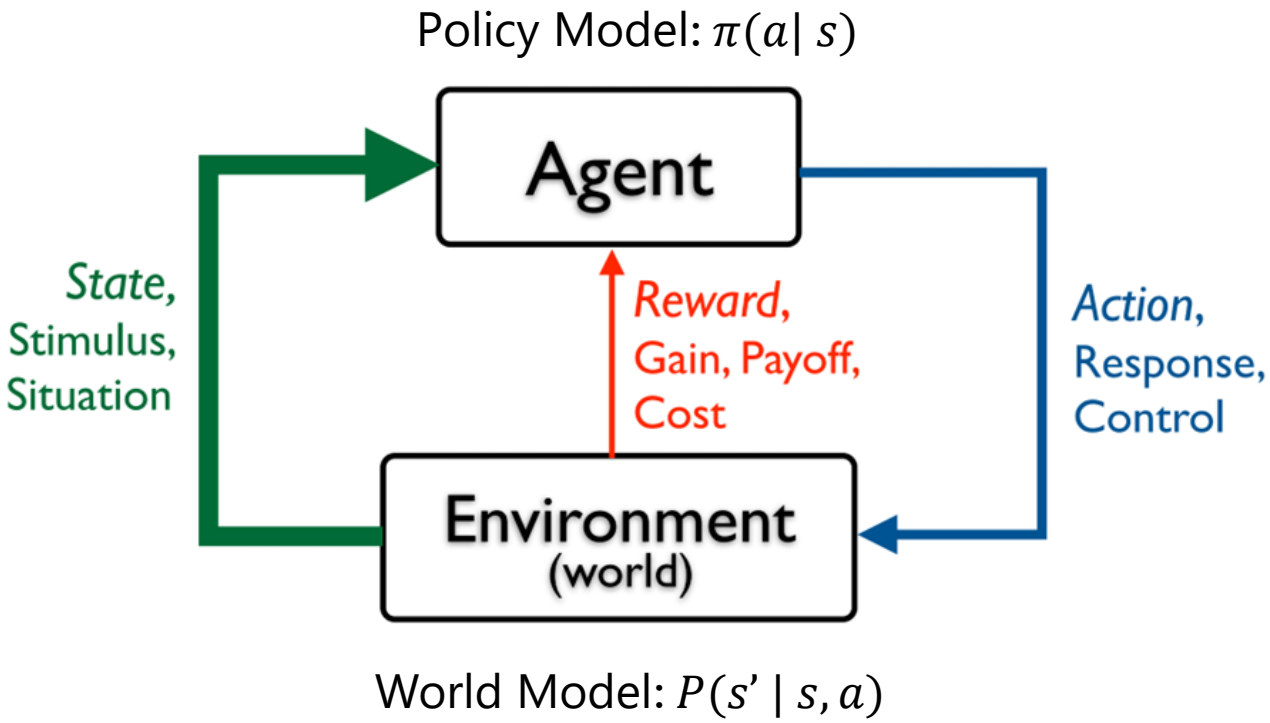
Towards Interactive World Simulator

Tianyu He
Microsoft Research Asia

4/9/2025

Microsoft

World Model



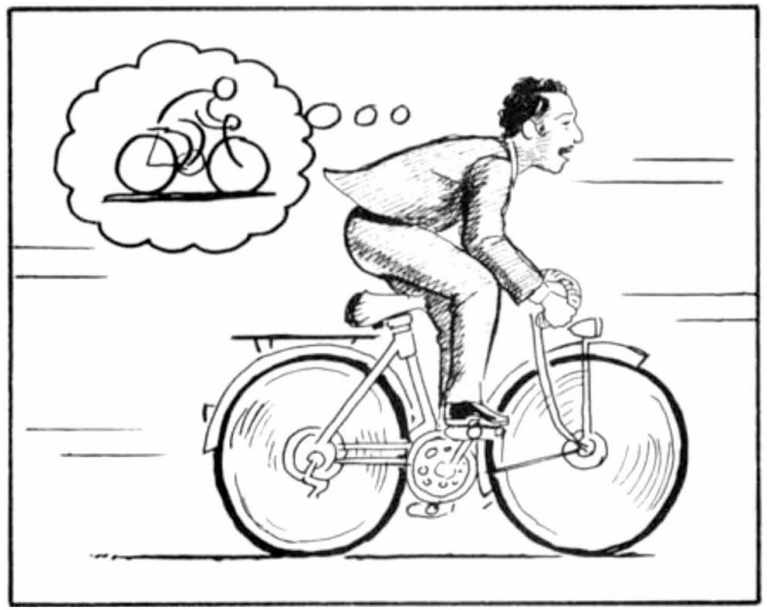


Figure 1. A World Model, from Scott McCloud's *Understanding Comics.* (McCloud, 1993; E, 2012)

interactive !

