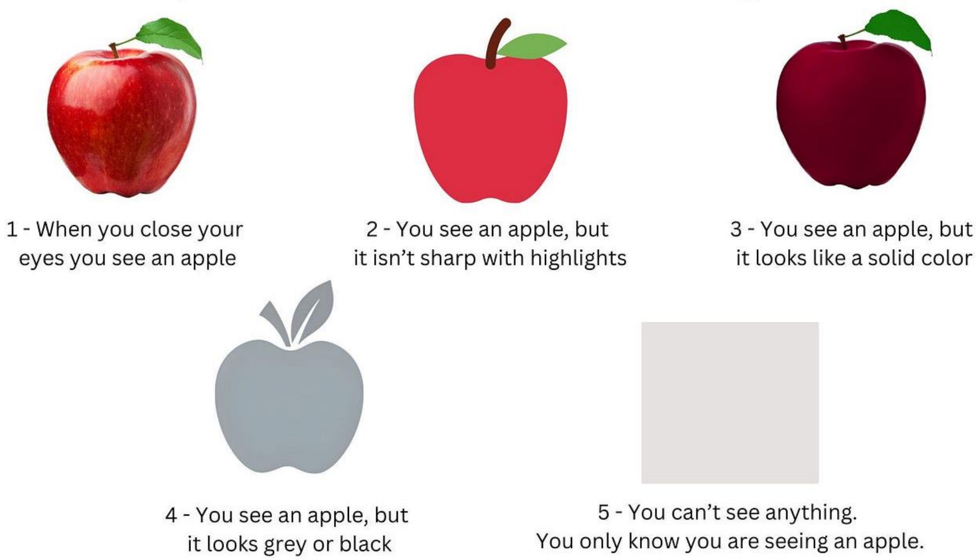
Generative (Mental) World Explorer

Jieneng Chen
PhD candidate, Siebel Scholar Class 2025
Dept. Computer Science, Johns Hopkins University
04/04/2025

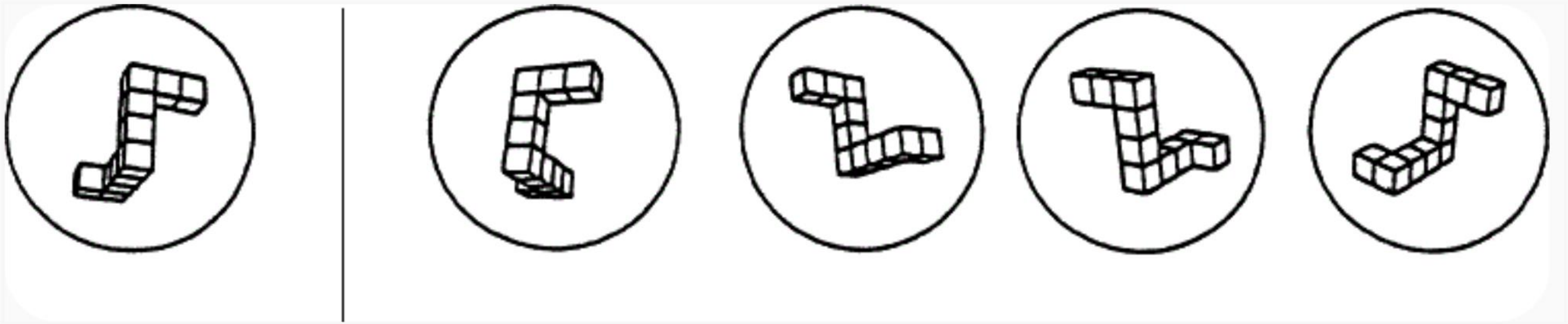
GenEx

Mental imagery

- Simple test: **Close your eyes and visualize an apple.**
- How vivid?



Mental rotation



Which image is the same as the original image, aside from its orientation?

https://www.proprofs.com/quiz-school/story.php?title=mental-rotation-task

Mental imagery

The experience of “seeing” (or otherwise sensing)

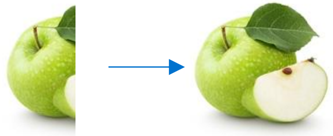
in the **mind’s eye** without direct external sensory input.

Mental imagery

Why do different people visualize apples differently in their mind's eye?

Gestalt Psychology

- **Holism: the whole is greater than the sum of its parts.**
  - Structure & Part-whole relationship;
  - How the world emerges from the integration of its parts.



- **Law of past experience**:
  - Our perception of a apple is not solely derived from its shape, color, or size as sensory inputs; it also incorporates our past experiences and impressions of flowers.
  - Together, these elements form our holistic perception of the apple.

