Towards Physics-Consistent Efficient Visual World Models

Jianfei Cai

Department of Data Science & AI

Monash University

What is World Model?

Gemini:

***World model** is an **internal representation** of how the world works. It is a system's "mental map" that allows it to **predict what will happen next based on current observations and potential actions**.*

Our focus: Physics-Consistent Efficient Visual World Model

Text Conditioned Visual Generation

T2I **T2V** **T-to-3D**

What are the problems?

Precision, physically plausible, efficiency, camera control, editing, interactive …

IT2I **I2V / IT2V** **I-to-3D / IT-to-3D** **X-to-4D**

Outline

Efficiency: T2I, T2Pano

- ***T-Stitch*** *(ICLR'25)*
- *PanFusion (CVPR'24 Highlight)*
- *Para: Personalized T2I (ICLR'25 Spotlight)*
- *VQVA (CVPR'26)*

Geometry consistency: I-to-3D, 3DGenAI

- ***MVSplat (ECCV'24 oral), MVSplat360*** *(NeurIPS'24)*
- *PanSplat (CVPR'25), PanFlow (AAAI'26)*
- ***MotionCrafter*** *(CVPR'26 highlight)*
- ***Ov3R*** *(CVPR'26 highlight)*

Camera Control

- *UCPE (CVPR'26)*

Audio Generation

- ***Omni2Sound*** *(CVPR'26 highlight)*
- *CineDub*

Physics consistency: T2V

- ***VLIPP (ICCV'25)***

World-model: memory

- ***I3DM***

VLIPP: Towards Physically Plausible Video Generation with Vision and Language Informed Physical Prior

Xindi Yang, Baolu Li*, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma,*
Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, Xu Jia

ICCV 2025

Project Page: https://madaoer.github.io/projects/physically_plausible_video_generation/
Code: https://github.com/Madaoer/VLIPP

Motivation

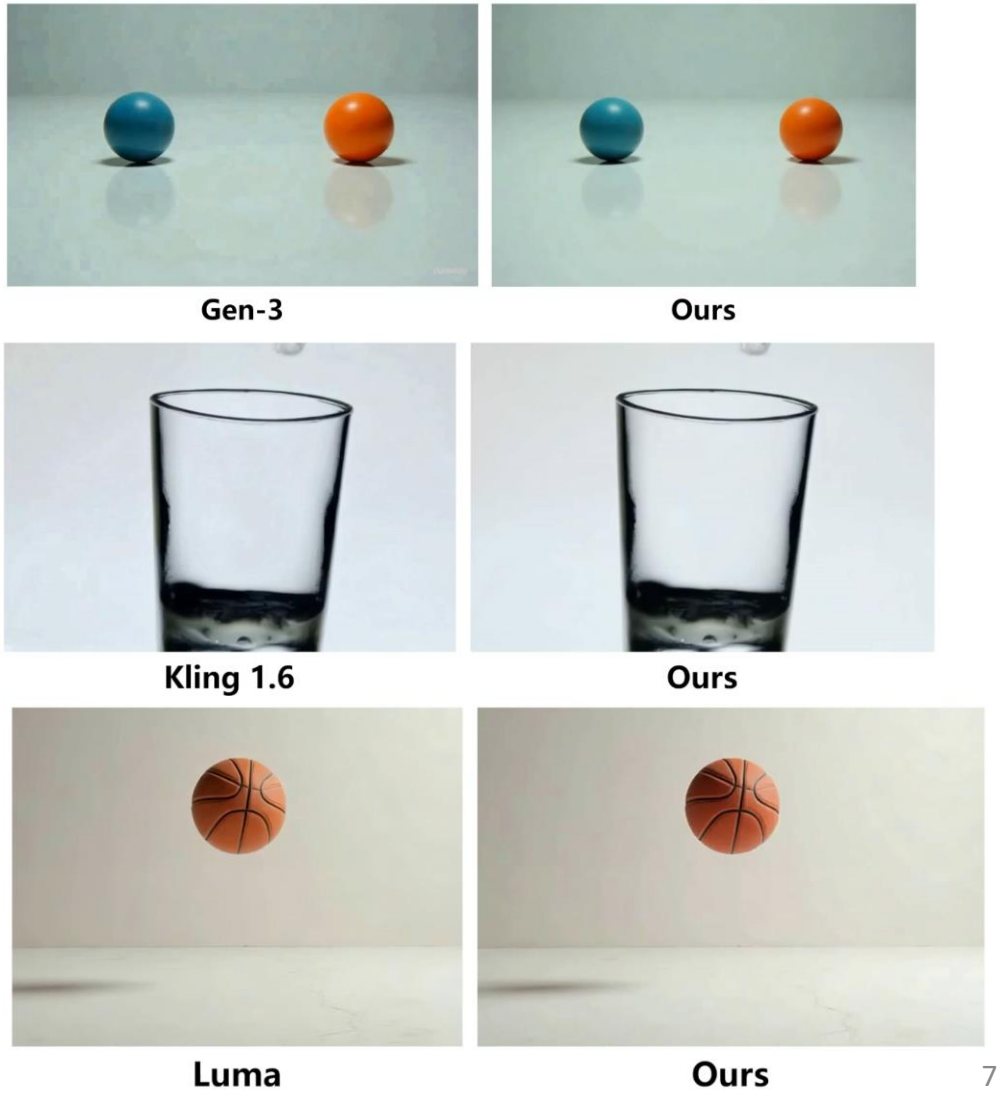
- VDMs can generate realistic videos.

Struggle with understanding the physical laws and generating videos accordingly.

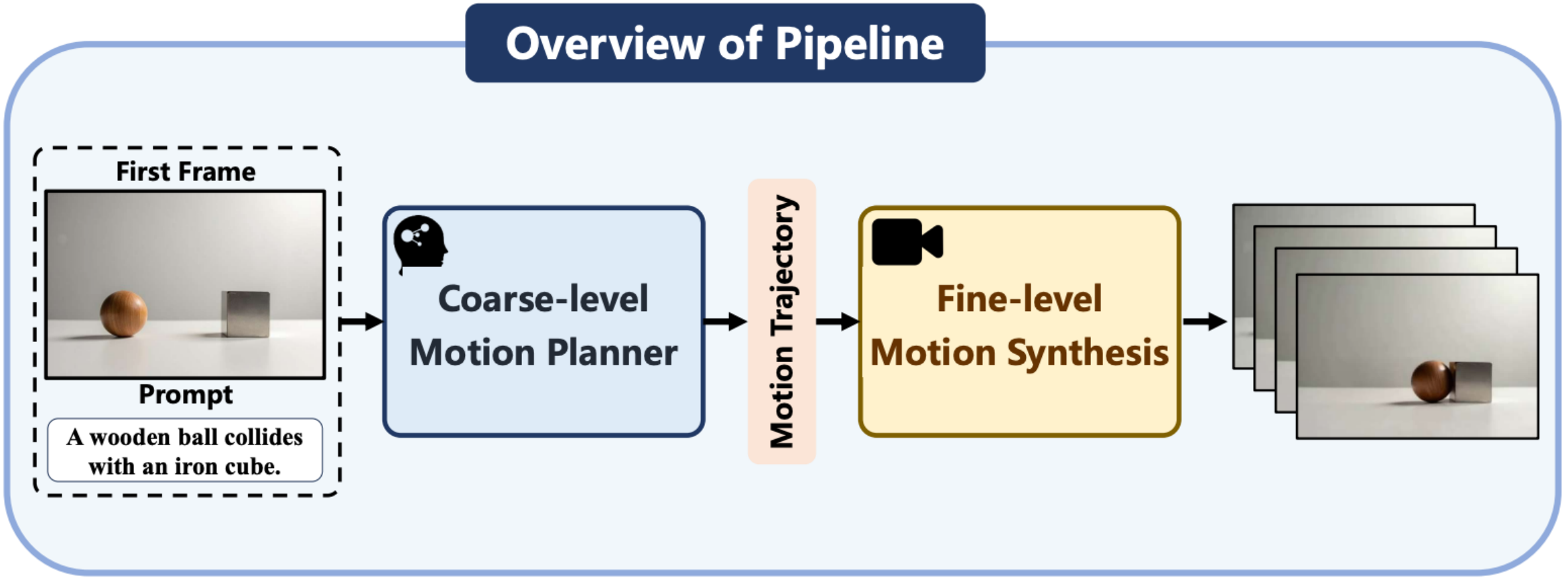
Motivation

- LLMs exhibit some capability in understanding and reasoning about physical phenomena.

- Incorporating LLM's physics prior as condition into VDM, enabling the generation of physically plausible motion.



Method

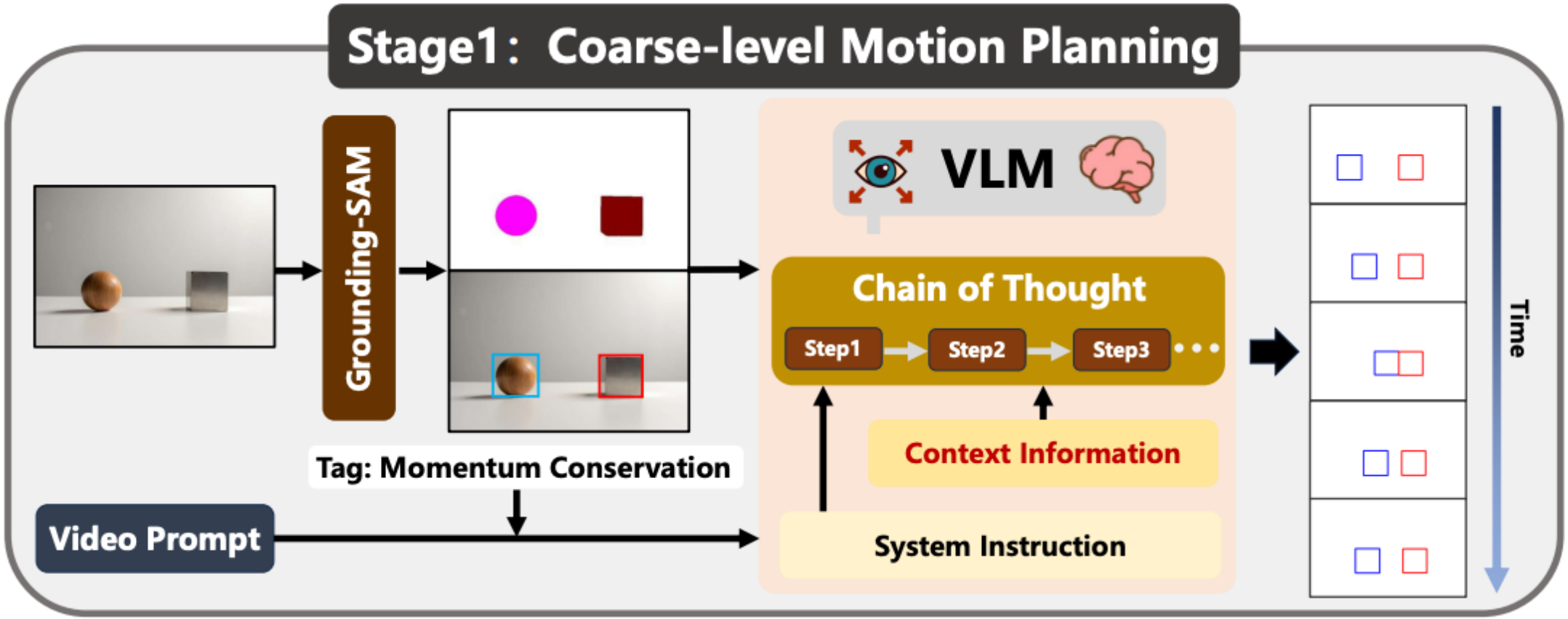

Overview of Pipeline
First Frame
Prompt
A wooden ball collides with an iron cube.
Coarse-level Motion Planner
Motion Trajectory
Fine-level Motion Synthesis