David Ha, Jürgen Schmidhuber. World models. arXiv:1803.10122.

Outline

1 How to represent the visual world?

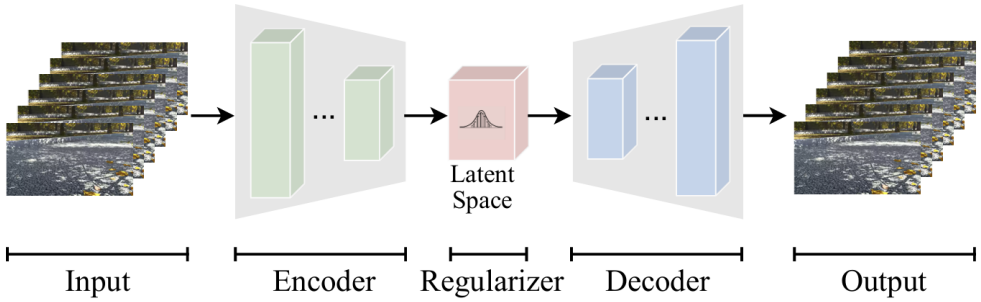
- We introduce VidTok, A cutting-edge family of video tokenizers.

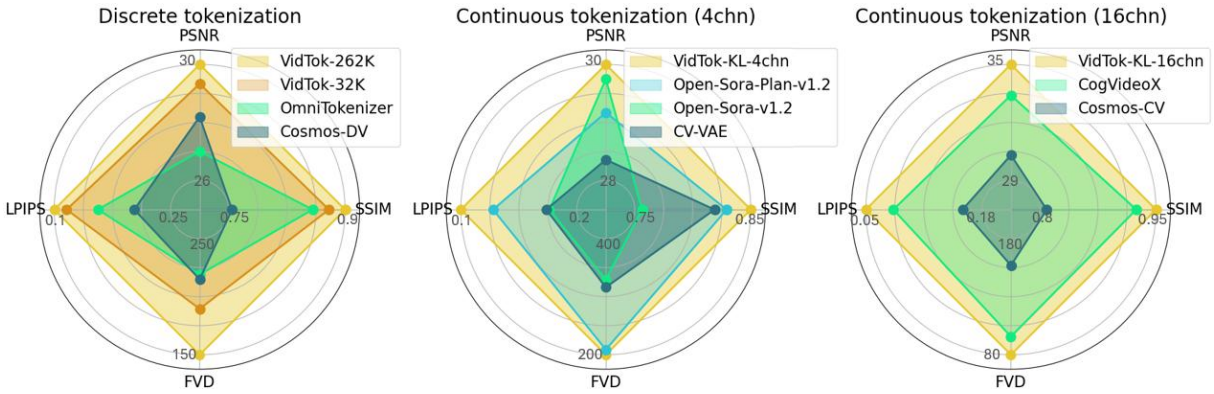
2 How to enable interactive visual world modeling?

- Interact with **action**: autoregressive world model on MineCraft.
- Interact with **latent action**: human-to-robot cross-embodiment generalization.
- Interact with **video demonstration**: zero-shot video imitation in real-world.
- Interact with **camera viewpoint**: explicit world model with underlying 3D structure.

VidTok

A cutting-edge family of video tokenizers that excels in both continuous and discrete tokenizations.





- **Efficient Architecture**

  Separate spatial and temporal sampling reduces computational complexity without sacrificing quality.

- **Advanced Quantization**

  Finite Scalar Quantization (FSQ) addresses training instability and codebook collapse in discrete tokenization.

- **Enhanced Training**

  A two-stage strategy—pre-training on low-res videos and fine-tuning on high-res—boosts efficiency. Reduced frame rates improve motion dynamics representation.

Tang et al. VidTok: A Versatile and Open-Source Video Tokenizer. arXiv:2412.13061.

VidTok

Leading Reconstruction Performance.

