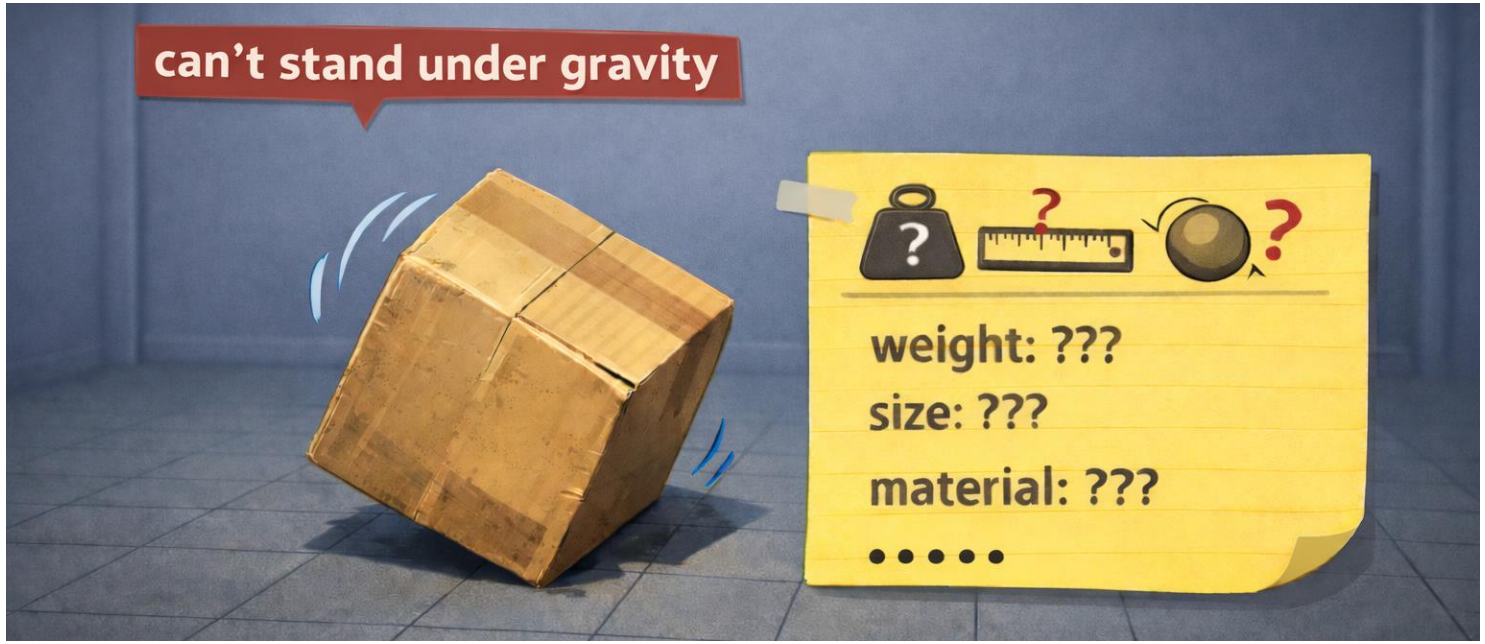
PhysX-Anything: Simulation-Ready Physical 3D Assets from Single Image

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, Ziwei Liu

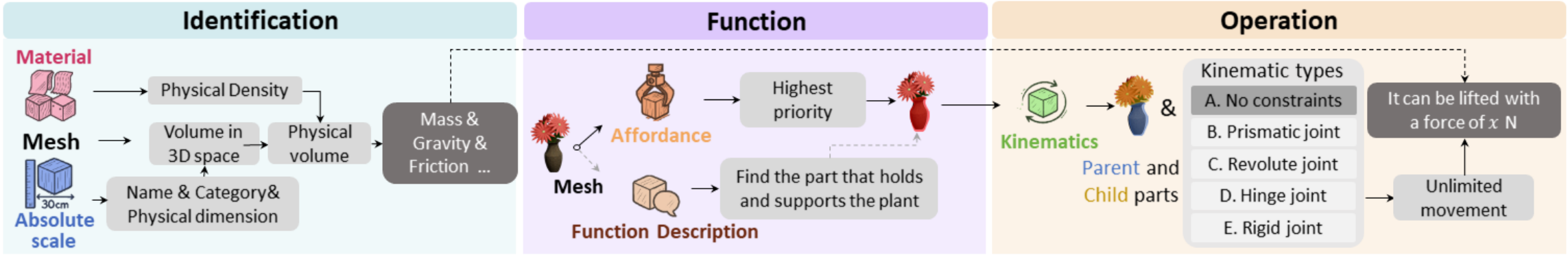
CVPR 2026

Challenges

- Modelling of Physical Attributes
  - Existing 3D generation primarily emphasizes geometries and textures while neglecting **physical-grounded modeling**, hampering their real-world application in physical domains like simulation and embodied AI.



Definition of Physical Properties



Identification – Determining the basic nature of the object

Absolute scaling and material (Young's modulus, Poisson's ratio, and density)

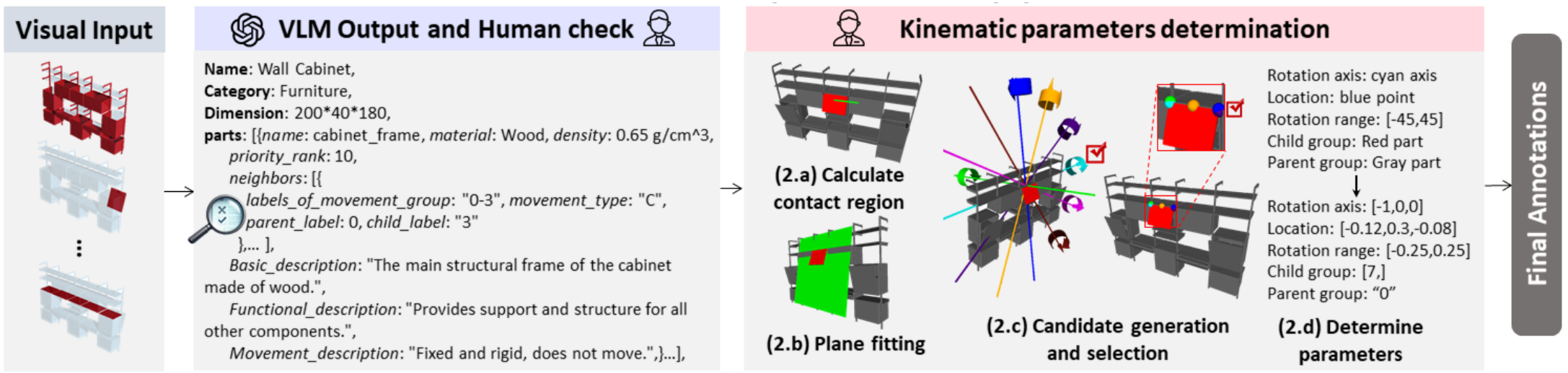
Function – Understanding its potential applications

Functional affordance analysis and function descriptions

Operation – Detailed usage methodologies

Kinematic Parameter

Human-in-the-loop Annotation Pipeline



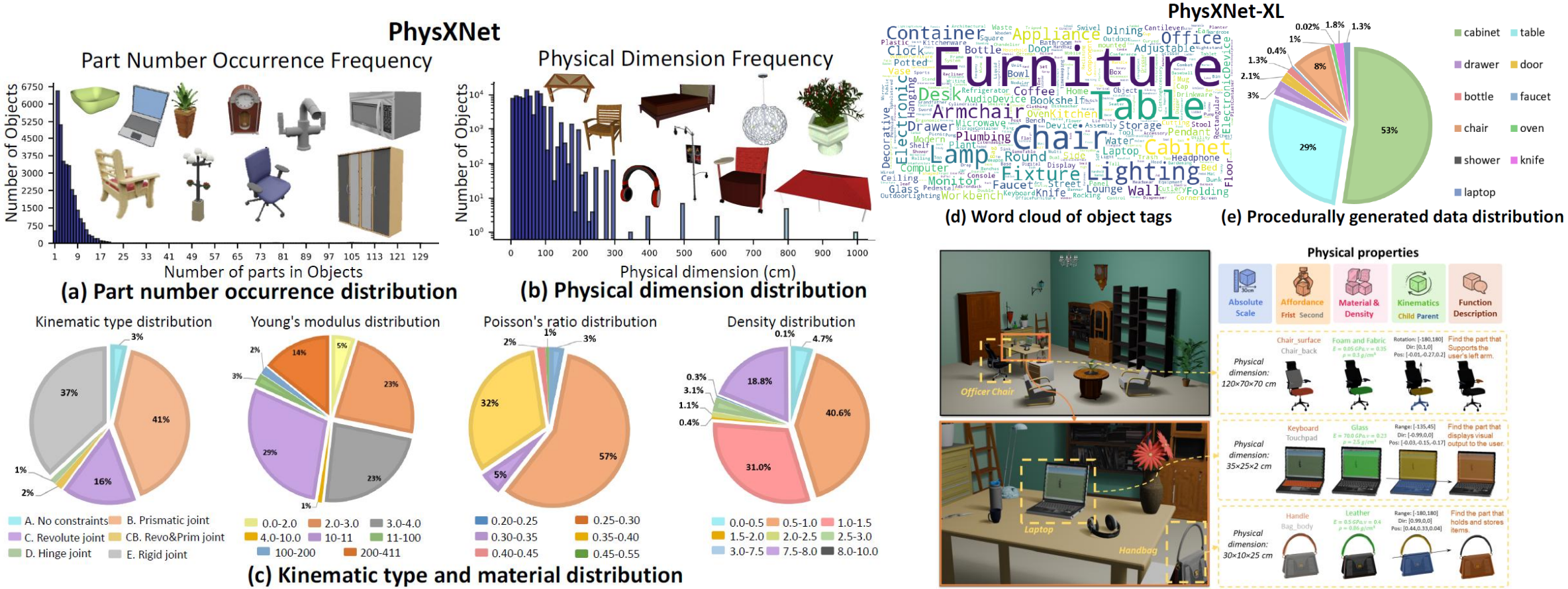
Preliminary Data Acquisition

GPT-4o → Human Check

Kinematic Parameter Determination

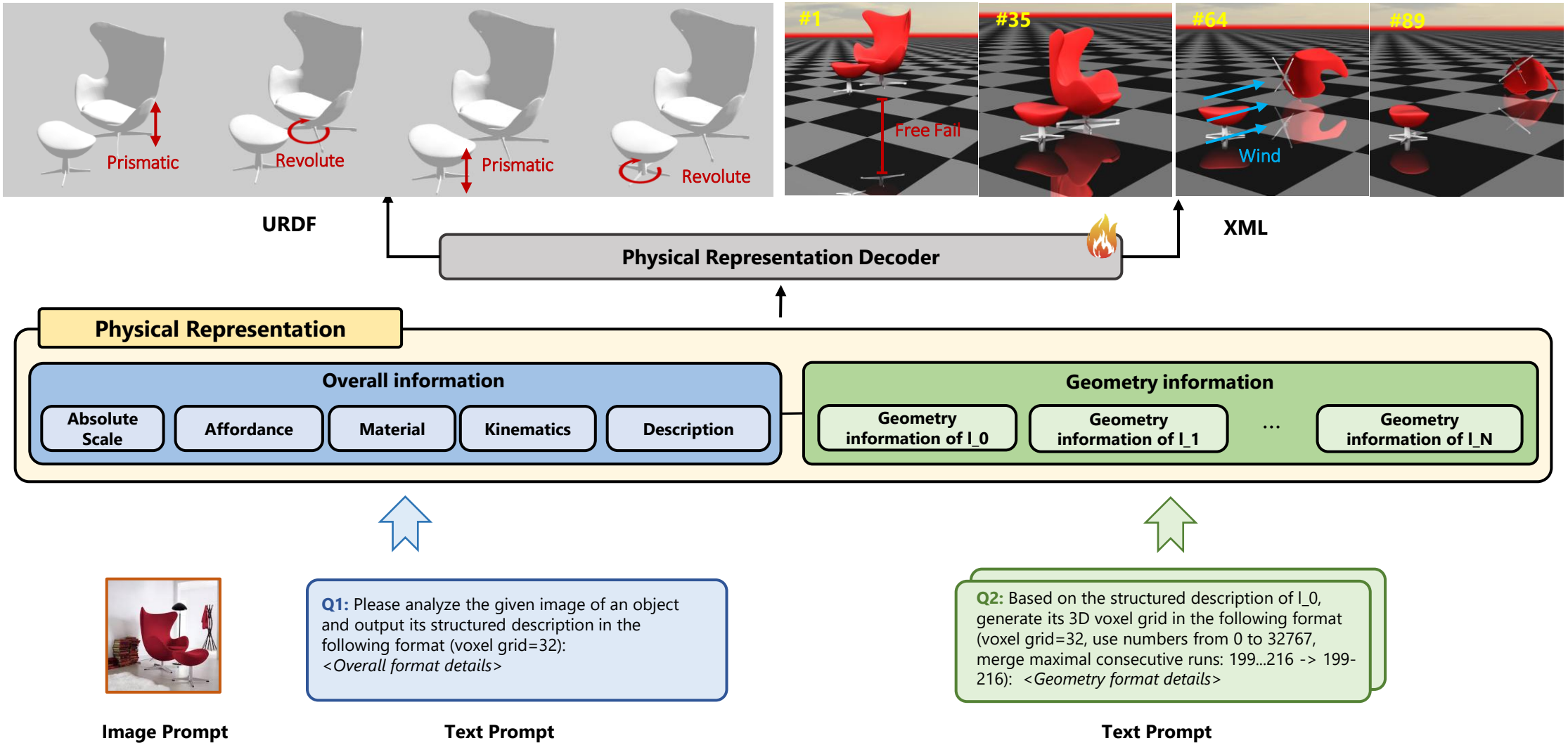
Contact Region → Plane Fitting → Candidate Generation and Selection → Kinematic Parameter Determination

PhysXNet & PhysXNet-XL

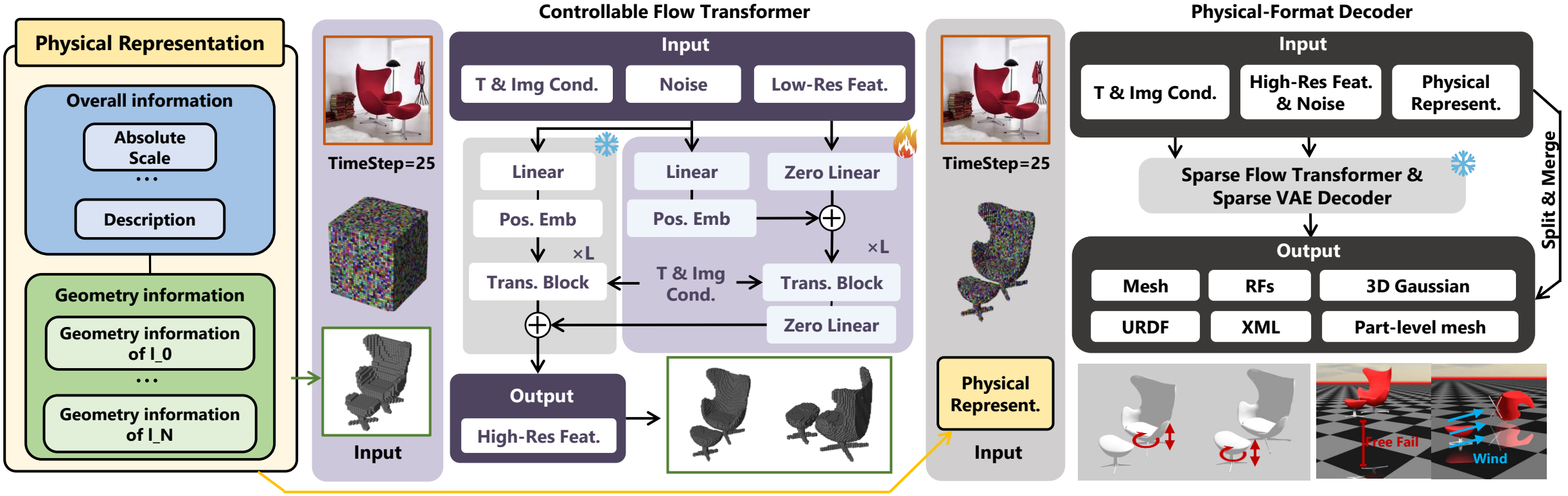


We propose PhysXNet and PhysXNet-XL – the first comprehensive physical 3D dataset with over 26K and 6M richly annotated 3D objects

Key Idea: Interactive 3D Object as Physical Code

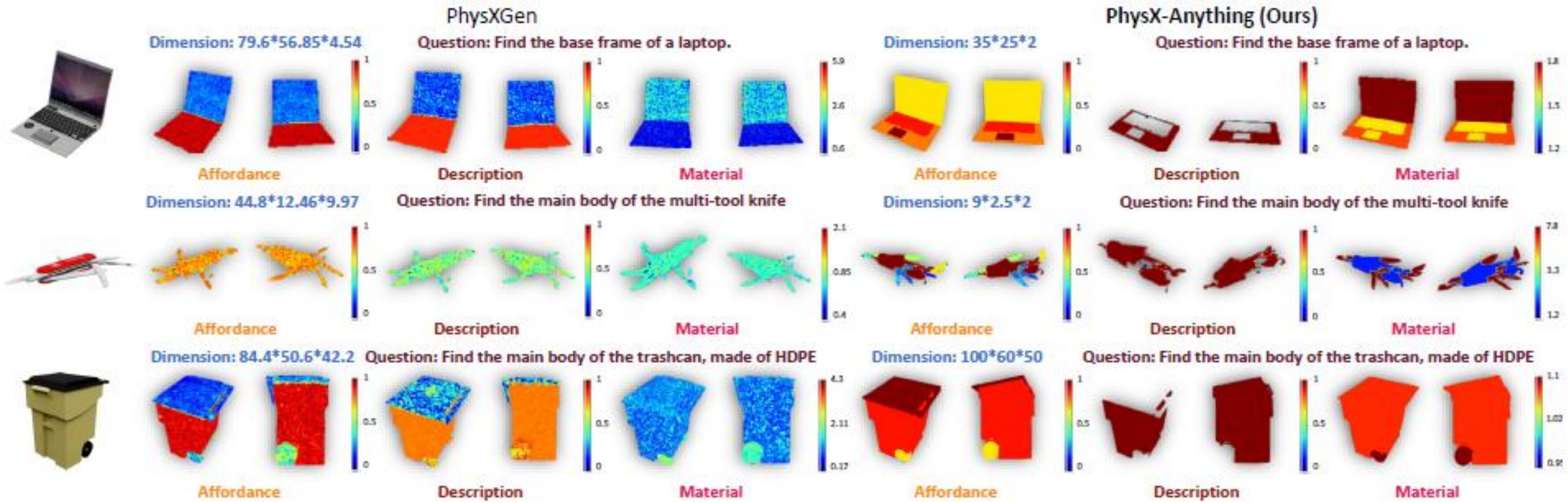


Method: Physical Representation Decoder



Results: Qualitative Comparisons

Results: Qualitative Comparisons



Results: Quantitative Comparisons

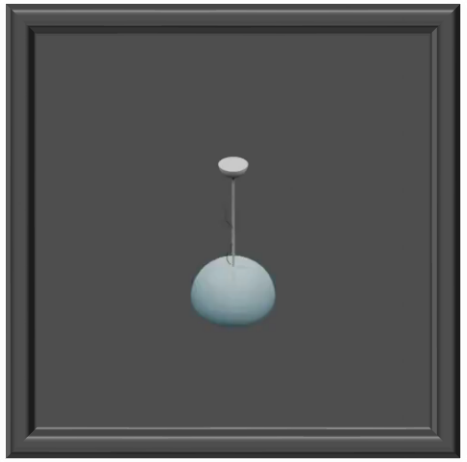
Evaluations on PhysX-Mobility

Methods	Geometry			Physical Attributes				
	PSNR ↑	CD ↓	F-score ↑	Absolute scale ↓	Material ↑	Affordance ↑	Kinematic parameters (VLM) ↑	Description ↑
URDFormer [11]	7.97	48.44	43.81	–	–	–	0.31	–
Articulate-Anything [16]	16.90	17.01	67.35	–	–	–	0.65	–
PhysXGen [3]	20.33	14.55	76.3	43.44	6.29	9.75	0.71	12.89
PhysX-Anything (Ours)	**20.35**	**14.43**	**77.50**	**0.30**	**17.52**	**14.28**	**0.83**	**19.36**

In-the-wild Evaluations

Methods	Geometry (Human) ↑	Physical Attributes (Human)				
		Absolute scale ↑	Material ↑	Affordance ↑	Kinematic parameters ↑	Description ↑
URDFormer [11]	0.21	–	–	–	0.23	–
Articulate-Anything [16]	0.53	–	–	–	0.37	–
PhysXGen [3]	0.61	0.48	0.43	0.34	0.32	0.33
PhysX-Anything (Ours)	**0.98**	**0.95**	**0.84**	**0.94**	**0.98**	**0.96**