Method

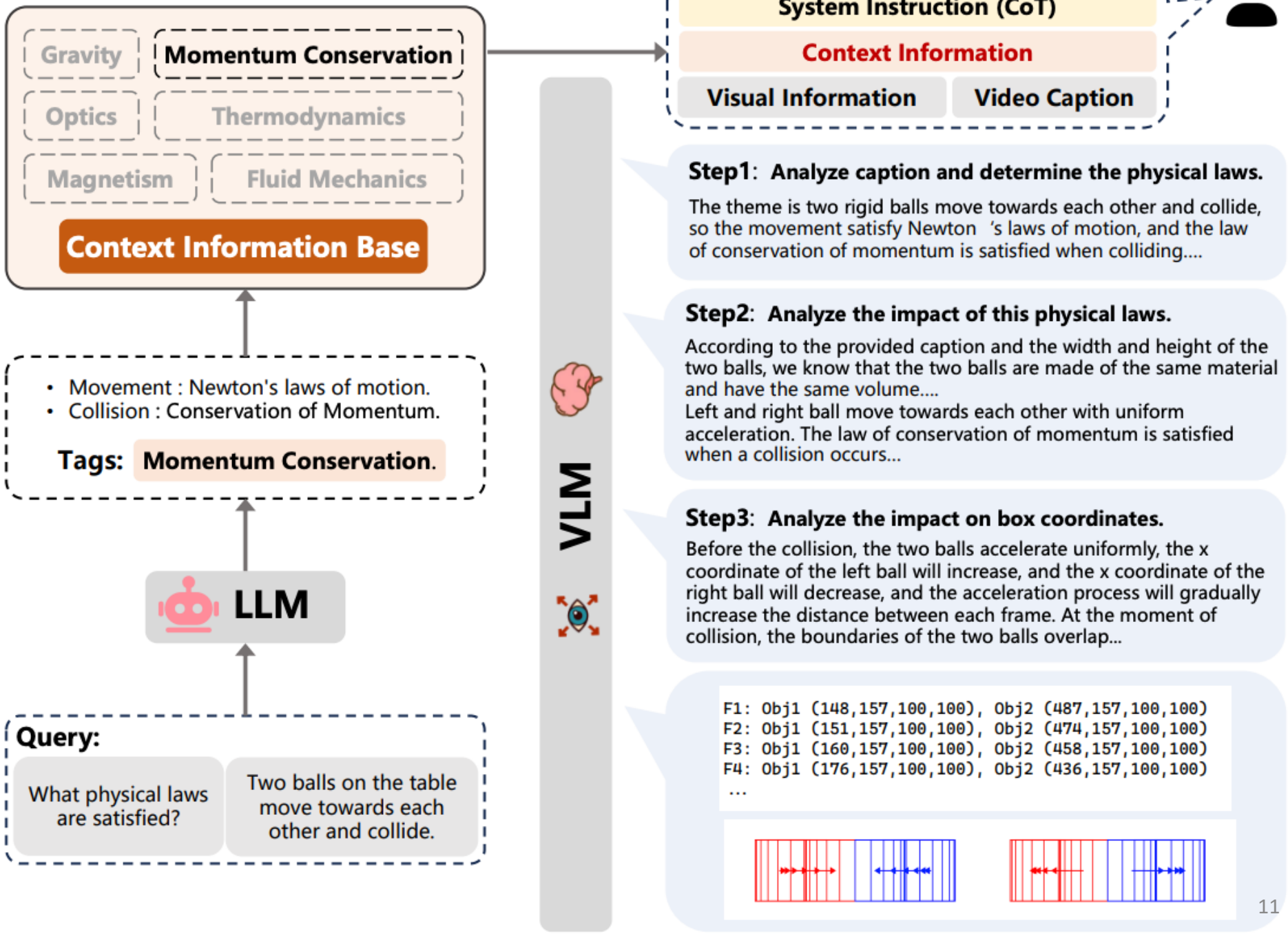
What is a good motion representation to bridge VLM and VDM?

Method

How to plan a coarse-grained, physically plausible motion trajectory?

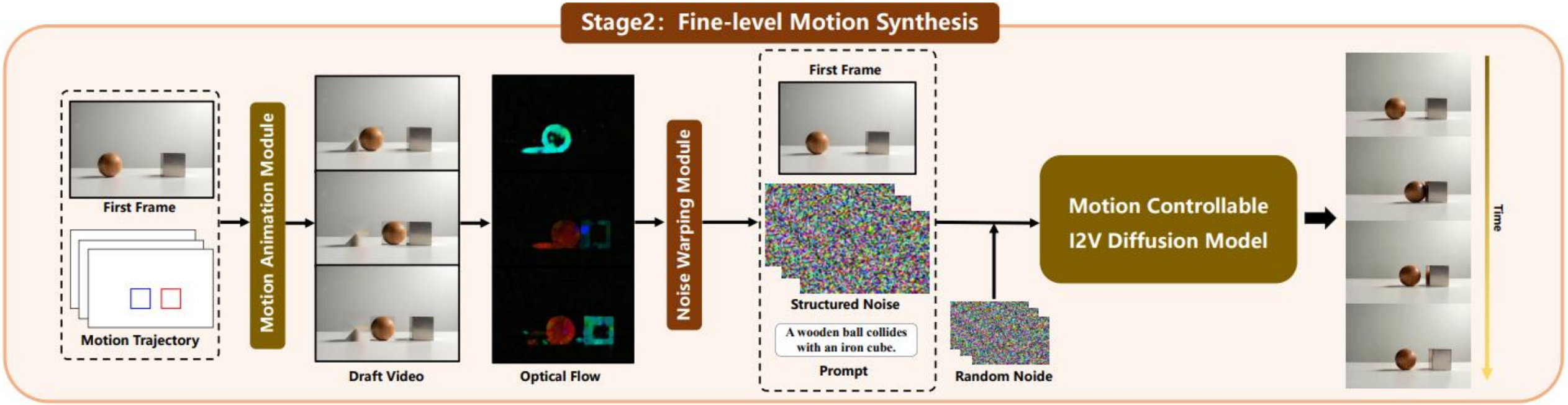


CoT reasoning in VLM



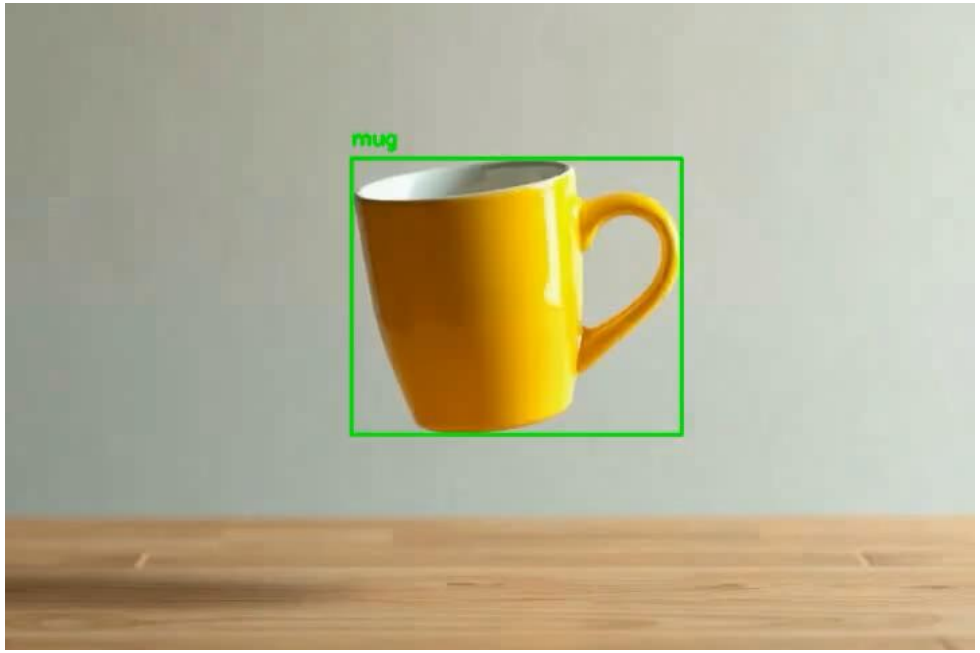
Method

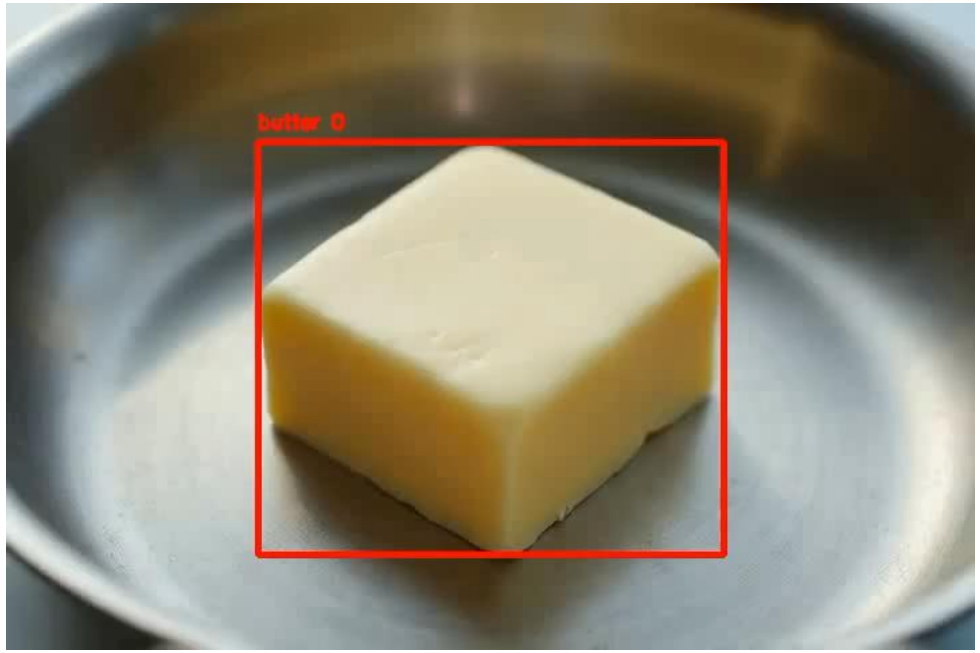
- ***How to incorporate the planned motion into VDM?***



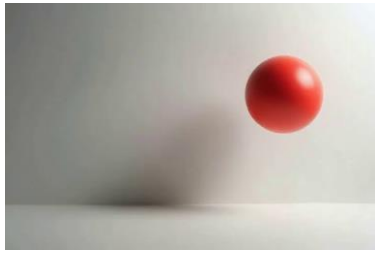
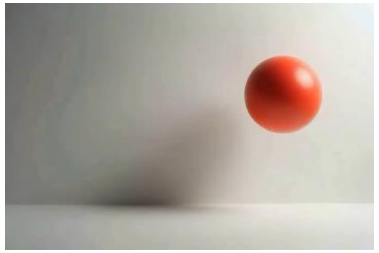
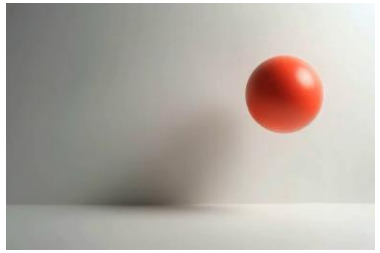
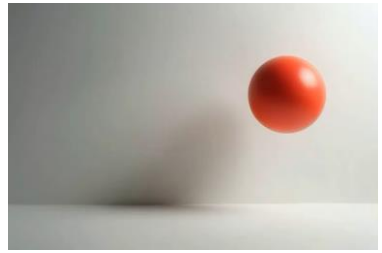
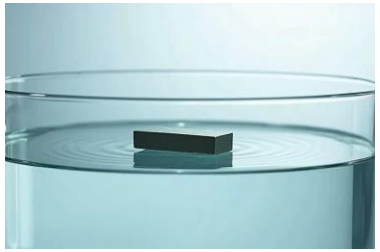
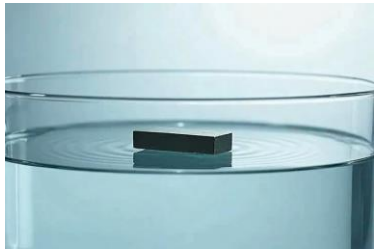
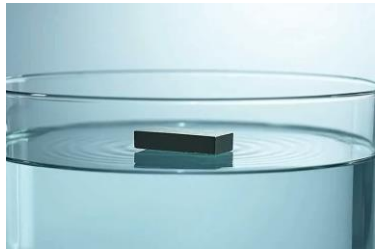
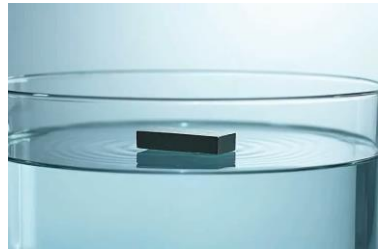
Results

- Using 2D bounding boxes as a coarse motion prior can guide video diffusion models to generate physically plausible videos.