Generative Prior ⬅ Gestalt Psychology

- Holism: the whole is greater than the sum of its parts.
- Law of past experience
- The era of **generative priors**:

  learn the visual commonsense (e.g., **holism**)

  from huge amount of data (**past experience**),

  encoding high-level structural regularities, as parameters (**deep learning**)

Generative Priors ➡ Mental World Models

Novel view from observed view

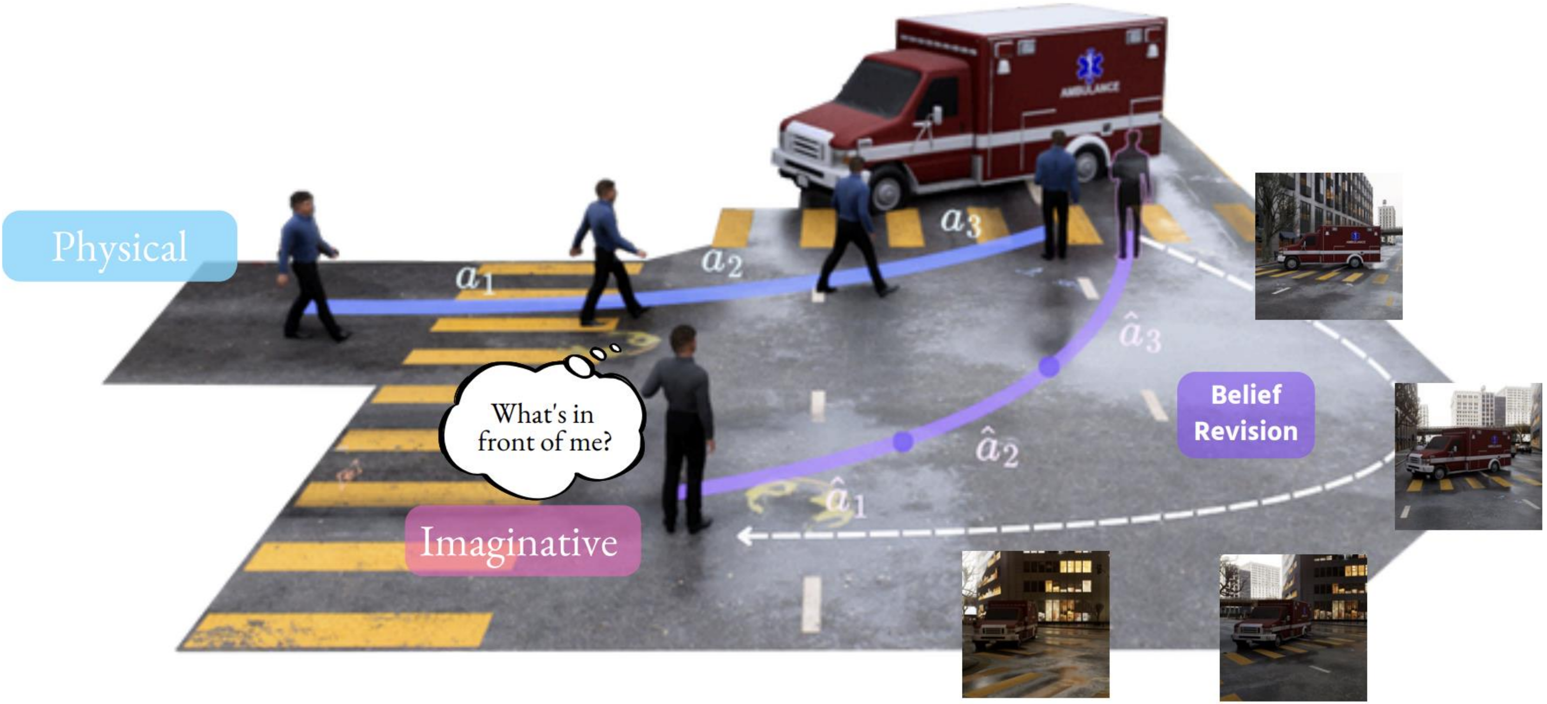
Part-to-Whole Relationship

Johns Hopkins Whiting School of Engineering

Why Mental Imagery/Models Matter?

- Planning with partial observation is challenging.
- Humans can imagine unseen parts of the world through a mental exploration and revise their beliefs with imagined observations.
- Such updated beliefs can allow them to make more informed decisions, without necessitating the physical exploration of the world at all times.

