Exploring Video World Models: Memory, Space, and Perception

Xingang Pan

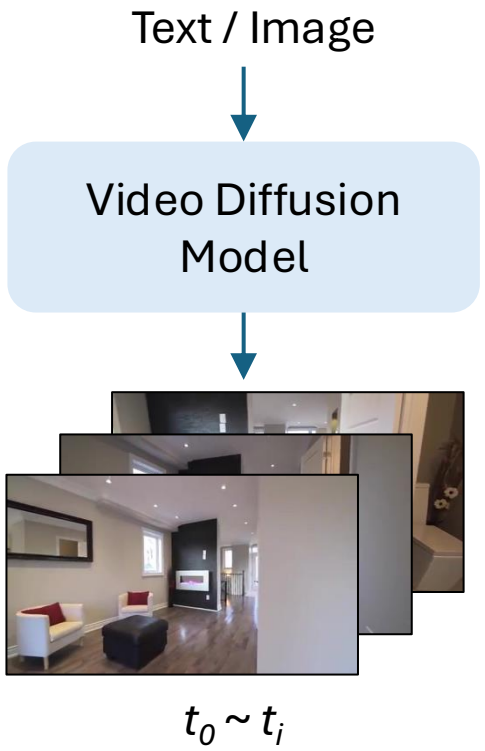
Assistant Professor

Nanyang Technological University

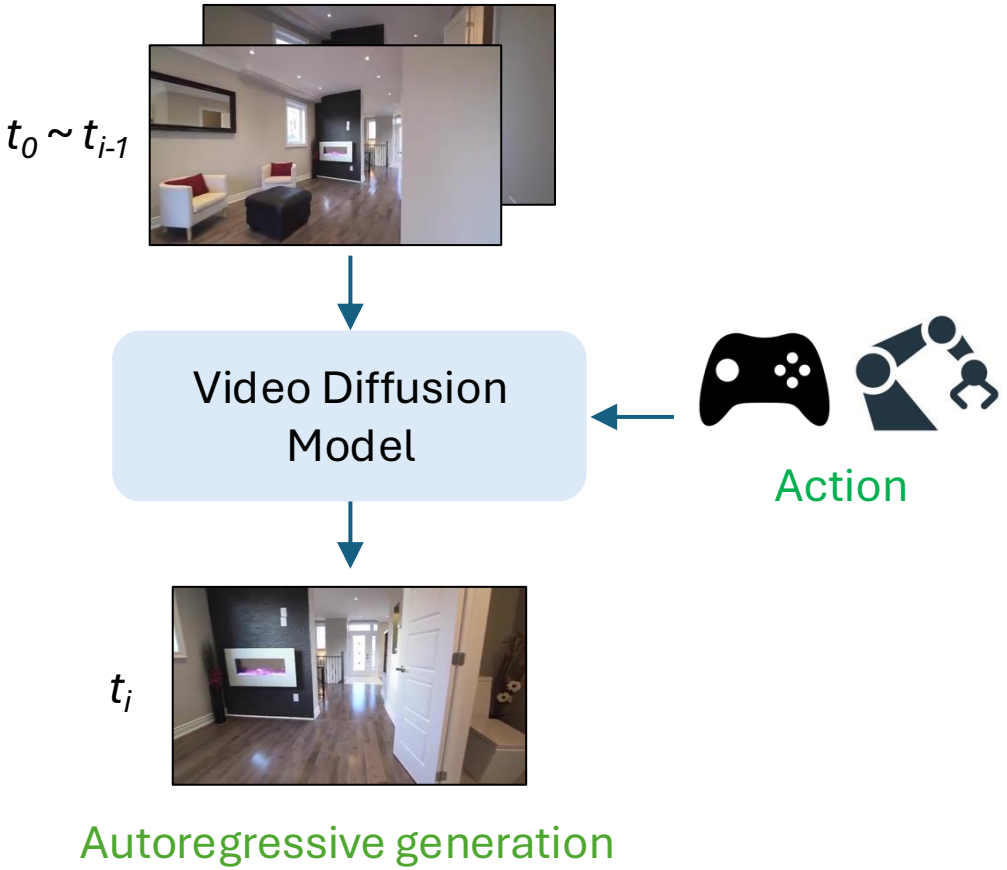
1st Workshop on Video World Models

Video World Models

Vanilla Video Diffusion Models

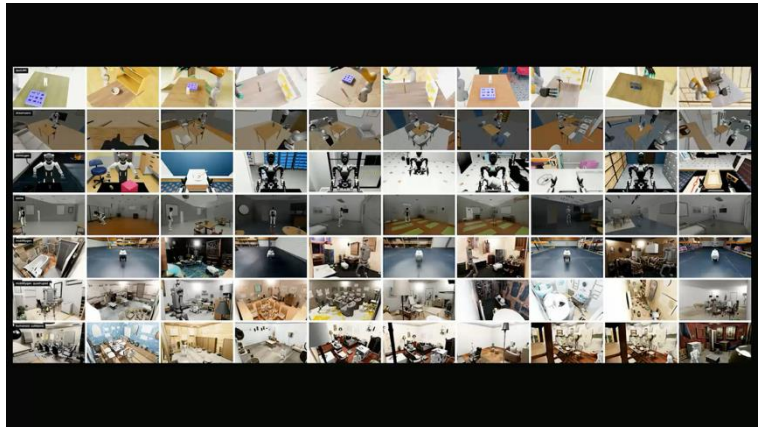


Video World Models

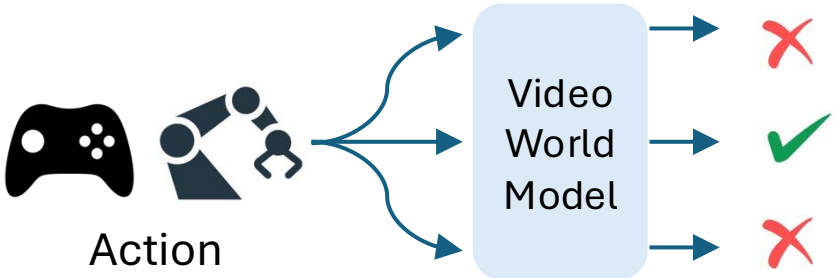


Video World Models

Simulation / Playable Environment



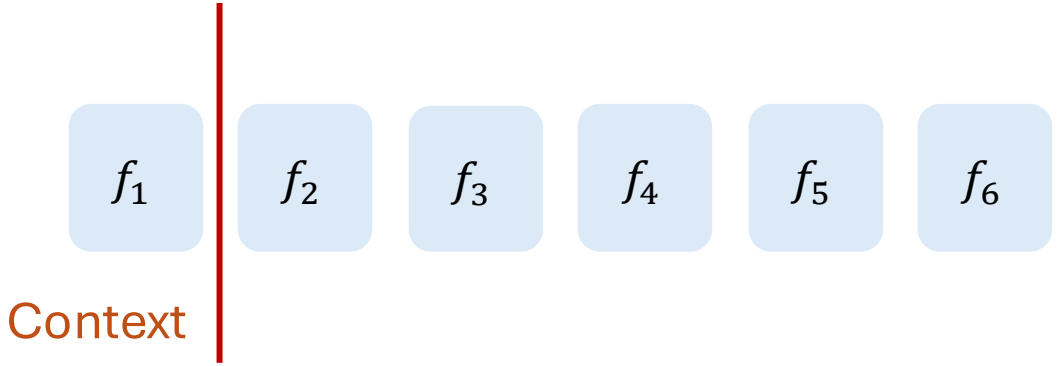
Create Synthetic Data



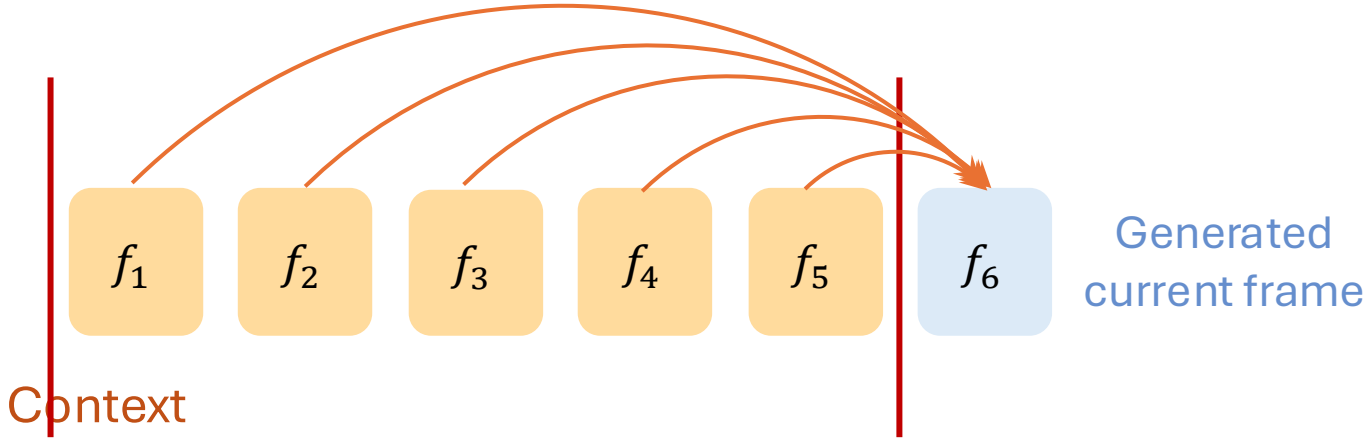
Reward / Planning

Build Policy Models (e.g., WAM)