Experimental Results (I2V)

Prompt	Input Frame	Ours	CogVX	LTX
A rubber ball falls under gravity, accelerating as it descends. Upon hitting the ground, it undergoes an elastic collision.				
A piece of iron is gently placed on the surface of the water in a tank filled with water				
A whiskey glass on a wooden table in tavern, and amber-colored whiskey is poured into it from a bottle.				
A timelapse captures the reaction as concentrated sulfuric acid is poured onto a cotton ball.				

Experimental Results

Model	Mechanics(↑)	Optics(↑)	Thermal(↑)	Material(↑)	Average(↑)
CogvideoX-T2V-5B	0.43	0.55	0.40	0.42	0.45
LTX-Video-T2V	0.35	0.45	0.36	0.38	0.39
OpenSora	0.43	0.50	0.44	0.37	0.44
PhyT2V	0.49	0.61	0.49	0.47	0.52
LLM-Grounding Video Diffusion	0.32	0.41	0.26	0.24	0.31
CogvideoX-I2V-5B	0.48	0.69	0.43	0.41	0.52
SVD-XT	0.46	0.68	0.48	0.41	0.52
LTX-Video-I2V	0.47	0.65	0.46	0.37	0.50
SG-I2V	0.52	0.69	0.51	0.39	0.54
Ours	**0.55**	**0.71**	**0.60**	**0.53**	**0.60**

Table 1. Quantitative results of VDMs on PhyGenBench.

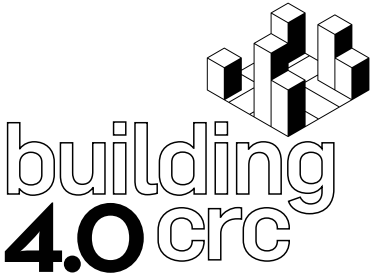
Model	S.M.(↑)	F.D.(↑)	Optics(↑)	Magnetism(↑)	Thermodynamics(↑)	Average(↑)
Cogvideo-I2V-5B	30.4	29.8	16.7	13.3	8.5	27.1
SVD-XT	21.9	20.5	6.8	8.4	**17.1**	19.1
LTX-Video-I2V	30.2	29.8	15.9	13.2	8.4	26.8
SG-I2V	34.6	31.2	15.9	13.1	8.4	29.7
Ours	**42.3**	**34.1**	**16.9**	**13.4**	8.8	**34.6**

Table 2. Quantitative results of physically plausible video generaion on Physics-IQ Benchmark. S.M. refers to Solid Mechanics, and F.D. refers to Fluid Dynamics.

Unified Camera Positional Encoding for Controlled Video Generation

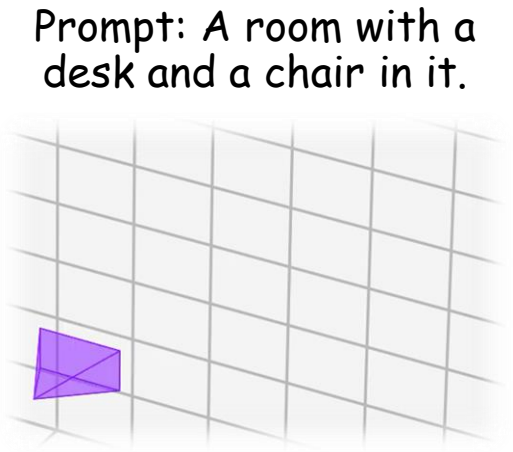
Cheng Zhang[1,2] Boying Li[1] Meng Wei[1]
Yan-Pei Cao[3] Camilo Cruz Gambardella[1,2] Dinh Phung[1] Jianfei Cai[1]

[1]Monash University [2]Building 4.0 CRC [3]VAST



Video Generation as World Model

At its core: Camera-controllable video generation ***as world being seen***!



Input: Camera Pose

CameraCtrl

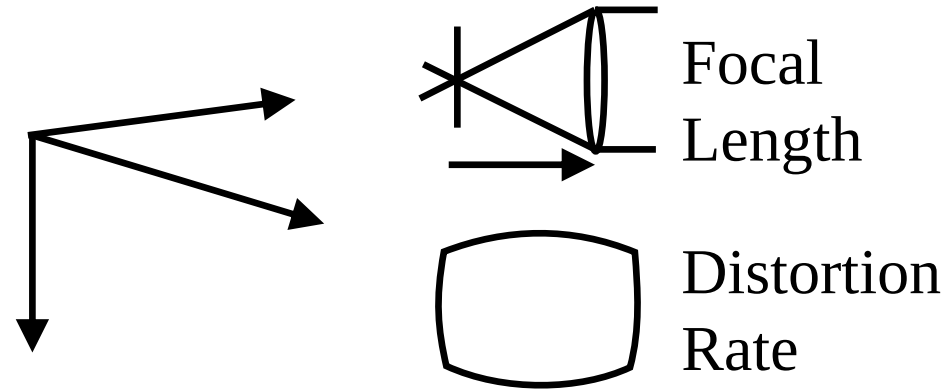
AC3D

What's missed:

Precise & diverse real-world camera geometries with varying intrinsics & distortions.

Video Generation as World Model

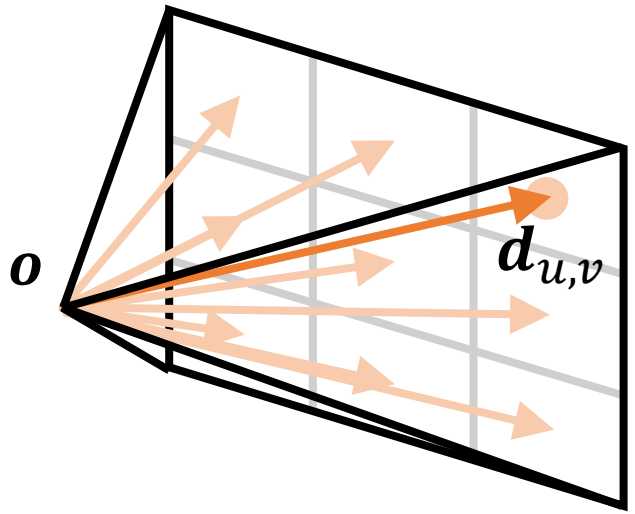
Two current baselines that rely on **absolute camera encodings**:



$\mathbf{T}^{\mathrm{wc}} \in \mathrm{SE}(3)$

ψ: model-specific parameter set

Direct Parameterization



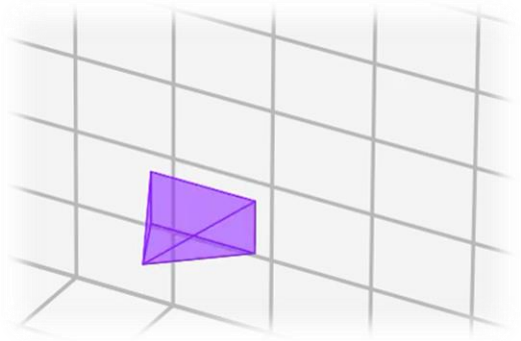
$$(\boldsymbol{d}_{u,v}, \boldsymbol{m}_{u,v}) = (\boldsymbol{d}_{u,v}, \boldsymbol{o} \times \boldsymbol{d}_{u,v})$$

Plücker Encoding

The ‘Absolute’ Trap: Why Current Method Fail

Direct generalization:

Failed motion control



ReCamMaster

(Direct parameterization)

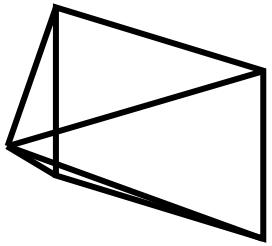
Wan CameraCtrl

(Plücker Encoding)

The ‘Absolute’ Trap: Why Current Method Fail

Direct generalization:

Failed lens control



Pinhole Camera

$\mathrm{xFoV} = 98°, \xi = 0.0$

ReCamMaster

(Direct parameterization)

Wan CameraCtrl

(Plücker Encoding)

The ‘Absolute’ Trap: Why Current Methods Fail

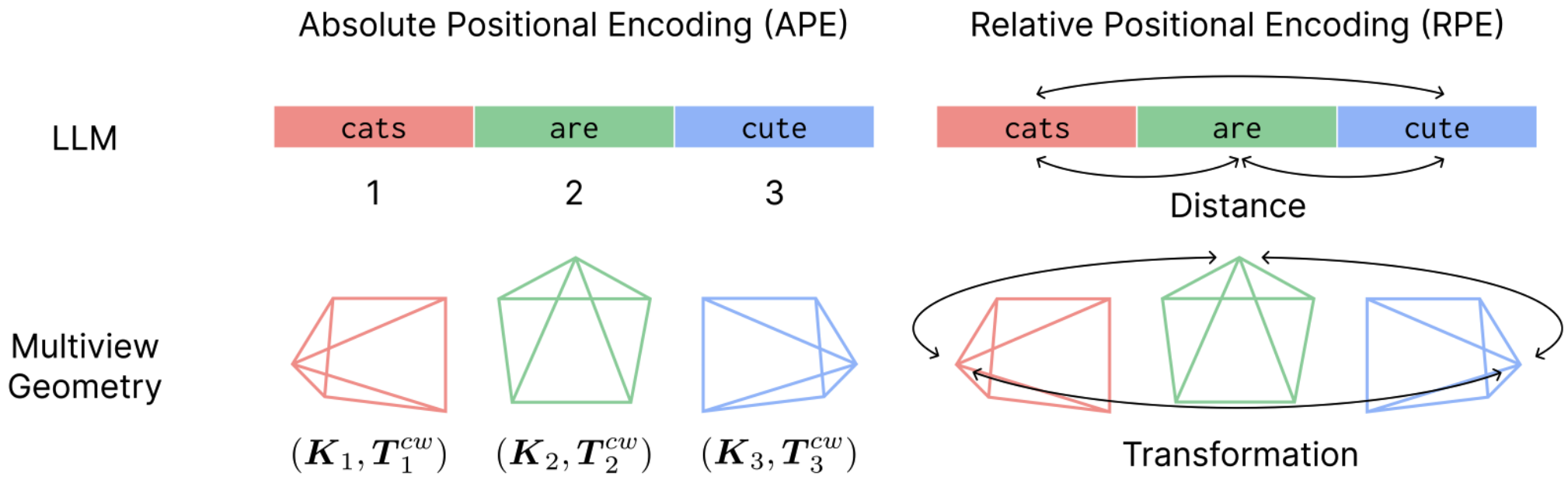
While encoding full camera information, they lack robustness:

- **Poor Generalization**: Fails to adapt to different datasets, coordinate set-ups, trajectories, camera lenses
- **Limited Controllability**: Inaccurate camera motion and lens control

Not Agent sees in the world!
The Root Cause: Coordinate Sensitivity

Previous Solution: PRoPE

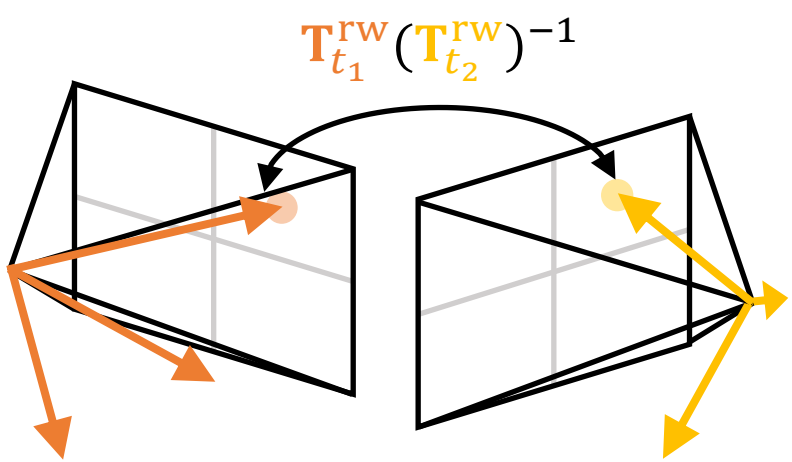
PRoPE (Projective Positional Encoding) Moves from absolute coordinates to relative relationships:



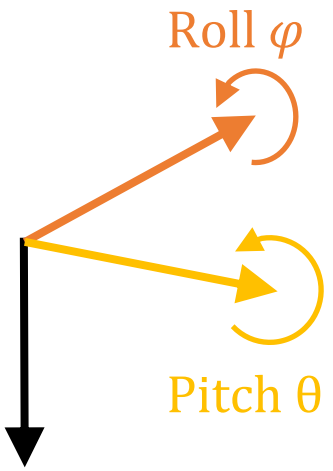
PRoPE limitation: linear projective transformations by assuming a perfect Pinhole Camera.