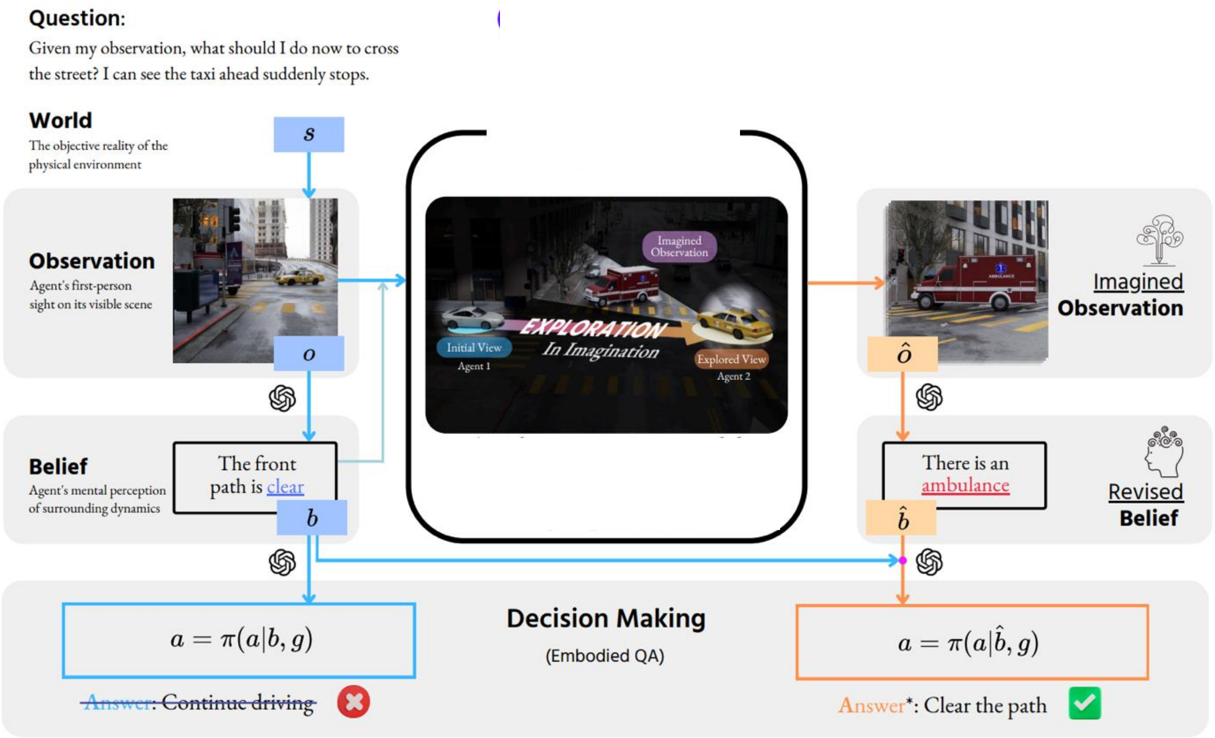
Mental Exploration



Mental Exploration Enhances Decision Making

Youtube:
https://www.youtube.com/watch?v=cf4apIcnPtU&ab_channel=CenterforLanguage%26SpeechProcessing%28CLSP%29%2CJHU

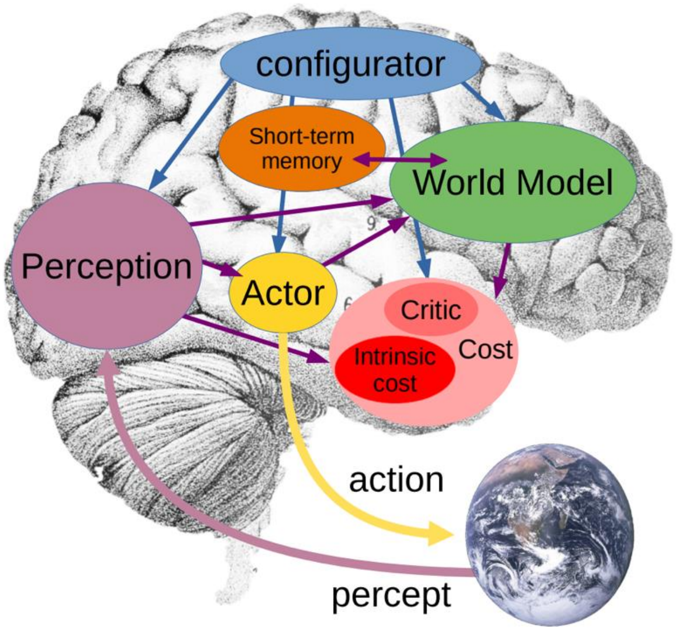
Mental Exploration Enhances Decision Making