https://github.com/microsoft/VidTok

Method	Regularizer	Param.	MCL-JCV				WebVid-Val			
			PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓
MAGVIT-v2*	LFQ - 262, 144	-	26.18	-	0.104	-	-	-	-	-
OmniTokenizer	VQ - 8, 192	51M	26.93	0.841	0.165	232.7	26.26	0.883	0.112	48.46
Cosmos-DV	FSQ - 64, 000	101M	28.07	0.820	0.212	227.7	29.39	0.840	0.170	57.97
Ours-FSQ	FSQ - 32, 768	157M	29.16	0.854	0.117	196.9	31.04	0.883	0.089	45.34
Ours-FSQ	FSQ - 262, 144	157M	**29.82**	**0.867**	**0.106**	**160.1**	**31.76**	**0.896**	**0.080**	**38.17**
CV-VAE	KL - 4chn	182M	28.56	0.823	0.163	334.2	30.79	0.863	0.116	70.39
Open-Sora-v1.2	KL - 4chn	393M	29.44	0.844	0.164	350.7	31.02	0.866	0.137	112.34
Open-Sora-Plan-v1.2	KL - 4chn	239M	29.07	0.839	0.131	201.7	30.85	0.869	0.101	44.76
Ours-KL	KL - 4chn	157M	**29.64**	**0.852**	**0.114**	**194.2**	**31.53**	**0.878**	**0.087**	**36.88**
CogVideoX	KL - 16chn	206M	33.76	0.930	0.076	93.2	36.22	0.952	0.049	15.30
Cosmos-CV	AE - 16chn	101M	31.27	0.886	0.149	153.7	33.04	0.904	0.107	23.85
Ours-KL	KL - 16chn	157M	**35.04**	**0.942**	**0.047**	**78.9**	**37.53**	**0.961**	**0.032**	**9.12**

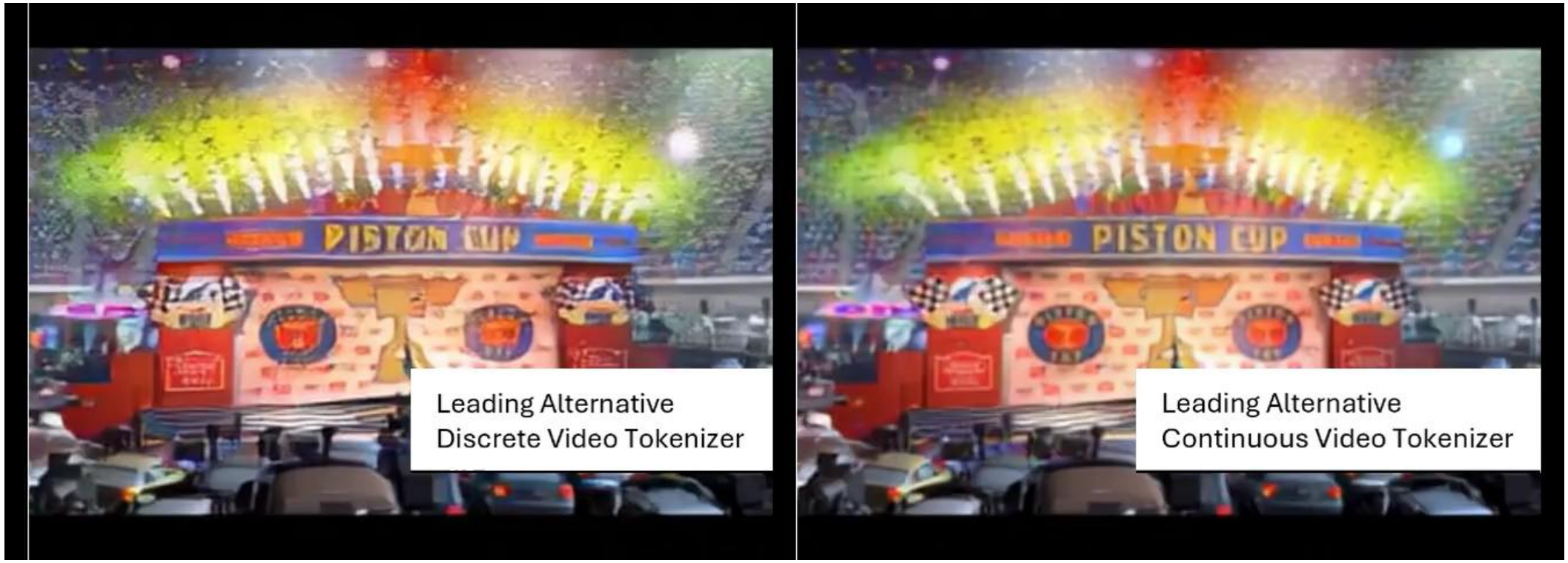
VidTok, trained on a large-scale video dataset, outperforms previous models across all metrics, including PSNR, SSIM, LPIPS, and FVD.

Tang et al. VidTok: A Versatile and Open-Source Video Tokenizer. arXiv:2412.13061.