Unified Camera Positional Encoding (UCPE)

Decouple



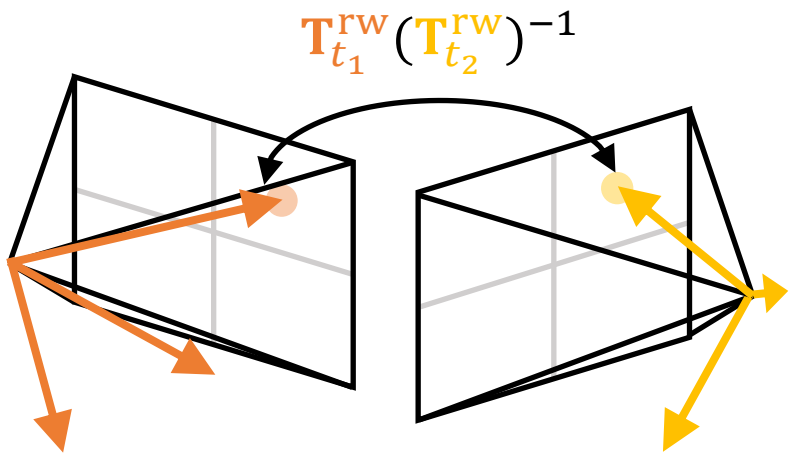
Relative Ray Encoding
Relative transformations between local ray coordinates



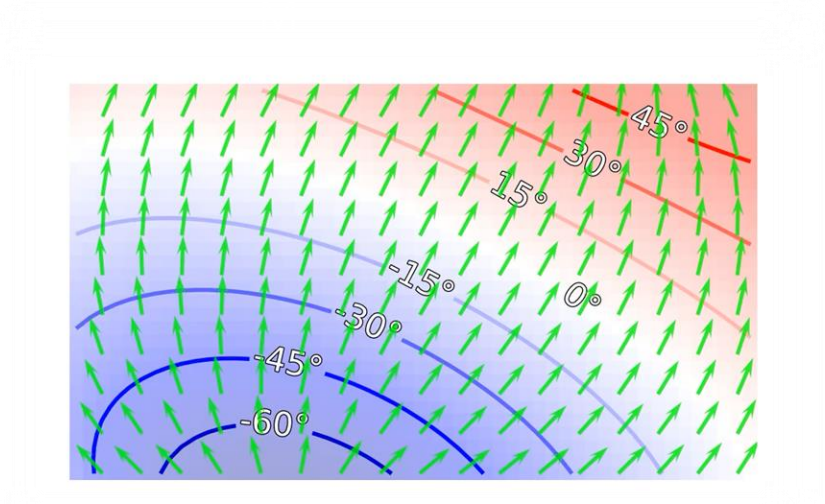
Absolute Orientation
Resolves the ambiguity of the first frame orientation

Unified Camera Positional Encoding (UCPE)

Decouple

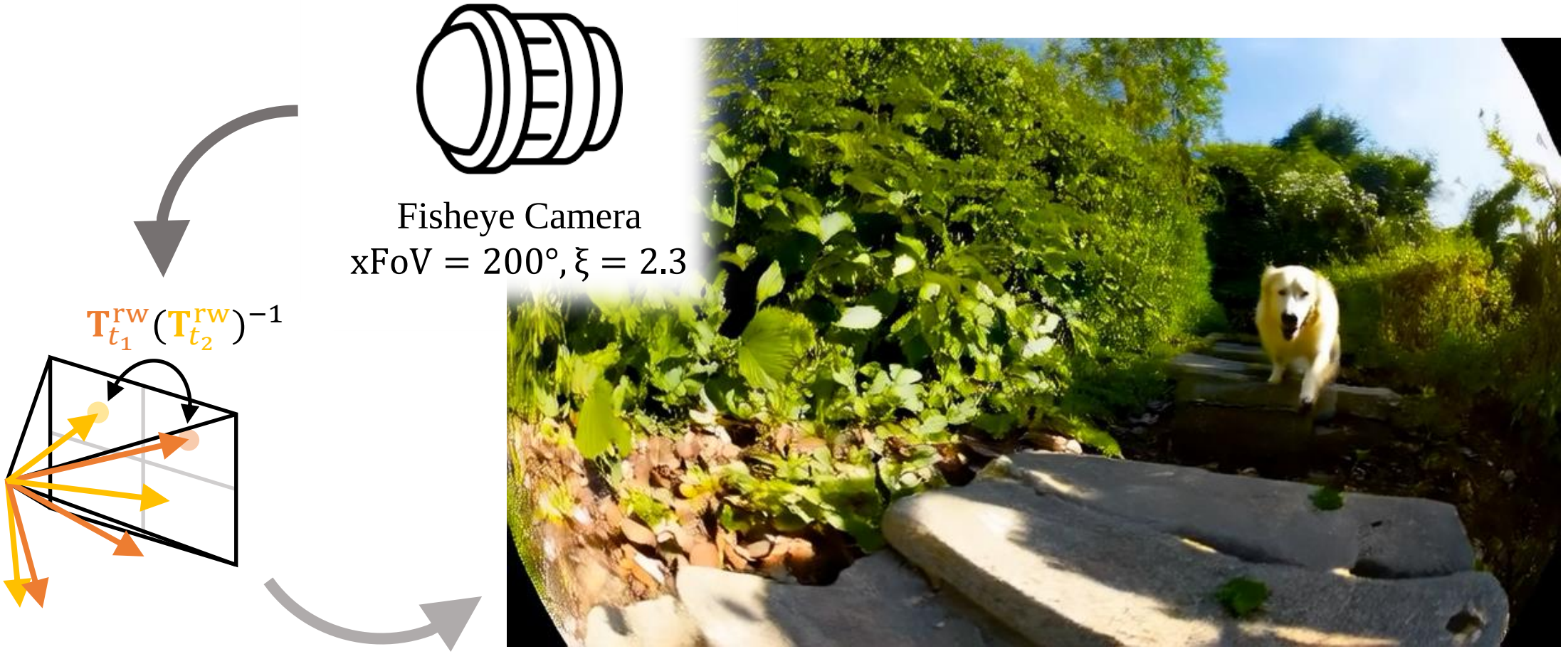


Relative Ray Encoding
Relative transformations between local ray coordinates



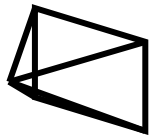
Absolute Orientation Encoding
1-dim Latitude Map + 2-dim Up Map

Relative Ray Encoding: Lens Control

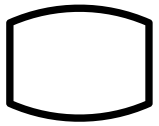


*We demonstrate with the Unified Camera Model (UCM) using distortion ξ, but the method is not restricted to UCM.

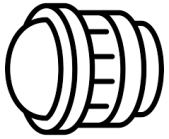
Relative Ray Encoding: Lens Control



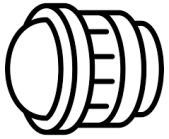
Pinhole Camera
$\text{xFoV} = 100°, \xi = 0.0$



Wide-angle Camera
$\text{xFoV} = 140°, \xi = 0.8$



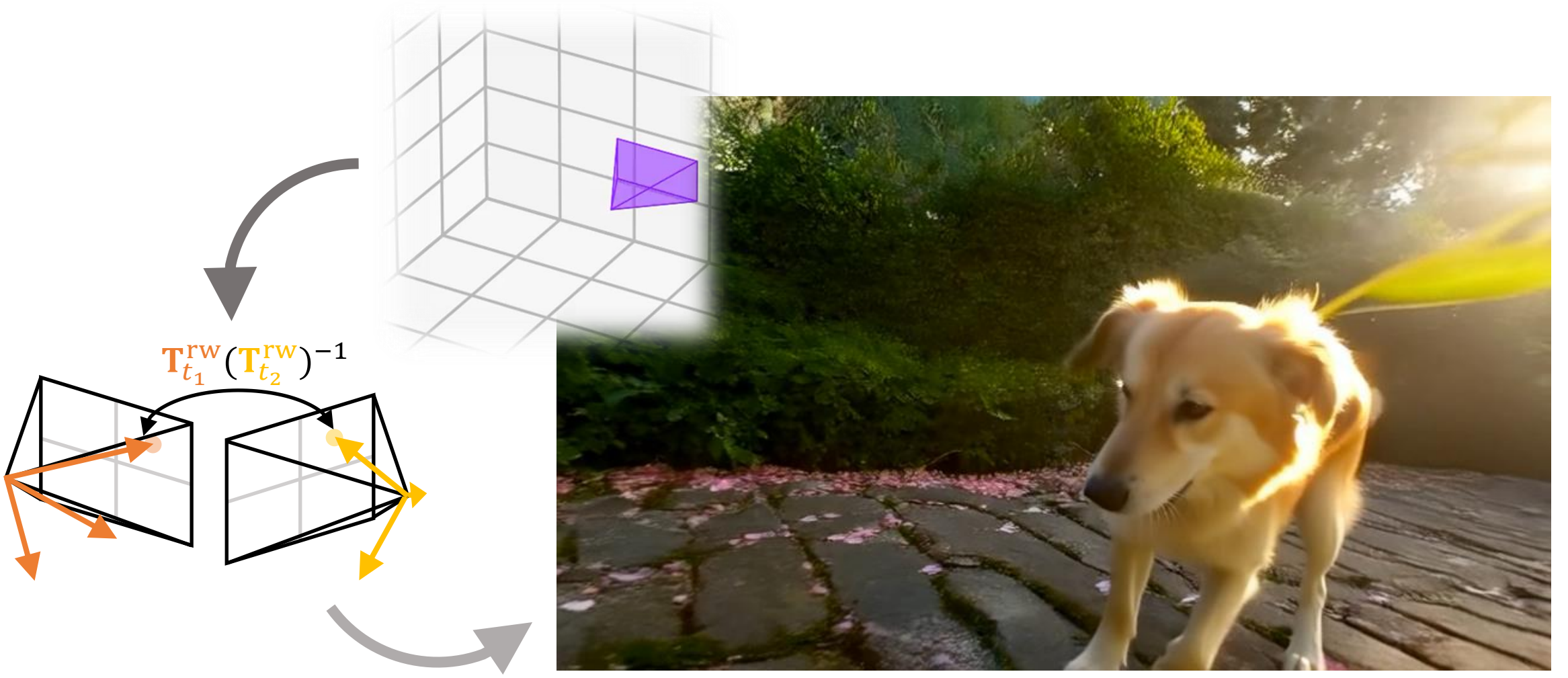
Fisheye Camera
$\text{xFoV} = 180°, \xi = 2.0$



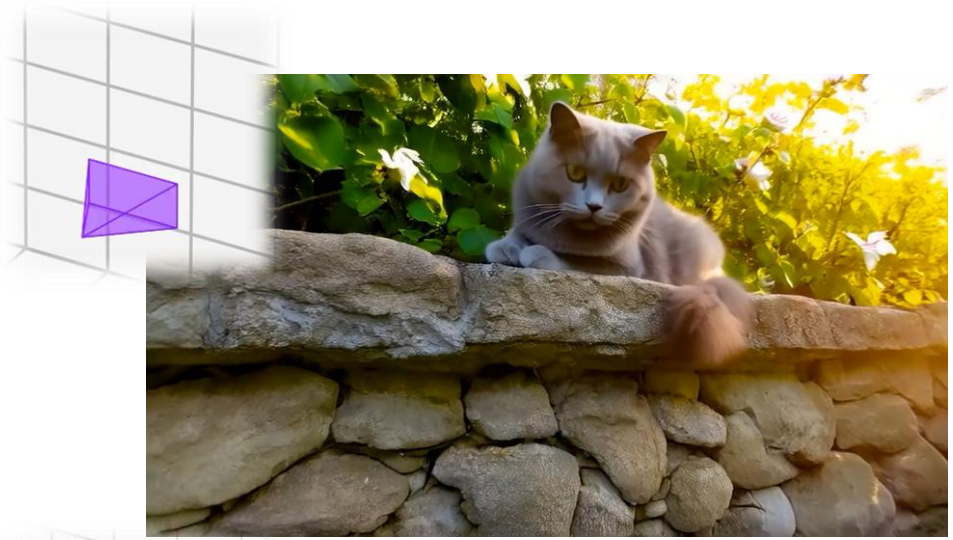
Fisheye Camera
$\text{xFoV} = 200°, \xi = 2.3$