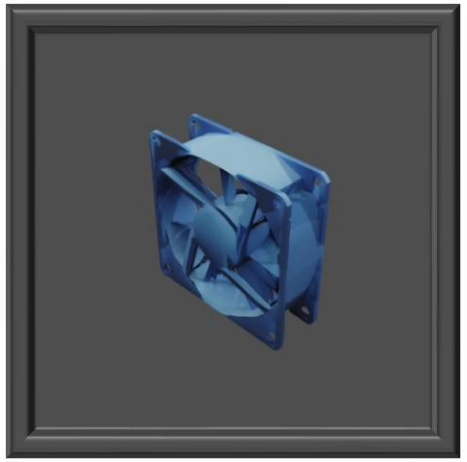
Results: Simulated Objects



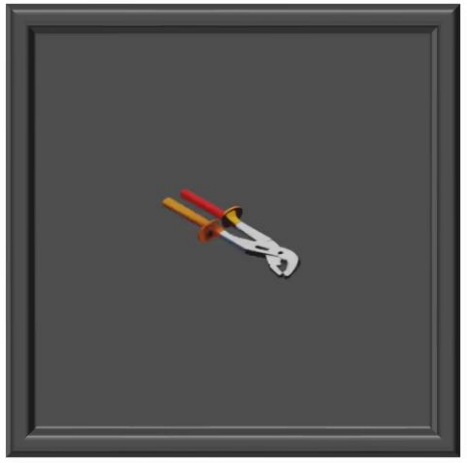


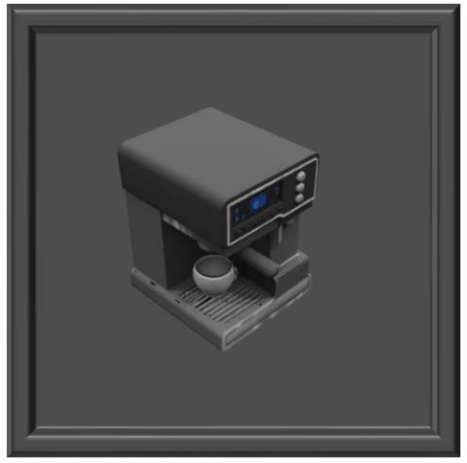



Results: Quantitative Comparisons



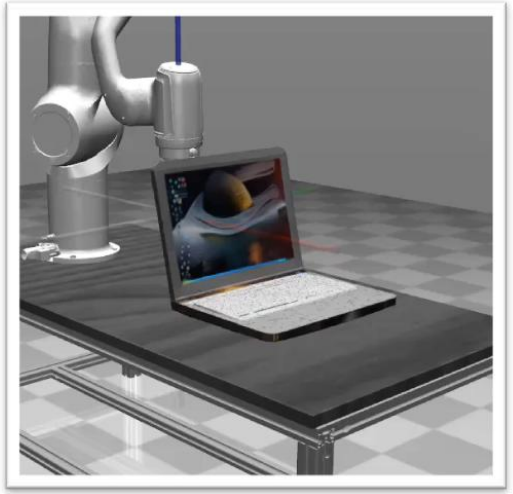




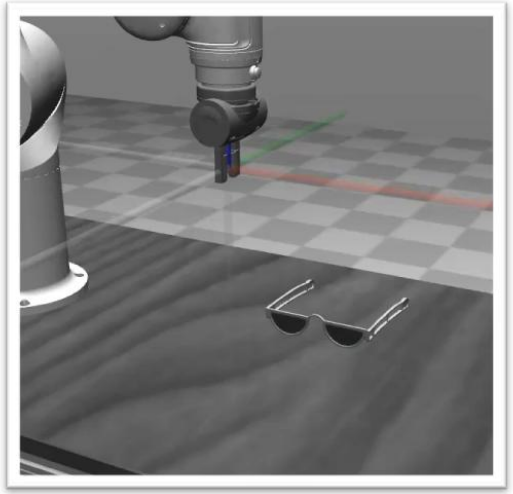



Results: Simulated Interactions

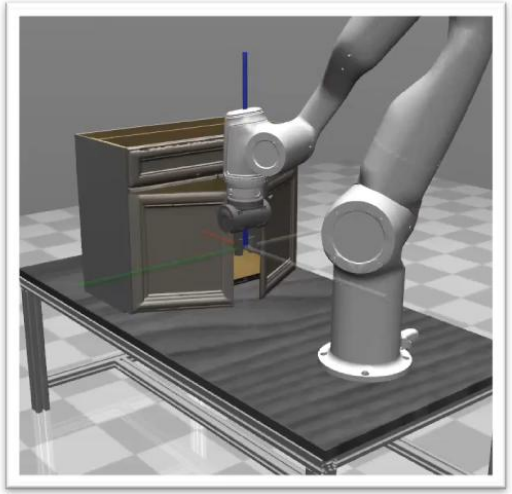
Laptop Closing



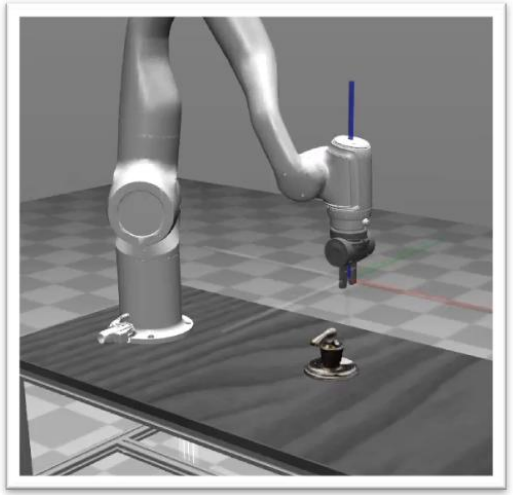
Eyeglass Temple Folding



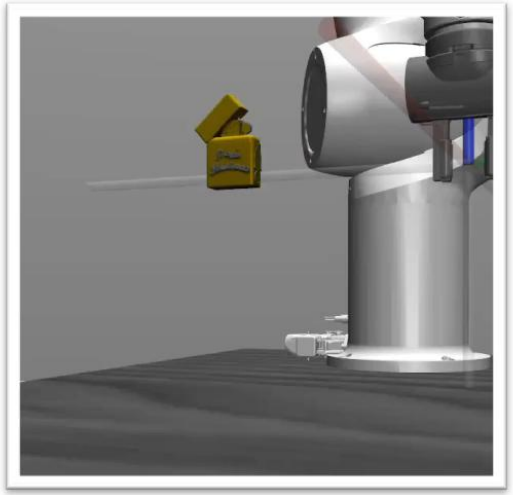
Door Opening and Closing



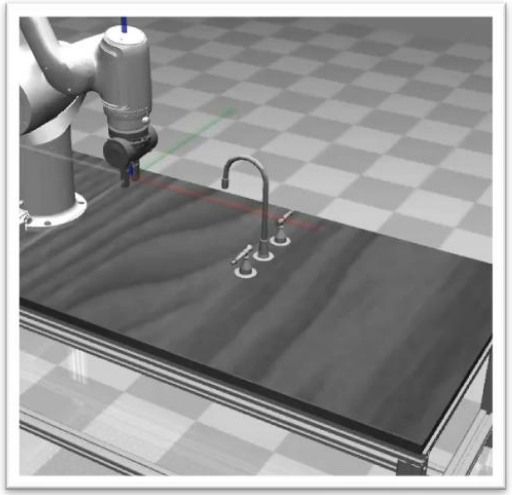
Handle Manipulation



Lighter Snapping Open



Faucet Switch Manipulation



Be Physical: IGGT

lifuguan/IGGT_official

IGGT: Instance-Grounded Geometry Transformer for Semantic 3D Reconstruction

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, and Ziwei Liu

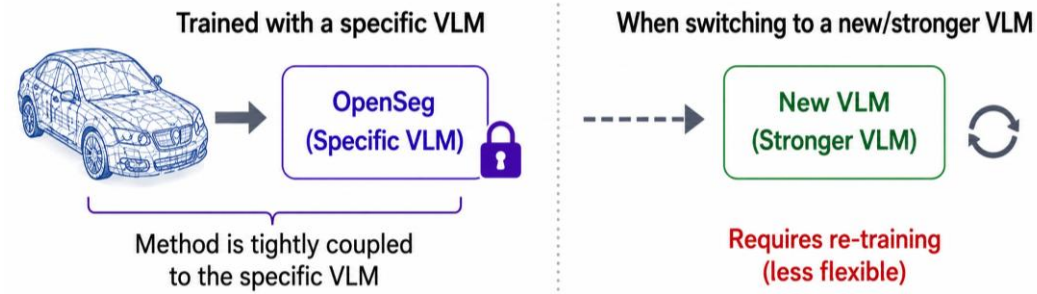
ICLR 2026

Challenges

Model-side Challenges

1 Language Alignment Hurts Geometry

2 Tight Coupling with Specific VLMs



Application-side Challenges

Methods	2D Obj. / Sem. Segmentation	3D Obj. / Sem. Segmentation	2D / 3D Obj. Tracking	3D Grounding
OpenSeg	✓	–	–	–
Feat-3DGS	–	✓	–	–
LSM	✓	✓	–	–
SAM2-Track	–	–	✓	–
IGGT (Ours)	✓	✓	✓	✓

No unified method supports all four applications.

Core Challenge

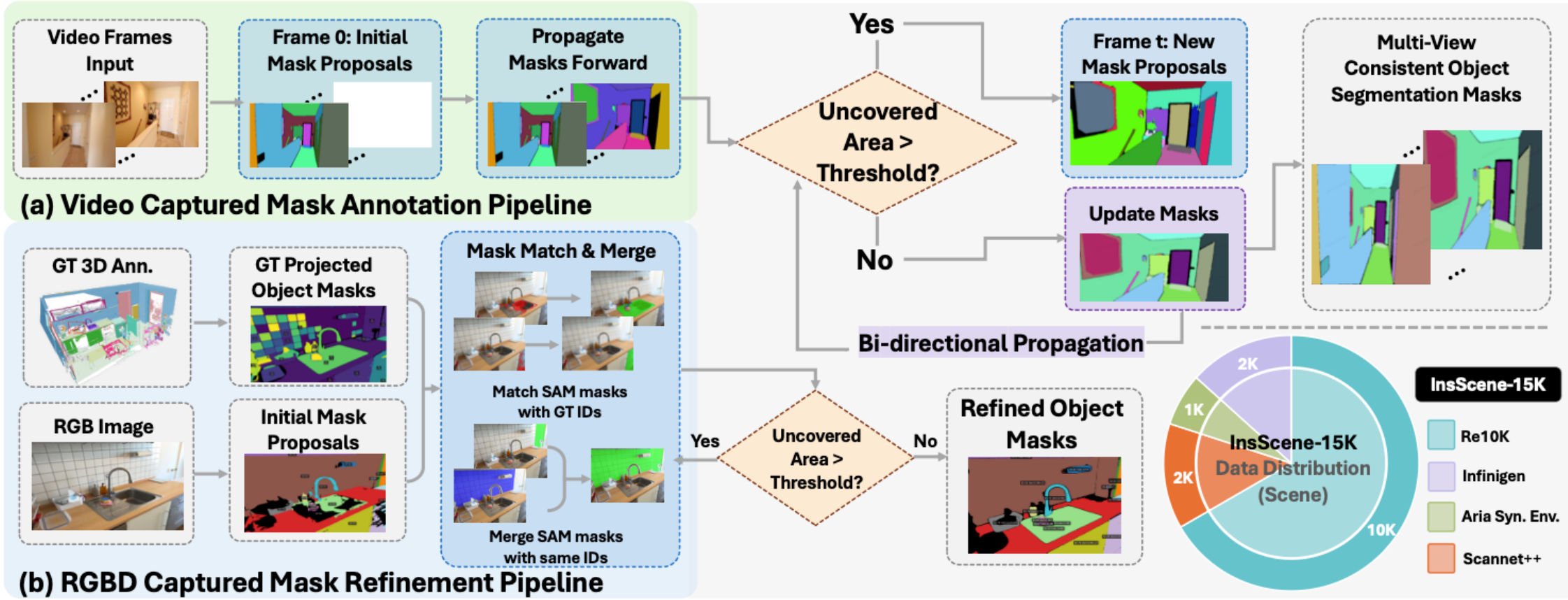
How can we build a unified 3D representation that preserves geometry, is not tied to a single VLM, and supports all four downstream applications?

Key Idea: Learning 3D Instance-level Semantics with Geometry

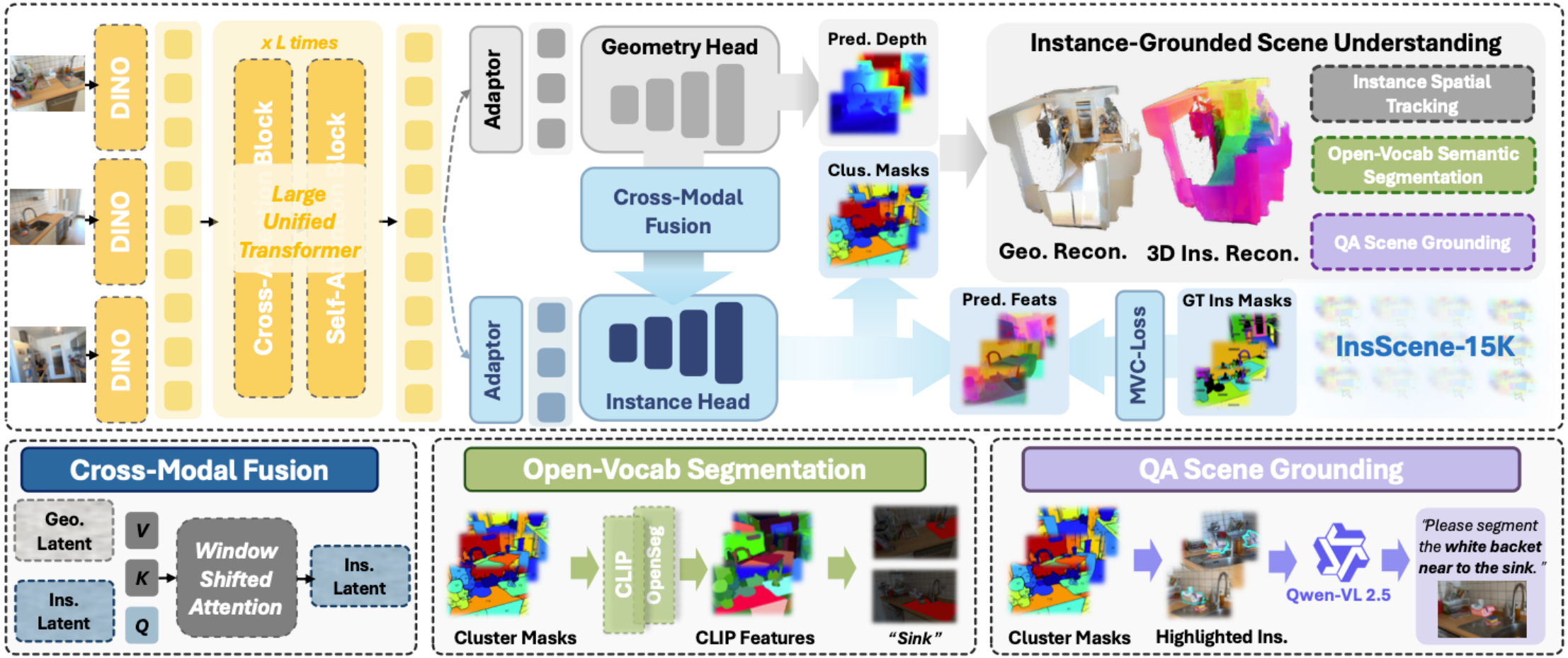
Instance Grounded Geometry Transformer

For Semantic 3D Reconstruction

Step 1 : Scalable Data Curation Pipeline



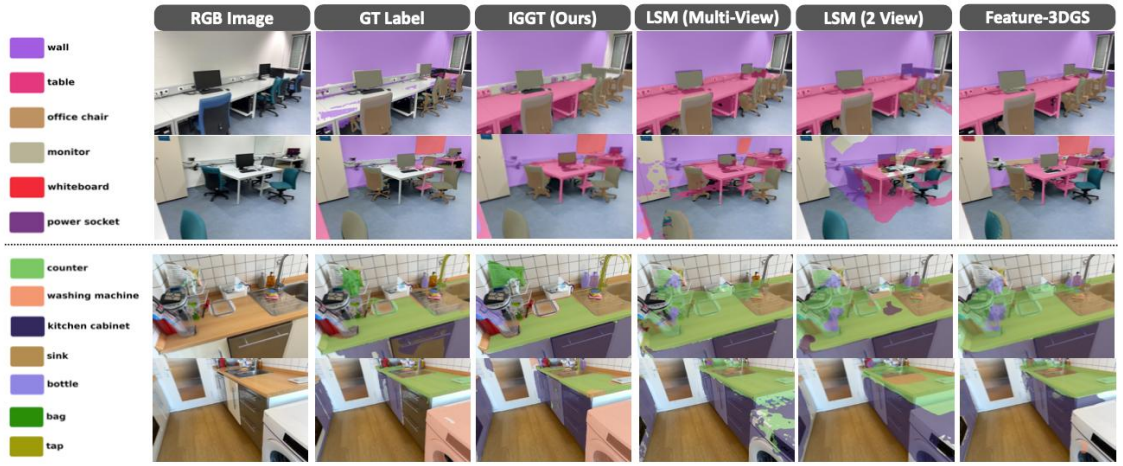
Step II: End-to-End IGGT Training & Inference Procedure



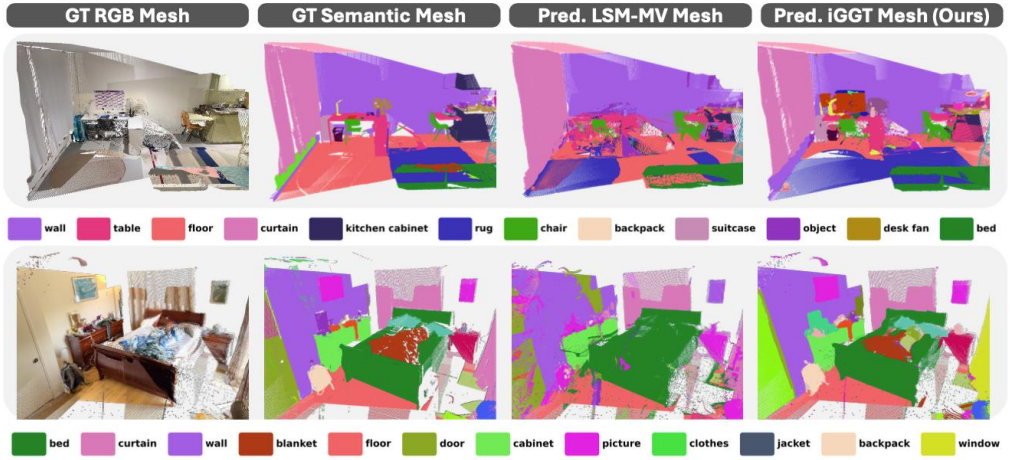
Gallery: Annotated InsScene-15K

Gallery: Inference Results of Our IGGT

Gallery: Board Downstream Applications

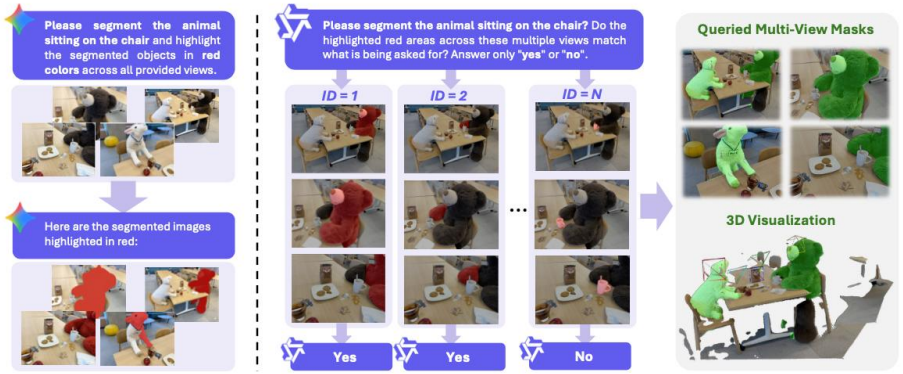


(a) 2D Open-Vocabulary Segmentation



(b) 3D Open-Vocabulary Segmentation

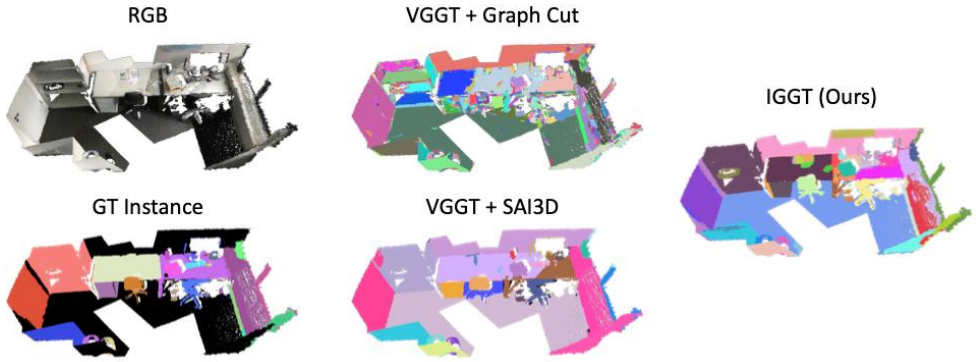
(c) 3D Object Segmentation and Tracking



(a) Vanilla Gemini 2.5 Pro (b) Ours Instance-Grounded Scene Understanding + Qwen-VL 2.5

(d) Boosting VLM Spatial Understanding

Performance: Quantitative Results



Method	AP	AP_{50}	AP_{25}	Time
VGGT + Graph Cut	3.42	9.30	30.86	10.65min
VGGT + SAI3D	14.94	31.06	50.07	12.22min
Ours	12.25	24.93	47.55	2.52min

Class-Agnostic 3D Mask Segmentation Results

Model	Capability			MV Ins. Mat.		Recon. Metric		Open-Vocab. Semantic Segment		
	Recon.	Understand	Mat.	T-mIoU↑	T-SR↑	Abs. Rel↓	τ↑	2D mIoU↑	2D mAcc↑	3D mIoU↑
LSeg	✗	✓	✗	-	-	-	-	58.11	65.76	-
OpenSeg	✗	✓	✗	-	-	-	-	42.33	68.06	-
NeRF-DFF	✓	✓	✗	-	-	7.99	36.53	45.40	65.29	12.29
Feature-3DGS	✓	✓	✗	-	-	6.48	41.63	57.69	63.26	23.42
LSM (2 Views)	✓	✓	✗	-	-	4.22	58.65	53.07	53.86	-
LSM (Multi-Views)	✓	✓	✗	-	-	3.17	64.81	53.40	59.50	35.37
SpaTracker+SAM	✗	✗	✓	26.43	38.57	-	-	-	-	-
SAM2*	✗	✓	✓	53.74	71.25	-	-	-	-	-
VGGT	✓	✗	✗	-	-	**1.84**	83.60	-	-	-
Ours	✓	✓	✓	**69.41**	**98.66**	1.90	**83.71**	**60.46**	**81.84**	**39.68**

Evaluation on ScanNet++ Dataset

Model	MV Ins. Mat.		Recon. Metric		Open-Vocab. Semantic Segment		
	T-mIoU↑	T-SR↑	Abs. Rel↓	τ↑	2D mIoU↑	2D mAcc↑	3D mIoU↑
LSeg	-	-	-	-	22.61	34.42	-
OpenSeg	-	-	-	-	13.92	48.13	-
Feature-3DGS	-	-	5.92	41.64	22.47	33.14	10.59
LSM (2 Views)	-	-	4.22	74.02	17.76	26.95	-
LSM (Multi-Views)	-	-	2.96	83.28	17.88	27.84	15.17
SpaTracker+SAM	16.15	23.68	-	-	-	-	-
SAM2*	44.16	57.89	-	-	-	-	-
VGGT	-	-	2.75	85.41	-	-	-
Ours	**73.02**	**98.90**	**2.61**	**85.66**	**31.31**	**70.78**	**20.14**

Evaluation on ScanNet Dataset

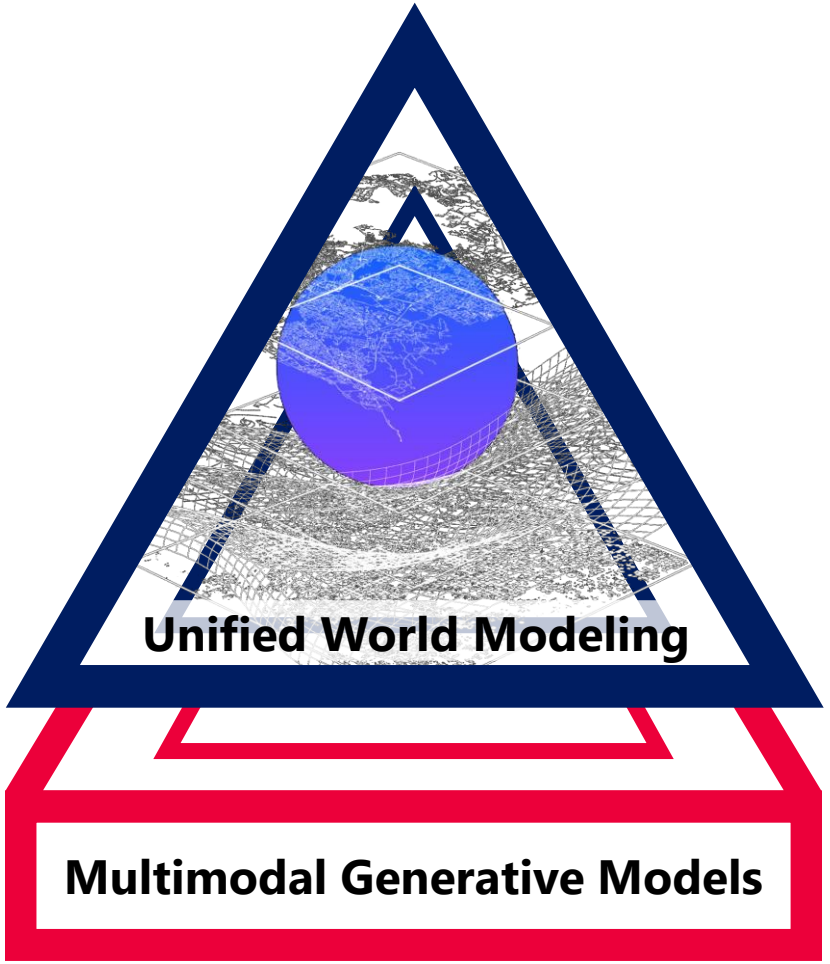
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model Dynamic Scenes



Be Actionable

How to Interact with the Physical World

Be Dynamic: 4DNeX

3DTopia/4DNeX

4DNeX: Feed-Forward 4D Generative Modeling Made Easy

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, Ziwei Liu
arXiv 2508.13154

Challenges

- **Efficiency**
  - From per-scene optimization to feed-forward
- **Data**
  - Scaling 4D data in the wild
- **Representation**
  - Compact & structural
  - Unified RGB-XYZ modeling

Key Idea: Unified RGB-XYZ video diffusion

Input

4DNeX

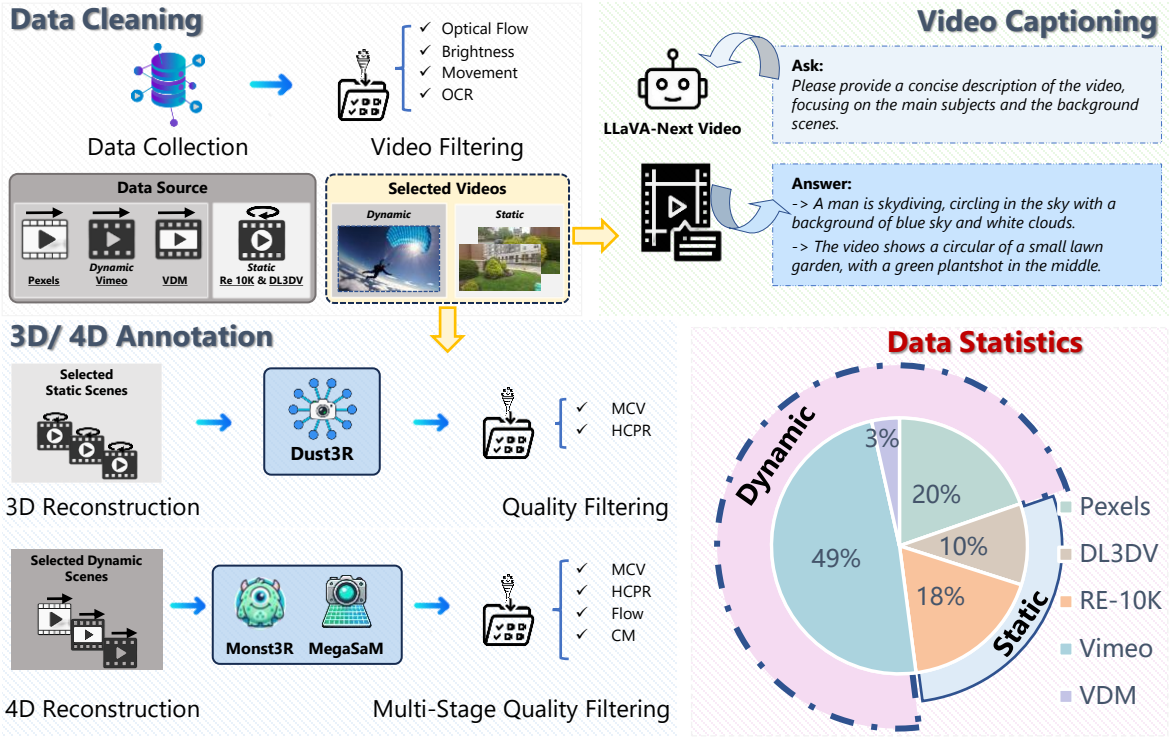
"The video showcases..."