World models: Computational Counterpart of Mental Models

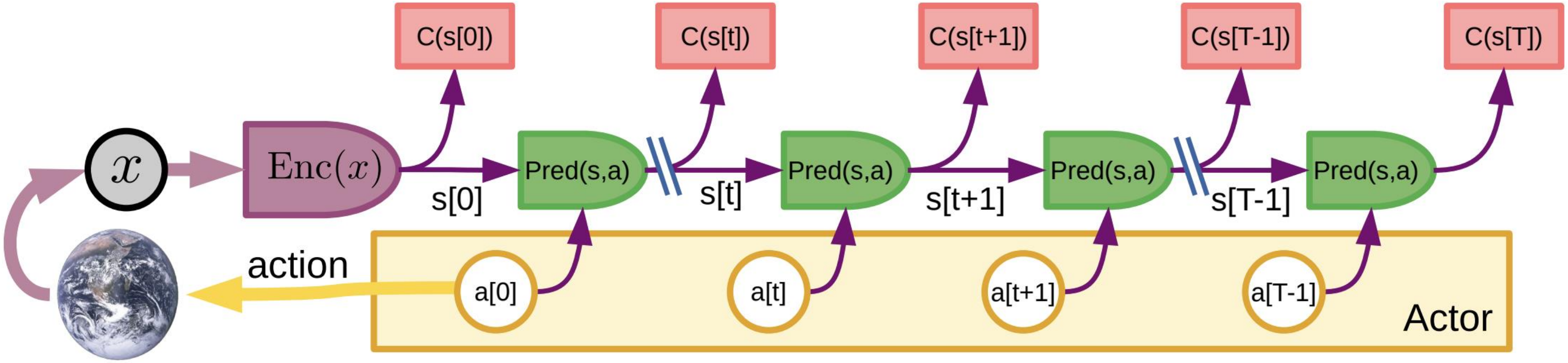
- **Definition**: multi-sensory neural networks that offer a **predictive distribution over "changes" in the world.**
  - $p(s_t|s_{t-1}, a_t)$
- **Functionality**: mimic human understanding and interaction by predicting **future world states** (e.g., the existence, properties and location of the objects in a scene) to help agents make informed decisions.

World Models



The world model module predicts possible future world states as a function of imagined actions sequences proposed by the actor

World Models



The world model recursively predicts an estimate of the world state sequence using $s[t+1] = \mathrm{Pred}(s[t], a[t])$

World Models Summary

- multi-sensory neural network that offer a **predictive distribution over "changes" in the world.**
  - $p(s_t|s_{t-1}, a_t)$

Engineering Mental World Models

- Develop **generative models** grounded in physical world.
- The models are capable of predicting world dynamics conditioned on actions.

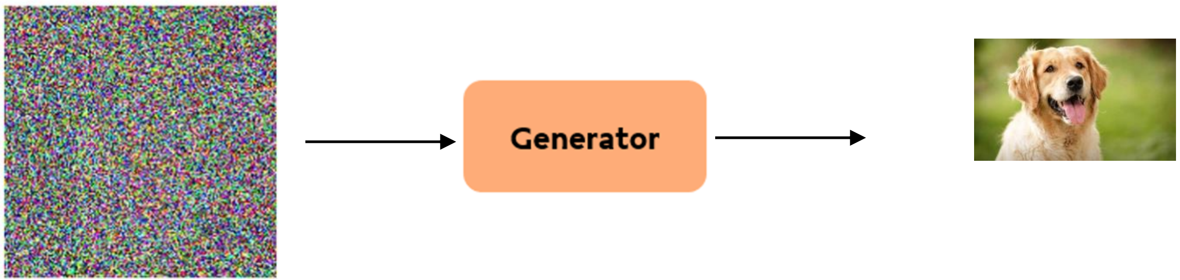
Inference

- **Gather** imagined observation from (interactively) imaginative exploration.
- **Planning** with imagined observation.