VidTok

Leading Reconstruction Performance.

https://github.com/microsoft/VidTok

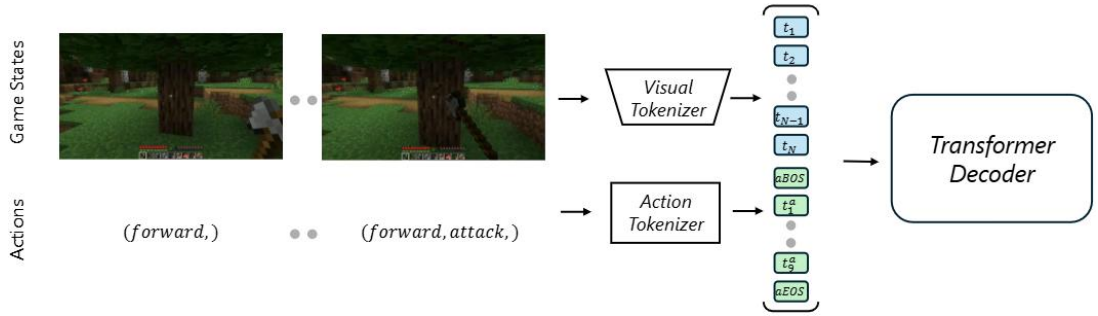


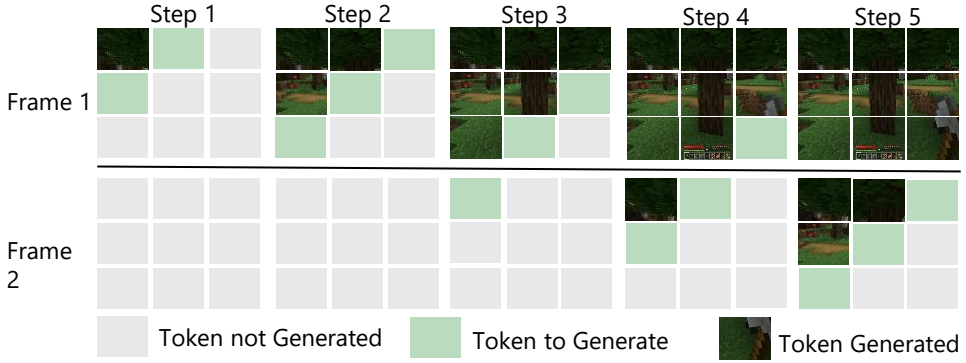
VidTok exhibits a distinct advantage in detail reconstruction fidelity and subjective viewing experience.

Tang et al. VidTok: A Versatile and Open-Source Video Tokenizer. arXiv:2412.13061.

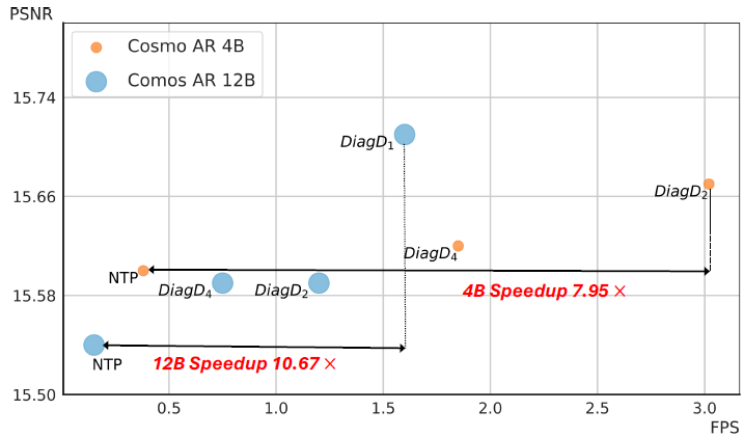
MineWorld

Action-conditioned autoregressive world model on Minecraft.





A Visual-Action model with image and action tokens concatenated interleaved.

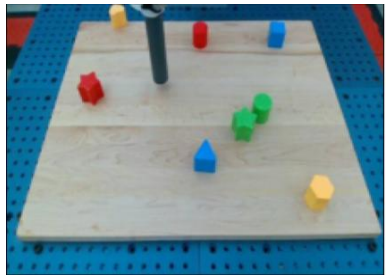


Ye et al. Autoregressive Video Generation with Diagonal Decoding. arXiv:2503.14070.

What’s next?

We observe key challenges when training world models in real-world:

- A lack of expensive interaction data.
- Different embodiments have different state space.
- Different embodiments have different action space.
- How to answer ‘what-if’ problem?