6D (RGB+XYZ) Video

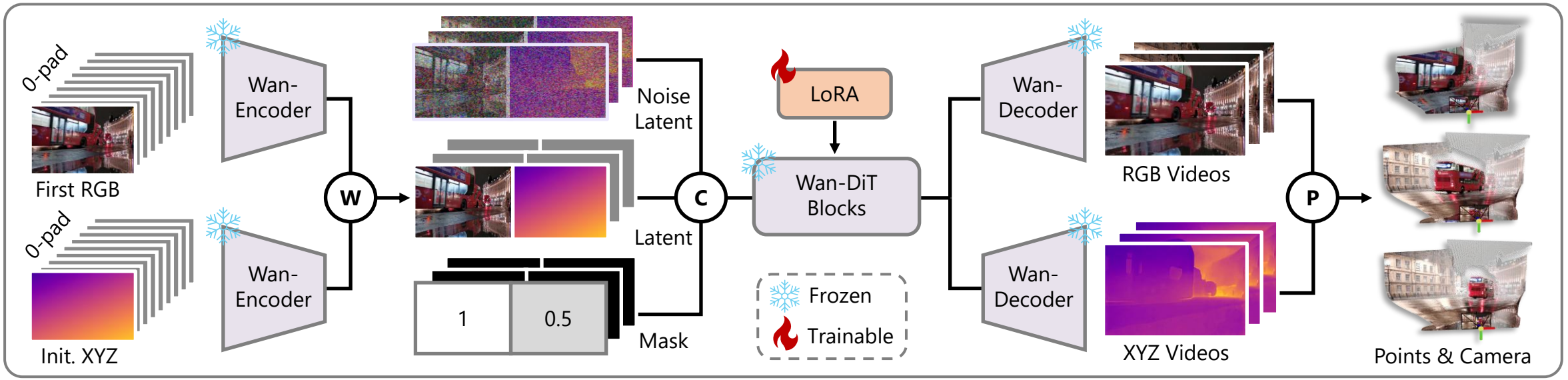
Dynamic Point Cloud

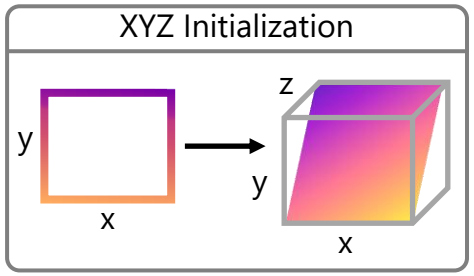
Novel-view Video

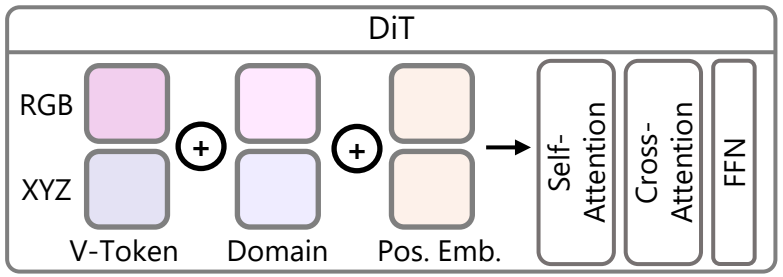
4DNeX-10M: a Large-Scale 4D Dataset

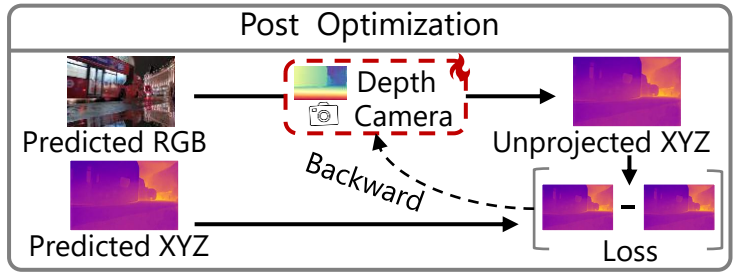


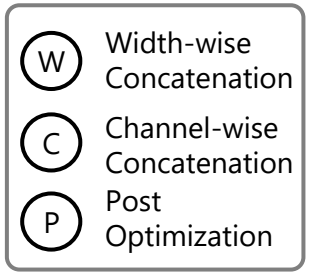
The Proposed Method: 4DNeX



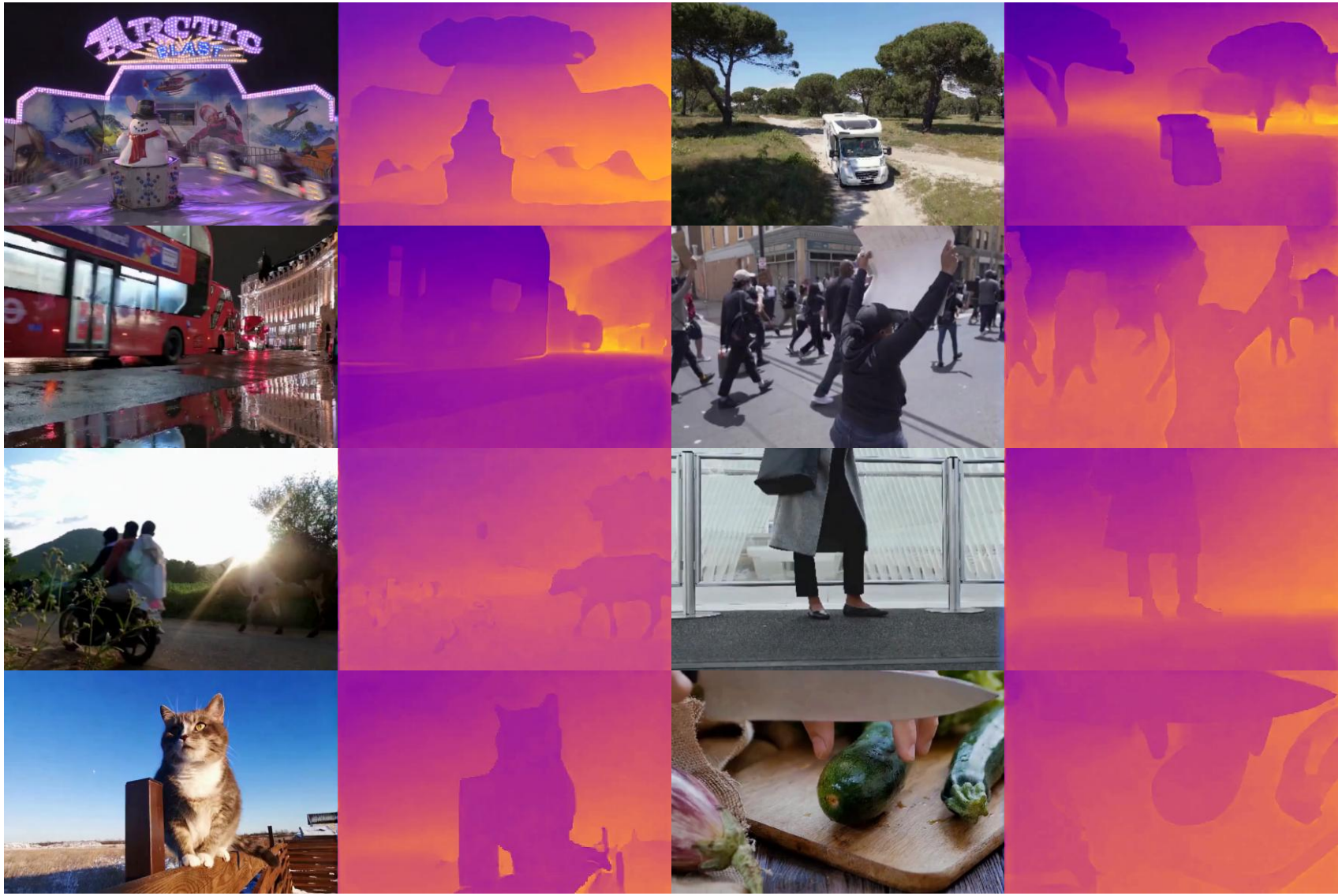




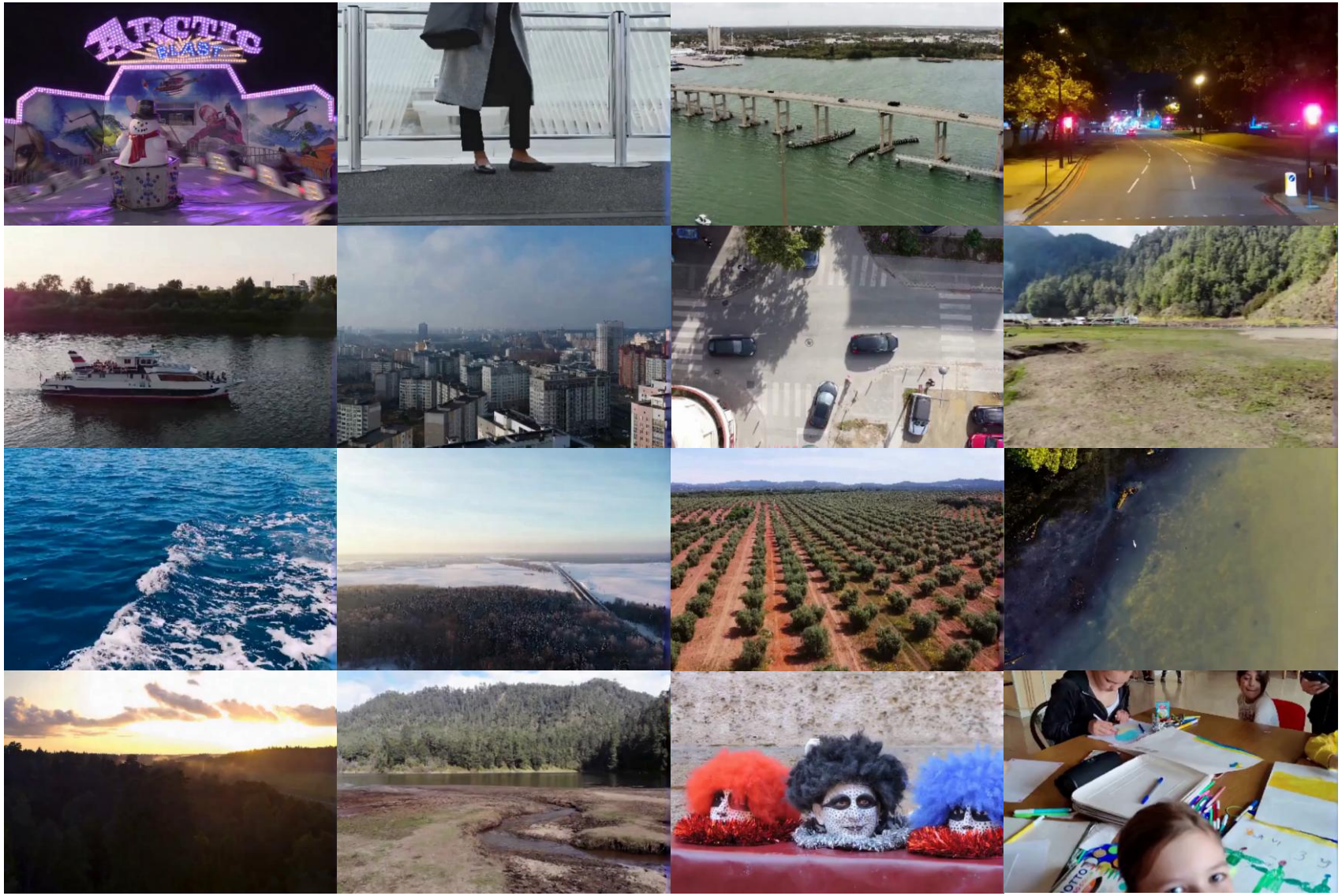




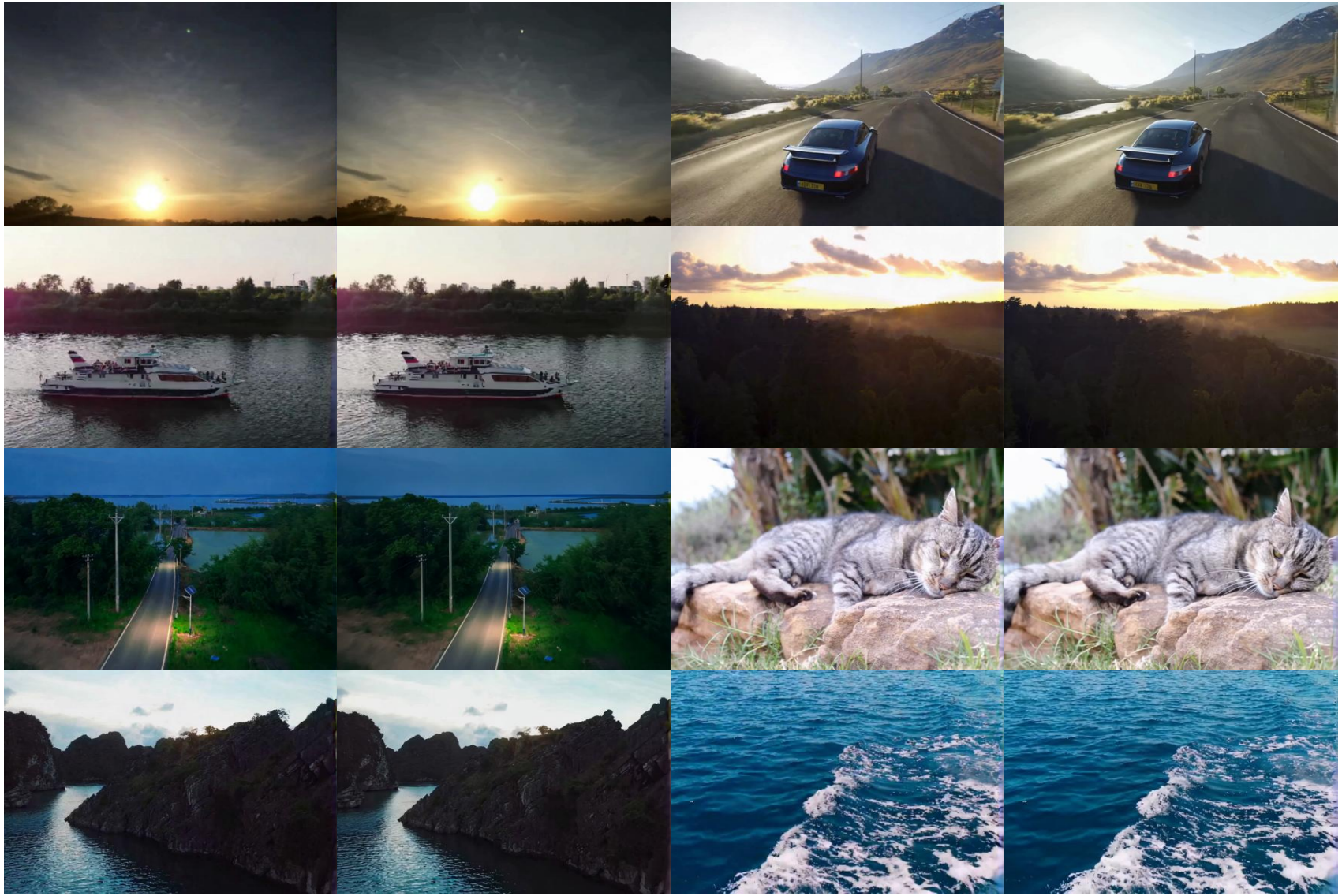
Results: 6D (RGB+XYZ) Video



Results: Dynamic Point Cloud



Results: Novel-View Video



Quantitative Comparisons

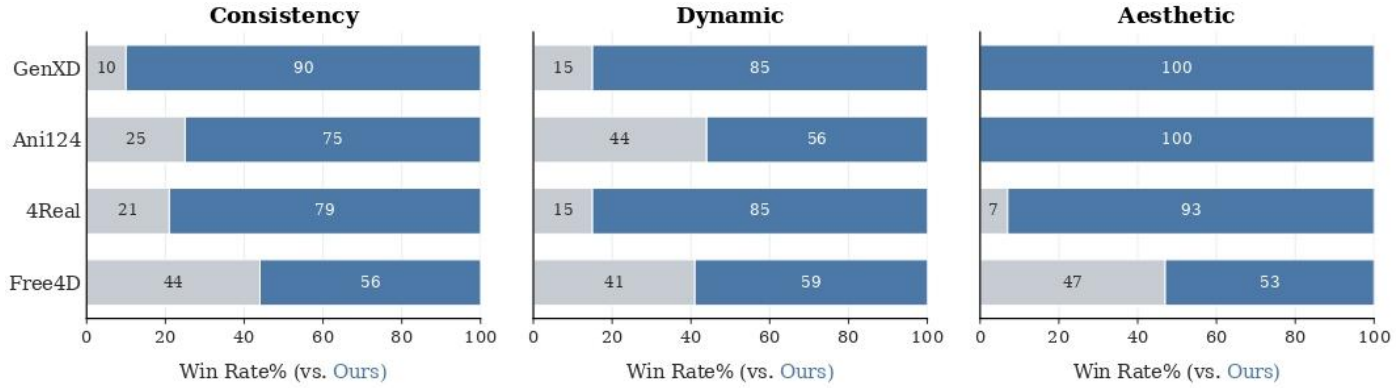
Geometric Evaluation Results (compare with video-based reconstruction methods)

Dataset	Method	Input	Type	AbsRel ↓	LogRMSE ↓	$\delta_{1.25}$ ↑	ATE ↓	RTE ↓	RRE ↓	CD ↓
Sintel [134]	MonST3R [21]	Video	Reconstruction	0.283	0.603	0.610	0.132	0.085	0.099	0.069
	Cut3R [24]	Video	Reconstruction	0.244	0.595	0.735	0.734	1.157	5.983	0.049
	Ours	Single Image	Generation	0.210	0.381	0.719	0.200	0.102	0.334	0.067
Bonn [135]	MonST3R [21]	Video	Reconstruction	0.137	0.274	0.900	0.081	0.033	0.512	0.058
	Cut3R [24]	Video	Reconstruction	0.166	0.353	0.910	0.300	0.490	1.374	0.055
	Ours	Single Image	Generation	0.172	0.245	0.763	0.219	0.096	0.665	0.062
DAVIS [136]	MonST3R [21]	Video	Reconstruction	0.365	0.707	0.652	0.162	0.165	0.172	0.068
	Cut3R [24]	Video	Reconstruction	0.352	0.842	0.647	0.896	1.554	2.887	0.066
	Ours	Single Image	Generation	0.135	0.227	0.832	0.274	0.106	0.304	0.061

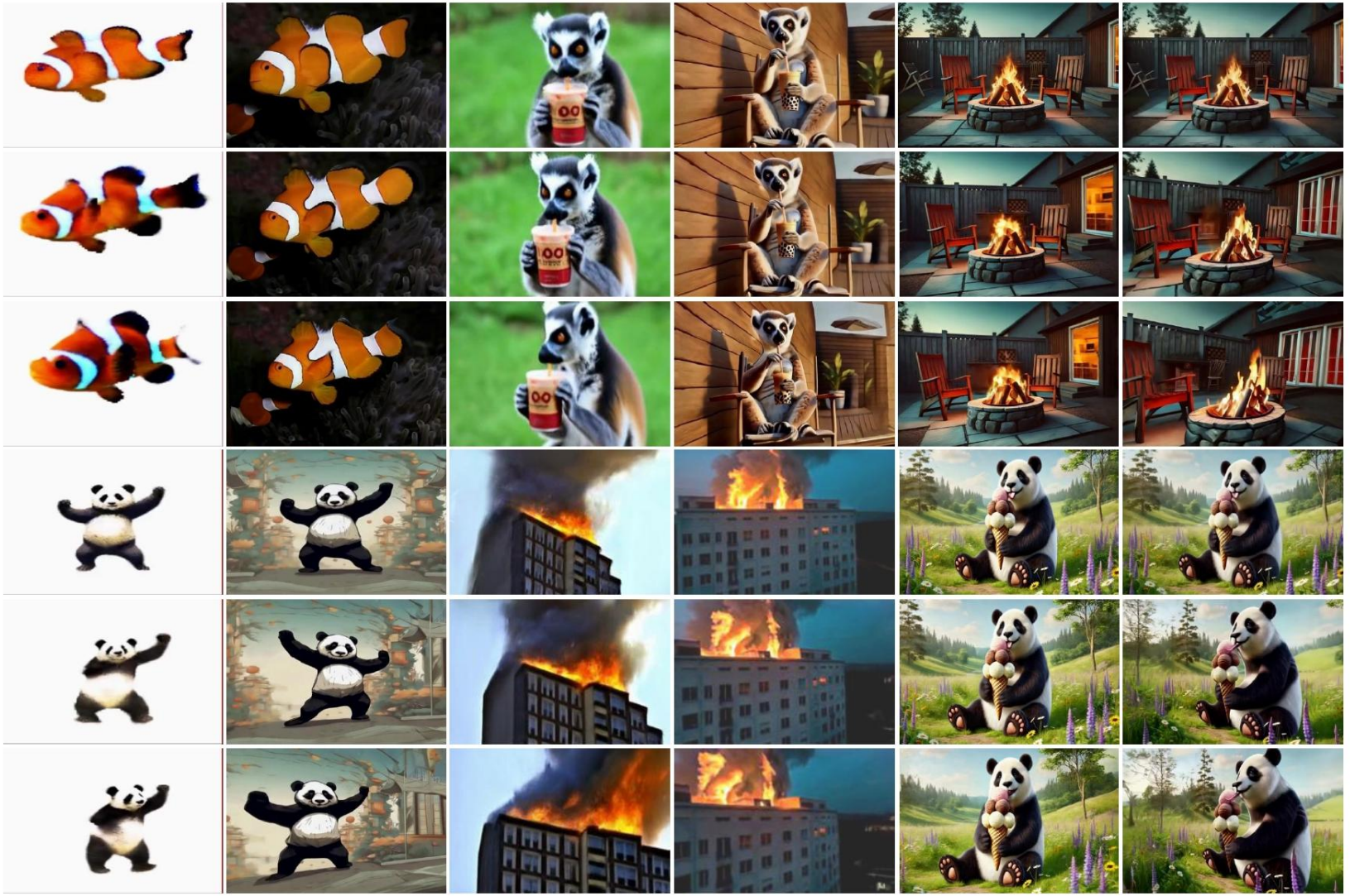
4D Generation Results on VBench

Method	Consistency↑	Dynamic↑	Aesthetic↑	Time↓
4Real [14]	95.7%	32.3%	50.9%	90min
Free4D [13]	96.0%	47.4%	64.7%	60min
Ours	96.4%	58.0%	59.5%	15min
Animate124 [17]	90.7%	45.4%	42.3%	\
Free4D [13]	96.9%	40.1%	60.5%	60min
Ours	97.2%	58.3%	53.0%	15min
GenXD [10]	89.8%	98.3%	38.0%	\
Free4D [13]	96.8%	100.0%	57.9%	60min
Ours	96.8%	100.0%	52.4%	15min