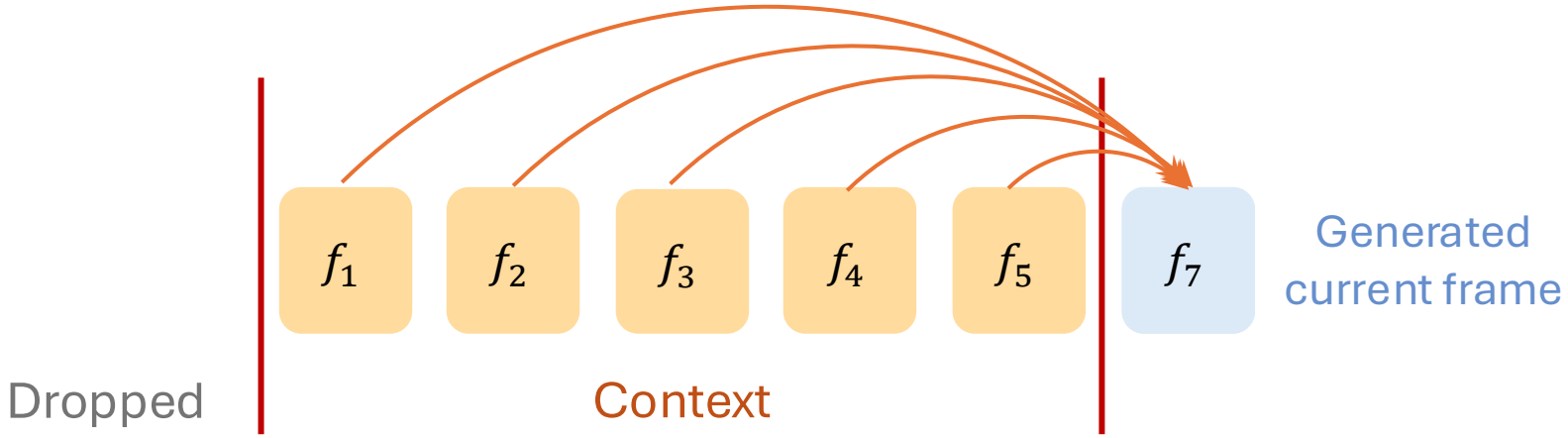
Autoregressive Generation



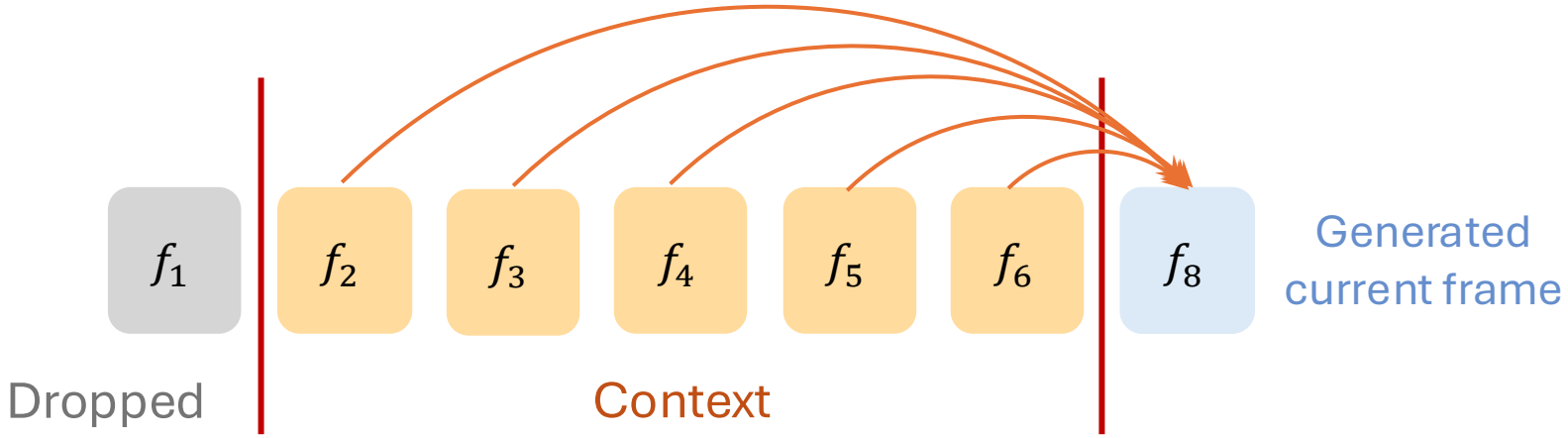
Autoregressive Generation



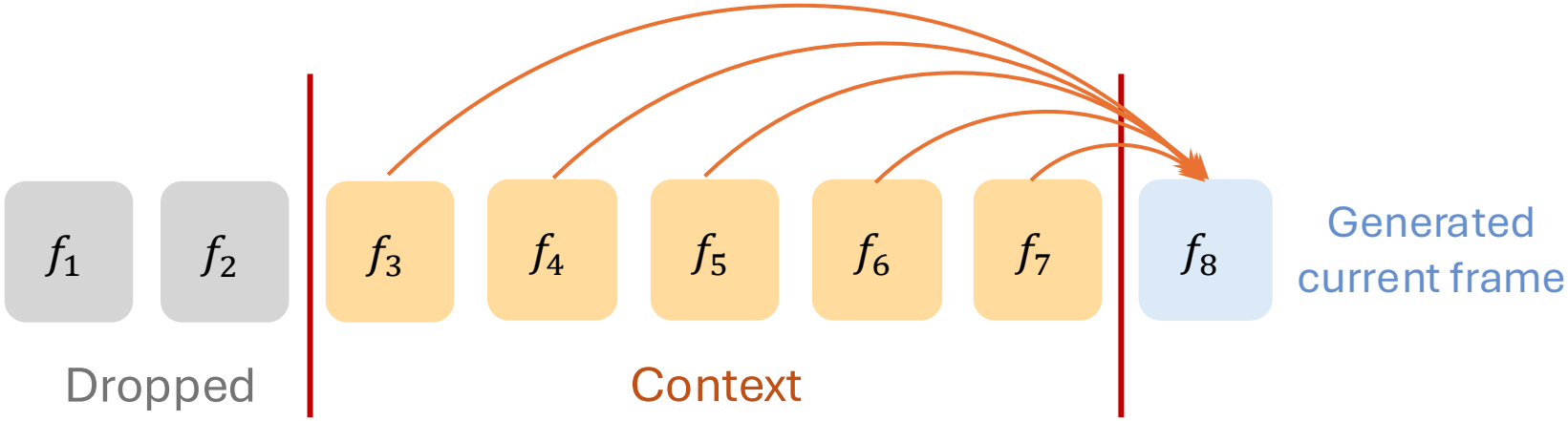
Autoregressive Generation

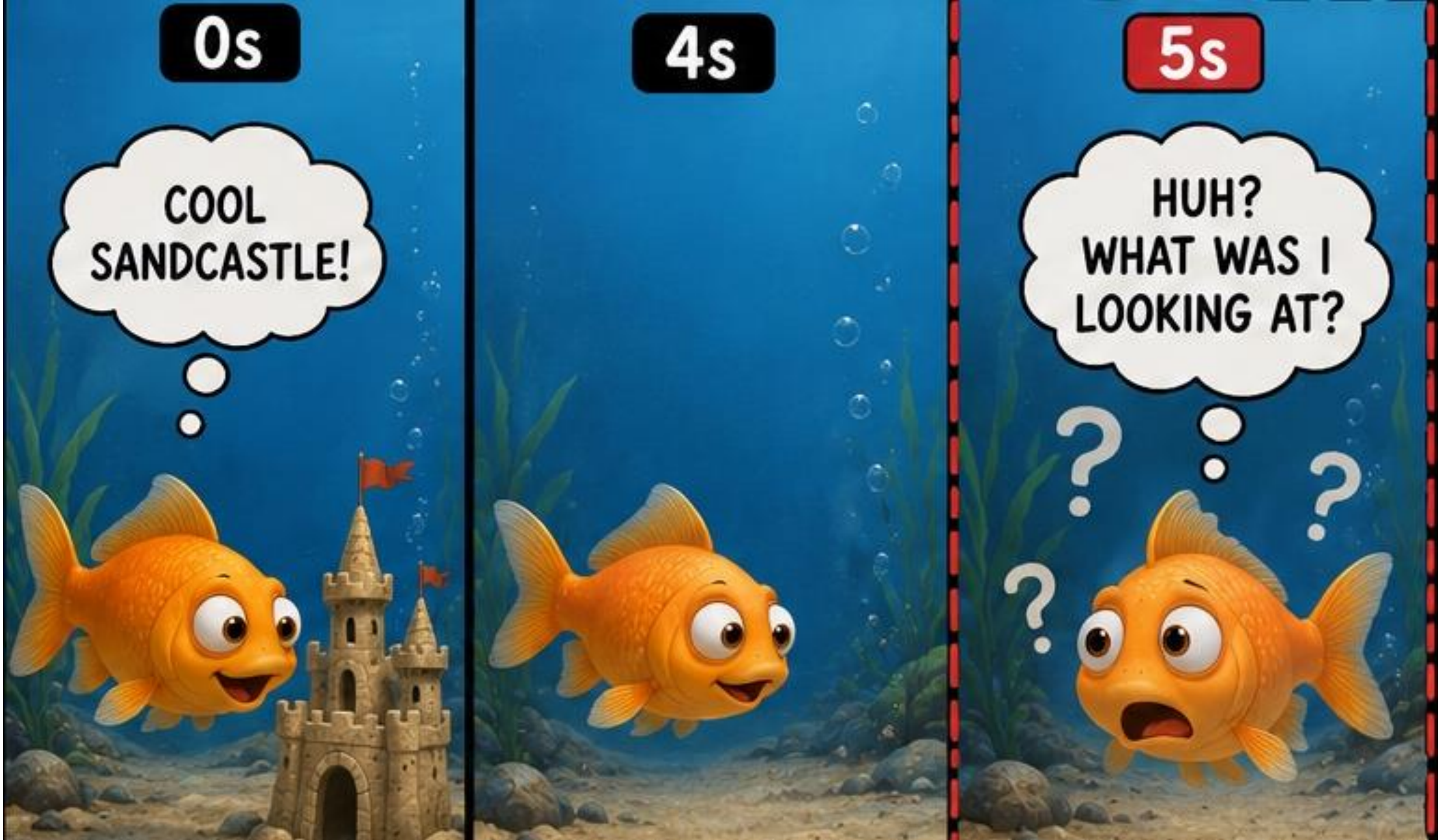


Autoregressive Generation



Autoregressive Generation




0s
COOL SANDCASTLE!
4s
5s
HUH? WHAT WAS I LOOKING AT?