Generative Models

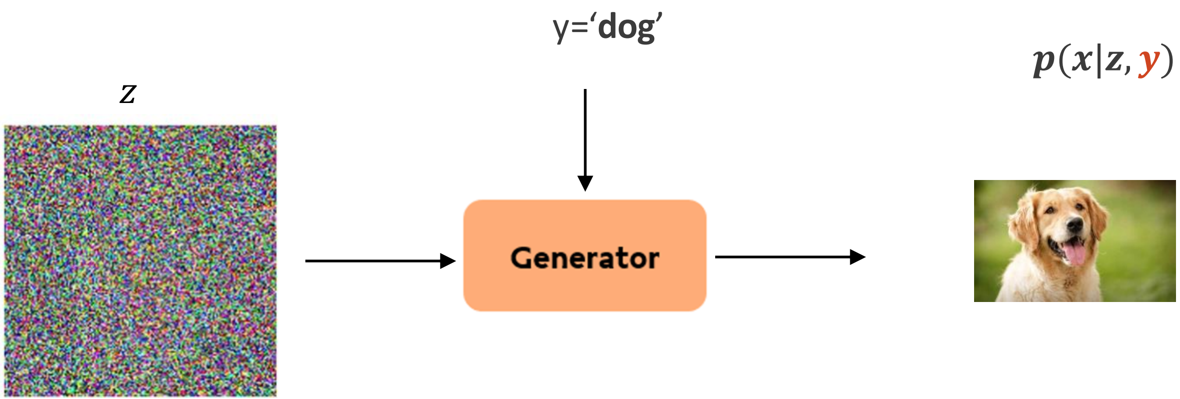
Generative Models

- z : a random variable sample from normal distribution
- $\boldsymbol{x}$: a predicted data, with the learnt distribution $\boldsymbol{P(x|z)}$
- Ideally, we expect the output X is a real image without corruption.



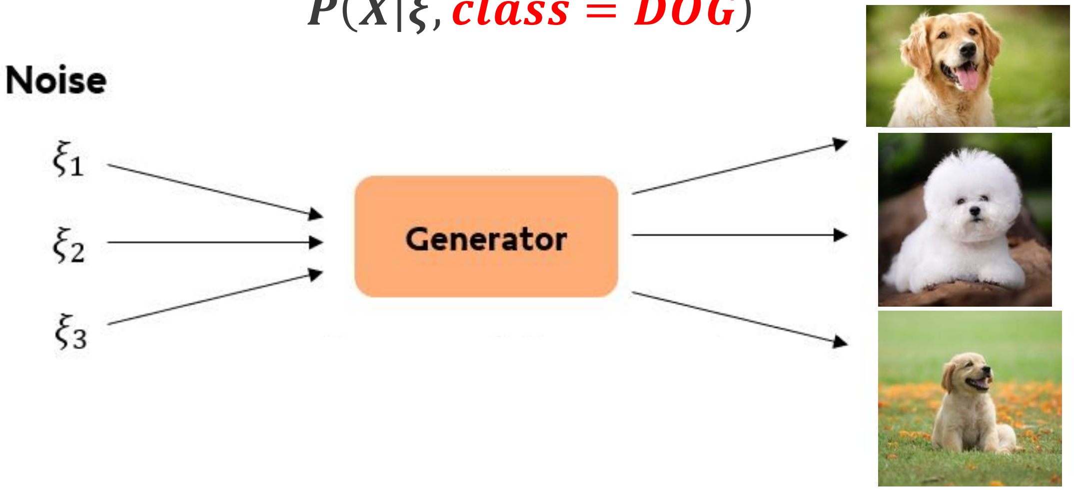
Conditional Generative Models

- z : a random variable sample from normal distribution
- $\boldsymbol{x}$: a predicted data, with the conditional distribution $\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})$
- Ideally, we expect the output X is a real image without corruption.



Class-Conditioned Generative Models

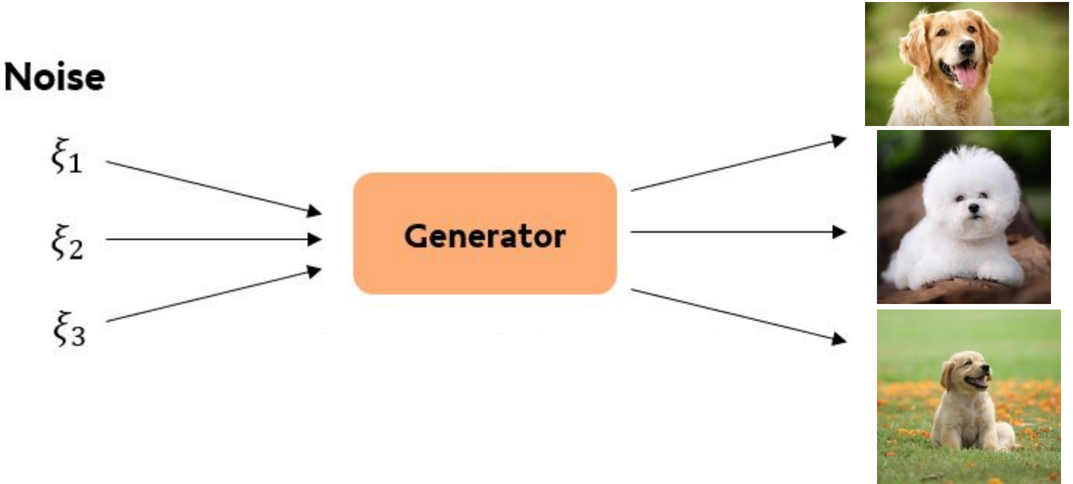
$$P(X|\xi, class = DOG)$$



Class-Conditioned Generative Models

- Class condition can have the same effect of text condition.
- Essentially, the class label of ‘DOG’ has the same meaning of the text ‘a photo of dog’.