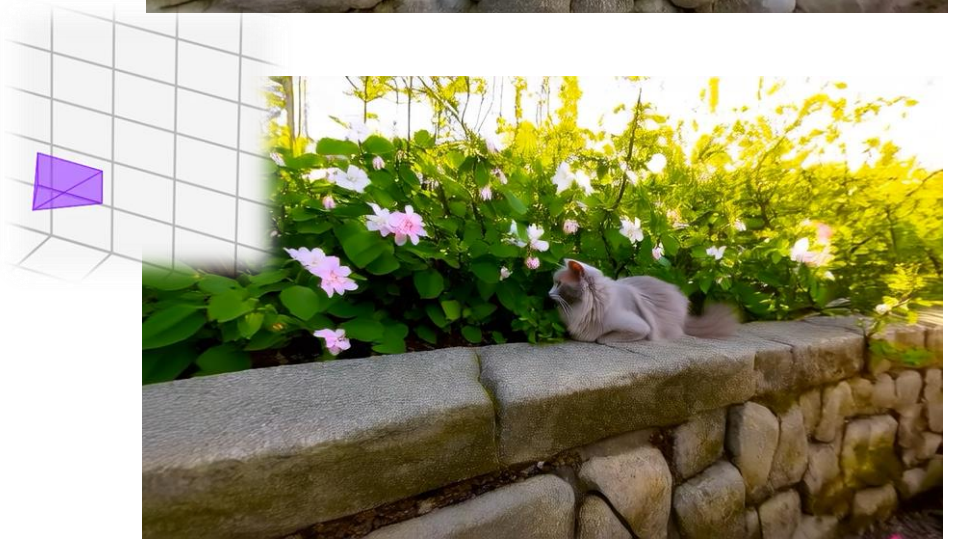
*We demonstrate with the Unified Camera Model (UCM) using distortion ξ, but the method is not restricted to UCM.