Inconsistency from Limited Context Window

Oasis, https://oasis-model.github.io/

Zeqi Xiao

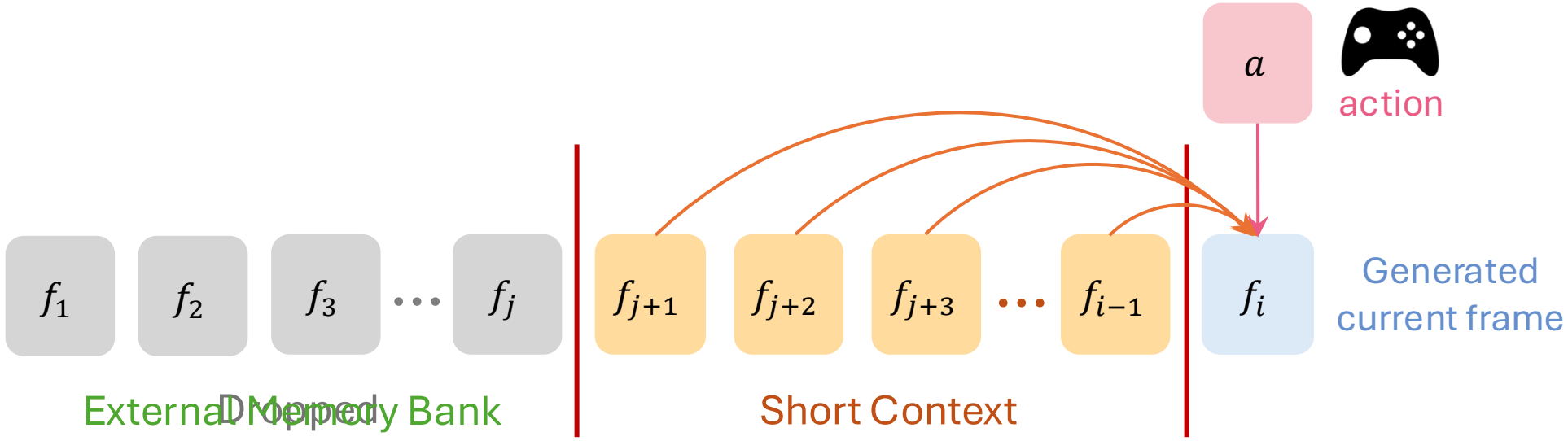
WorldMem: Long-term Consistent World Simulation with Memory

NeurIPS 2025

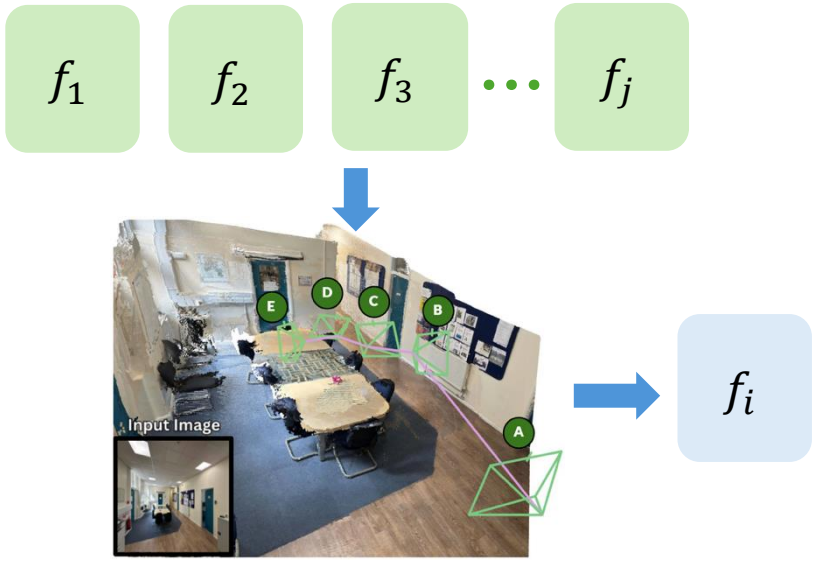
Zeqi Xiao[1] Yushi Lan[1] Yifan Zhou[1] Wenqi Ouyang[1]
Shuai Yang[2] Yanhong Zeng[3] Xingang Pan[1]

[1]S-Lab, Nanyang Technological University,
[2]Wangxuan Institute of Computer Technology, Peking University,
[3]Shanghai AI Laboratory

Motivation: Design External Memory



Geometry-aware memory

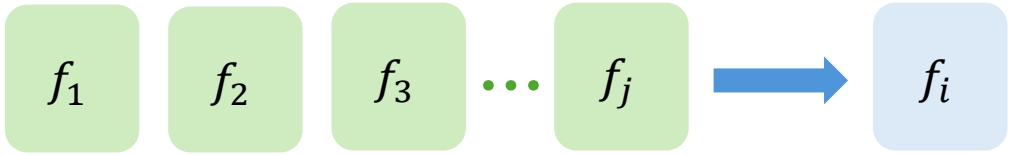


3D Reconstruction [1,2]

Restrict flexibility

- Modify or interact with the world is challenging
- Hard to model dynamic scene

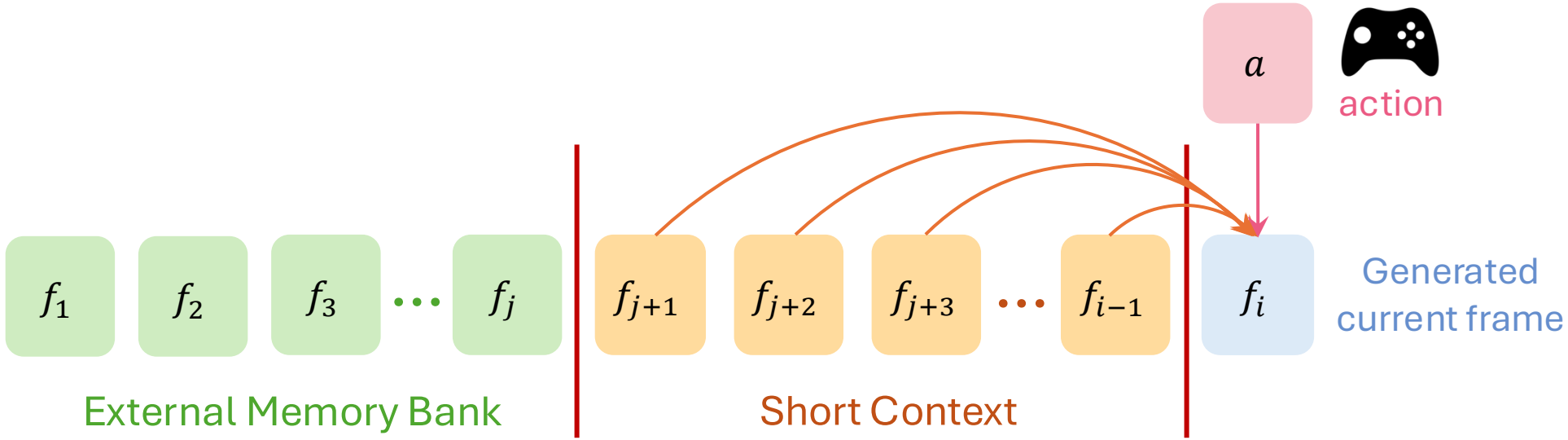
Geometry-free memory



- Less cumbersome design. Clean and scalable
- Easier to model dynamic scene
- Contain all details

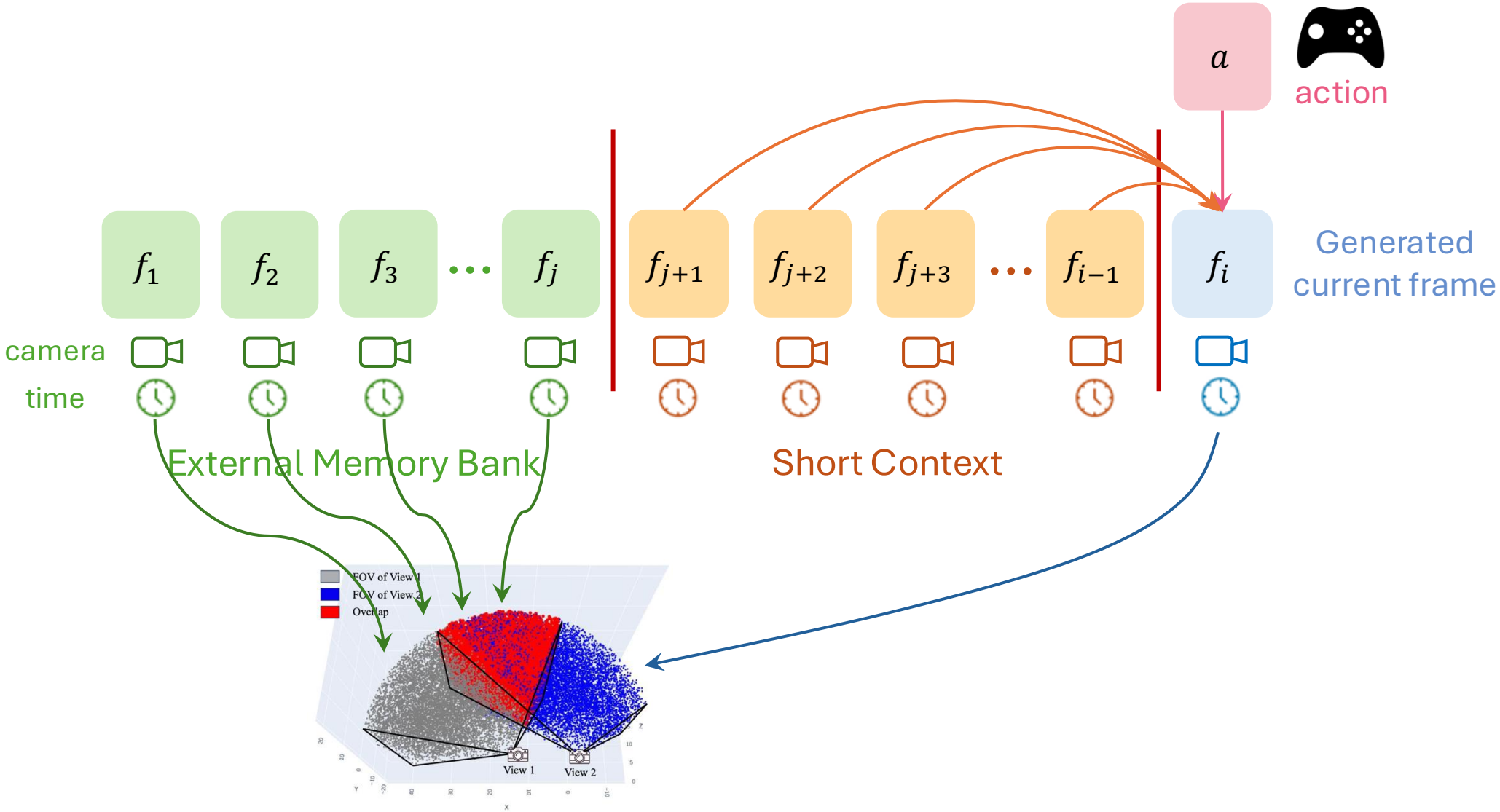
[1] ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis
[2] VMem: Consistent Interactive Video Scene Generation with Surfel-Indexed View Memory