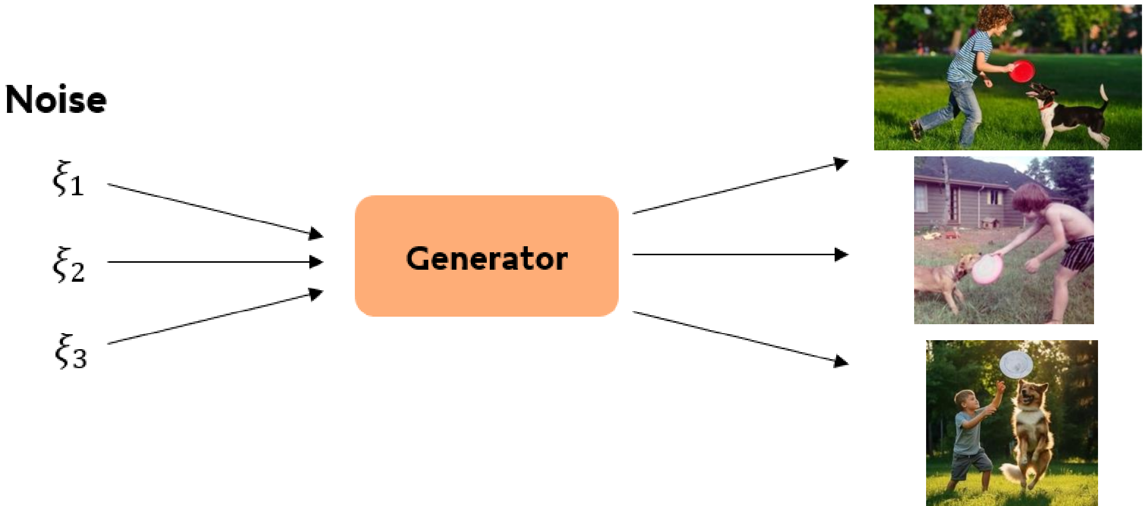
$$P(X|\xi, class = DOG) = P(X|\xi, 'A\ photo\ of\ dog')$$



Text-Conditioned Generative Models

- progressed from class-conditioned to text-conditioned approaches.

$$P(X|\xi, '\ a\ boy\ is\ playing\ frisbee\ with\ a\ dog')$$



Text-Conditioned Video Generative Models

OpenAI Sora

- The prediction is not limited to image!
- Video includes dynamics, and thus generating video is harder.

$$P(X|\xi, \text{'}\textbf{\textit{A kid throws a bubble into air}}\text{'})$$

$\xi \longrightarrow$ Generator $\longrightarrow$

Limitation of Sora-like Models

- 3D consistency?
- Physical commonsense?
- Interaction?

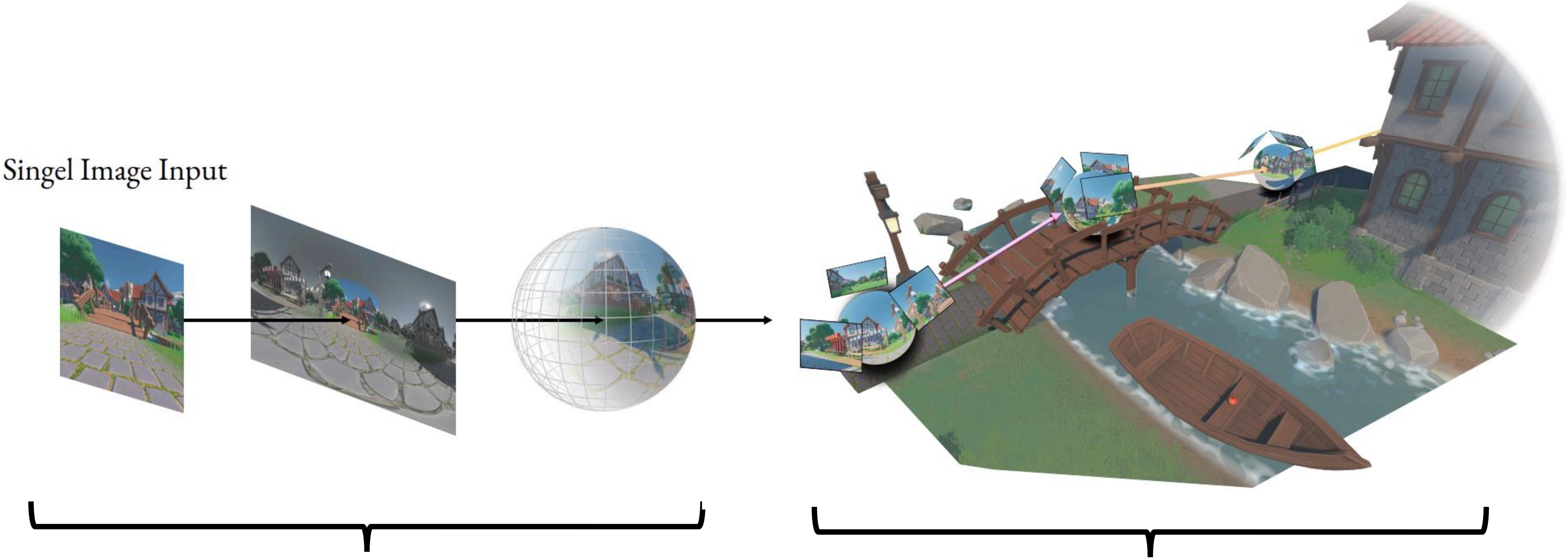
Ground Generative Models in the Physical world