IGOR

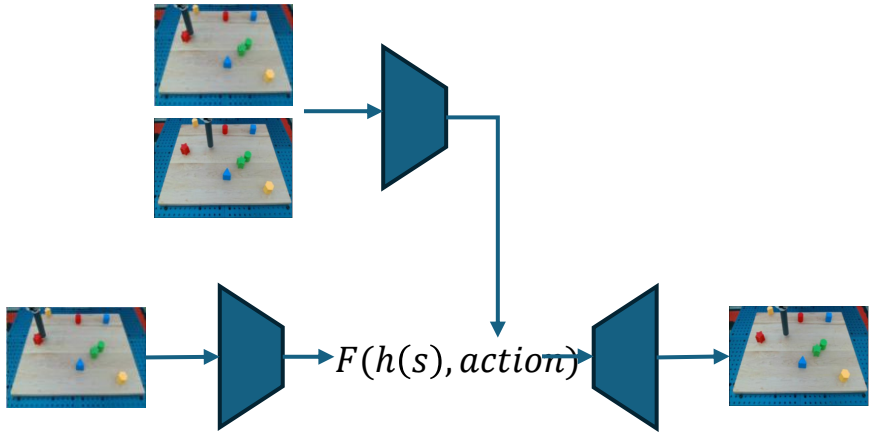
Human-to-robot cross-embodiment generalization.

We extract the semantic movement of humans manipulating various objects through latent actions, and transfer this movement to robot arms manipulating objects by applying these latent actions.

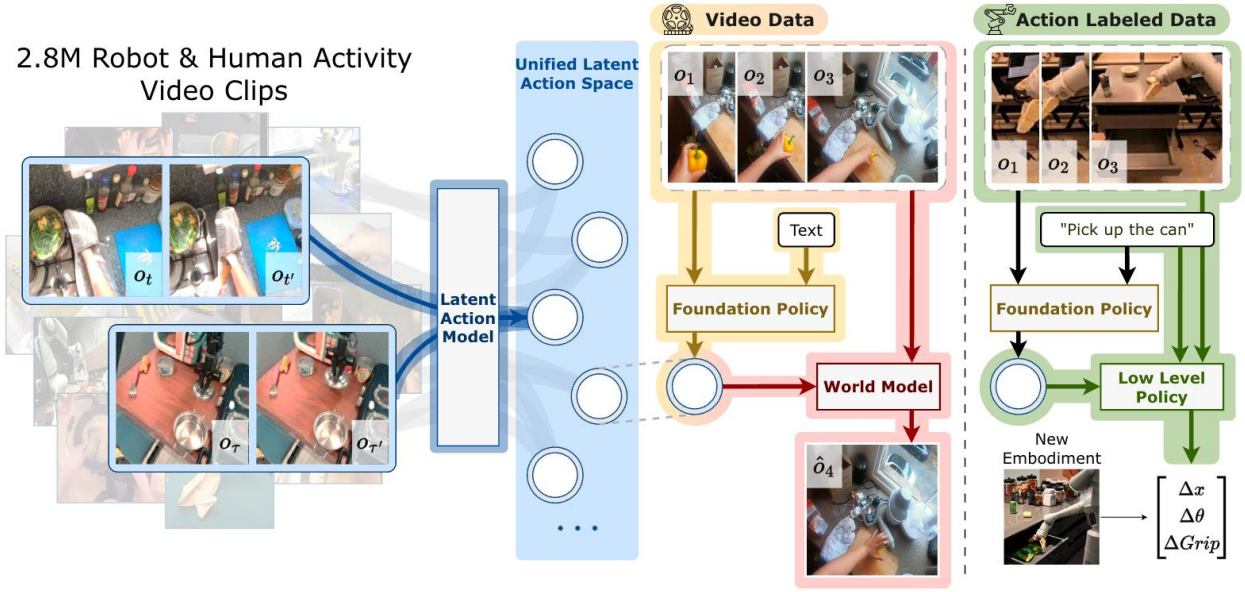
Chen et al. IGOR: Image-GOal Representations are the Atomic Building Blocks for Next-Level Generalization in Embodied AI. arXiv:2411.00785.

IGOR

Human-to-robot cross-embodiment generalization.



Latent Action Model



We compress visual changes between image and goal into latent actions.
Ideally, action tokens contain something like (object, how it moves)

Chen et al. IGOR: Image-GOal Representations are the Atomic Building Blocks for Next-Level Generalization in Embodied AI. arXiv:2411.00785.

IGOR

Human-to-robot cross-embodiment generalization.