Design External Memory

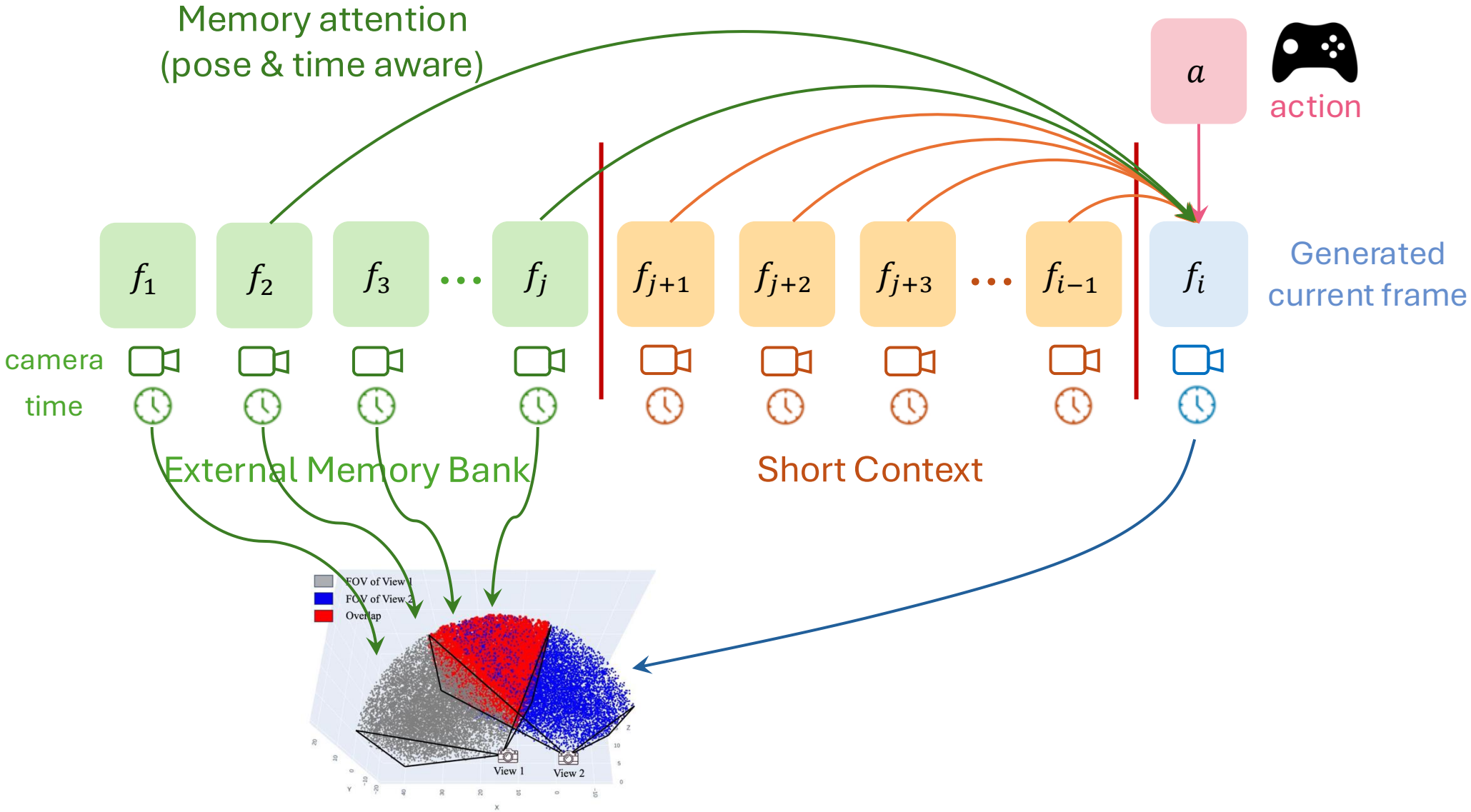


Which Memory Frame to Retrieve?

Design External Memory



Design External Memory



Experiments

Table 1. Evaluation on Minecraft benchmark [6]

Within context window			
Methods	PSNR ↑	LPIPS ↓	rFID ↓
Full Sequence	20.35	0.0691	13.87
Diffusion Forcing [3]	26.56	0.0094	13.88
Ours	**27.01**	**0.0072**	**13.73**
Beyond context window			
Methods	PSNR ↑	LPIPS ↓	rFID ↓
Full Sequence	/	/	/
Diffusion Forcing [3]	18.04	0.4376	51.28
Ours	**25.32**	**0.1429**	**15.37**

WorldMem maintains consistency over long trajectory.

Ours

GT

Experiments

Initial View

Revisited View

Initial View

Revisited View

By making timestamp as part of the state, we achieve consistent event.

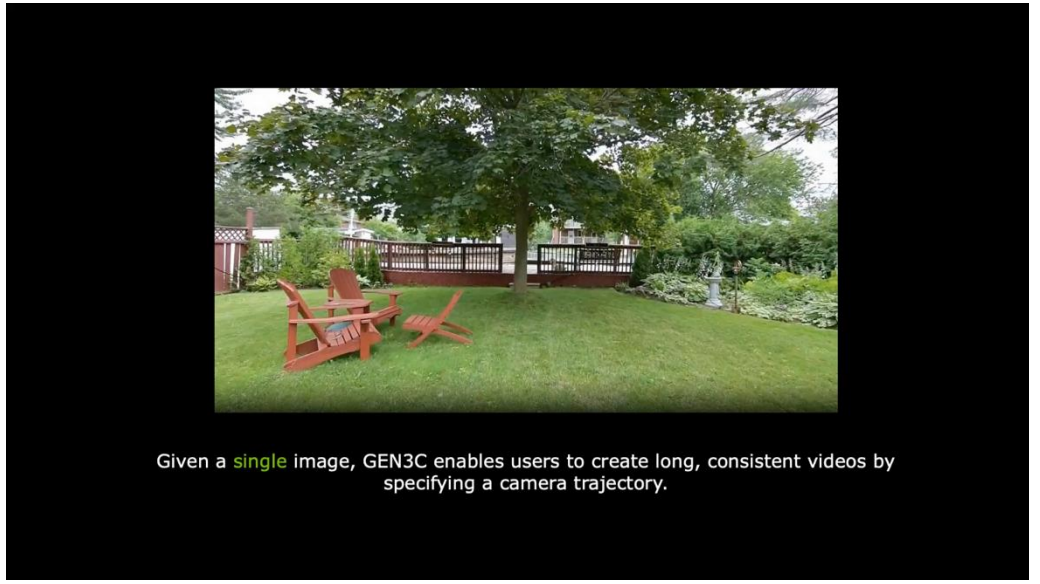
Memory

Video World Model

action

Notable Related Works

Context as Memory[1]



Gen3C[2]

[1] Jiwen Yu et al, Context as Memory: Scene-Consistent Interactive Long Video Generation with Memory Retrieval, SIGGRAPH Asia 2025
[2] Xuanchi Ren, Tianchang Shen, et al, GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control, CVPR 2025

Video World Model for Spatial Intelligence?