User Study



Qualitative Comparisons



Be Dynamic: Light-X

TQTQliu/Light-X

Light-X: Generative 4D Video Rendering with Camera and Illumination Control

Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, Ziwei Liu

arXiv 2512.05115

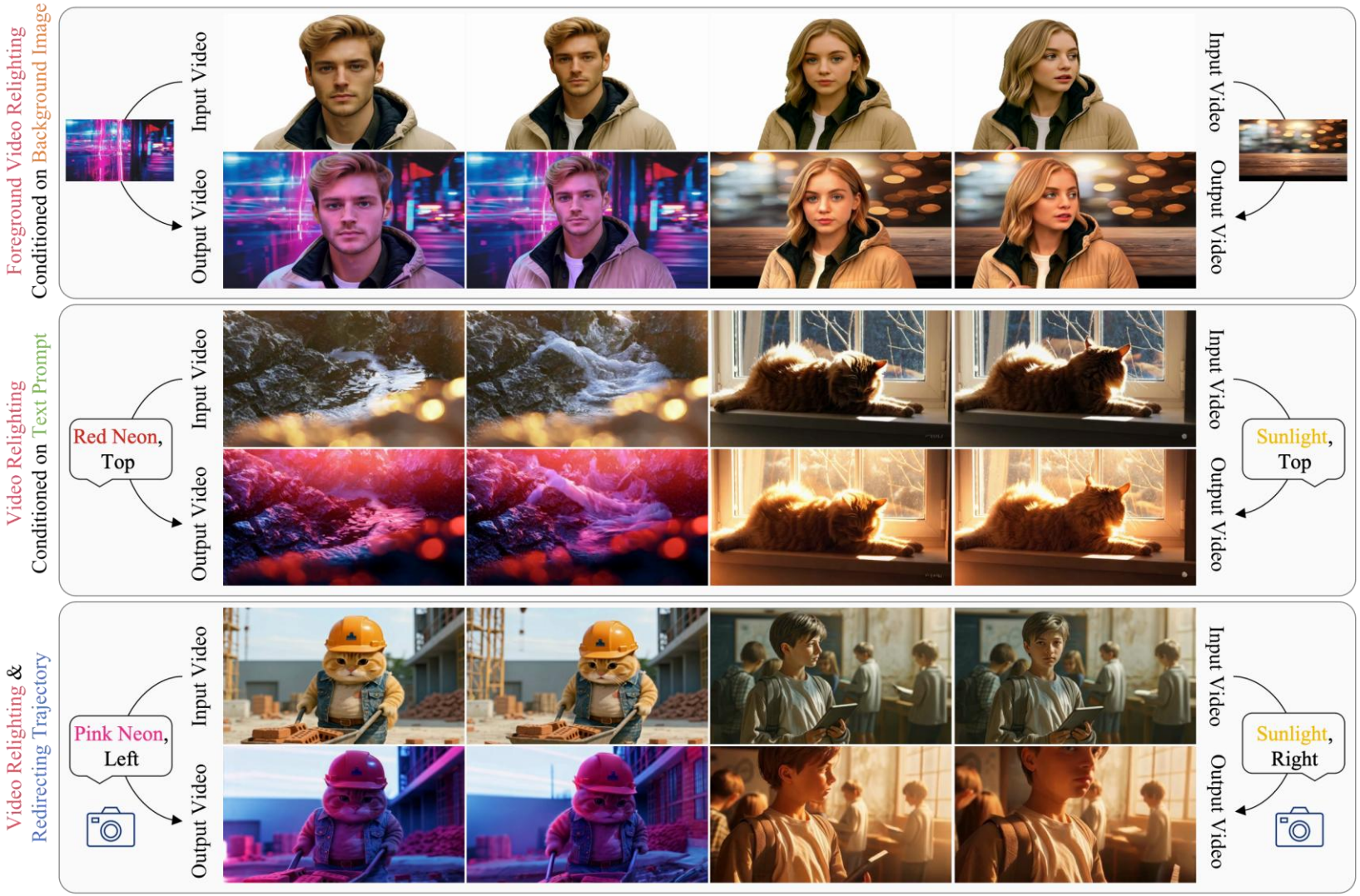
Challenges

- **Controllability**
  - Multi-dimension: camera, illumination & scene dynamics
  - Diverse lighting modality (text / HDR / background / reference)

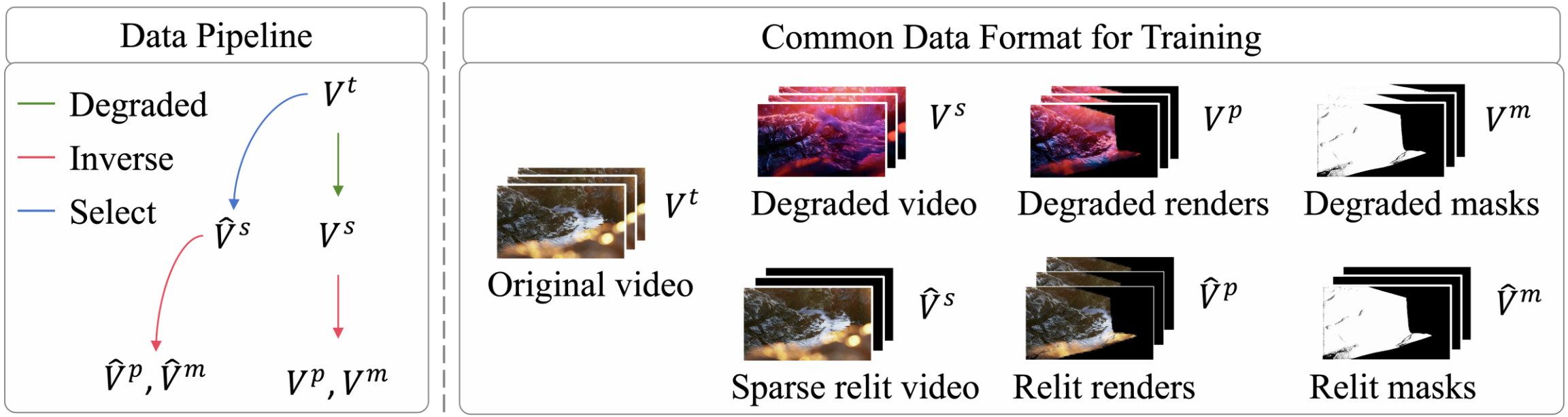
- **Data**
  - No paired multi-view × multi-illumination videos
  - Infeasible to capture in the wild

- **Dual nature**
  - deterministic geometry transformation vs. generative appearance (lighting & novel-view content)

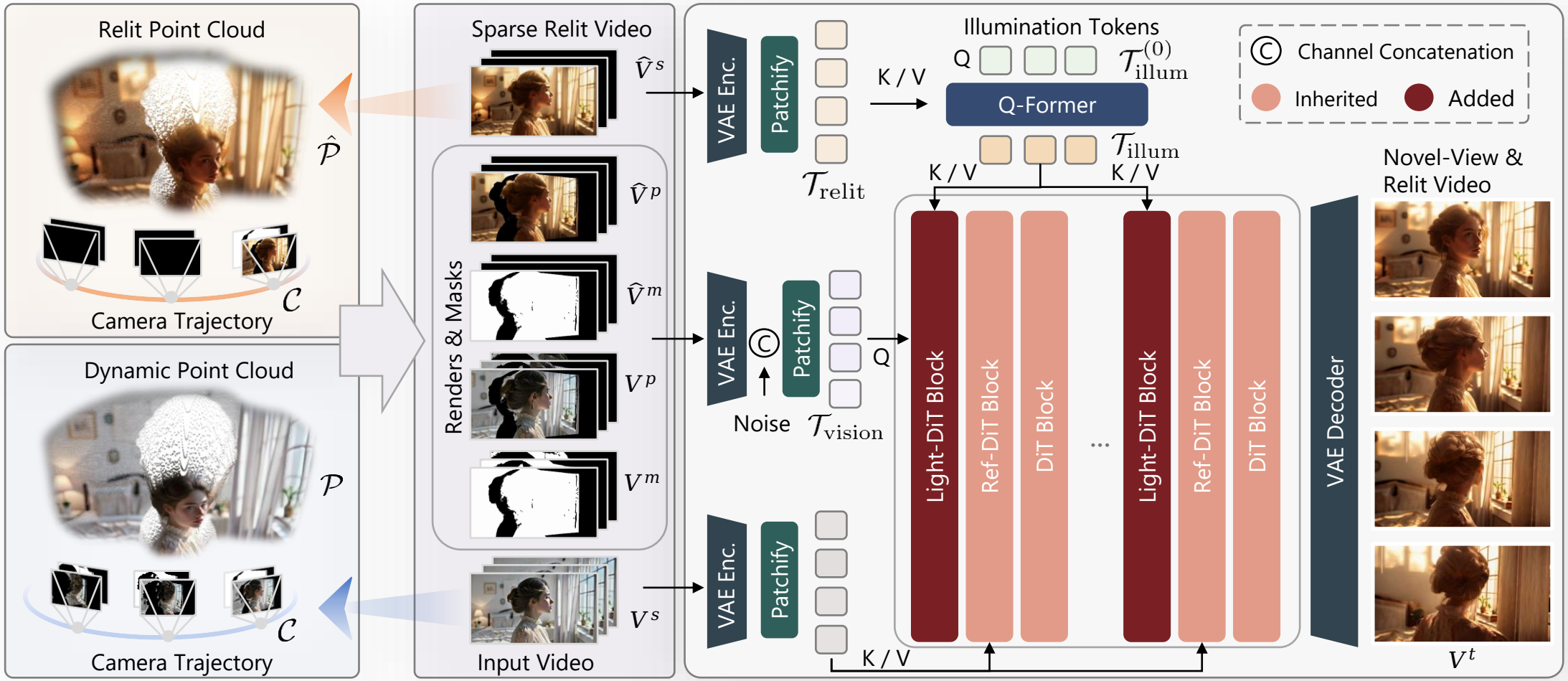
Light-X: Controlling Camera and Illumination in Video



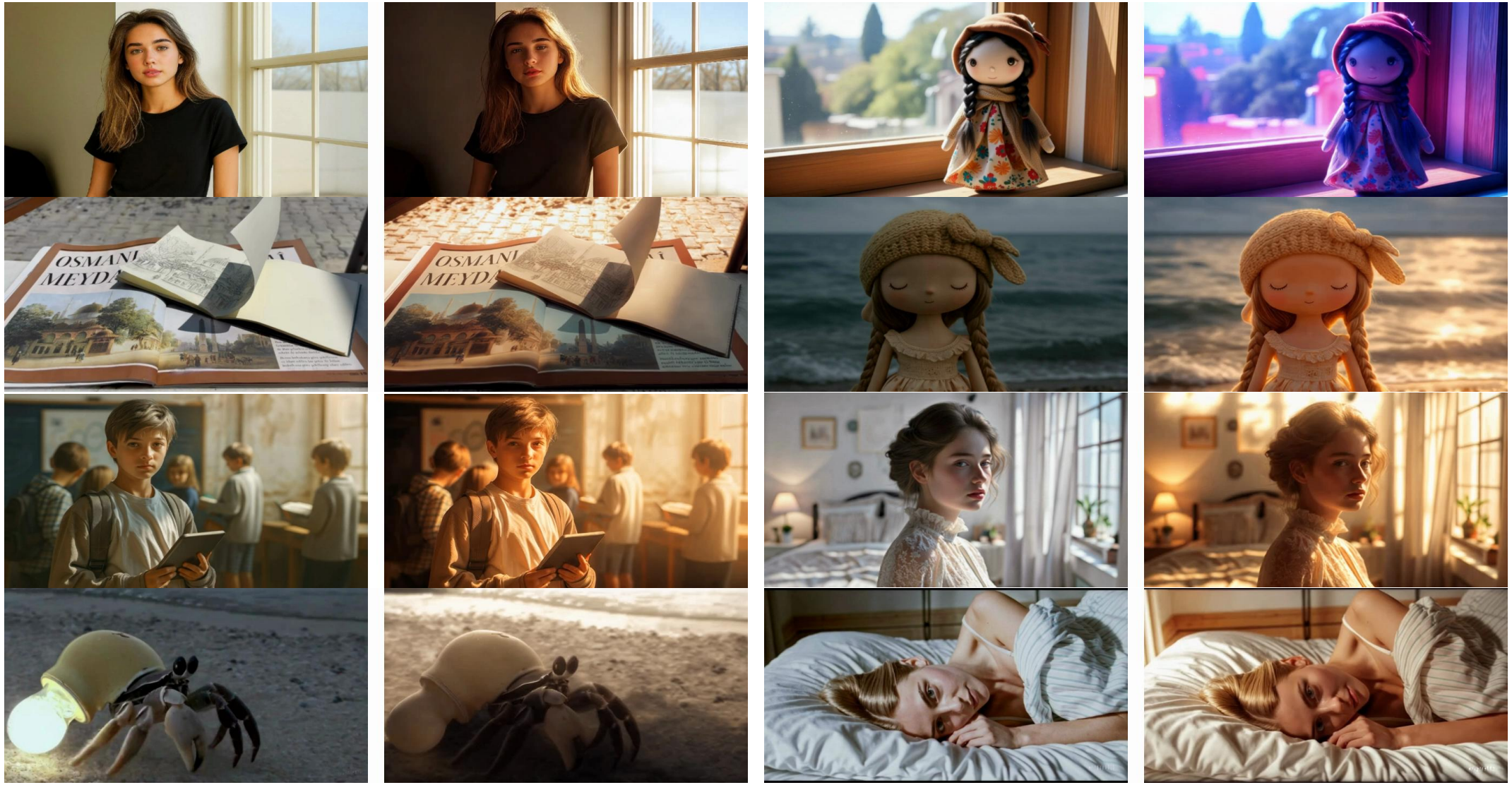
Light-Syn: Degradation-Based Data Synthesis



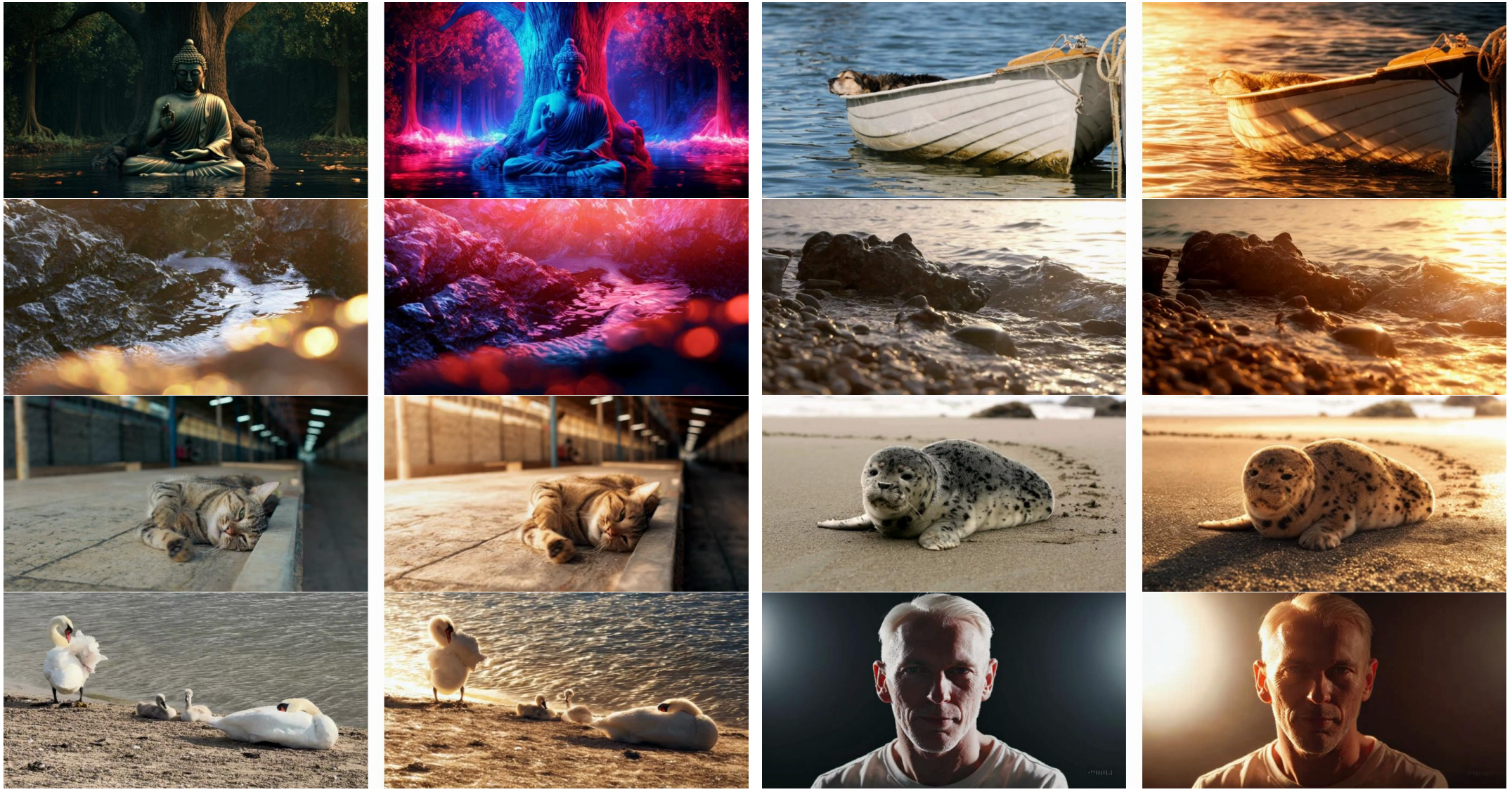
Light-X: Disentangled Camera–Illumination Diffusion



Results: Joint Camera + Illumination Control



Results: Video Relighting



Results: Diverse Lighting Conditions

Input Video

Lighting condition

Reference image

HDR map

Sunlight

Text

Output Video

Quantitative Comparison

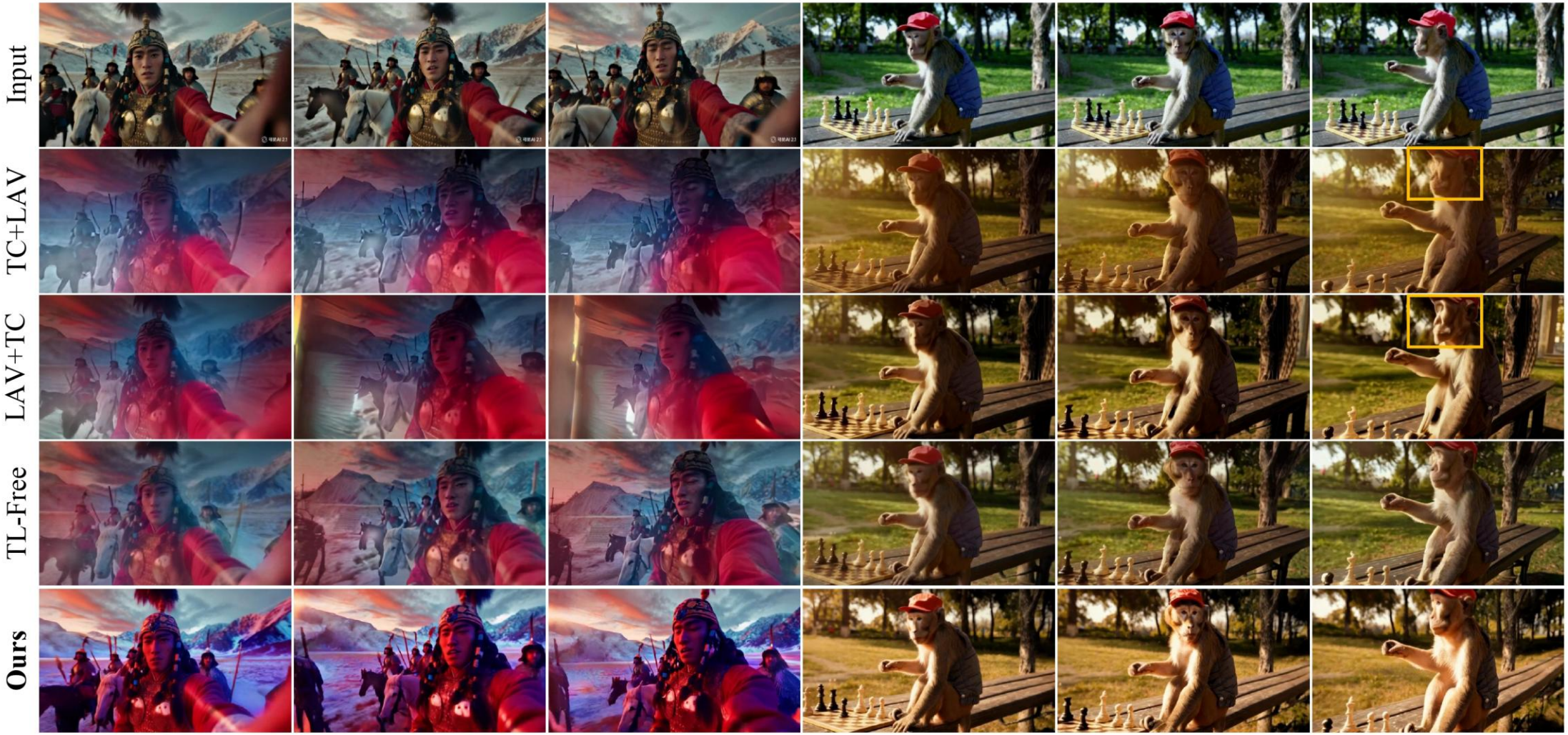
Illumination

Method	Image Quality		Video Smoothness		User Study (%, Ours)				Time ↓
	FID ↓	Aesthetic ↑	Motion Pres. ↓	CLIP ↑	RQ	VS	IP	4DC	
TC+IC-Light	/	0.573	6.558	0.976	89.3	91.7	88.3	88.5	3.25 min
TC+LAV	138.89	0.574	4.327	0.986	86.0	84.4	88.0	89.0	4.33 min
LAV+TC	144.61	0.596	5.027	0.987	85.1	89.3	88.8	87.5	4.33 min
TL-Free	122.73	0.595	3.356	0.987	88.0	89.2	88.2	88.2	5.50 min
Ours	**101.06**	**0.623**	**2.007**	**0.989**	/	/	/	/	1.83 min

Video Relighting

Method	Image Quality		Video Smoothness		User Study (%, Ours)		
	FID ↓	Aesthetic ↑	Motion Preservation ↓	CLIP ↑	RQ	VS	IP
IC-Light	/	0.645	0.374	0.987	81.8	91.7	88.0
Light-A-Video	76.05	0.619	0.296	0.990	85.5	87.1	88.0
Ours	**61.75**	**0.680**	**0.220**	**0.992**	/	/	/
RelightVid	86.94	0.635	0.230	0.988	81.8	87.1	87.3
Ours*	**56.60**	**0.682**	**0.199**	**0.990**	/	/	/