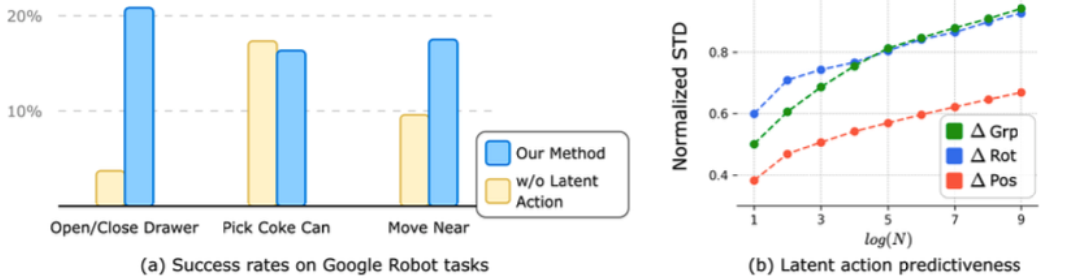
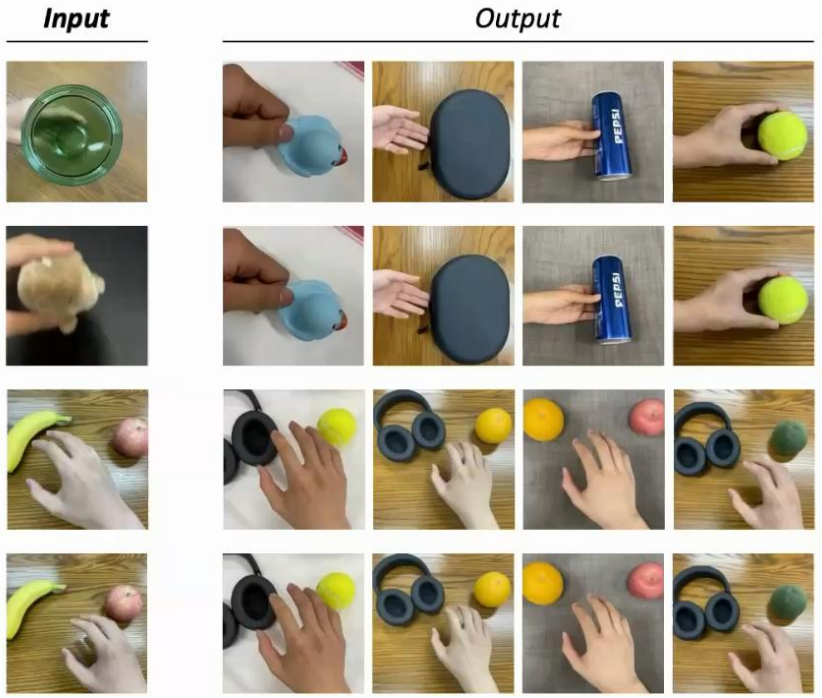


Figure 6: **(a).** Success rate of IGOR and the low-level policy trained from scratch methods on Google Robot tasks under SIMPLER simulator, finetuned on 1% data of RT-1. **(b).** Predictiveness of latent action on robot action. X-axis: $\log(N)$, where N is the number of nearest neighbours in latent action embedding. Y-axis: normalized standard derivation in action embedding with respect to movement actions (orange), rotation actions (blue), and gripper actions (green).



The latent actions contains information for predicting the real robot actions.

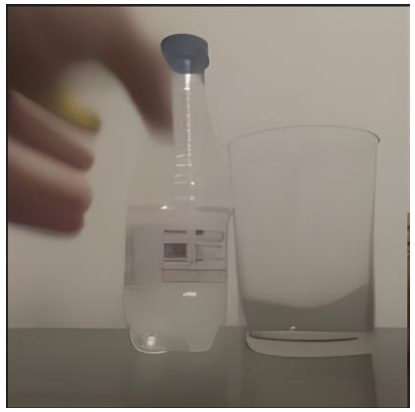
Chen et al. IGOR: Image-GOal Representations are the Atomic Building Blocks for Next-Level Generalization in Embodied AI. arXiv:2411.00785.

Video In-context Learning

Video as new interfaces to interact with the real-world.