- **Data collected** from 3D physical world (rather than Youtube video)
- **Action as condition** (rather than text)
  - ‘the agent is moving two meters forward’
- **Predict the world dynamics**

Generative World Explorer

ICLR
International Conference On
Learning Representations

Johns Hopkins
Whiting School
of Engineering



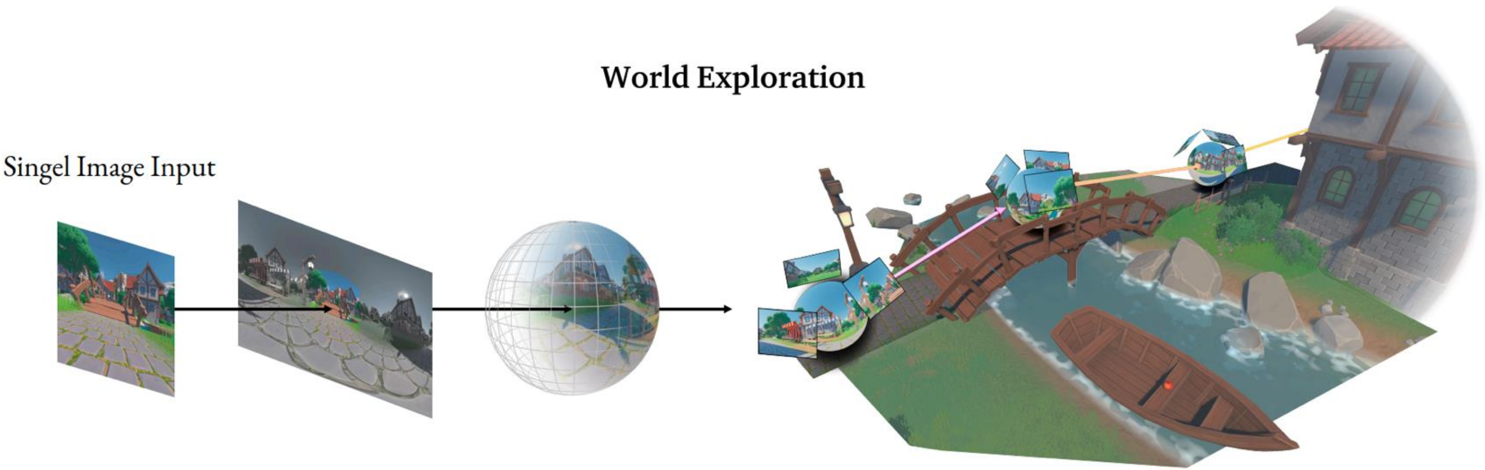
- **World initialization** (§2.2): Given the initial image i_0 and a language description l_0, the anchor 360° world view x_0 is sampled from:

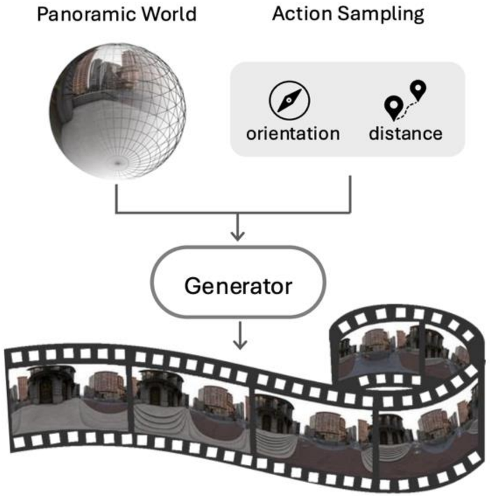
$$x_0 \sim p_{\theta_1}(x \mid i_0, l_0),$$