Qualitative Comparison



Applications

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Autonomous Driving

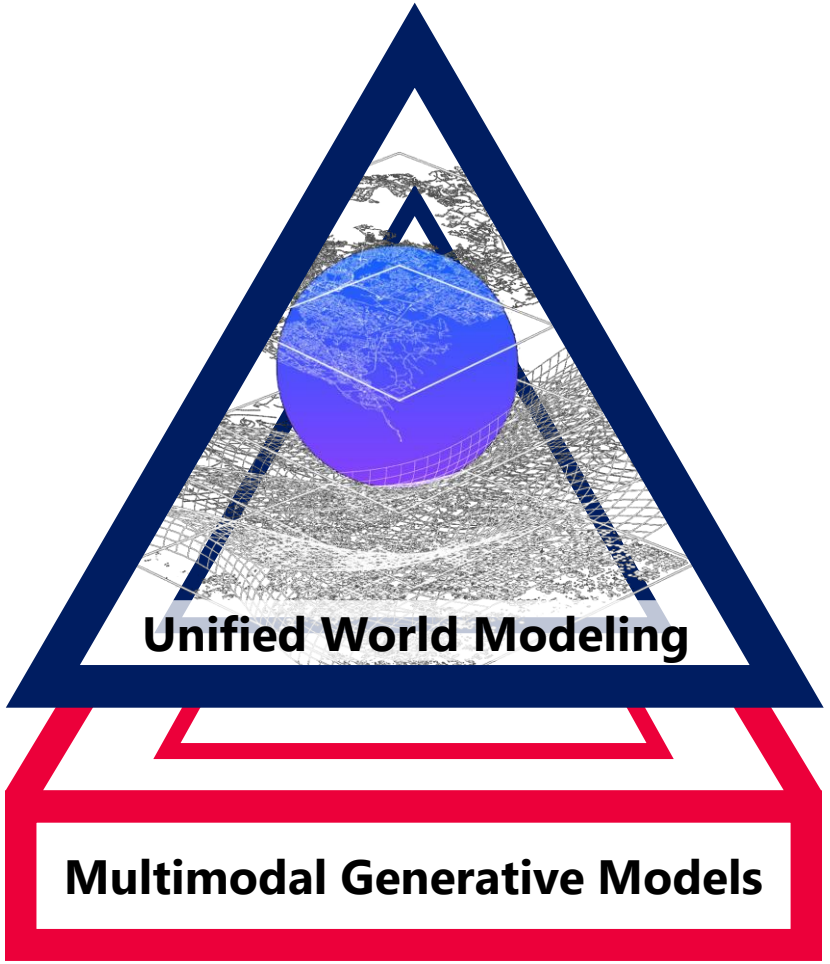
Embodied AI

Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model Dynamic Scenes



Be Actionable

How to Interact with the Physical World

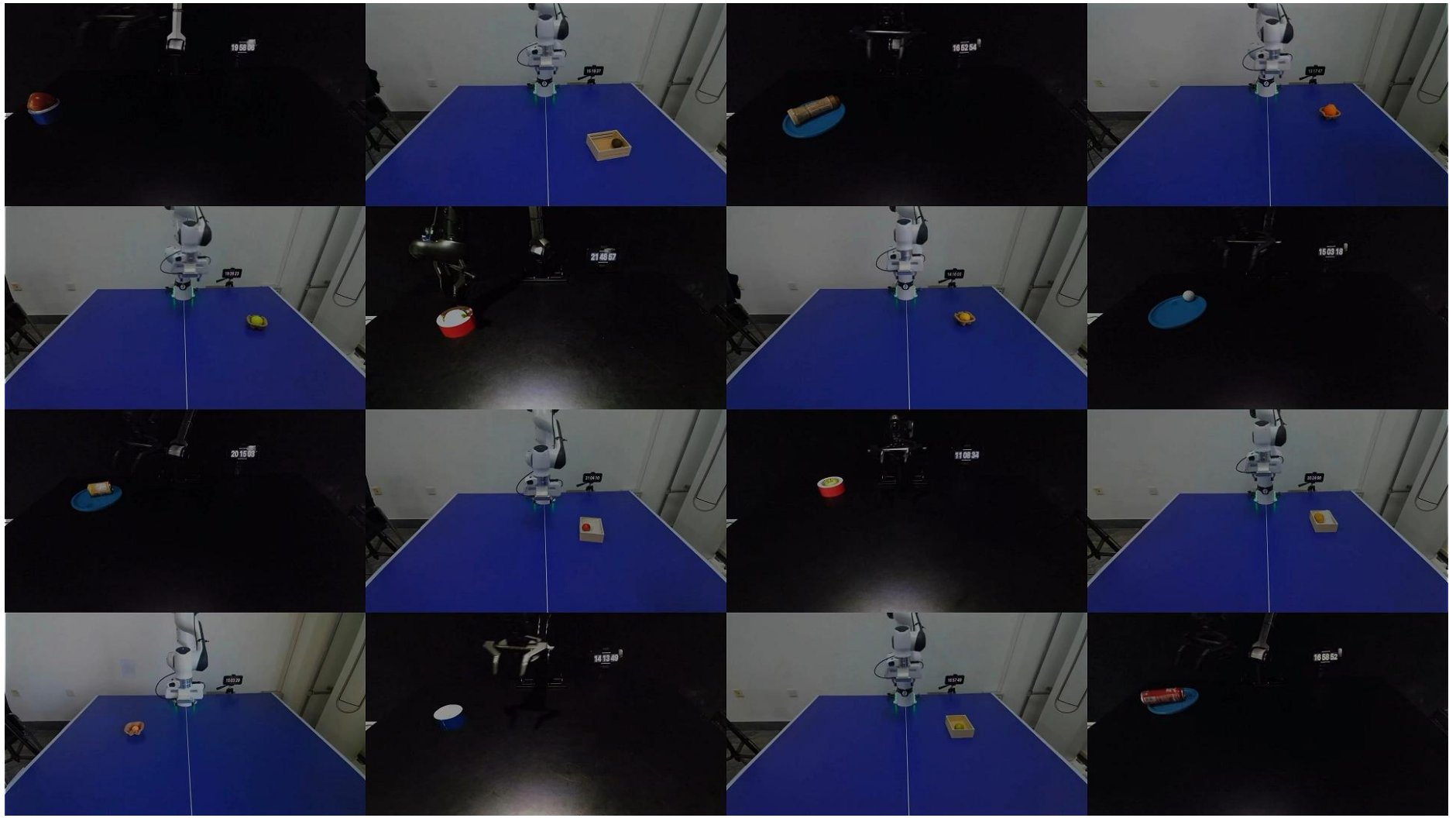
Be Actionable: DynamicVLA

 hzxie/DynamicVLA

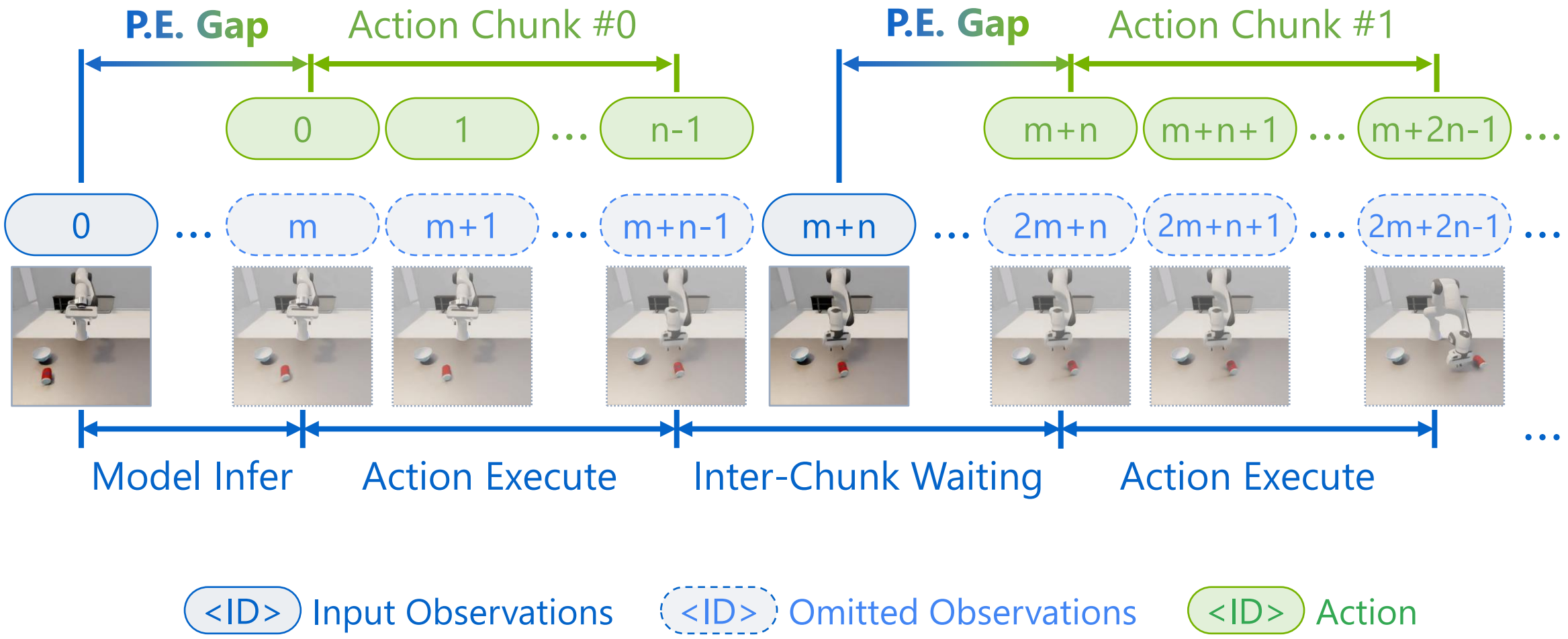
DynamicVLA: A Vision-Language-Action Model for Dynamic Object Manipulation

Haozhe Xie, Beichen Wen, Jiarui Zheng, Zhaoxi Chen, Fangzhou Hong, Haiwen Diao, Ziwei Liu
arXiv 2601.22153

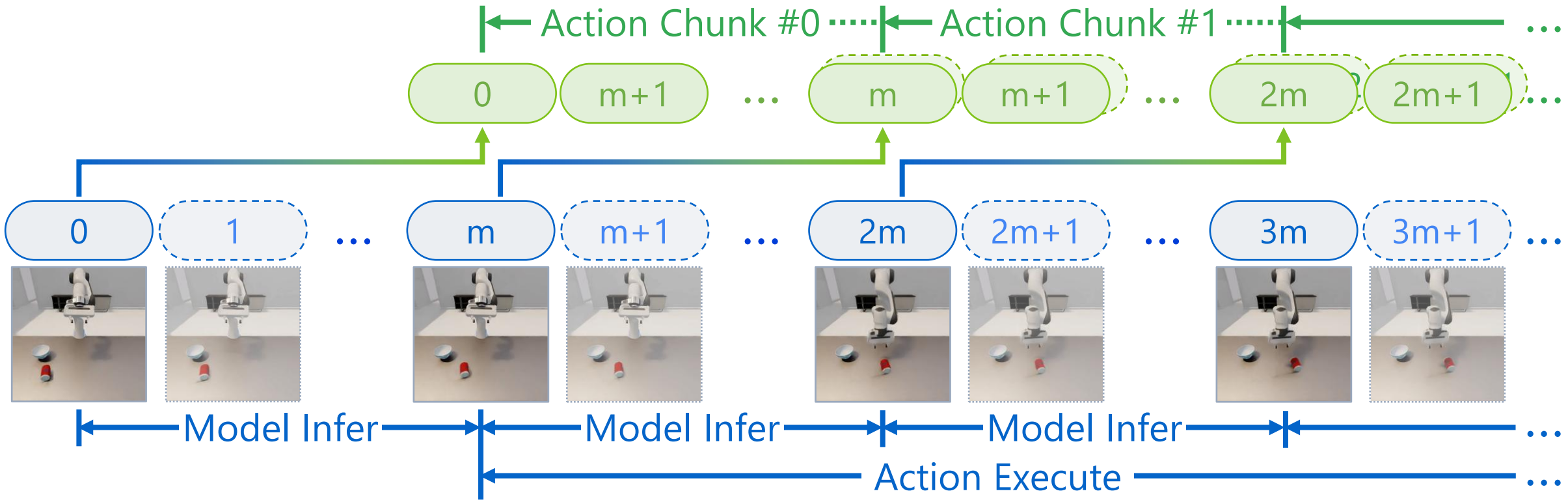
DynamicVLA – A Quick Glance



Latent Analysis for Current VLAs

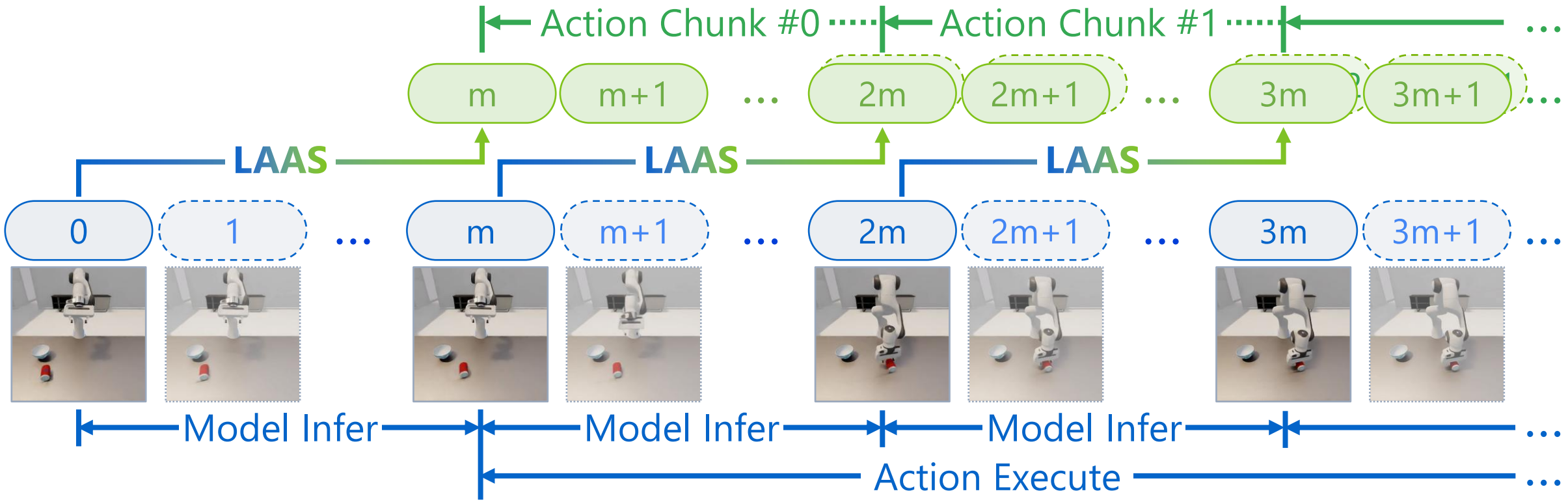


Contiguous Inference in DynamicVLA



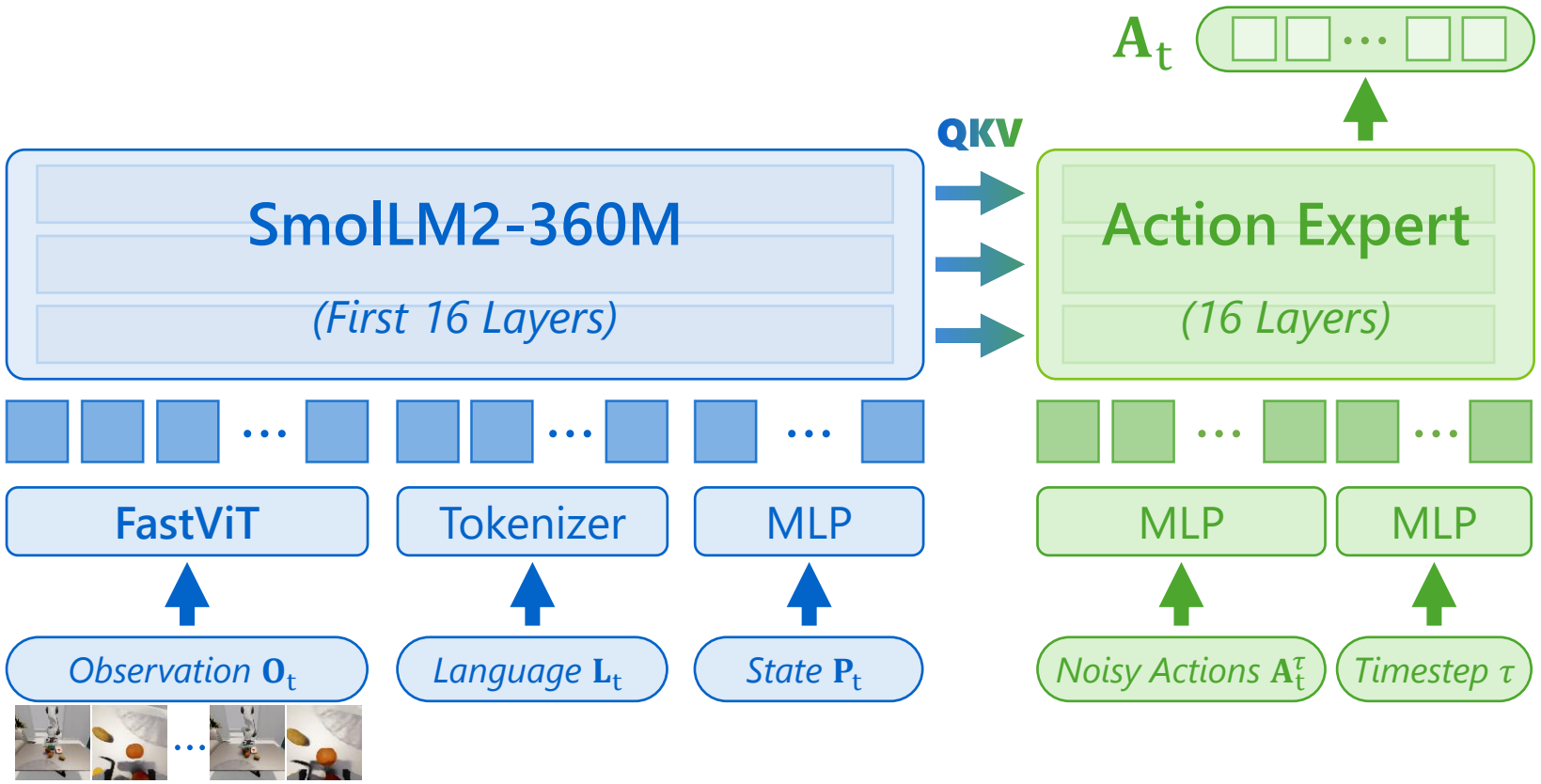
<ID> Input Observations <ID> Omitted Observations <ID> Action <ID> Omitted Action

Latent-aware Action Streaming in DynamicVLA



<ID> Input Observations <ID> Omitted Observations <ID> Action <ID> Omitted Action

Lightweight VLA Model



Automatic Data Collection

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Simulation

Objects and Dynamics

#Objects
206

Speed
0–1 m/s

Friction Coefficient
0.5-1.5

#Scenes
2824

Lighting
4000-8000K; 150-750 lm

Cameras
f = 2.3mm
25 FPS @ 480x360
front / side / wrist views

Scenes and Sensors

Isaac Sim Simulator

Object Position | Object Rotation | Object Velocity

Real-world

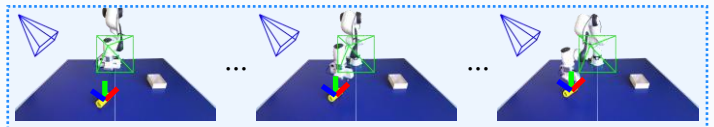
Objects

Third-Person Cameras
Microsoft Azure Kinect DK
25 FPS @ 1280x720
front / side views

Wrist Camera
Intel RealSense D435i
25 FPS @ 1280x720

Scenes and Sensors

Real-time 2D Object Tracking



Real-time 6D Object Pose Estimation

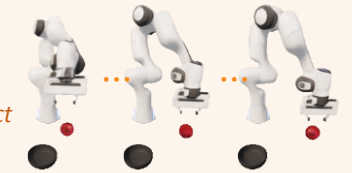
Real-world "Simulator"

Object Position | Object Rotation | Object Velocity

(S1) Approach Object

Actions
- *Move above object*

Transition
- *Gripper above object*



(S2) Grasp & Lift

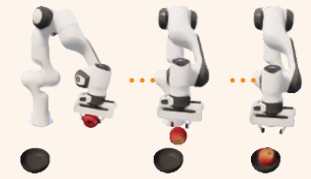
Actions
- *Close gripper*
- *Lift the object*

Transition
- *Object grasped*

(S3) Approach Target & Place

Actions
- *Move to the target*
- *Release object*

Transition
- *Object placed*



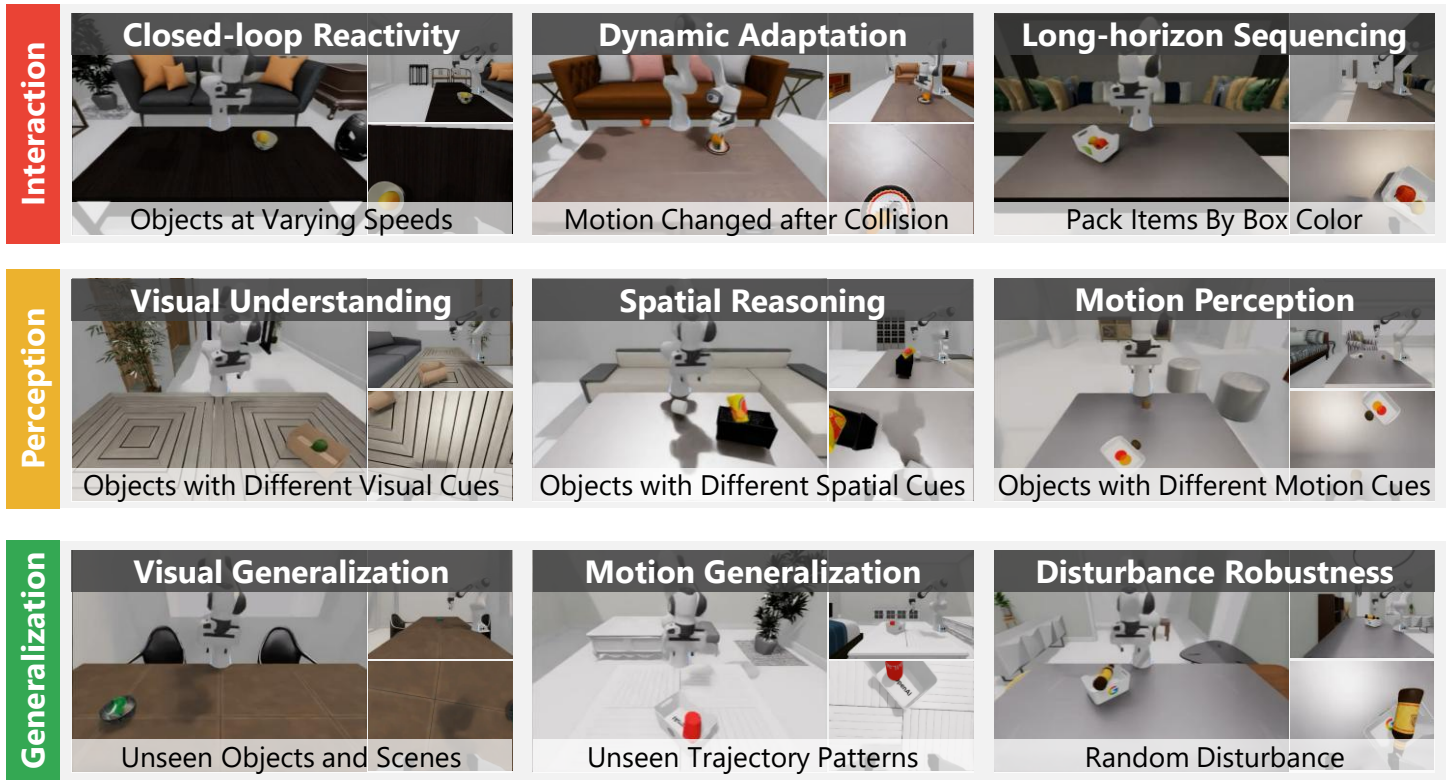
(S4) Reset

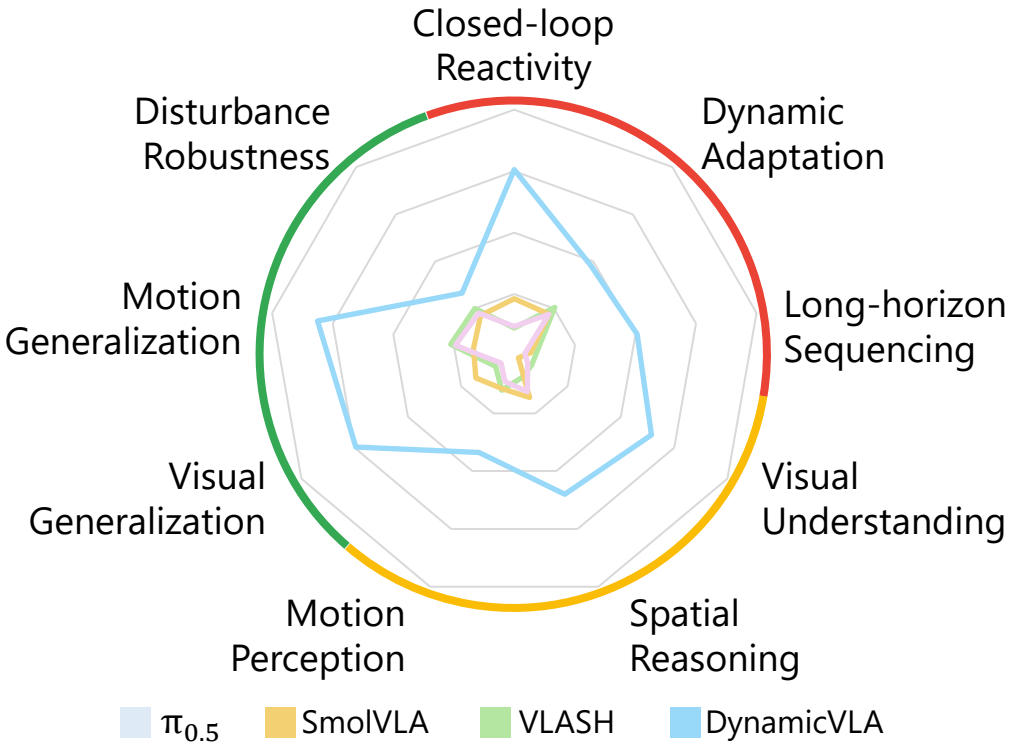
Actions
- *Return to home pose*

Object drop detected (return to S1)

Environment Setup | Object State Acquisition | State-machine Controller

Dynamic Object Manipulation Benchmark





(All actions shown above are generated by DynamicVLA)

Dynamic Object Manipulation Benchmark

Table 1: **Dynamic Object Manipulation Simulation Benchmark Results.** Average success rates (SR, %) are reported across nine evaluation sub-dimensions, organized under three categories: Interaction, Perception, and Generalization. In addition, overall average SR (%) and task completion time (Time, seconds) are reported. Each method is evaluated over 1,800 trials (10 scenes × 9 dimensions × 20 trials). All baseline models are fine-tuned on the DOM dataset using their official implementations and released pretrained weights. Best results are highlighted in bold.