Video World Model ← action

Video World Model ← *Find and go to the table in the living room*

- Understanding 3D structure
- Understanding semantic layout
- Planning towards goals
- Maintaining long-horizon temporal coherence

Zeqi Xiao

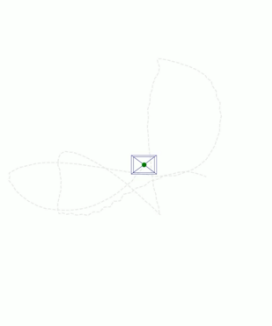
Video4Spatial: Towards Visuospatial Intelligence with Context-Guided Video Generation

CVPR 2026 Findings

Zeqi Xiao[1,2] Yiwei Zhao[1,*], Lingxiao Li[1], Yushi Lan[3], Ning Yu[4],
Rahul Garg[1], Roshni Cooper[1], Mohammad H. Taghavi[1] Xingang Pan[2,*]

[1]Netflix,
[2]Nanyang Technological University,
[3]University of Oxford
[4]Netflix Eyeline Studios

Customized Video Generation via Context



Noisy frames

First frame

Camera pose

Clean context frames

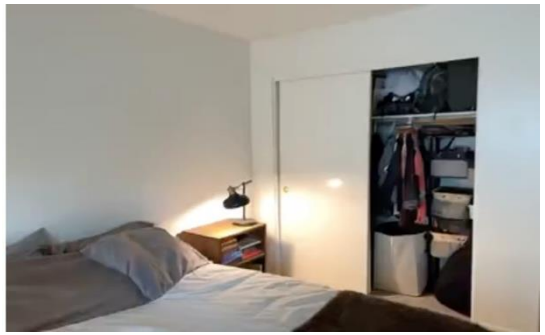
c_1 c_2 ... c_j f_1 f_2 ... f_i g

Instructions

Video Diffusion Model

c_1 c_2 ... c_j f_1 f_2 ... f_i

Denoised frames



Video Generation as Spatial Reasoner

The camera moves smoothly through a bedroom. Finally, it focuses on a guitar in the center of the frame.

First frame

Text

Noisy frames

Clean context frames

c_1 c_2 ... c_j f_1 f_2 ... f_i g

Instructions

Video Diffusion Model

c_1 c_2 ... c_j f_1 f_2 ... f_i

Denoised frames

*Red bbox is generated by the model

Experiments

Clean context frames

Video Diffusion Model

The camera moves smoothly through a bedroom. Finally, it focuses on a guitar in the center of the frame.

The camera moves smoothly through a bedroom. Finally, it focuses on a green plant in the center of the frame.

The camera moves smoothly through a bedroom. Finally, it focuses on a door in the center of the frame.

The camera moves smoothly through a bedroom. Finally, it focuses on a door in the center of the frame.

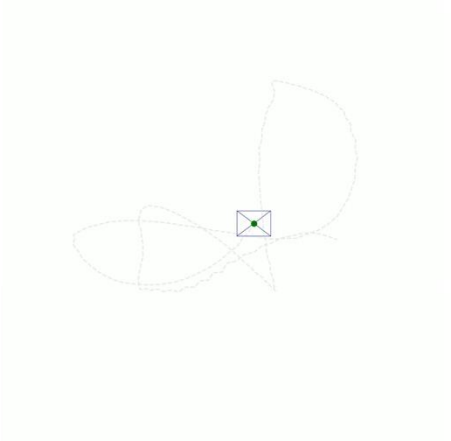
OOD Generalization

Clean context frames

Video Diffusion Model

Target object: Bicycle

Target object: Tree



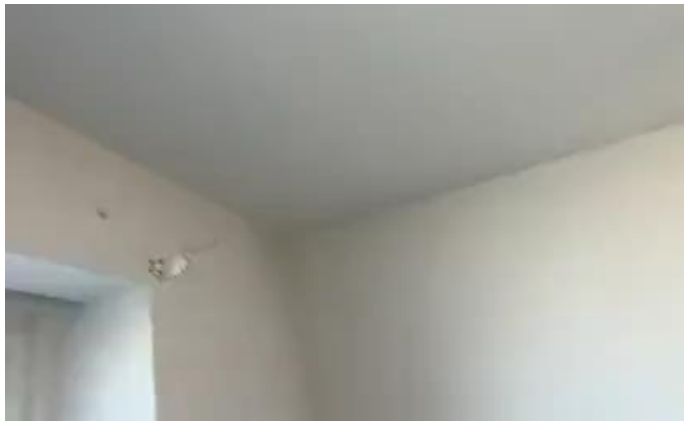
Camera pose instruction

Results

Failure Cases

Context

Failure case: temporal discontinuity



The camera moves smoothly through a kitchen. Finally, it focuses on a sink in the center of the frame.

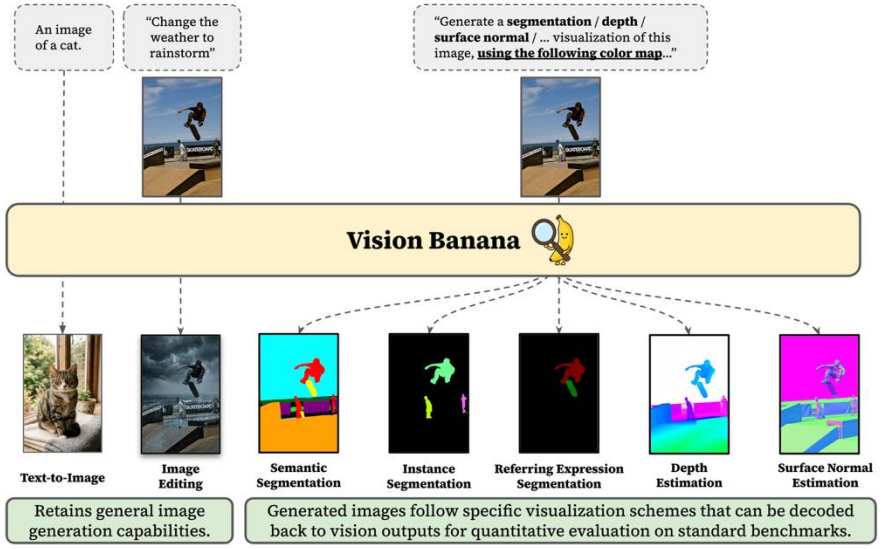
Context

Failure case: incorrect grounding

The camera moves from a desk area. Finally, it focuses on a suitcase in the center of the frame.

Rethinking Video Diffusion Models …

WAN[1] finetuned on Arctic dataset[2]



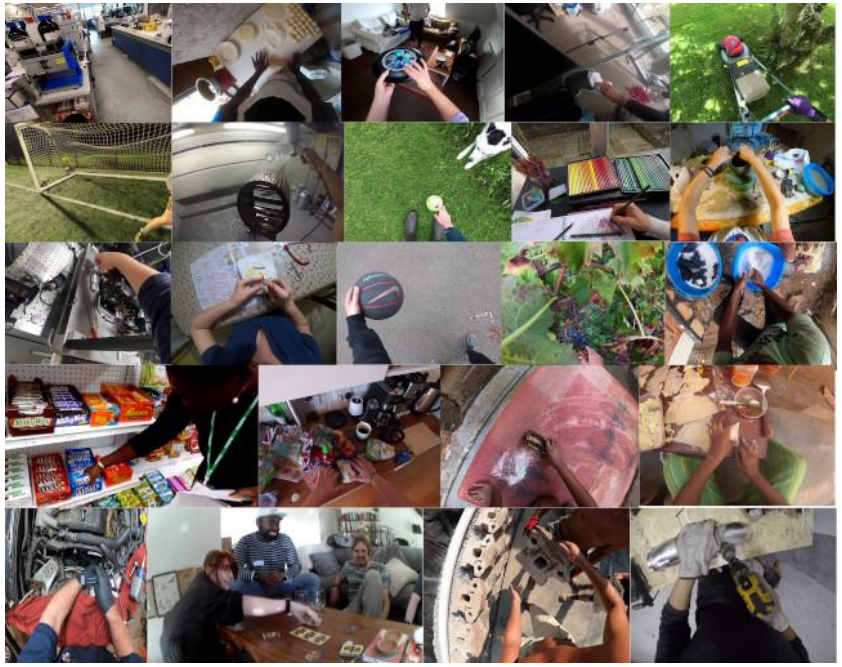
Vision Banana[3]

The fact that they can generate implies that they ***understand*** the

- **structure**
- **motion**
- **correspondence**

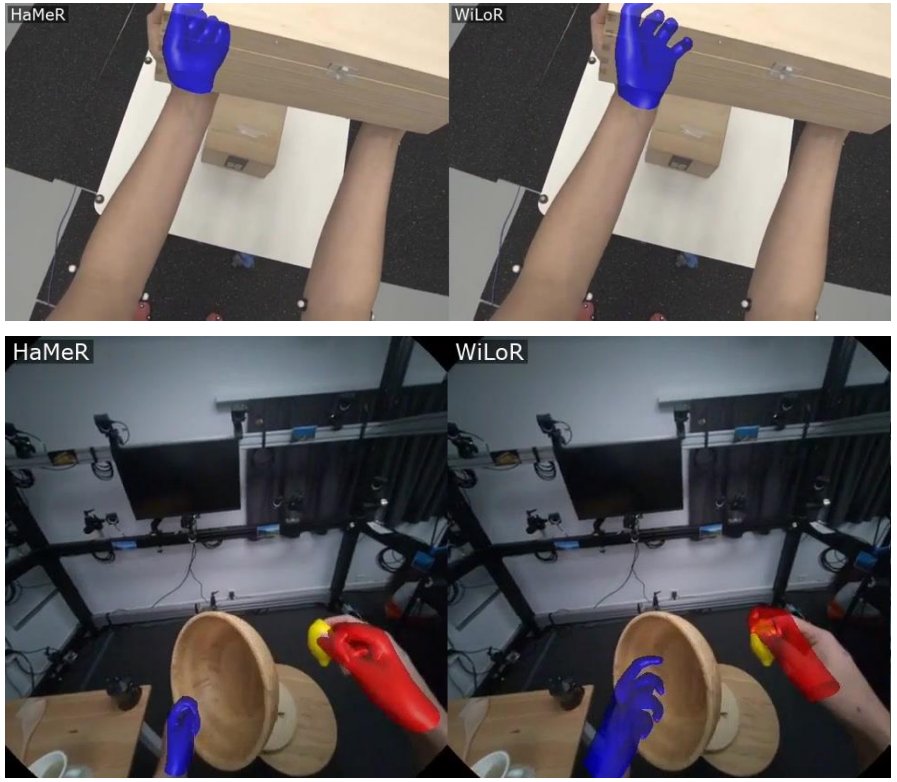
[1] Team Wan, Wan: Open and Advanced Large-Scale Video Generative Models, 2025
[2] Zicong Fan et al, ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation, CVPR 2023
[3] Valentin Gabeur, Shangbang Long, Songyou Peng et al, Image Generators are Generalist Vision Learners, arXiv preprint, 2026

On the other hand …



Ego4D [1]

Egocentric vision is increasingly important



Yet understanding hand motion is still challenging …

[1] Kristen Grauman et al, Ego4D: Around the World in 3,000 Hours of Egocentric Video, CVPR 2022

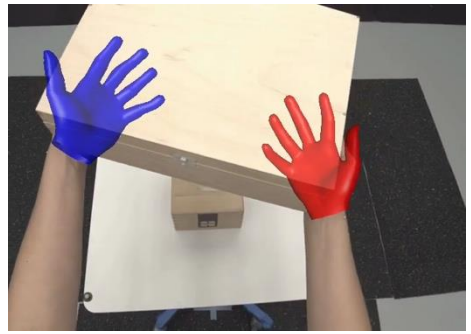
Leveraging VDMs for Hand Motion Capture?