A demonstration video

A new scene

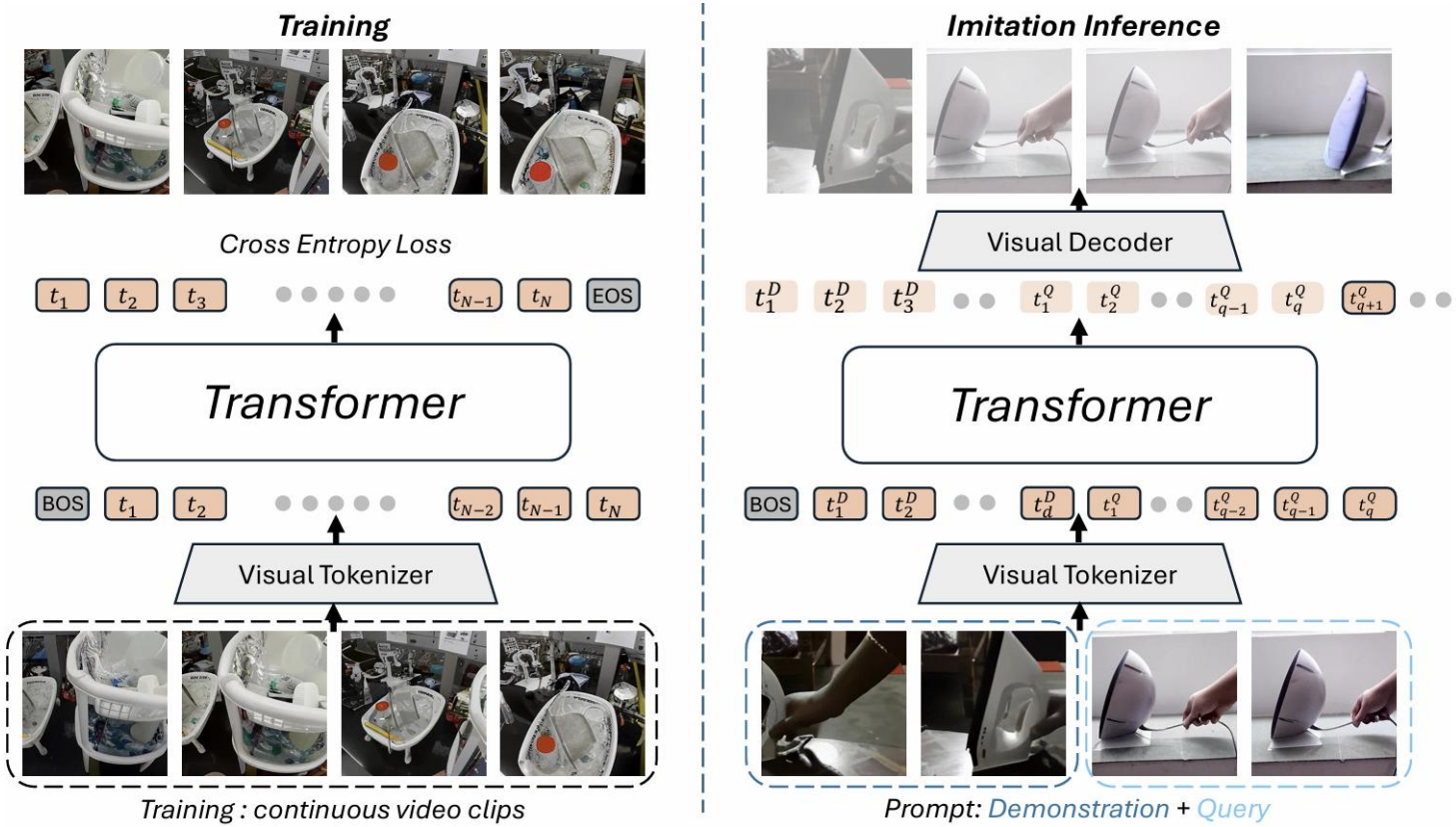


Following the same action shown in the demonstration, but **in the new scene**!

Zhang et al. Autoregressive Transformers are Zero-Shot Video Imitators. ICLR 2025.

Video In-context Learning

Video as new interfaces to interact with the real-world.



Just autoregressive training, and then the in-context learning capacity evolves in **zero-shot** manner!

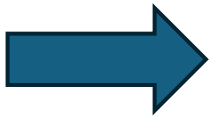
Zhang et al. Autoregressive Transformers are Zero-Shot Video Imitators. ICLR 2025.

Video In-context Learning

Different demo videos lead to different results on the given scene.

Demo video

Generation



Zhang et al. Autoregressive Transformers are Zero-Shot Video Imitators. ICLR 2025.

Explicit World Simulator

Obtaining explicit world representation from implicit priors (e.g., Sora).

Input: In a Magician's magical cabin alone in a serene forest, an alien walking on the floor, starting from the cabin’ s door to the mow near the bottom right corner.

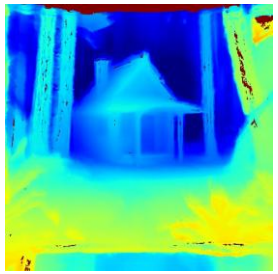
GPT-4

Decompose the complex query into sub-tasks.