Methods	Interaction			Perception			Generalization			Average	
	CR	DA	LS	VU	SR	MP	VG	MG	DR	SR ↑	Time ↓
Diffusion Policy [29]	0.50	0.50	0.00	1.00	0.00	0.00	1.00	0.50	0.00	0.38	10.89
ACT [16]	2.50	1.50	0.00	1.00	1.50	0.50	2.50	2.50	0.00	1.33	10.80
OpenVLA-OFT [54]	3.50	0.50	0.50	0.00	1.50	0.50	3.50	2.00	0.00	1.33	10.83
$\pi_{0.5}$ [55]	9.50	17.50	3.50	5.00	12.50	9.00	5.00	19.50	18.00	11.06	10.62
RTC [17]	12.00	15.00	4.00	7.50	13.50	8.50	6.50	14.00	20.00	11.22	10.63
SmolVLA [14]	18.50	17.50	5.50	1.50	14.50	11.50	14.50	13.50	17.00	12.67	10.65
GR00T-N1.5 [59]	10.50	12.00	4.00	9.50	13.50	14.00	14.50	19.50	20.00	13.05	10.56
VLA-Adapter-Pro [15]	21.00	15.50	6.00	6.50	16.50	10.50	15.00	18.50	13.00	13.61	9.98
VLASH [10]	9.00	20.50	7.50	6.50	7.50	12.00	7.00	21.00	20.00	12.33	10.60
DynamicVLA	**60.50**	**38.50**	**40.50**	**51.50**	**48.00**	**33.50**	**59.50**	**65.00**	**26.50**	**47.06**	**8.53**

CR: Closed-loop Reactivity; DA: Dynamic Adaptation; LS: Long-horizon Sequencing; VU: Visual Uderstanding; SR: Spatial Reasoning; MP: Motion Perception; VG: Visual Generalization; MG: Motion Generalization; DR: Disturbance Robustness

Experiments on Real-world Robots

Be Actionable: Kinema4D

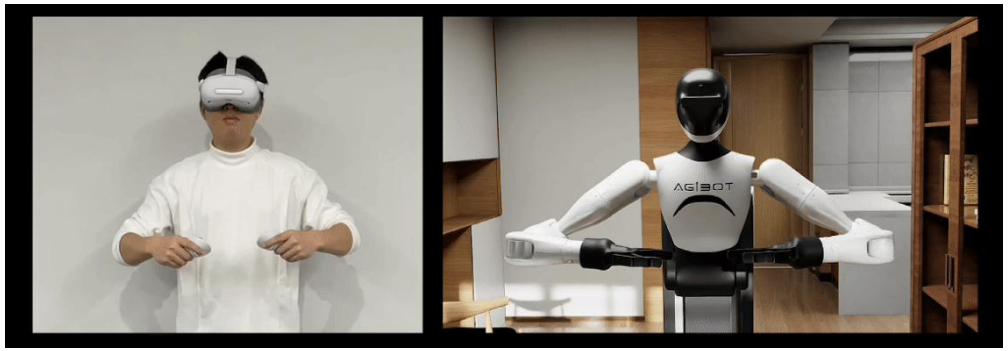
mutianxu/Kinema4D

Kinema4D: Kinematic4D World Modeling for Spatiotemporal Embodied Simulation

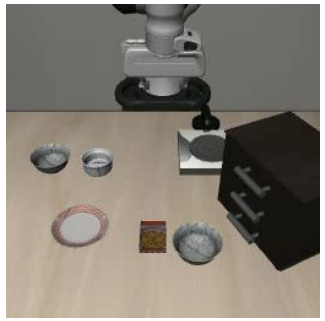
Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, Ziwei Liu
arXiv 2603.16669

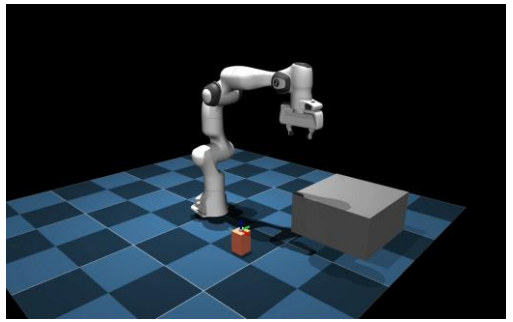
Background

Rolling out robot trajectories are **costly**, **unsafe**, and require **labor-intensive human supervision**
=> **restrict** **scalable robot learning** in the real world.

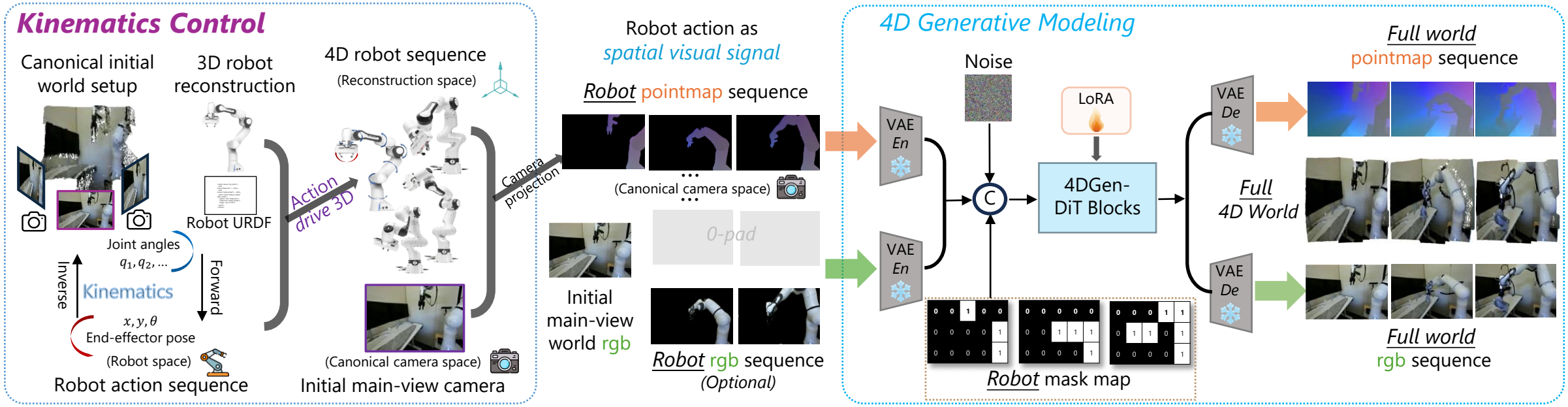


Existing **physical-engine** based simulators are **mostly synthetic**
=> **not visually realistic** and constrained by **predefined physical rules**, **hard to model** **real-world** dynamics.





Kinema4D: Precise 4D Control & Flexible 4D Generation



Result comparison

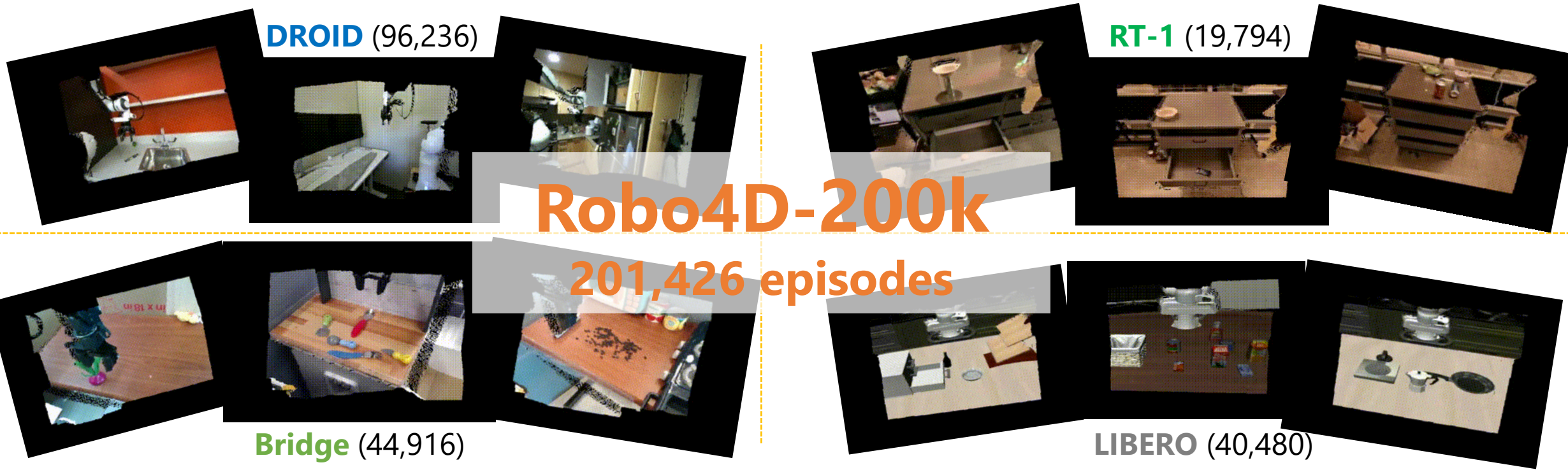
Method	Action	Output	PSNR↑	SSIM↑	L2_latent↓	FID↓	FVD↓	LPIPS↓
UniSim [ICLR'24]	Text	RGB	19.32	0.681	0.2120	32.3	153.2	0.175
IRA-Sim [ICCV'25]	Emb.	RGB	20.21	0.813	0.1722	25.2	126.0	0.135
Cosmos [arXiv'25]	Emb.	RGB	20.39	0.787	0.1935	27.1	113.4	0.110
EVAC [arXiv'25]	Emb.+2D	RGB	20.88	0.832	0.1896	29.3	122.0	0.150
ORV [CVPR'26]	Emb.+3D	RGB	19.45	0.790	0.2002	30.1	130.1	0.143
Ctrl-World [ICLR'26]	Emb.	RGB	21.03	0.803	0.1533	**24.9**	112.8	0.122
TesserAct [ICCV'25]	Text	**4D**	19.35	0.766	0.1911	29.5	120.3	0.158
Ours	**4D**	**4D**	**22.50**	**0.864**	**0.1380**	25.2	**98.5**	**0.105**

Outperform across various metrics of video generation

Method	CD-L1↓	CD-L1 (temp)↓	CD-L2↓	CD-L2 (temp)↓	F-Score↑	F-Score (temp)↑
TesserAct[ICCV'25]	0.0836	**0.0067**	0.0130	0.0008	0.2896	0.9523
Ours	**0.0479**	0.0074	**0.0077**	**0.0002**	**0.4733**	**0.9686**

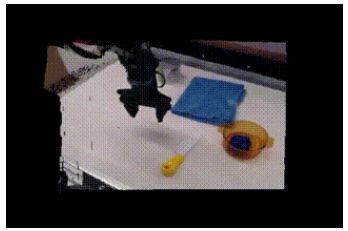
Outperform across various metrics of geometry

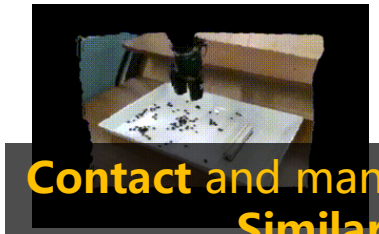
Dataset: Robo4D-200k



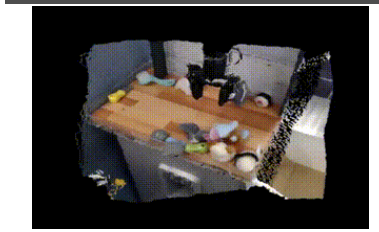
Comparison: Successful Simulation

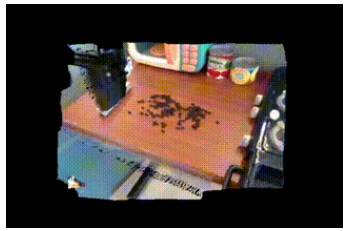
Ours



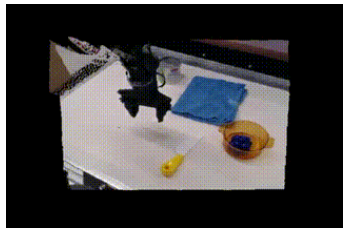


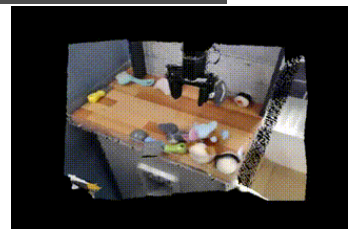
Contact and manipulate **same** object
Similar outcome

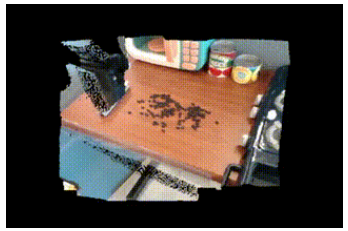




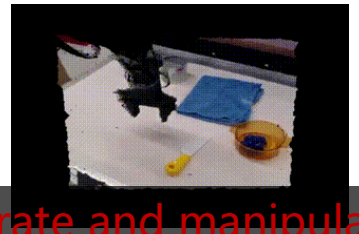
Ground Truth







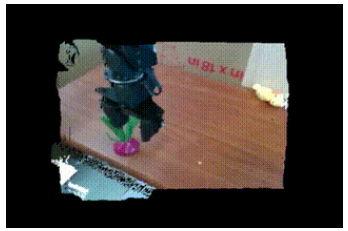
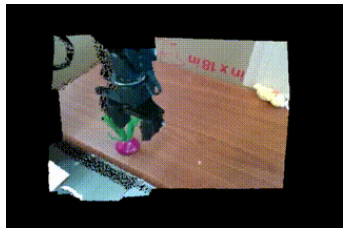
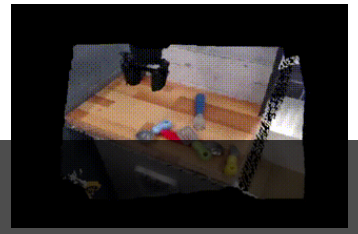
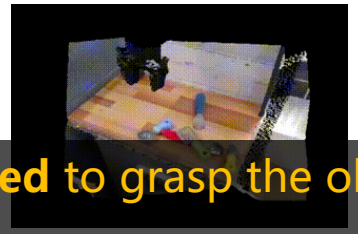
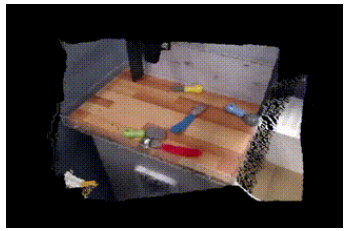
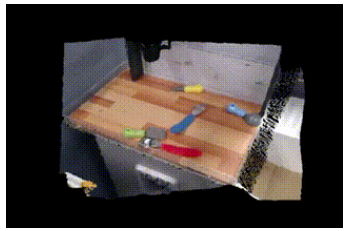
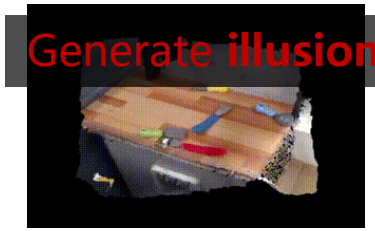
TesserAct [ICCV 2025]



Generate and manipulate
non-existing object

Manipulate **wrong** object
No contact but manipulate
Wrong outcome

Comparison: Failed Simulation

Ours	Ground Truth	TesserAct [ICCV 2025]
		
		
		
		

Still successfully grasp the object

All failed to grasp the object

Correctly interprets the **spatial gap** between the gripper and the object even when their **RGB textures overlap in 2D views**

Generate **illusion**

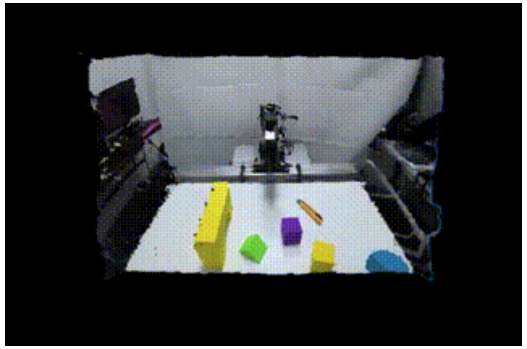
Policy Evaluation *(real-wold OOD test)*

All successful pick&place

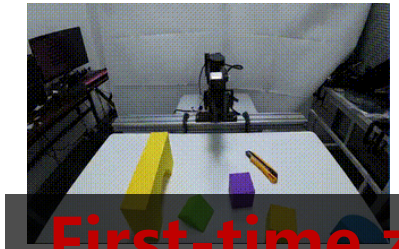
Ours

Ground Truth

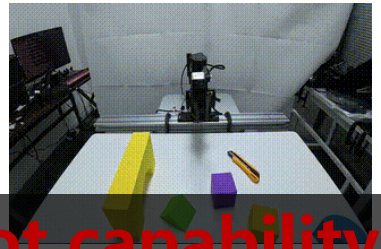
4D



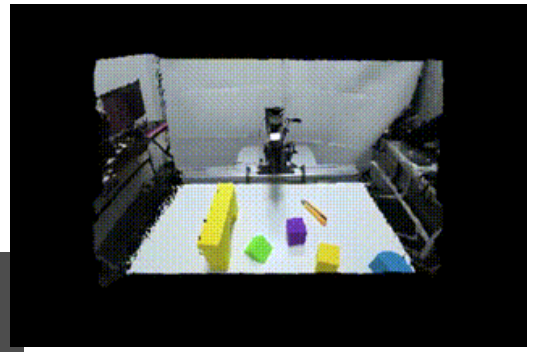
2D



2D



4D

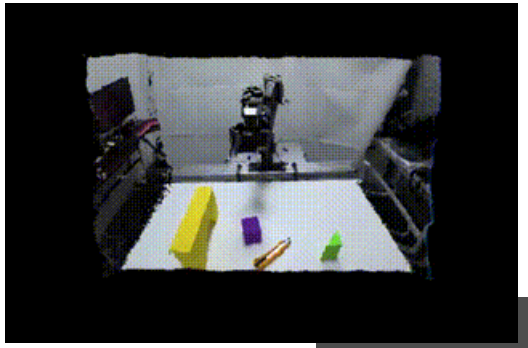


First-time zero-shot capability under rigorous OOD evaluation

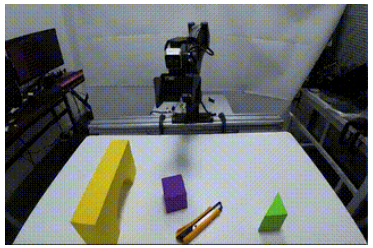
Ours

Ground Truth

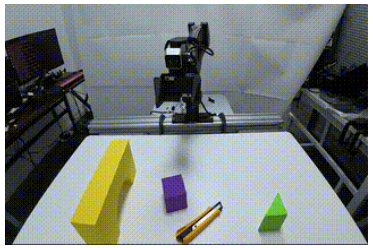
4D



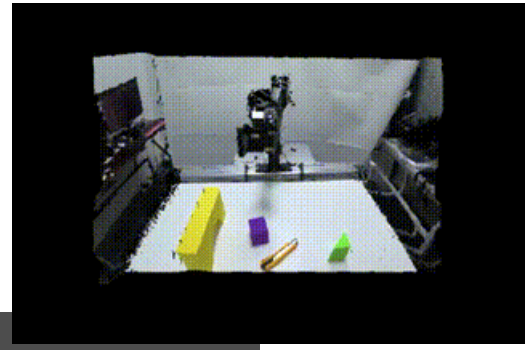
2D



2D



4D

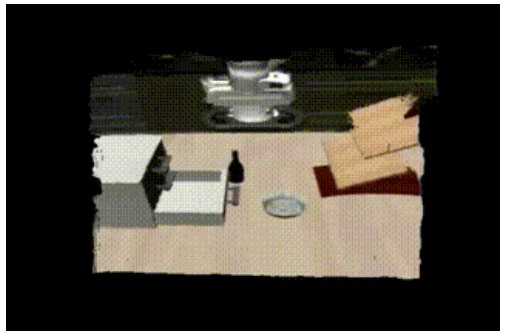


All failed to grasp the object

Correctly interprets the **spatial gap** between the gripper and the object even when their **RGB textures overlap in 2D views**

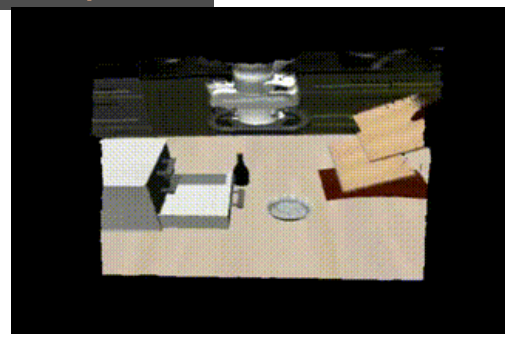
Policy Evaluation

Ours

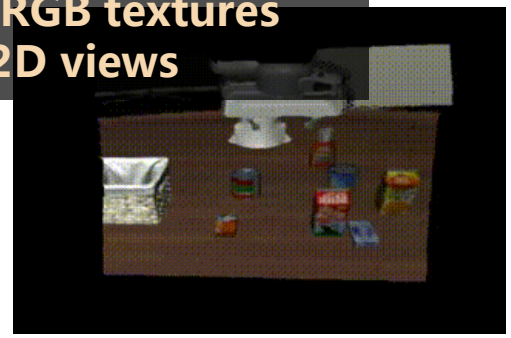
All successful pick&place

GT


All failed to grasp the object

Correctly interprets the **spatial gap** between the gripper and the object, even when their **RGB textures overlap in 2D views**