Egocentric Input

Video Diffusion Model

MANO Projector

4D Hand Mesh



Yuxi Wang

Chengkai Jin

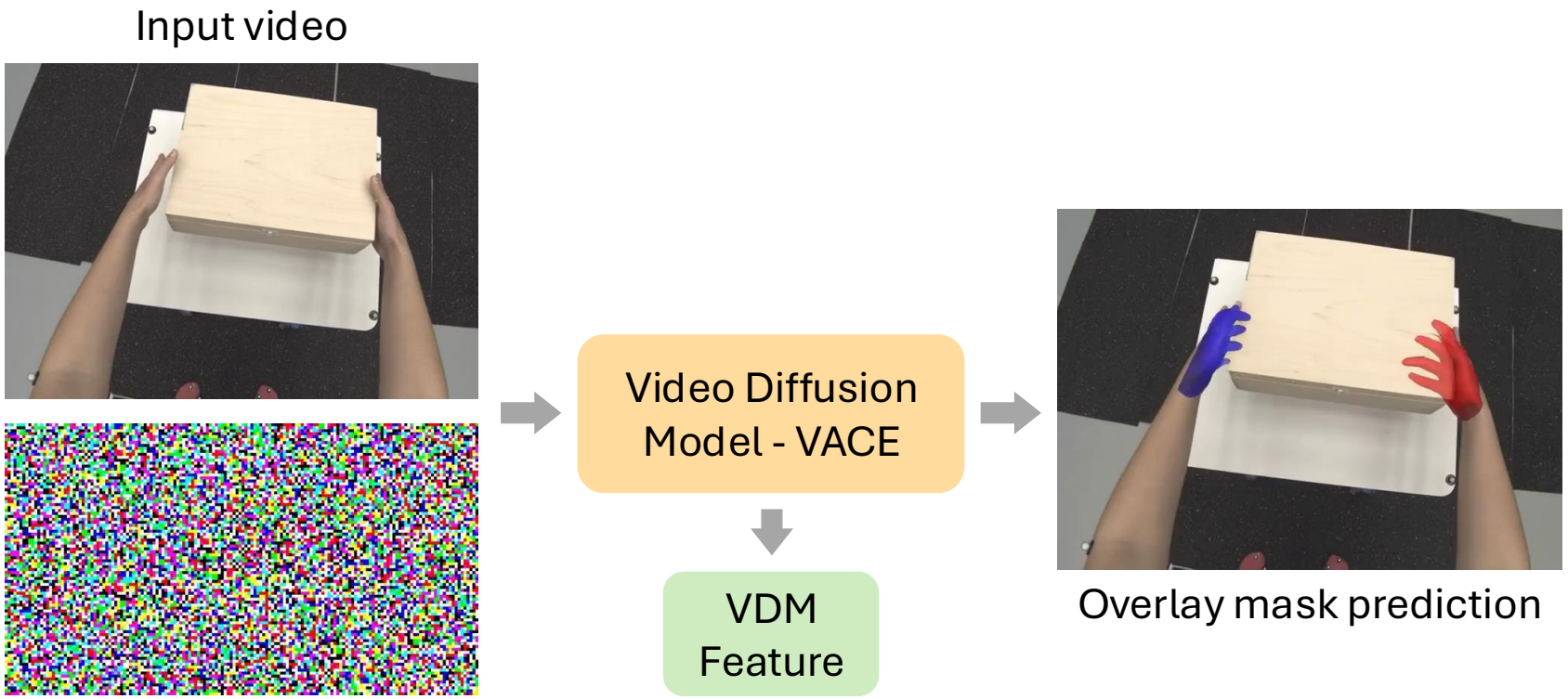
The Surprising Effectiveness of Video Diffusion Models for Hand Motion Reconstruction

To appear on arXiv soon

Yuxi Wang*[1], Chengkai Jin*[1], Yufei Liu [2], Wenqi Ouyang[1], Tianyi Wei[1], Zhiwei Zeng [1], Siyuan Huang [3], Zhiqi Shen[1], Xingang Pan[1]

[1]Nanyang Technological University,
[2] The Chinese University of Hong Kong
[3]ACE Robotics

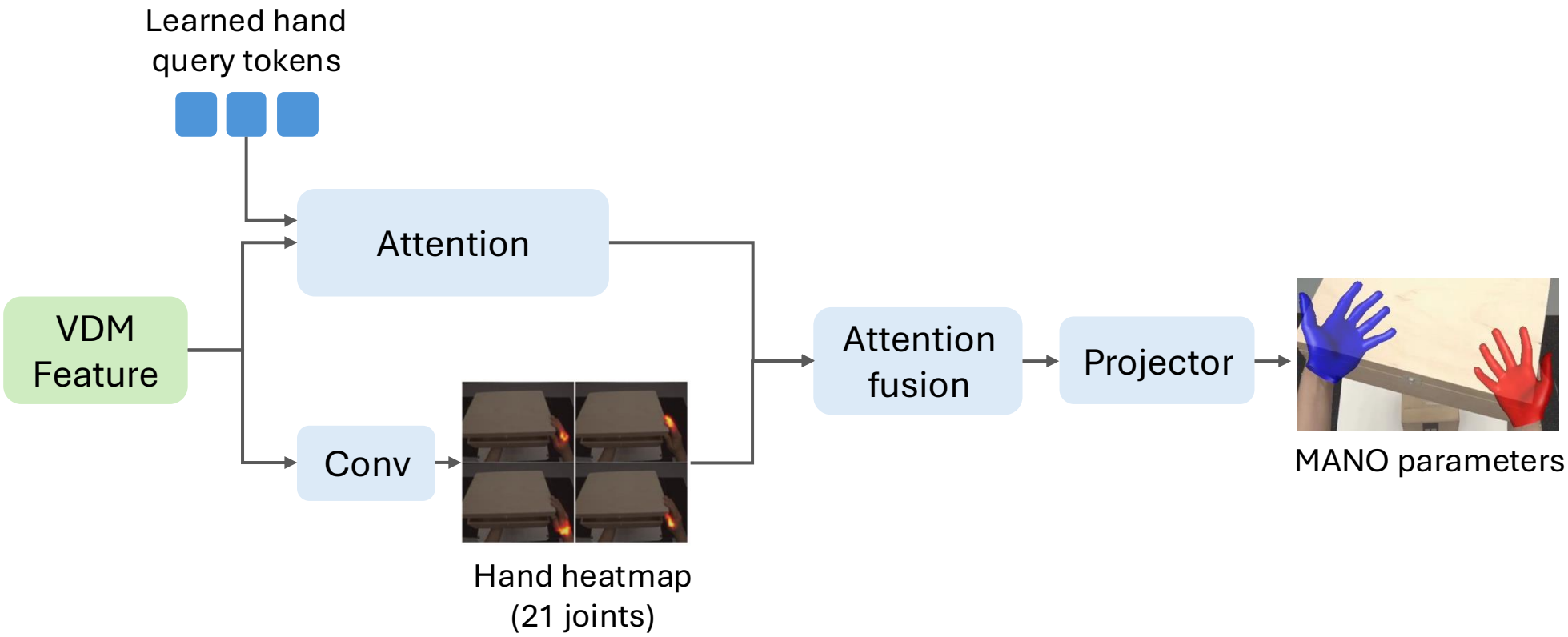
ViDiHand Stage 1: V2V overlay finetuning



Steer the video diffusion representation to focus more on hands

Zeyinzi Jiang et al, VACE: All-in-One Video Creation and Editing, ICCV 2025

ViDiHand Stage 2: MANO Decoder Training

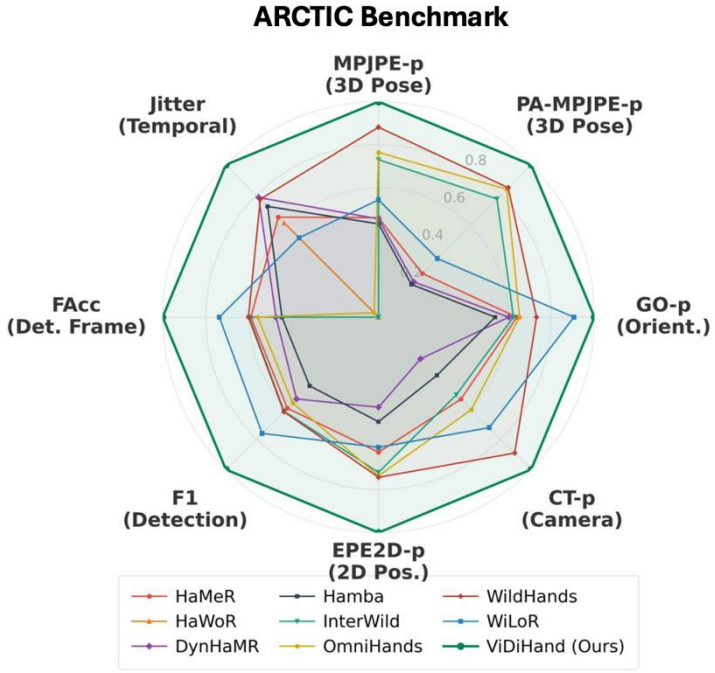


Quantitative Results: ARCTIC Benchmark

Training: ARCTIC, HOT3D

Testing:

	Method	Detection			3D Pose		Orient. & Position			Temporal
		FAcc	Recall	F1	MPJPE-p	PA-p	EPE-p	GO-p	CT-p	Jitter
ARCTIC	InterWild [16]	0.878	0.943	0.959	30.817	15.952	53.888	25.386	0.097	46.577
	HaMeR [18]	0.876	0.943	0.957	39.976	29.325	67.816	24.924	0.094	18.094
	Hamba [3]	0.833	0.912	0.942	40.965	31.300	88.839	27.822	0.110	15.026
	WildHands [20]	0.879	0.946	0.960	25.704	13.941	50.517	22.320	0.058	12.972
	WiLoR [19]	0.919	0.951	0.974	37.173	26.646	71.216	17.358	0.075	23.978
	Dyn-HaMR [31]	0.841	0.917	0.951	40.172	30.849	98.813	26.006	0.121	12.506
	HaWoR [33]	0.700	0.818	0.895	55.677	37.182	160.912	43.320	0.149	19.735
	OmniHands [14]	0.866	0.949	0.954	29.674	14.203	51.505	24.580	0.087	45.312
	ViDiHand (Ours)	**0.997**	**0.999**	**0.999**	**21.668**	**9.821**	**12.407**	**14.642**	**0.047**	**3.183**



- Frame Accuracy 0.997 - **27x** lower error rate than SOTA (WiLoR)
- PA-p 9.82 mm - **1.42x** lower than SOTA (WildHands)
- EPE-p 12.4 px - **4.07x** lower than SOTA (WildHands)
- Jitter 3.18 mm - **3.93x** lower than SOTA (Dyn-HaMR)

Quantitative Results: ARCTIC Benchmark

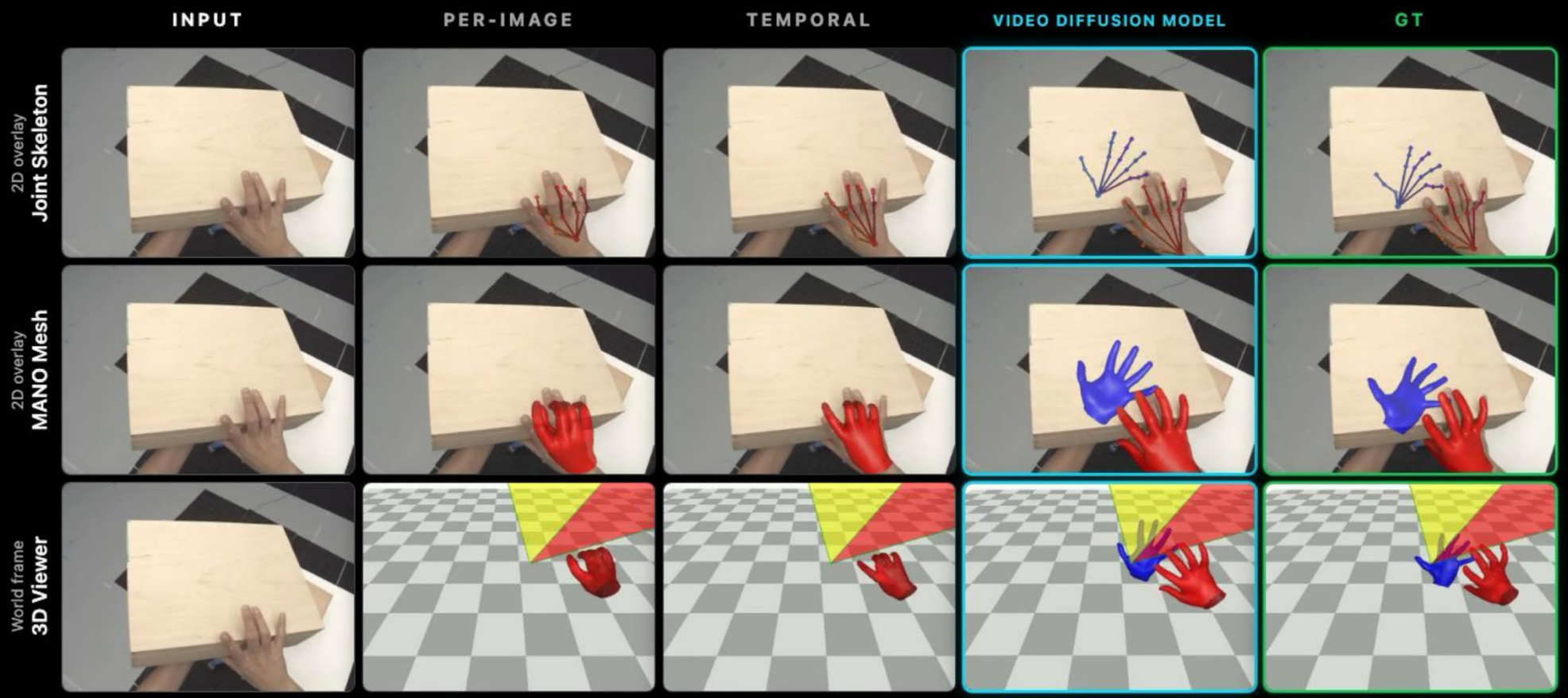
Evaluation on HOT3D and HOI4D, HOI4D is not included in training

HOT3D	InterWild [16]	0.669	0.881	0.868	77.168	24.811	71.482	58.501	0.213	101.164
	HaMeR [18]	0.692	0.904	0.883	67.593	36.241	60.801	49.633	0.102	23.206
	Hamba [3]	0.632	0.829	0.853	71.054	43.342	108.502	56.535	0.128	18.111
	WildHands [20]	0.655	0.863	0.844	52.791	28.946	111.438	53.933	0.157	22.885
	WiLoR [19]	0.827	0.898	0.937	44.825	35.079	69.881	25.750	0.098	17.784
	Dyn-HaMR [31]	0.558	0.761	0.755	82.865	51.921	205.428	50.700	0.583	47.483
	HaWoR [33]	0.348	0.499	0.655	80.146	74.957	332.311	79.350	0.262	23.806
	OmniHands [14]	0.649	0.895	0.868	63.281	22.682	68.437	49.120	0.133	69.510
	ViDiHand (Ours)	**0.948**	**0.974**	**0.983**	**21.514**	**11.383**	**14.953**	**15.829**	**0.040**	**3.741**
HOI4D	InterWild [16]	0.731	0.922	0.864	53.072	22.909	80.549	41.743	0.228	98.866
	HaMeR [18]	0.730	0.923	0.864	48.875	33.215	81.717	33.636	0.187	20.197
	Hamba [3]	0.709	0.885	0.849	51.698	37.395	117.801	37.466	0.204	21.740
	WildHands [20]	0.730	0.924	0.864	45.623	23.601	82.246	45.654	0.159	18.615
	WiLoR [19]	0.962	0.965	0.972	41.603	27.767	43.335	25.603	0.116	17.735
	Dyn-HaMR [31]	0.749	0.861	0.844	52.826	40.172	151.902	40.306	0.258	18.146
	HaWoR [33]	0.868	0.863	0.918	58.442	39.665	144.229	43.209	0.140	28.098
	OmniHands [14]	0.655	0.937	0.834	44.255	18.689	70.662	34.392	**0.108**	24.212
	ViDiHand (Ours)	**0.984**	**0.991**	**0.990**	**30.090**	**13.960**	**24.460**	**23.420**	0.117	**4.010**

Qualitative Comparison

We compare against the strongest per-image (HaMeR) and temporal (HaWoR) baselines, plus ground truth.