Scene Generation

a Magician's magical cabin alone in a serene forest

text to 3D scene model



Actor Generation

an alien

text to 3D avatar model

Motion Generation

walking on the floor

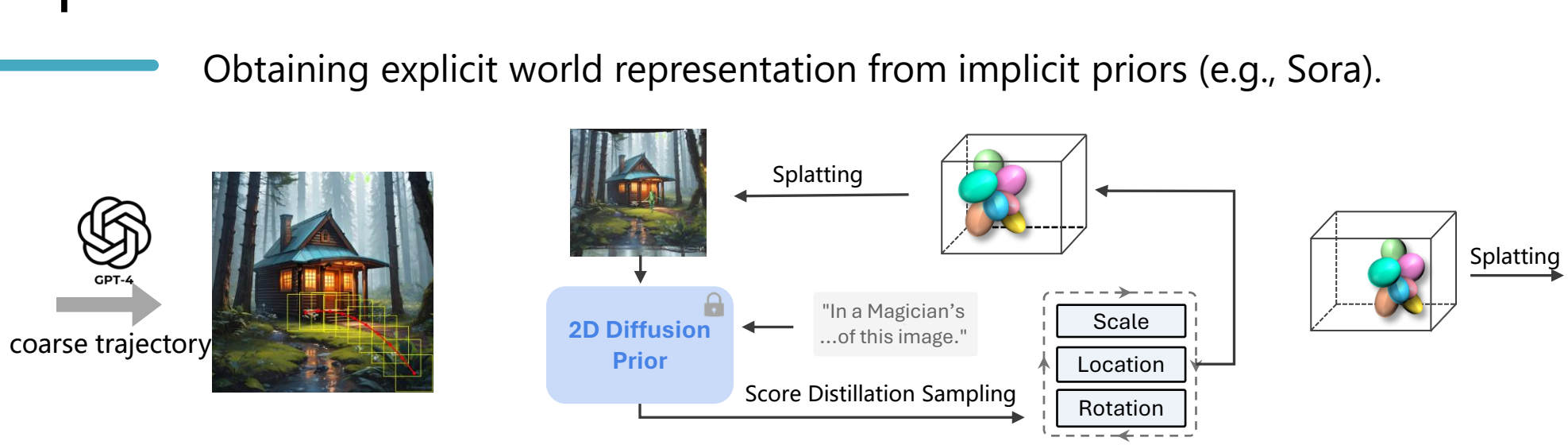
text to 3D motion model

LLM as Director, 3D as Structural Representation

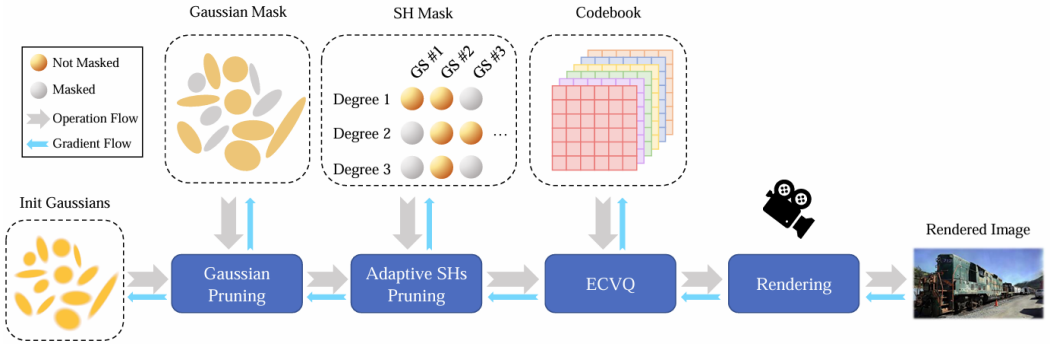
Zhu et al. Compositional 3D-aware Video Generation with LLM Director. NeurIPS 2024.

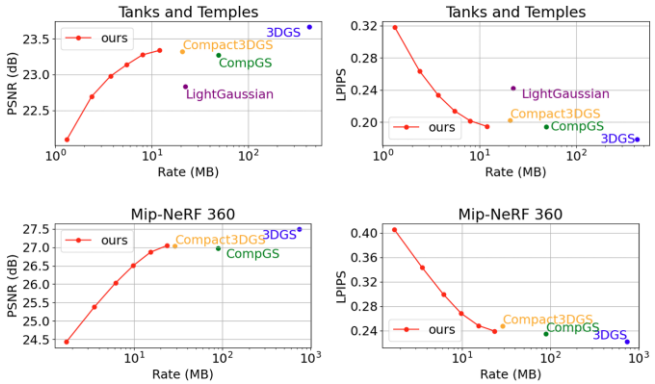
Explicit World Simulator

Obtaining explicit world representation from implicit priors (e.g., Sora).



Compose different concepts into one using priors from pre-trained LLM and diffusion models.





Zhu et al. Compositional 3D-aware Video Generation with LLM Director. NeurIPS 2024.
Wang et al. End-to-End Rate-Distortion Optimized 3D Gaussian Representation. ECCV 2024.

Achieve 40x compression for 3D Gaussian Splatting

Thanks

Tianyu He
Microsoft Research Asia