Be Actionable: WorldLens

worldbench/ WorldLens

WorldLens: Full-Spectrum Evaluations of Driving World Models in Real World

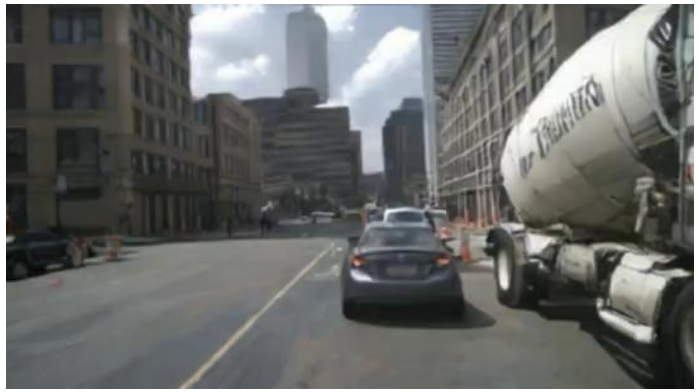
Ao Liang*, Lingdong Kong*†, Tianyi Yan*, Hongsi Liu*, Yu Yang*, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, Dongyue Lu, Youquan Liu, Guangfeng Jiang, Linfeng Li, Xiangtai Li, Long Zhuo, Lai Xing Ng, Benoit R. Cottereau, Changxin Gao, Liang Pan, Wei Tsang Ooi, Ziwei Liu

CVPR 2026 Oral Presentation

Challenges

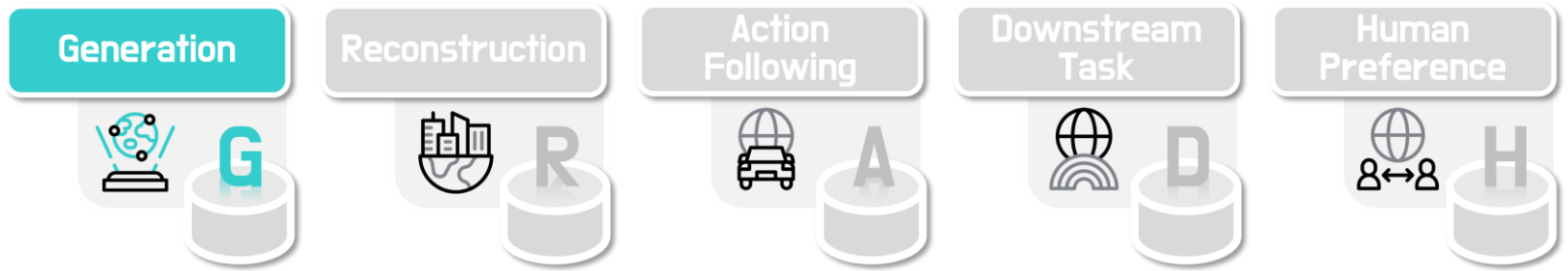
- Traditional benchmarks are fundamentally **appearance-driven** (pixel-level aesthetics), which leaves them blind to the core requirements of world modeling






WorldLens: Benchmarking World Models

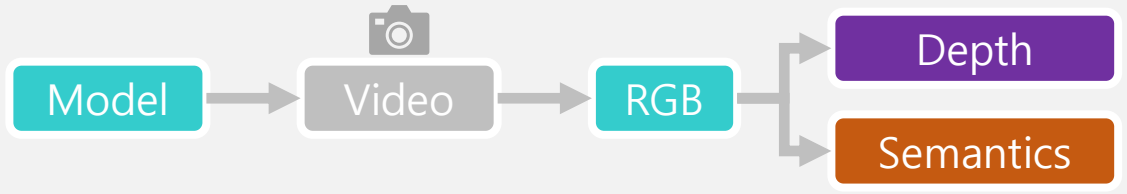
WorldLens: Benchmarking World Models



Generation

Objective:

Measure whether a model can synthesize **visually realistic**, **temporally stable**, and **semantically consistent** scenes.

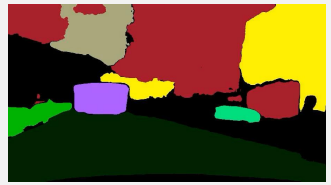


WorldLens: Benchmarking World Models

Generation | Reconstruction | Action Following | Downstream Task | Human Preference

G | R | A | D | H

Generation



WorldLens: Benchmarking World Models

Generation | Reconstruction | Action Following | Downstream Task | Human Preference

G | R | A | D | H

Generation

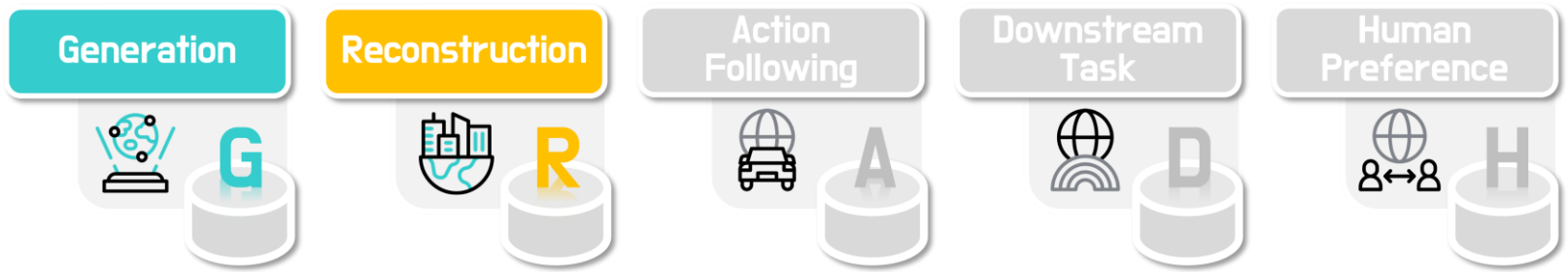
WorldLens: Benchmarking World Models

Generation | Reconstruction | Action Following | Downstream Task | Human Preference

G R A D H

Generation

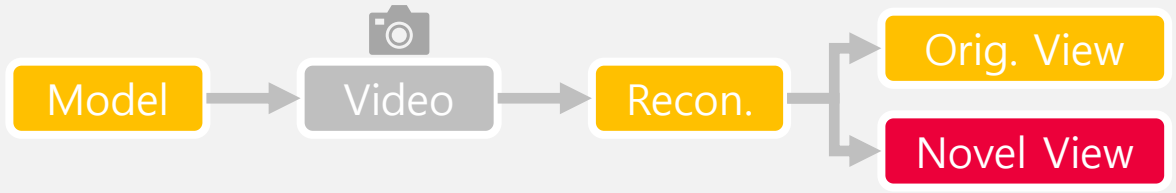
WorldLens: Benchmarking World Models



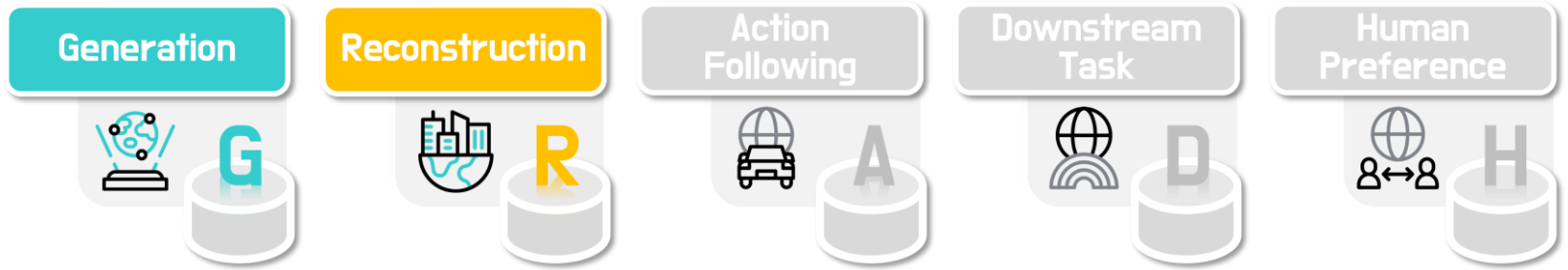
Reconstruction

Objective:

Probe whether generated videos can be reprojected into a **coherent 4D scene** using differentiable rendering.

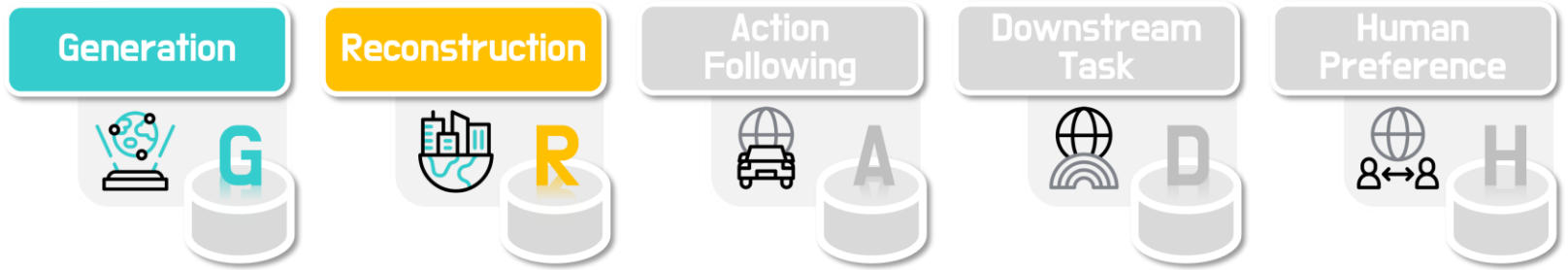


WorldLens: Benchmarking World Models

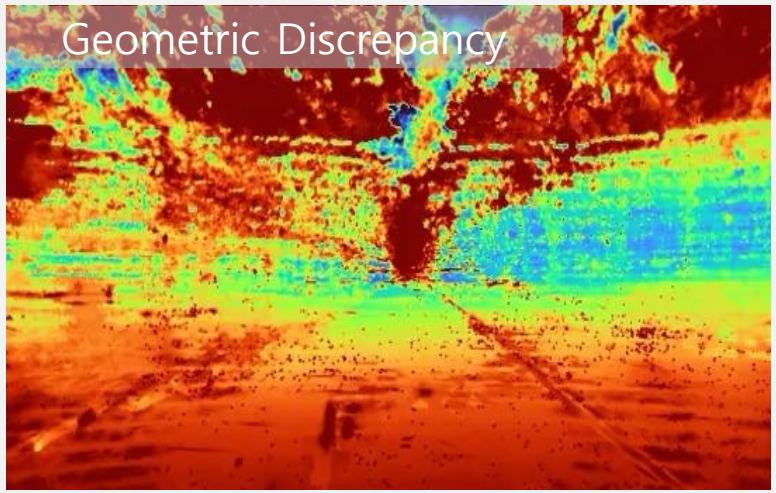


Reconstruction

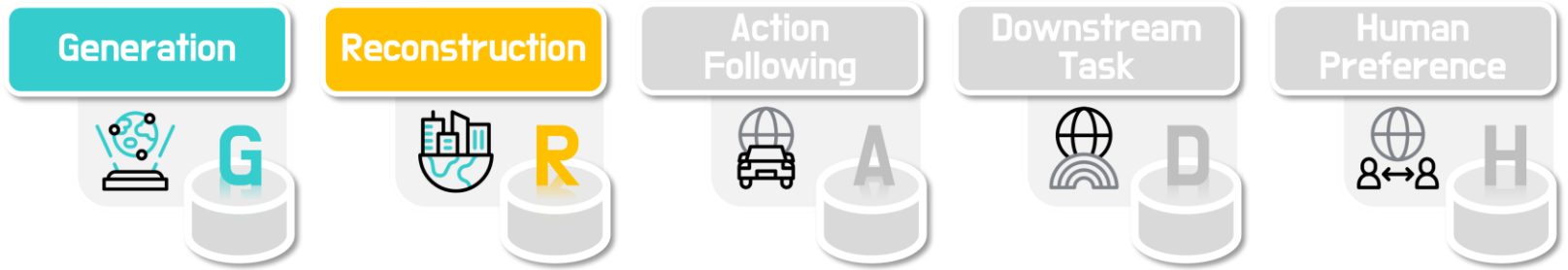
WorldLens: Benchmarking World Models



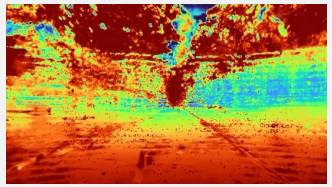
Reconstruction

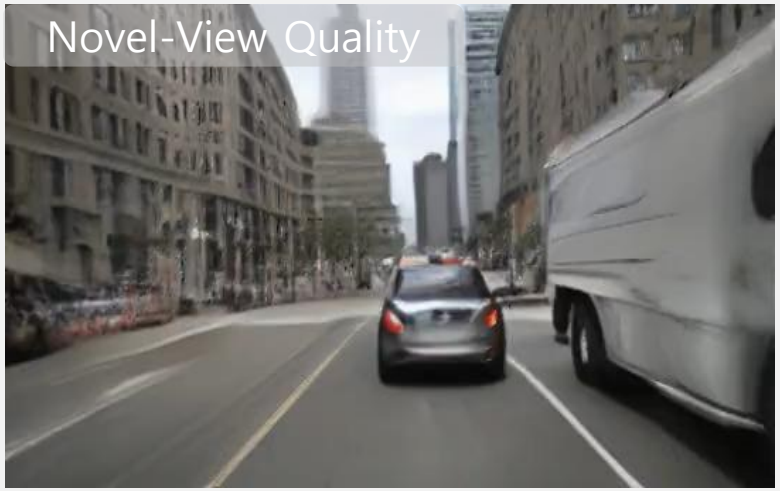


WorldLens: Benchmarking World Models



Reconstruction



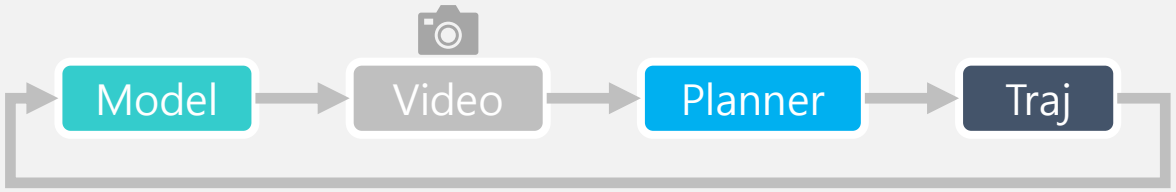


WorldLens: Benchmarking World Models

Action-Following

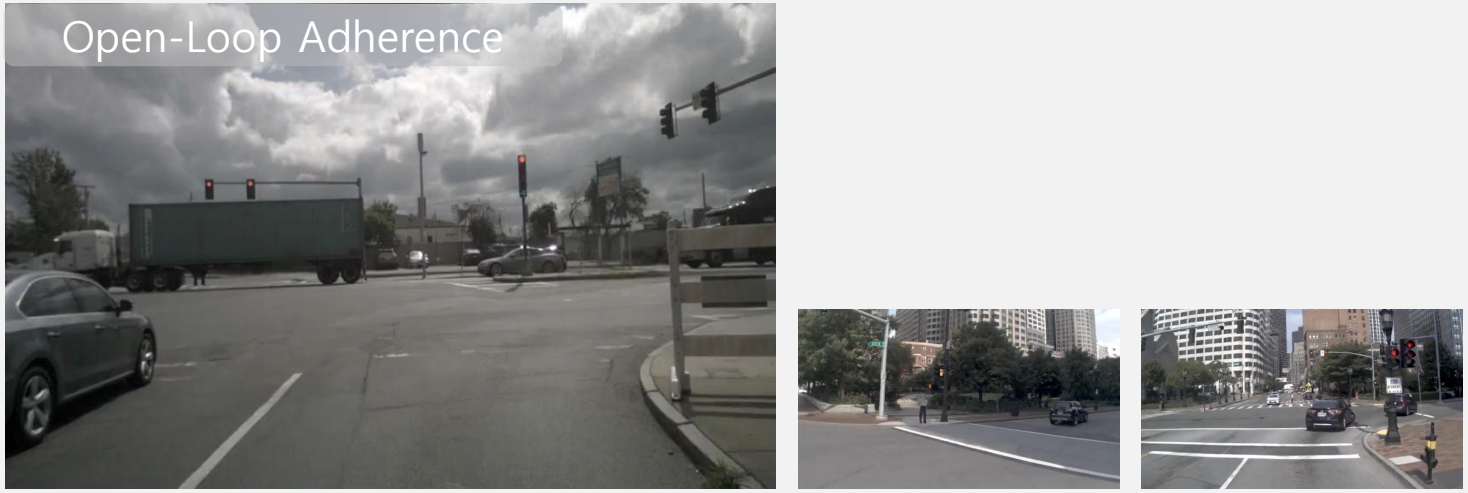
Objective:

Test if a pre-trained action planner can **operate safely** inside the generated world using an open- / closed-loop simulator.



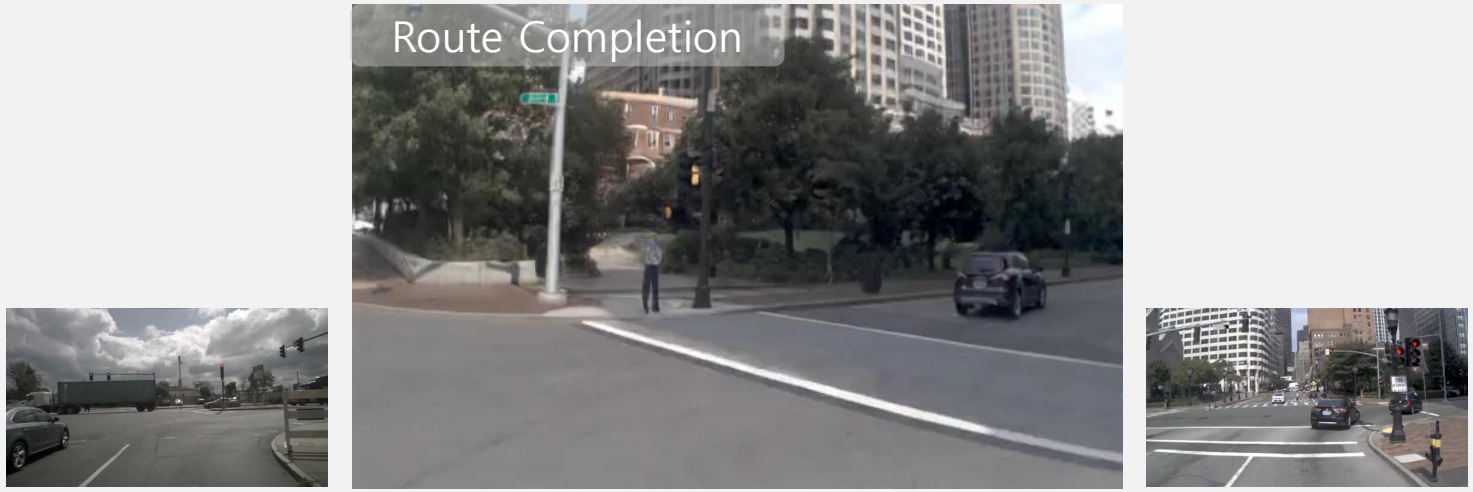
WorldLens: Benchmarking World Models

Action-Following



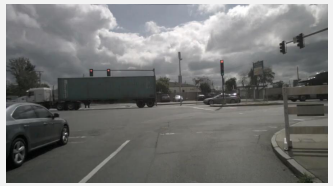
WorldLens: Benchmarking World Models

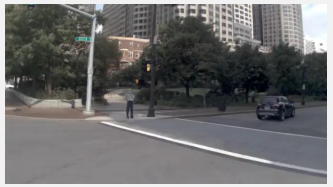
Action-Following

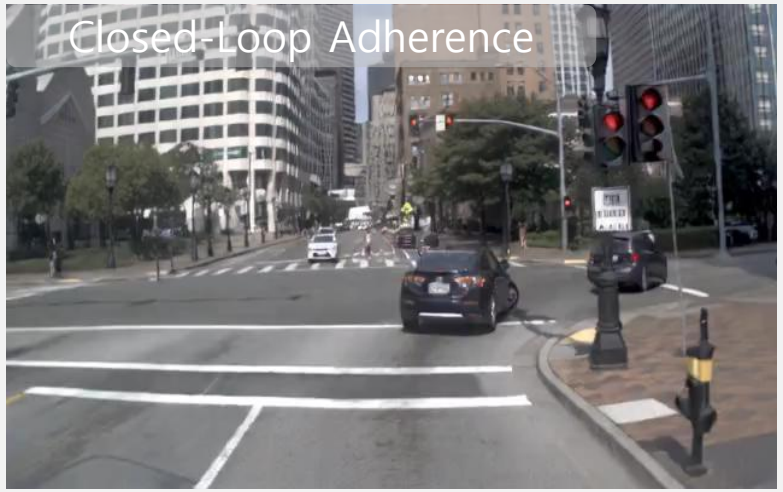


WorldLens: Benchmarking World Models

Action-Following







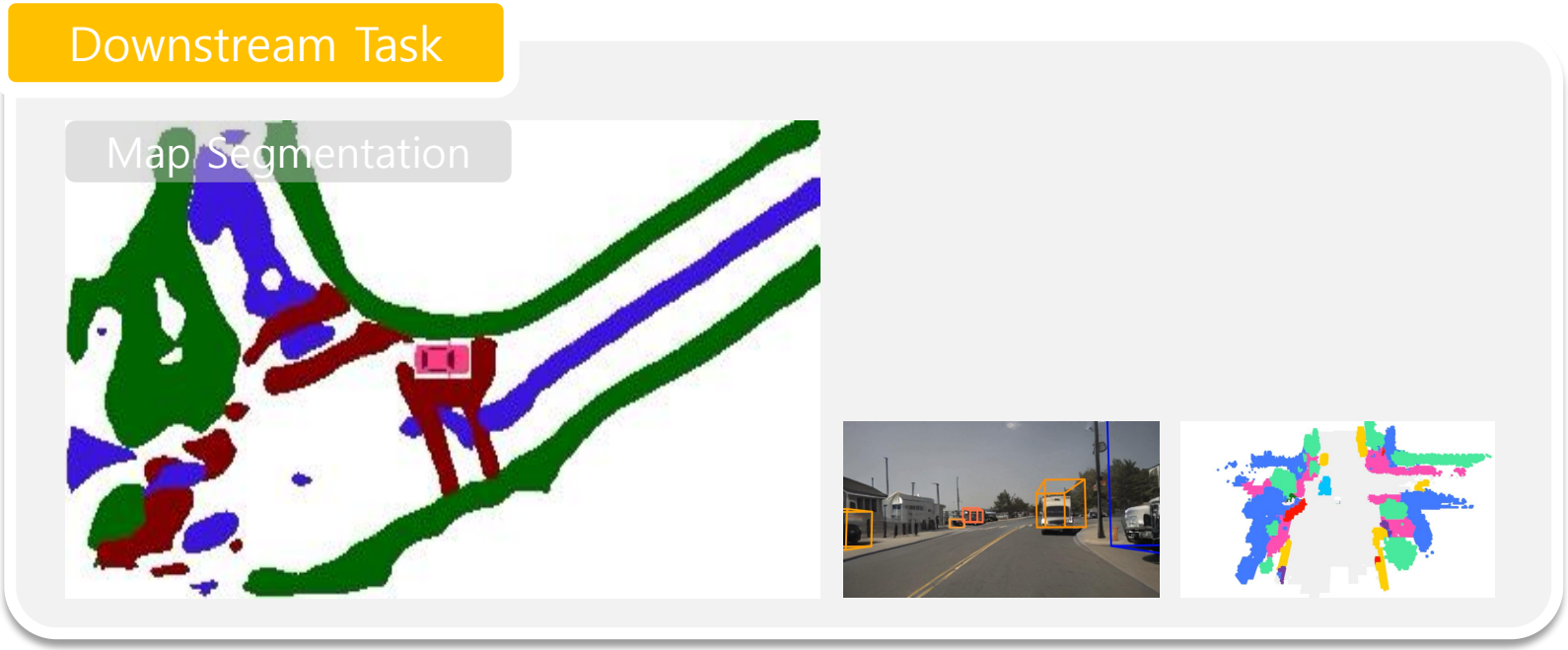
WorldLens: Benchmarking World Models

Downstream Task

Objective:

Test if a pre-trained action planner can **operate safely** inside the generated world using an **open- / closed-loop** simulator.