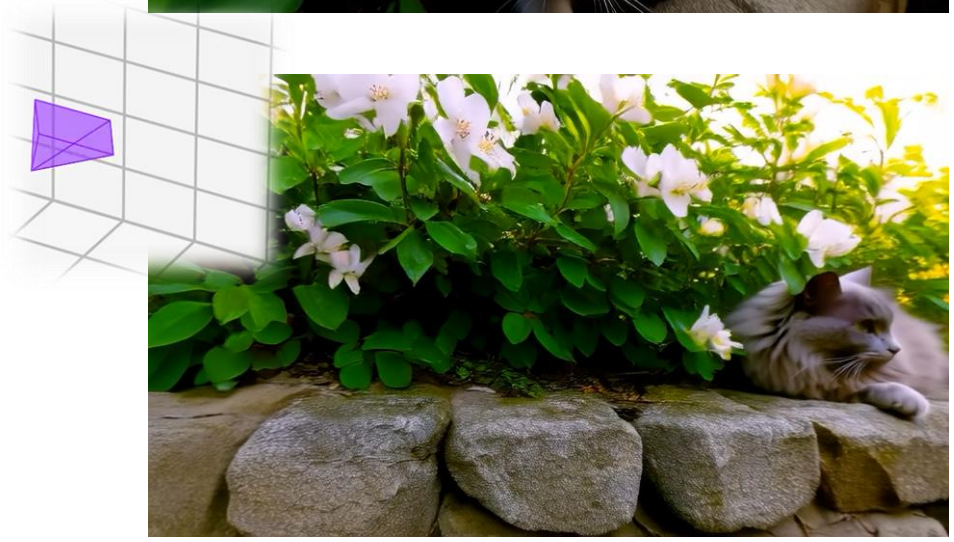
Relative Ray Encoding: Pose Control



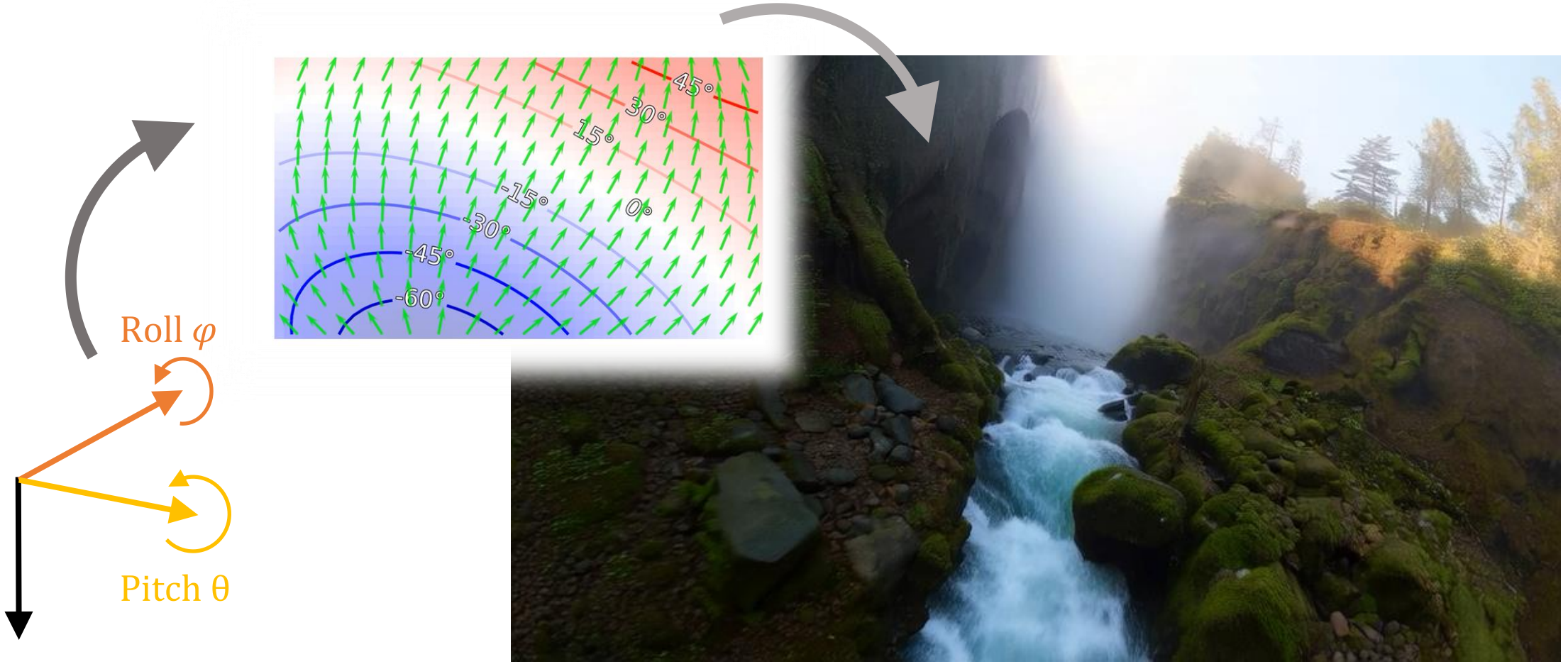
Relative Ray Encoding: Pose Control



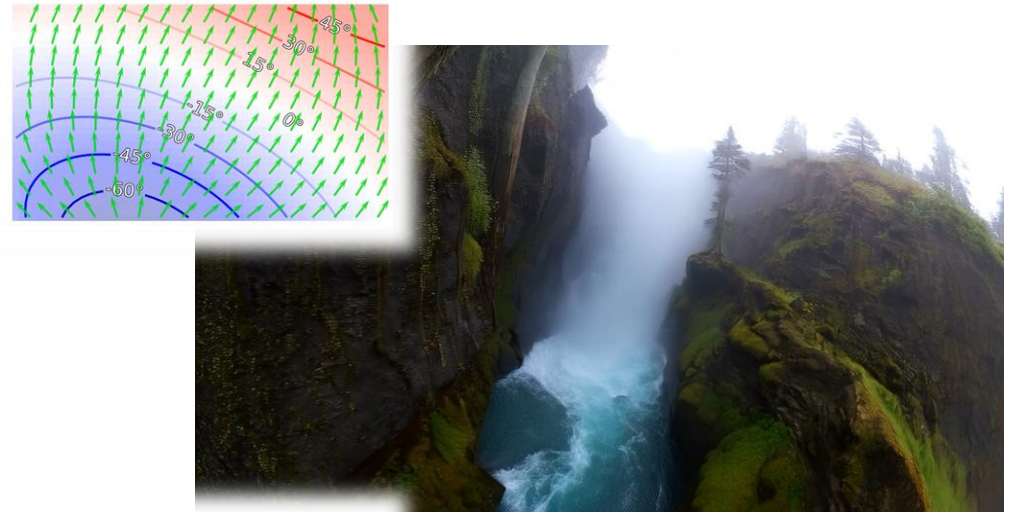


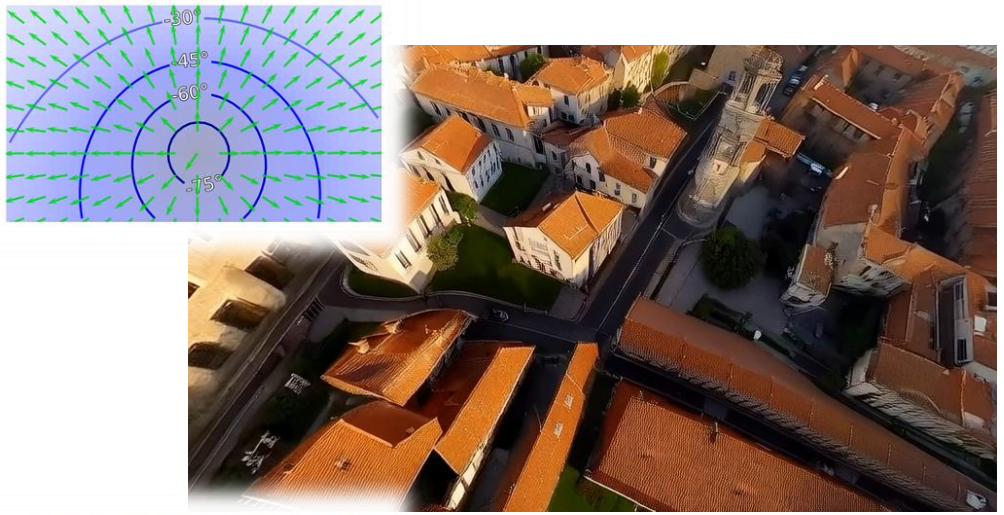


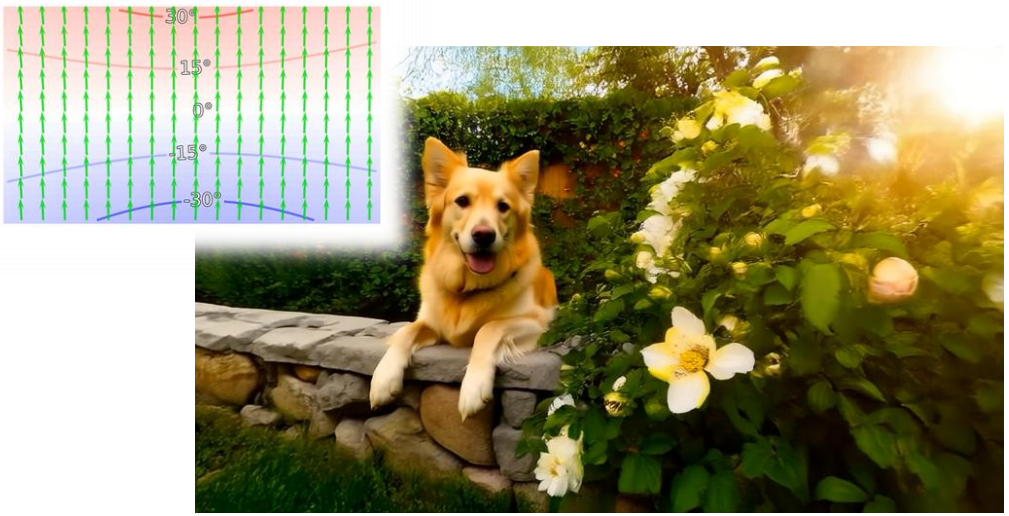
Absolute Orientation Encoding

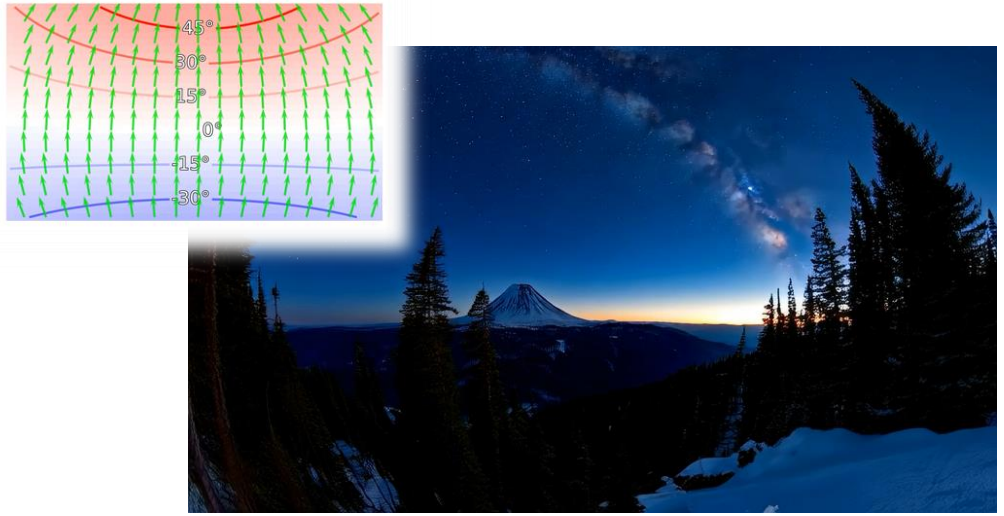


Absolute Orientation Encoding

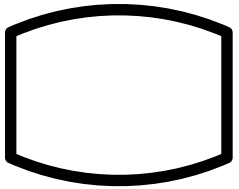






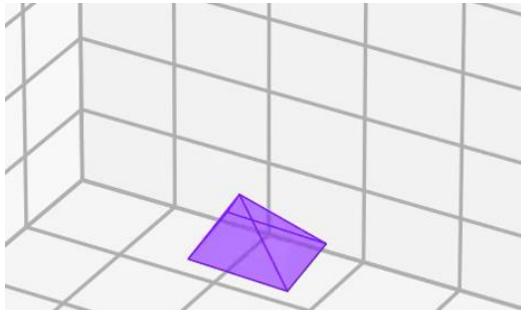


Unified Camera Positional Encoding (UCPE)

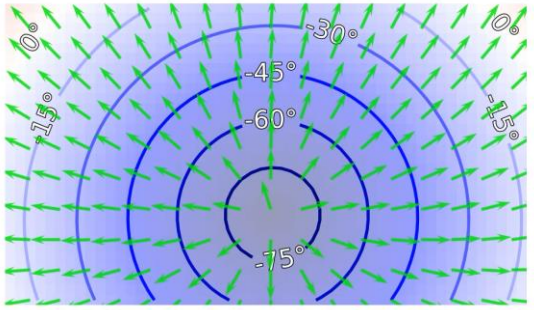


Wide-angle Camera
$\mathrm{xFoV} = 160°, \xi = 1.5$

Lens

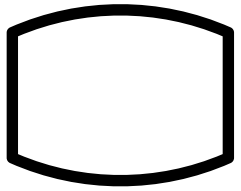


Pose

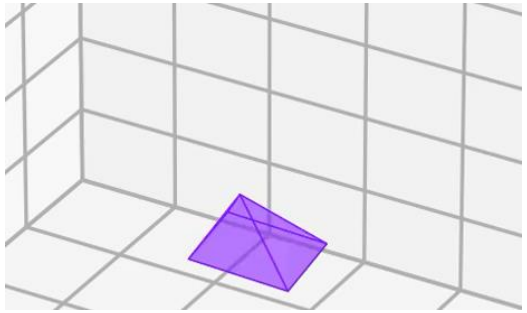


Orientation

Unified Camera Positional Encoding (UCPE)



Wide-angle Camera
$\text{xFoV} = 160°, \xi = 1.5$



Construct a large-scale synthetic video dataset spanning **diverse intrinsics, distortions, and camera motions**.




Usage of UCPE

How to integrate UCPE into networks?

- Relative Ray Encoding: Linear transform tokens like **RoPE**

$$\boldsymbol{q}_i = \mathbf{W}_q \boldsymbol{x}_i$$
$$\boldsymbol{k}_j = \mathbf{W}_k \boldsymbol{x}_j$$

Usage of UCPE

How to integrate UCPE into networks?

- Relative Ray Encoding: Linear transform tokens like **RoPE**

$$\boldsymbol{q}_i = \mathbf{R}_i \mathbf{W}_q \boldsymbol{x}_i$$
$$\boldsymbol{k}_j = \mathbf{R}_j \mathbf{W}_k \boldsymbol{x}_j$$

Usage of UCPE

How to integrate UCPE into networks?

- Relative Ray Encoding: Linear transform tokens like **RoPE**

$$\begin{aligned} \boldsymbol{q}_i &= \mathbf{R}_i \mathbf{W}_q \boldsymbol{x}_i \\ \boldsymbol{k}_j &= \mathbf{R}_j \mathbf{W}_k \boldsymbol{x}_j \end{aligned} \longrightarrow \quad \boldsymbol{q}_i^\top \boldsymbol{k}_j = (\mathbf{R}_i \mathbf{W}_q \boldsymbol{x}_i)^\top (\mathbf{R}_j \mathbf{W}_k \boldsymbol{x}_j)$$

Usage of UCPE

How to integrate UCPE into networks?

- Relative Ray Encoding: Linear transform tokens like **RoPE**

$$\begin{aligned} \boldsymbol{q}_i &= \mathbf{R}_i \mathbf{W}_q \boldsymbol{x}_i \\ \boldsymbol{k}_j &= \mathbf{R}_j \mathbf{W}_k \boldsymbol{x}_j \end{aligned} \longrightarrow \quad \boldsymbol{q}_i^\top \boldsymbol{k}_j = (\mathbf{W}_q \boldsymbol{x}_i)^\top \mathbf{R}_i{}^\top \mathbf{R}_j (\mathbf{W}_k \boldsymbol{x}_j)$$

Usage of UCPE

How to integrate UCPE into networks?

- Relative Ray Encoding: Linear transform tokens like **RoPE**

$$\begin{aligned} \boldsymbol{q}_i &= \mathbf{R}_i \mathbf{W}_q \boldsymbol{x}_i \\ \boldsymbol{k}_j &= \mathbf{R}_j \mathbf{W}_k \boldsymbol{x}_j \end{aligned} \longrightarrow \quad \boldsymbol{q}_i^\top \boldsymbol{k}_j = (\mathbf{W}_q \boldsymbol{x}_i)^\top \mathbf{R}_{ij} (\mathbf{W}_k \boldsymbol{x}_j)$$

Usage of UCPE

How to integrate UCPE into networks?

- **Relative Ray Encoding**: Linear transform tokens like RoPE

$$\begin{aligned} \boldsymbol{q}_i &= \mathbf{T}_i^\top \mathbf{W}_q \boldsymbol{x}_i \\ \boldsymbol{k}_j &= \mathbf{T}_j^{-1} \mathbf{W}_k \boldsymbol{x}_j \end{aligned} \longrightarrow \quad \boldsymbol{q}_i^\top \boldsymbol{k}_j = (\mathbf{W}_q \boldsymbol{x}_i)^\top \mathbf{T}_{ij} (\mathbf{W}_k \boldsymbol{x}_j)$$

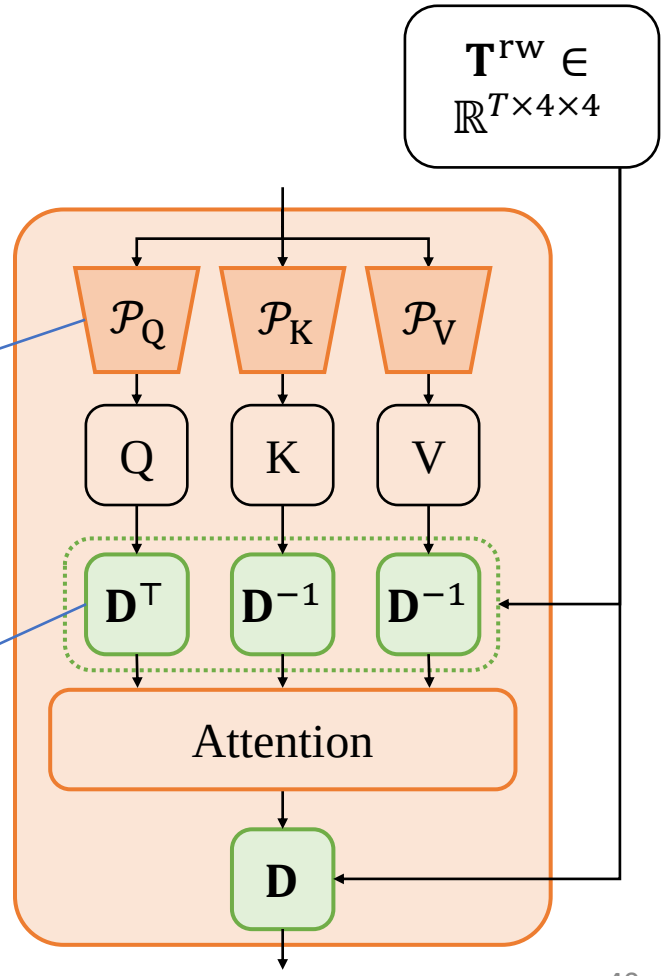
Usage of UCPE

How to integrate UCPE into networks?

- Relative Ray Encoding: Linear transforms tokens like RoPE

One attention head is enough
(Only 0.5% trainable parameters)

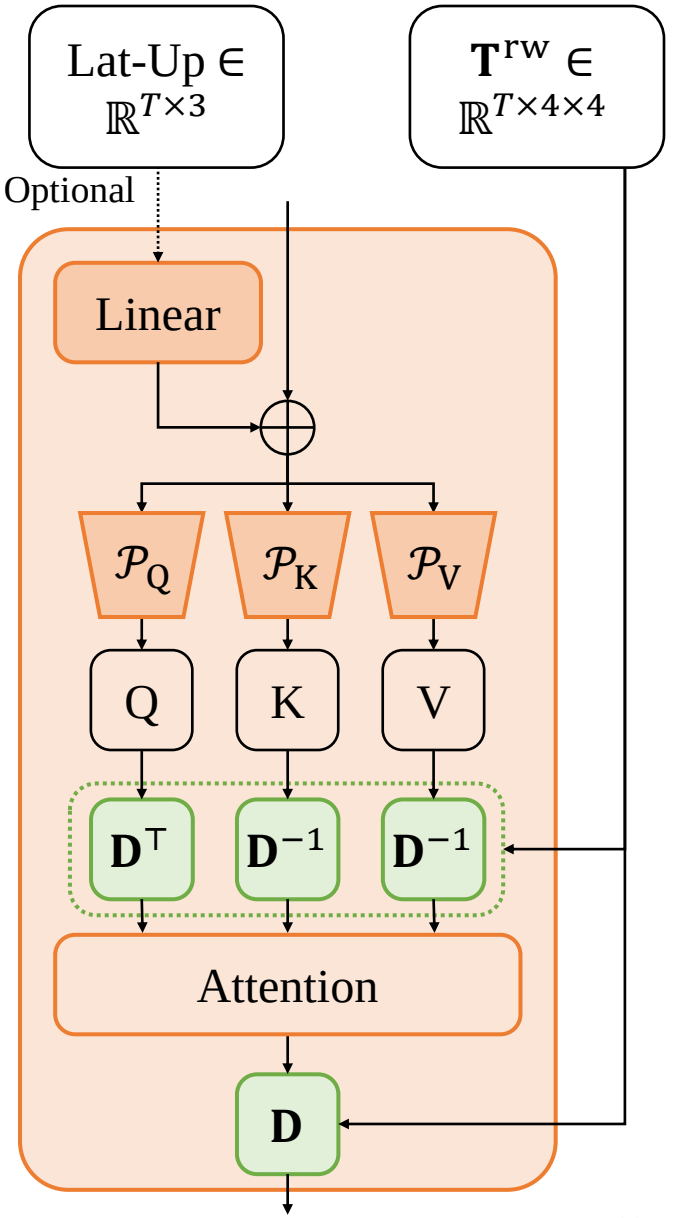
Block diagonal matrix of $\mathbf{T}^{\mathrm{rw}}$