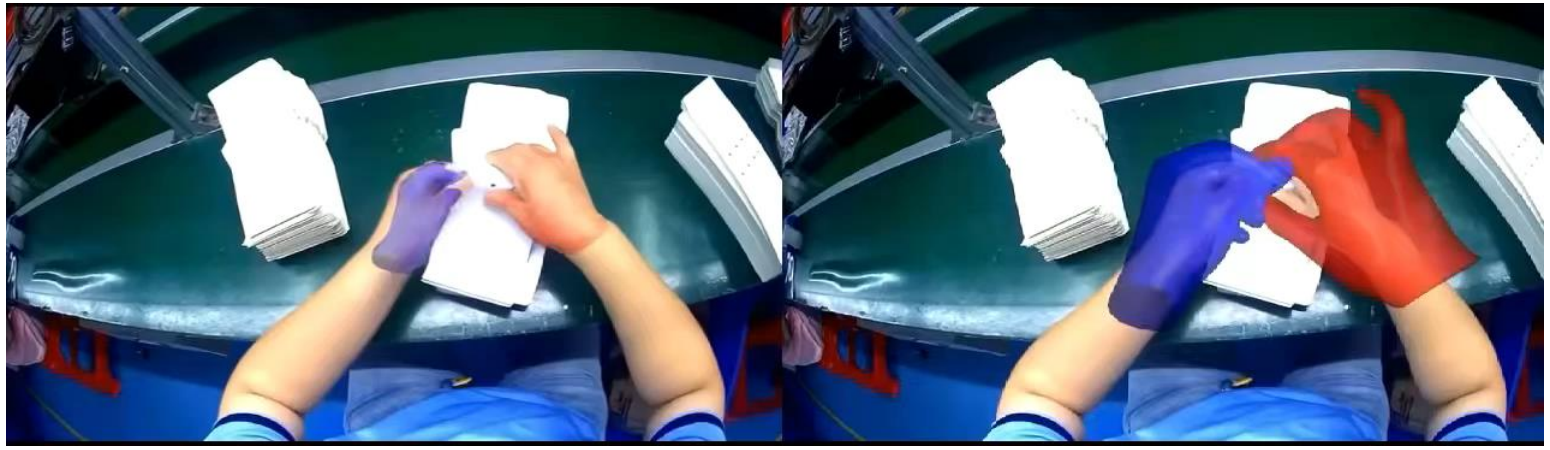
Hand-object occlusion · ARCTIC



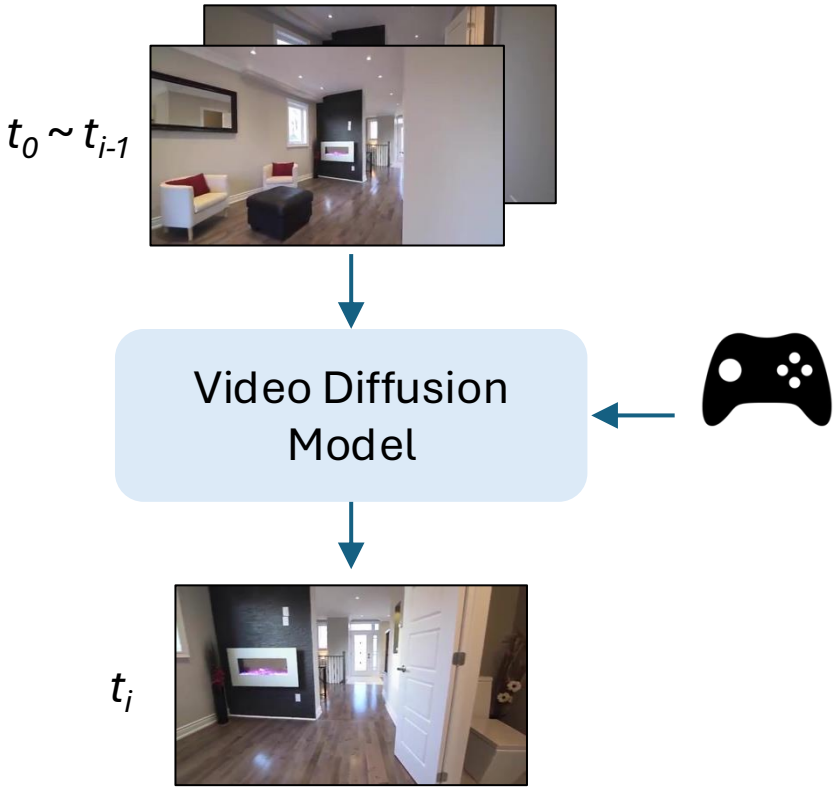
Failure Cases

Predicted overlay

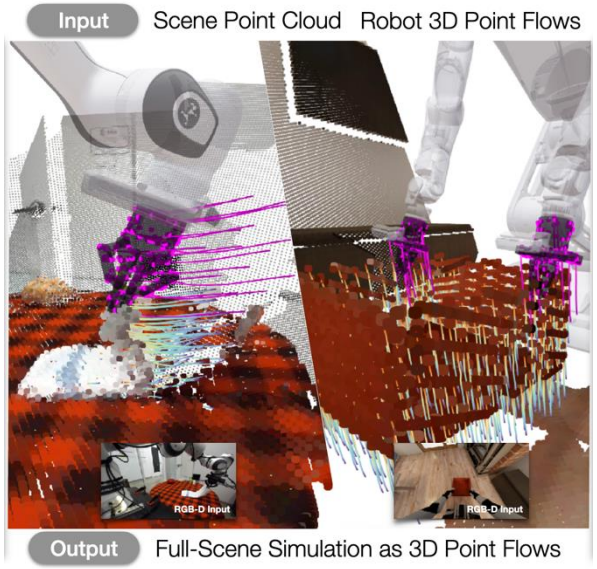
Predicted hand mesh



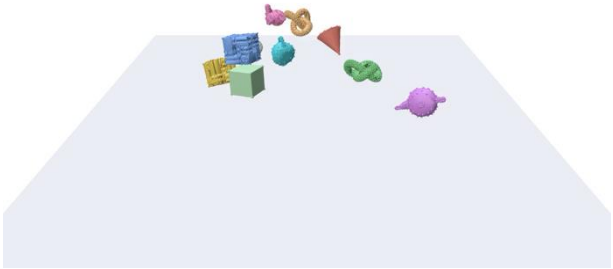
2D Video World Models



3D World Models



PointWorld [1]



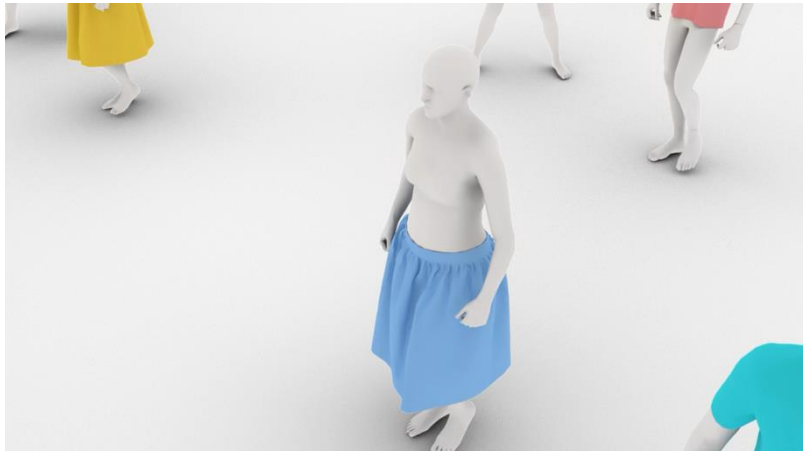
RigidFormer [2]

[1] Wenlong Huang et al, PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation, CVPR 2026
[2] Zhiyang Dou et al, RigidFormer: Learning Rigid Dynamics with Transformers, arXiv preprint, 2026

ClothTransformer

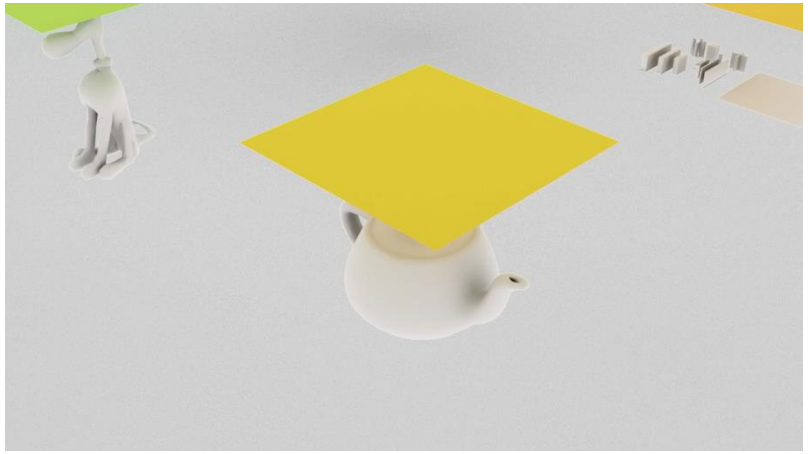
Yu Zhang

We construct a **~493.4k-frame penetration-free dataset** spanning three scenarios:



Garment

Grasp



Collision

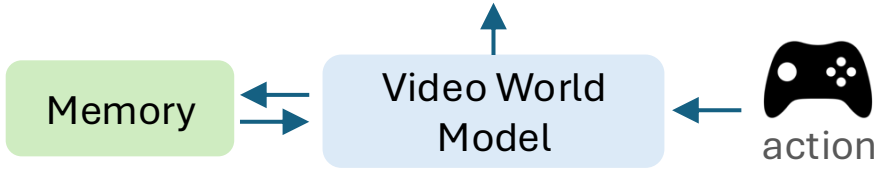
Zhang Yu et al, ClothTransformer: Unified Latent-Space Transformers for Scalable Cloth Simulation, arXiv preprint, 2026

ClothTransformer

Yu Zhang

Zhang Yu et al, ClothTransformer: Unified Latent-Space Transformers for Scalable Cloth Simulation, arXiv preprint, 2026

Summary



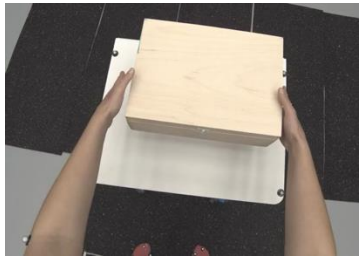
World model with memory mechanism

context

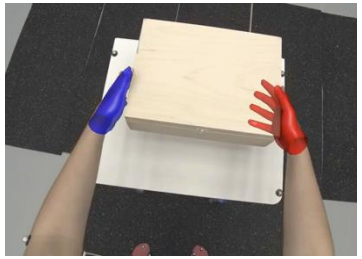
Video Diffusion Model

The camera focuses on a guitar in the center of the frame.

VDM for spatial reasoning



Video Diffusion Model



VDM for hand motion capture

Acknowledgement


Zeqi Xiao

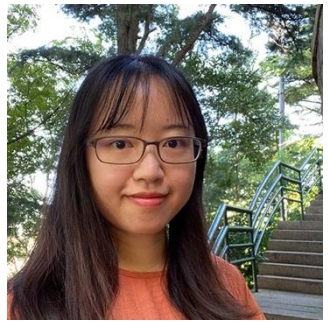

Yuxi Wang


Chengkai Jin


Yu Zhang

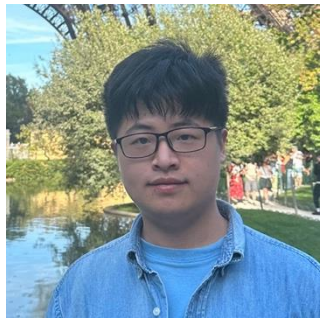

Yushi Lan


Yifan Zhou


Wenqi Ouyang


Zhexin Liang


Shuai Yang


Tianyi Wei


Yiwei Zhao


Lingxiao Li


Ning Yu

Xudong Xu

Q & A