WorldLens: Benchmarking World Models



WorldLens: Benchmarking World Models

Generation G

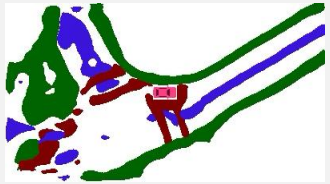
Reconstruction R

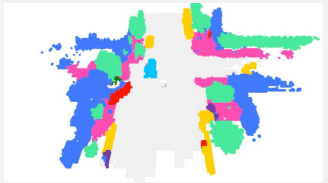
Action Following A

Downstream Task D

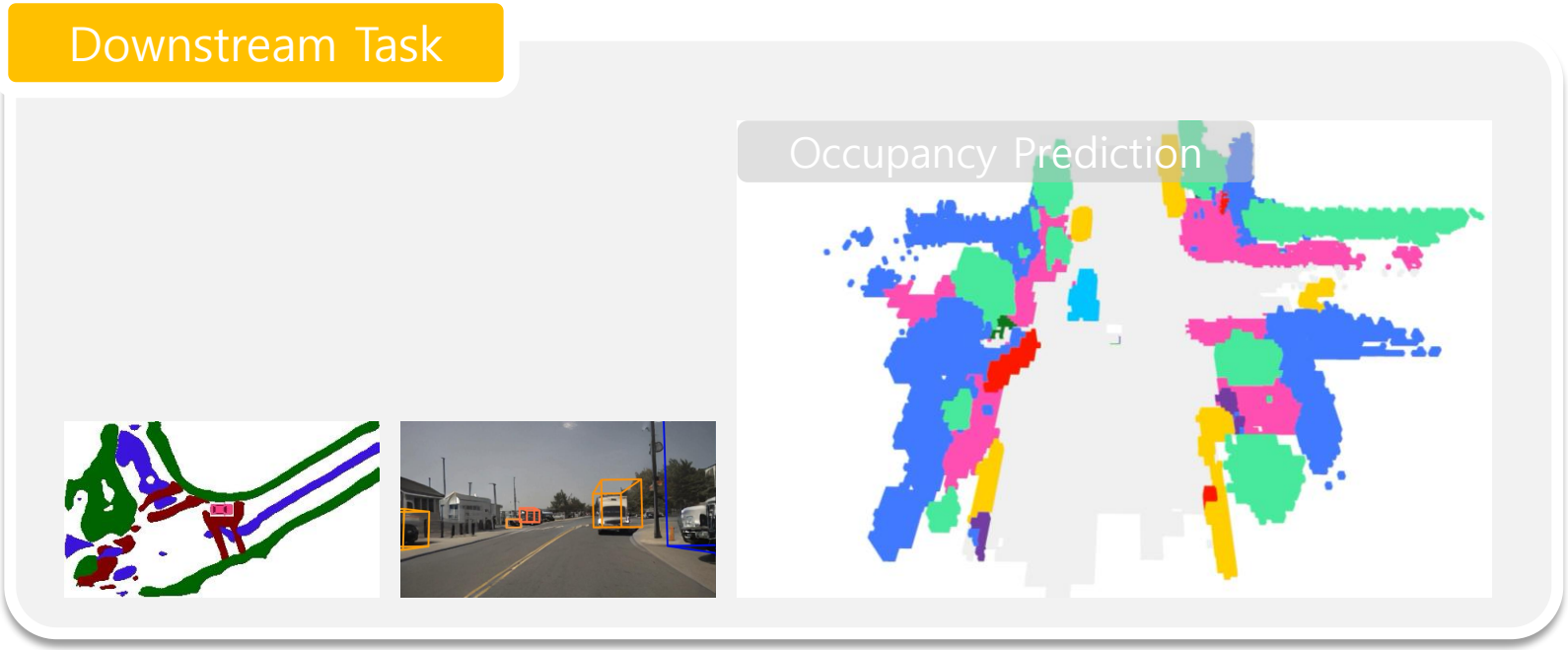
Human Preference H

Downstream Task





WorldLens: Benchmarking World Models

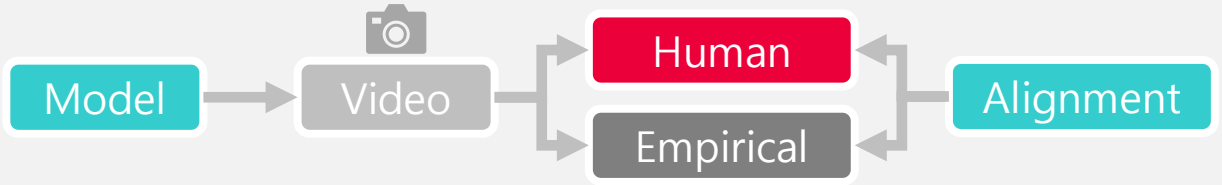


WorldLens: Benchmarking World Models

Human Preference

Objective:

Capture **subjective scores**, e.g., physical plausibility and behavioral safety, through **large-scale human annotations**.



WorldLens: Benchmarking World Models

Generation G

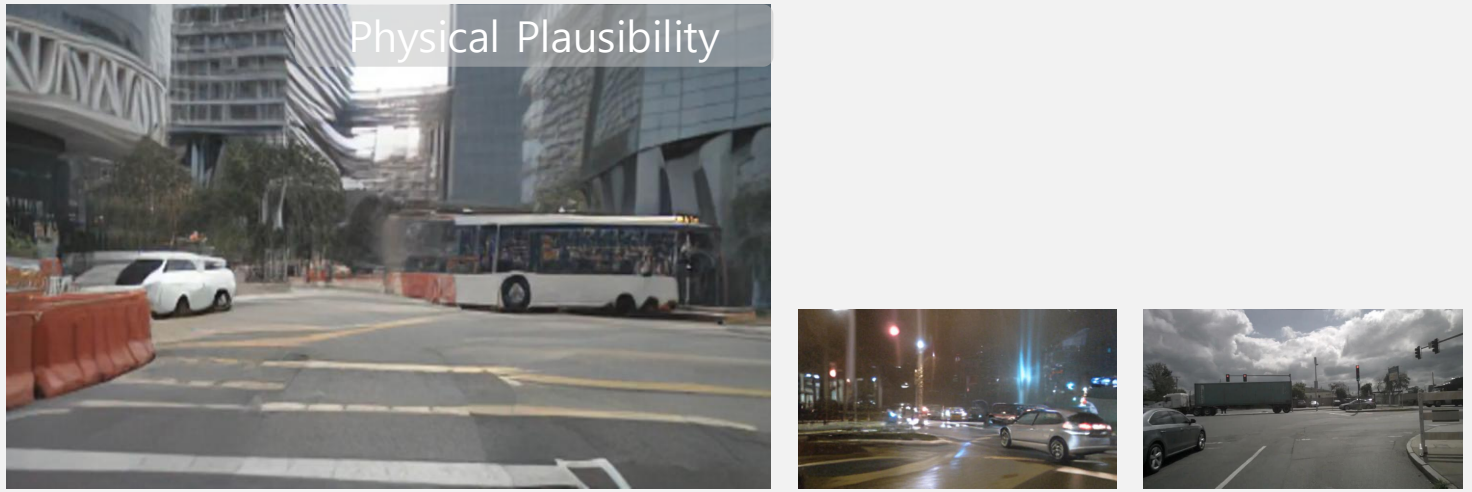
Reconstruction R

Action Following A

Downstream Task D

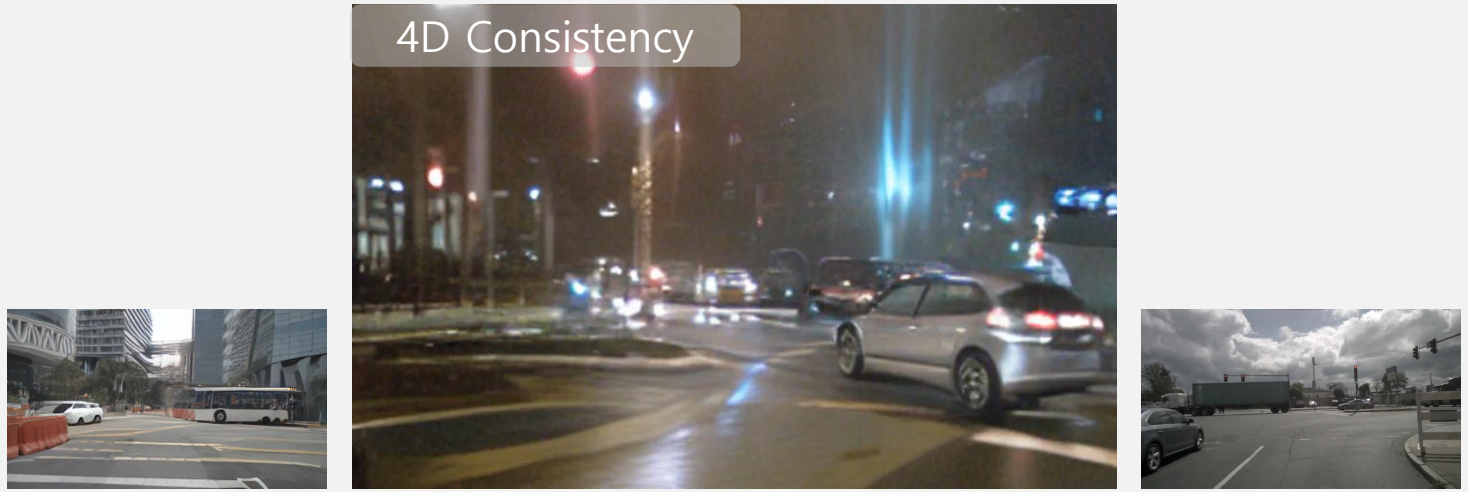
Human Preference H

Human Preference



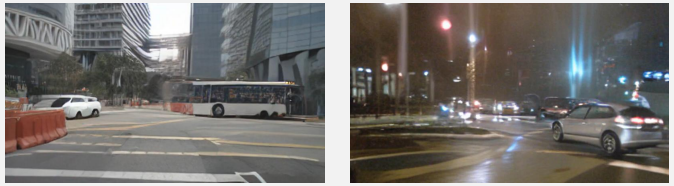
WorldLens: Benchmarking World Models

Human Preference



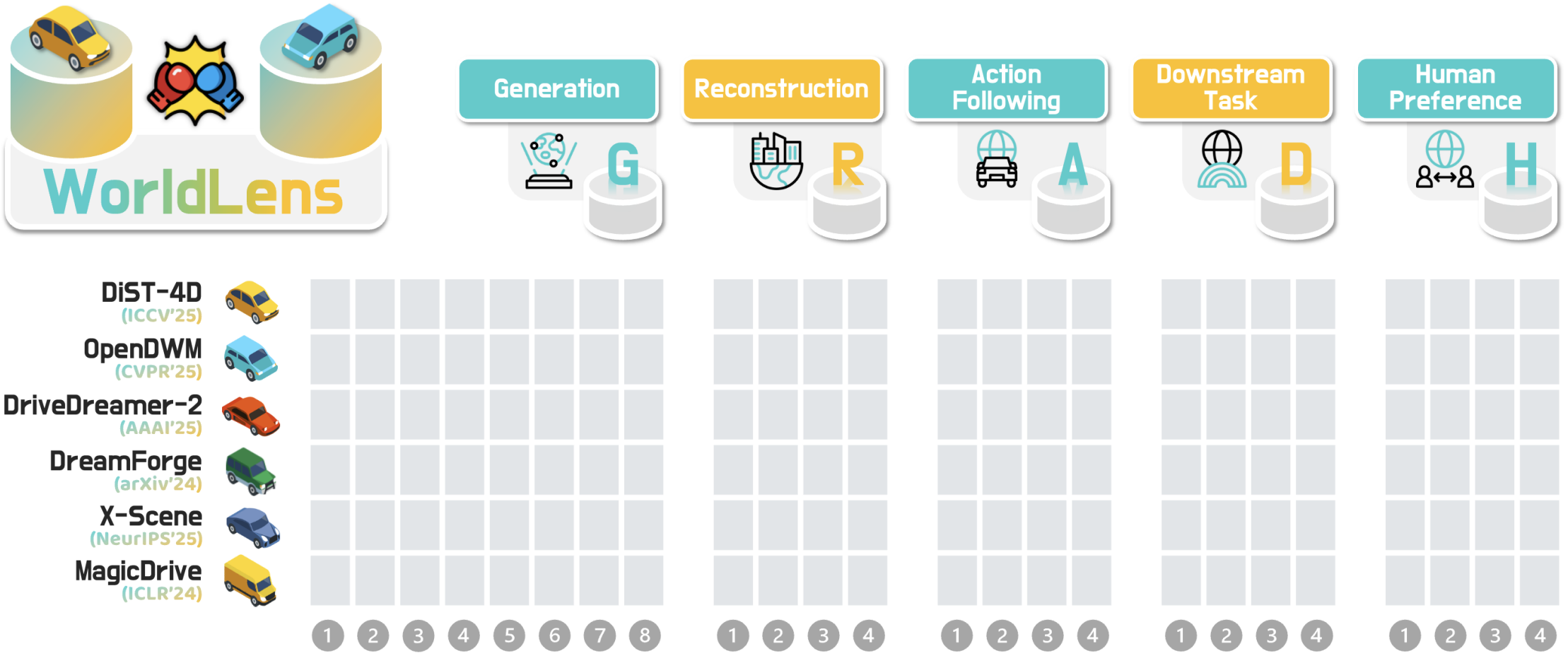
WorldLens: Benchmarking World Models

Human Preference



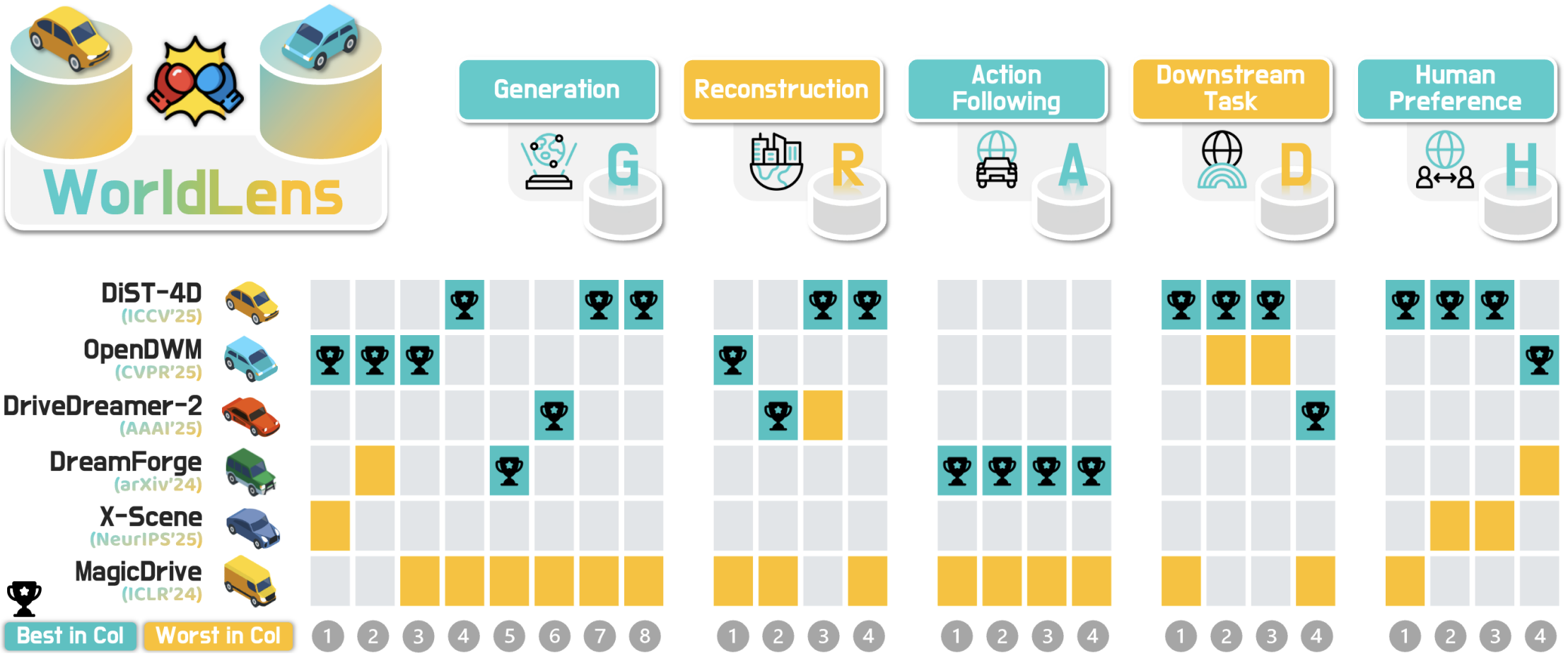
WorldLens: Benchmarking World Models

WorldLens: Benchmarking World Models

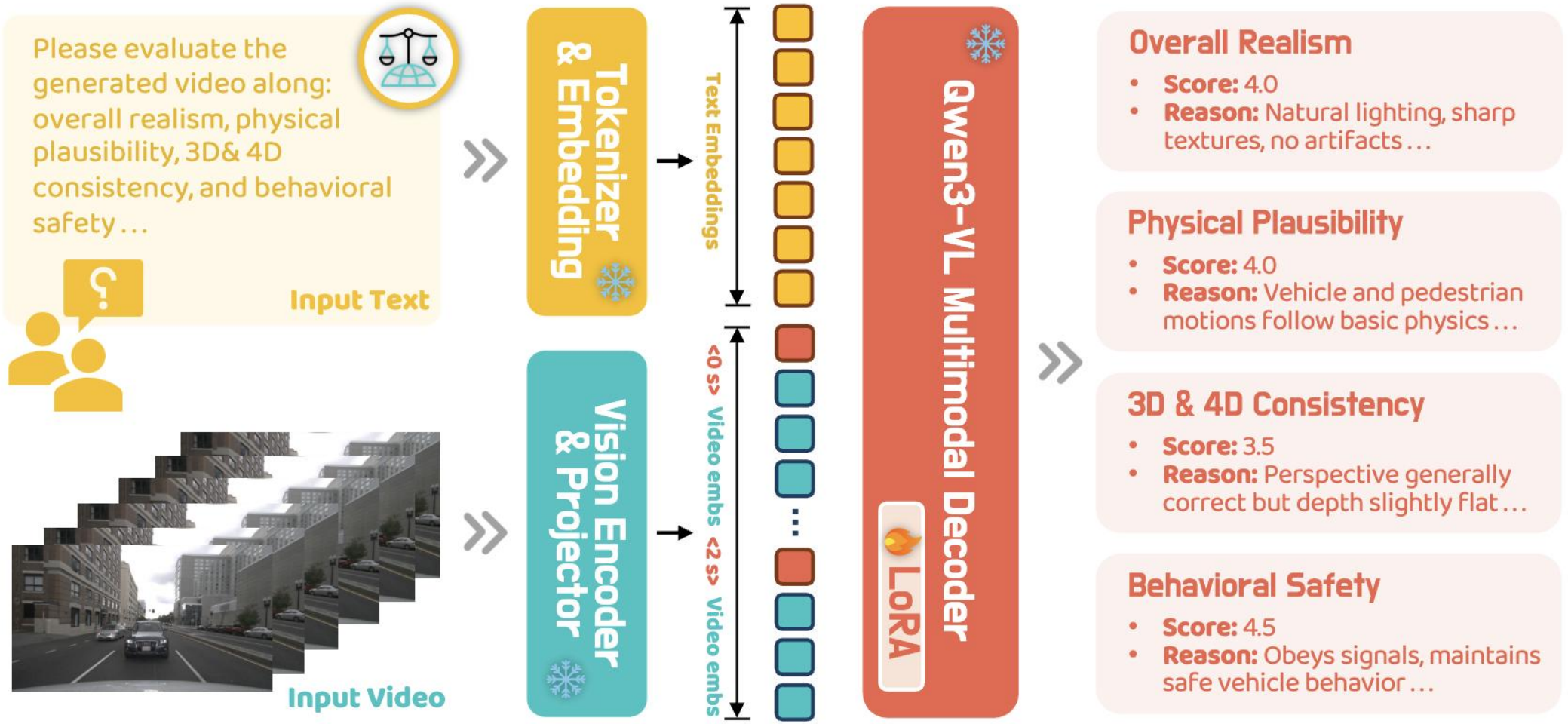


WorldLens: Benchmarking World Models

WorldLens: Benchmarking World Models



WorldLens-Agent



WorldLens-Agent: Use Cases

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Physical Plausibility

The video exhibits a failure in physical plausibility. The ego's perspective **inexplicably lifts off the ground** and ascends into the air, appearing to fly over the guardrails and terrain . . .

Rating: 1 / 10

WorldLens-Agent: Use Cases

Behavioral Safety

The ego-vehicle **drives along the sidewalk** and passes through sidewalk objects, which represents **clearly unsafe** and **non-compliant behavior**. Such violations make the vehicle . . .

Rating: 2 / 10

WorldLens-Agent: Use Cases

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Physical Plausibility

The scene's geometry and lighting remain broadly consistent, with stable road structure. However, the **cow's movement lacks physical realism**, its body glides laterally with minimal limb articulation . . .

Rating: 6 / 10

WorldLens-Agent: Use Cases

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Behavioral Safety

The ego-vehicle **fails to stop** at a clearly visible red light, proceeding through the intersection despite the signal being solidly red. This constitutes a **serious traffic violation** and, narrowly avoiding a collision . . .

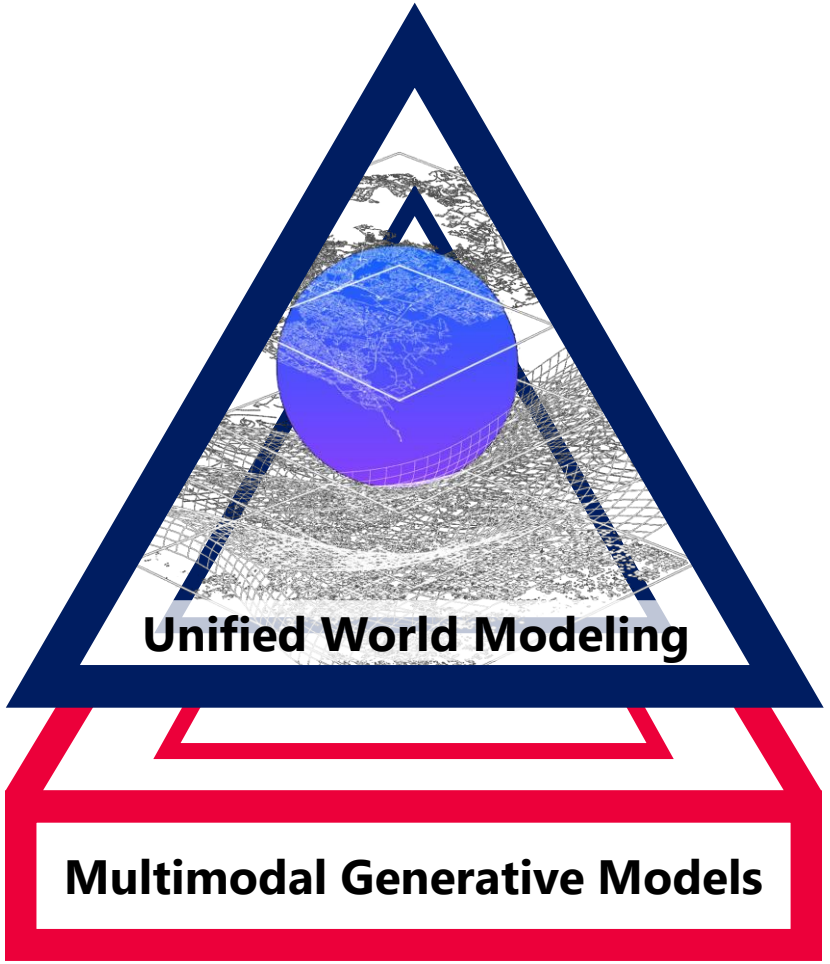
Rating: 1 / 10

Be Physical

How to Model Material and Illumination

Be Dynamic

How to Model Dynamic Scenes



Be Actionable

How to Interact with the Physical World

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Thank You

Ziwei Liu 刘子纬

Nanyang Technological University