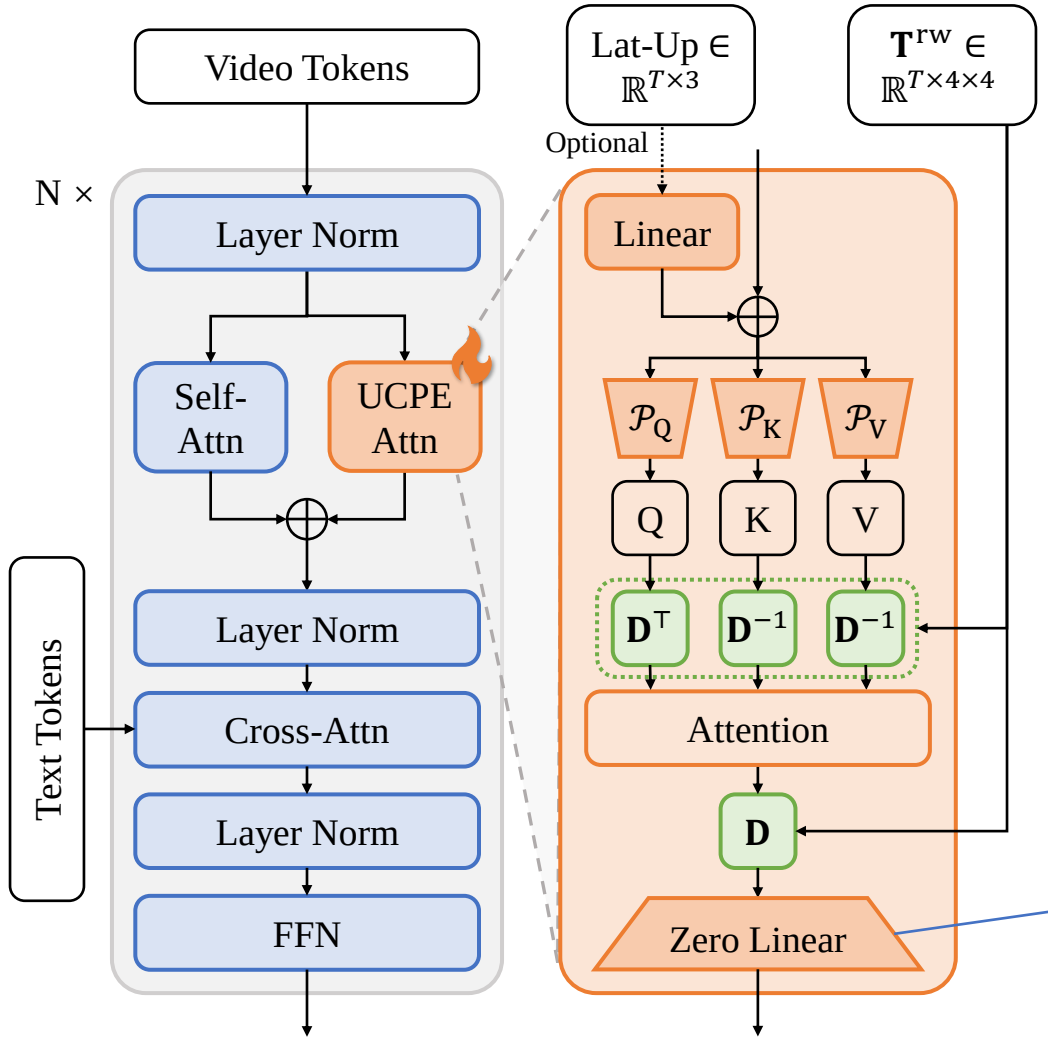
Usage of UCPE

How to integrate UCPE into networks?

- Relative Ray Encoding: Linear transforms tokens like RoPE
- Absolute Positional Encoding: Linear project and add as bias before attention



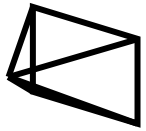
Integrated into a self-attention layer parallel with the original one (Similar to LoRA).



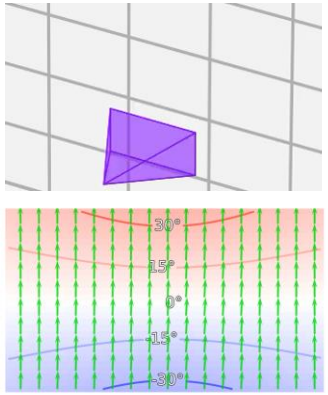
Zero-initialized Linear layer to keep the the pretrained priors at initialization

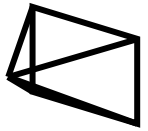
More Results

UCPE supports a wide range of applications, from cinematic content creation to autonomous driving and embodied AI.

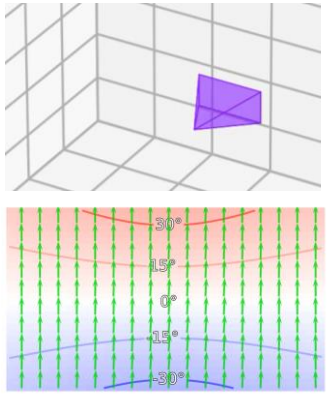


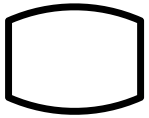
Pinhole Camera
$\text{xFoV} = 100°, \xi = 0.0$



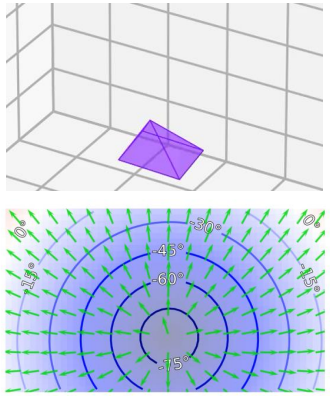


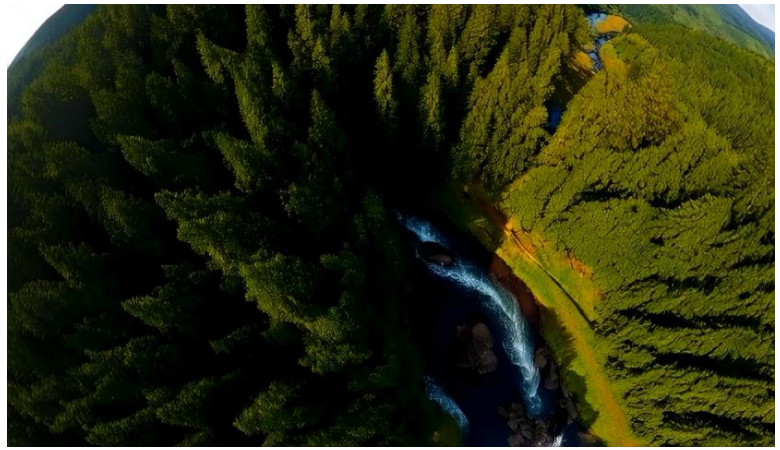
Pinhole Camera
$\text{xFoV} = 100°, \xi = 0.0$

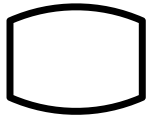




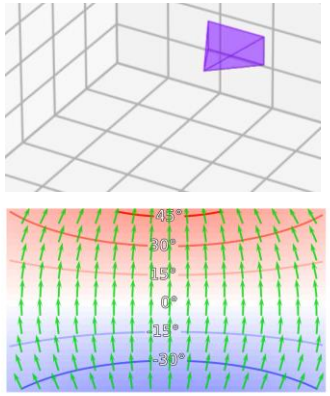
Wide-angle Camera
$\text{xFoV} = 160°, \xi = 1.5$



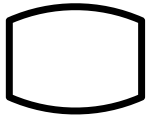




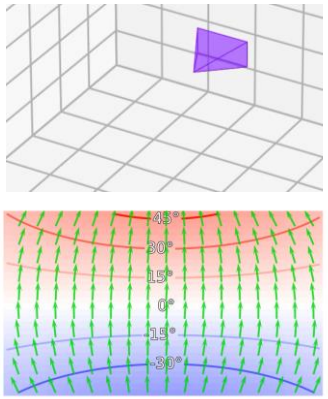
Wide-angle Camera
$\text{xFoV} = 160°, \xi = 1.5$



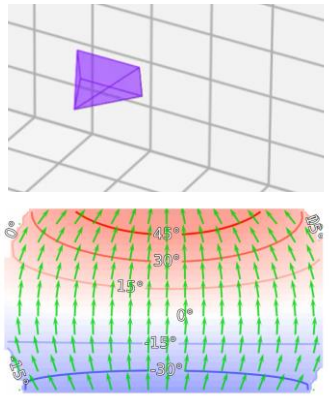
UCPE supports a wide range of applications, from cinematic content creation to autonomous driving and embodied AI.

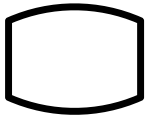


Wide-angle Camera
$\text{xFoV} = 160°, \xi = 1.5$

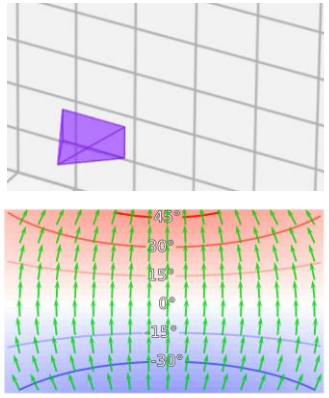


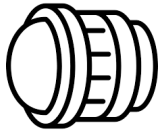
Fisheye Camera
$\text{xFoV} = 200°, \xi = 2.3$



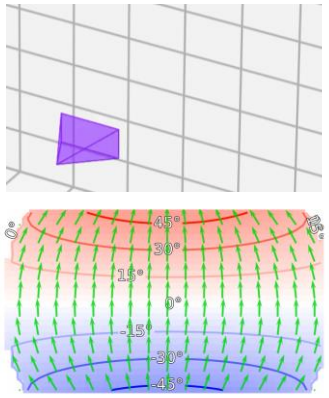


Wide-angle Camera
$\text{xFoV} = 160°, \xi = 1.5$



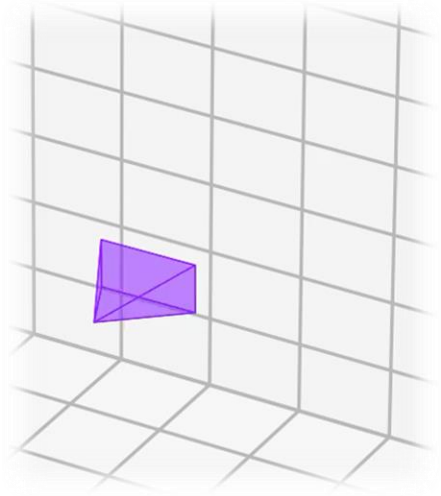


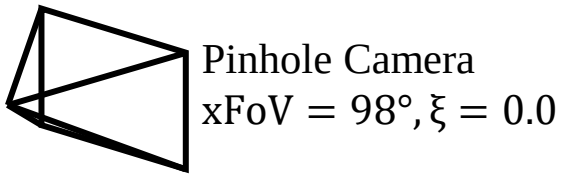
Fisheye Camera
$\text{xFoV} = 200°, \xi = 2.3$

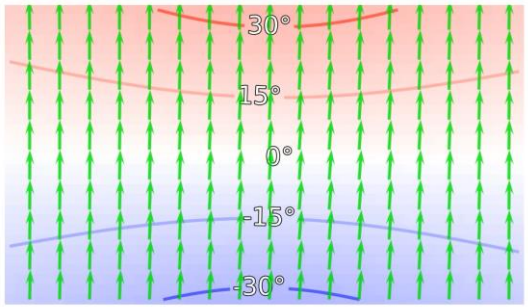


Comparison on
Our Synthetic Dataset

The video begins with a view of an urban street scene…







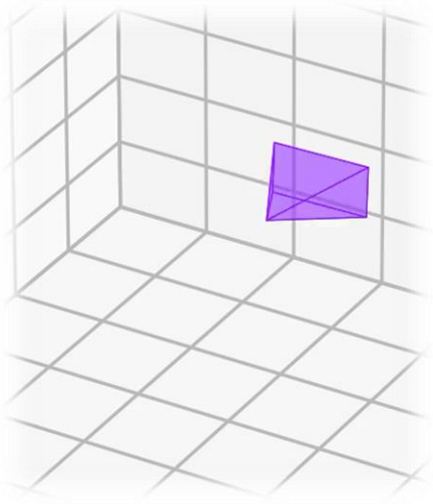
Inputs

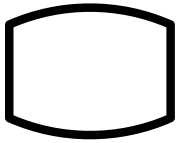
ReCamMaster

Wan CameraCtrl

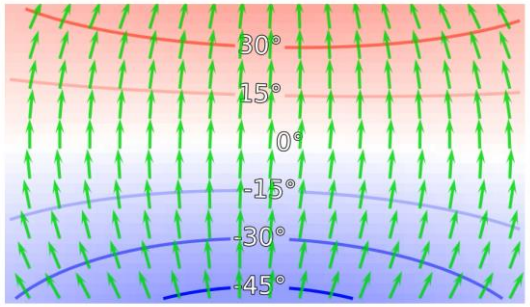
UCPE

The video showcases an outdoor event under a clear blue sky…



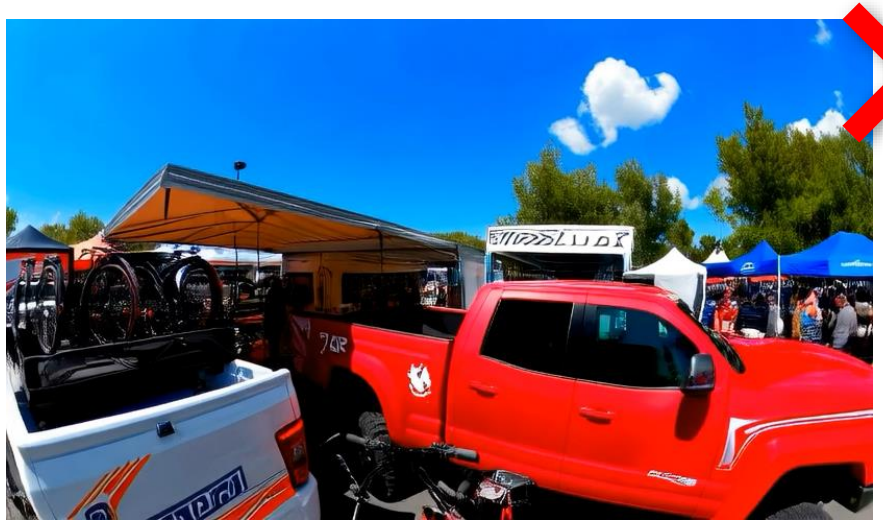


Wide-angle Camera
$\text{xFoV} = 160°, \xi = 1.5$



Inputs

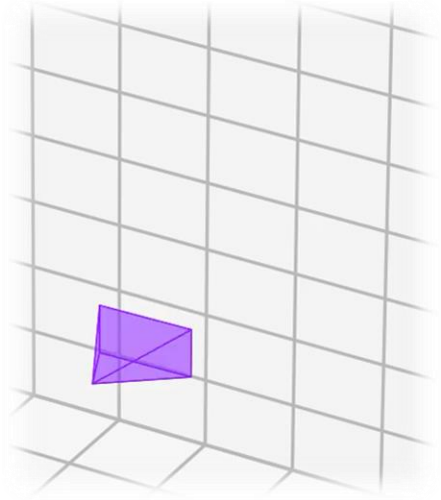
ReCamMaster

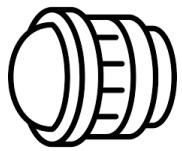


Wan CameraCtrl

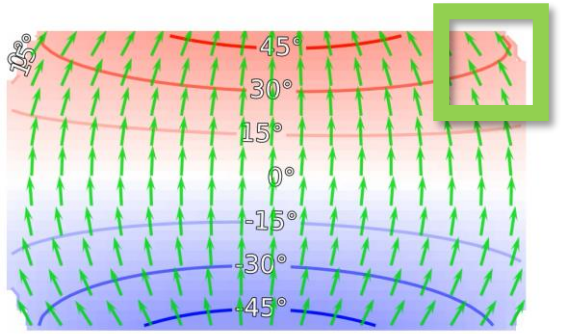
UCPE

The video showcases a vibrant and sunny outdoor setting…



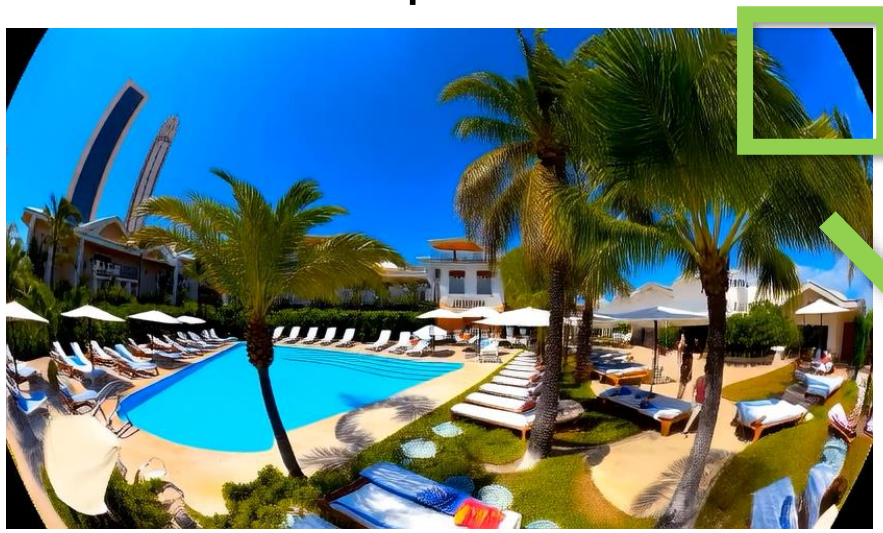


Fisheye Camera
$xFoV = 195°, \xi = 2.13$



Inputs

ReCamMaster



Wan CameraCtrl

UCPE

Quantitative Comparison

(1) Better than other absolute encodings

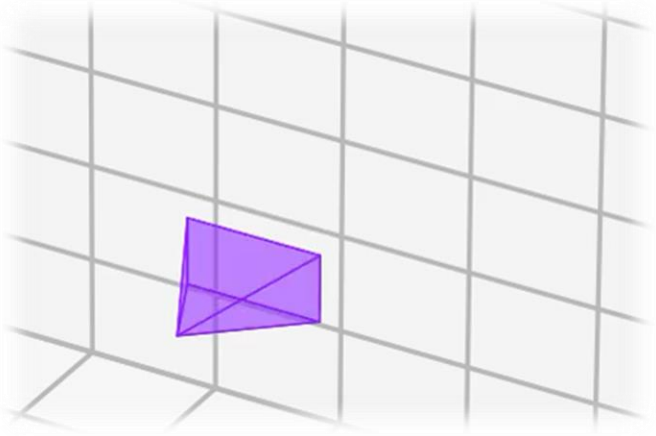
(2) Only 0.5% trainable parameters

Settings	Method	Camera Lens Control			Absolute Orientation		Relative Camera Pose Control			Video Generation Quality			Trainable Params
		FoV (°)↓	k_1↓	k_2↓	Pitch (°)↓	Roll (°)↓	RotErr (°)↓	TransErr↓	CamMC↓	FVD↓	FID↓	CLIP↑	
w/o Absolute Orientation Control	ReCamMaster	10.25	0.210	0.143	10.00	7.42	10.89	31.44	37.38	555.54	69.91	24.86	354M
	Wan CameraCtrl	10.05	0.222	0.142	9.35	7.31	17.04	35.09	46.10	593.10	67.83	25.05	1.5B
	UCPE	9.62	0.174	0.120	8.03	6.64	4.29	13.46	15.94	569.31	66.22	25.11	35.5M
w/ Absolute Orientation Control	ReCamMaster	10.04	0.183	0.136	6.62	5.29	9.23	28.95	33.88	605.83	67.07	24.84	354M
	Wan CameraCtrl	9.86	0.230	0.162	6.25	6.01	17.92	39.16	50.32	554.43	65.73	25.01	1.5B
	UCPE	8.22	0.129	0.102	4.35	3.74	4.12	15.21	17.59	495.14	63.37	25.12	35.6M
w/ Absolute Orientation Control & Ablation Study	1/2-dim (128 × 6)	8.39	0.170	0.110	4.11	3.93	3.69	14.03	16.06	534.44	64.88	25.05	141M
	1/4-dim (128 × 3)	8.47	0.149	0.101	3.94	4.08	3.43	14.26	16.02	512.85	62.86	25.09	71.0M
	1/8-dim (192 × 1)	8.22	0.129	0.102	4.35	3.74	4.12	15.21	17.59	495.14	63.37	25.12	35.6M
	1/12-dim (128 × 1)	8.96	0.151	0.107	3.91	3.89	5.13	14.54	17.84	487.54	62.98	25.06	23.8M
	Pre-Attn	8.47	0.145	0.108	4.26	3.96	4.03	15.77	17.85	502.73	63.62	25.07	35.6M
	Post-Attn	8.91	0.147	0.116	3.95	3.92	4.68	17.47	20.00	515.32	64.65	25.00	35.6M
	PRoPE	8.84	0.151	0.113	4.18	3.70	5.35	17.52	20.58	516.59	65.00	25.03	35.6M
	GTA	8.80	0.157	0.117	4.21	3.69	5.27	17.07	20.14	497.19	64.91	25.04	35.6M

(3) Better than other relative encodings

Comparison on RealEstate10k Dataset

An aerial view of a lake surrounded by houses.



Camera Pose

Severe artifacts

CameraCtrl

Unbalanced framing

AC3D

Not moving

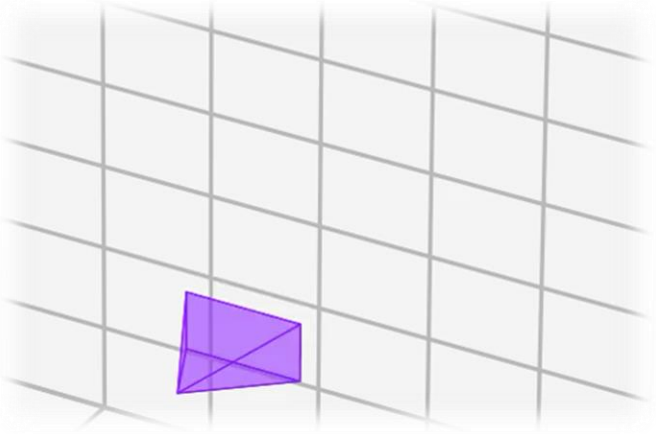
ReCamMaster

Wrong direction

Wan CameraCtrl

UCPE

A view of a dining room table and chairs.



Camera Pose

Poor composition

CameraCtrl

Less details

AC3D

Undesired distortion

ReCamMaster

Undesired distortion

Wan CameraCtrl

UCPE

Quantitative Comparison

Method	Relative Camera Pose Control			Q-Align Scores		
	RotErr (°)↓	TransErr↓	CamMC↓	Image Quality↑	Image Aesthetic↑	Video Quality↑
ReCamMaster	1.10	5.64	6.15	0.9492	0.5185	0.9720
Wan CameraCtrl	2.22	7.42	8.67	0.9822	0.5691	0.9885
CameraCtrl	1.17	3.96	4.59	0.6877	0.3306	0.7338
AC3D	0.62	2.11	2.43	0.7699	0.3651	0.8211
UCPE	0.56	1.25	1.58	0.9480	0.4686	0.9694

(4) Direct generalize to unseen data and achieve SoTA camera control

Take home messages

- Joint VLM and VDM can generate physically plausible videos.
- UCPE enables precise&diverse camera control for video generation.

Match what agents see in the world

Future directions

- *Moving world models from being seen to being experienced!*

- *What is a good representation for world models?*
- *How to ensure long-term consistency?*
- *How to do multi-modal world model?*
- *How to provide multi-agent world model?*
- *How to unify world model with LLMs?*
- *How to make use of world model in Agentic AI?*

- *Don’t forget the social impact: deepfake …*

Thank you!