- **World transition** (§2.3): Given the chosen action a_t, the next world view $\mathbf{x}_t$ is sampled from:

$$\mathbf{x}_t = (x_t^0, x_t^1, \ldots, x_t^S) \sim p_{\theta_2}(\mathbf{x} \mid x_{t-1}^S, a_t),$$

where θ_2 is a 360° panoramic video generator, $t = 1, \ldots, T$, and $x_0^S := x_0$.



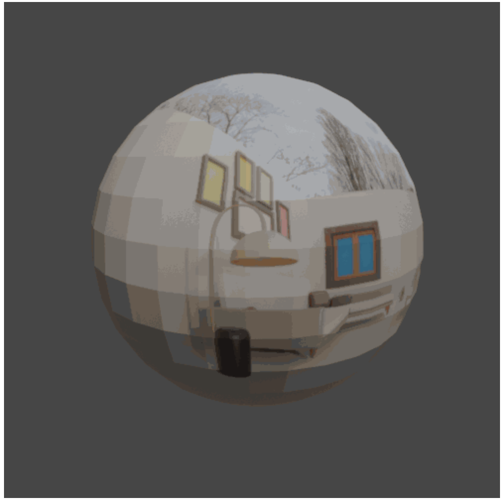


- **World transition** (§2.3): Given the chosen action a_t, the next world view $\mathbf{x}_t$ is sampled from:

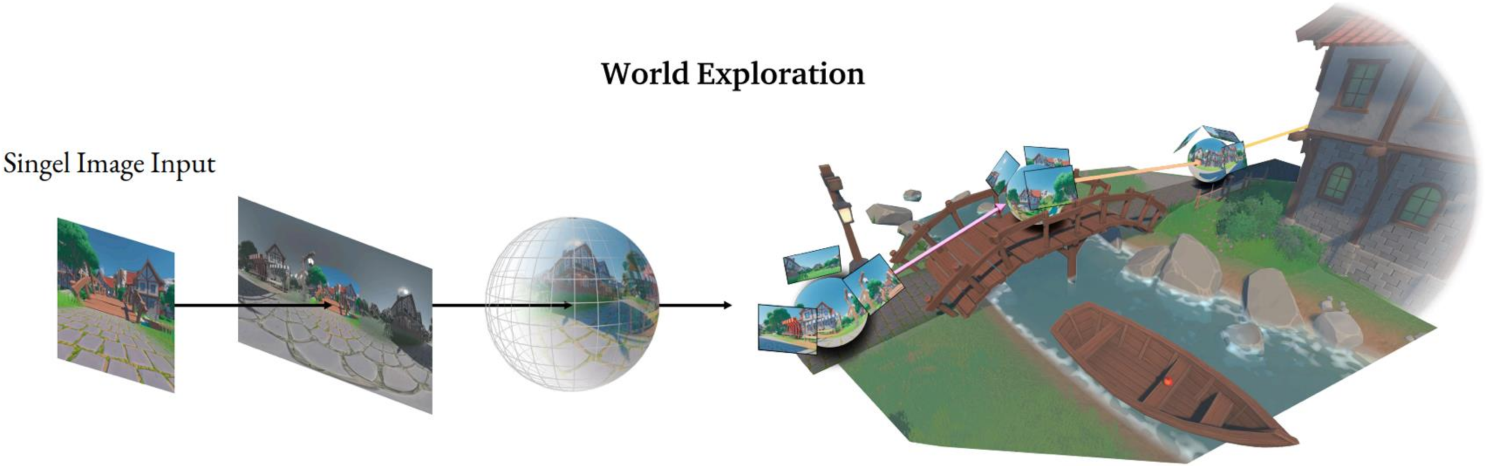
$$\mathbf{x}_t = (x_t^0, x_t^1, \dots, x_t^S) \sim p_{\theta_2}(\mathbf{x} \mid x_{t-1}^S, a_t),$$

where θ_2 is a 360° panoramic video generator, $t = 1, \dots, T$, and $x_0^S := x_0$.

Exploration at Any Direction



Forward Exploration (using diffusion model)



- **World transition** (§2.3): Given the chosen action a_t, the next world view $\mathbf{x}_t$ is sampled from:

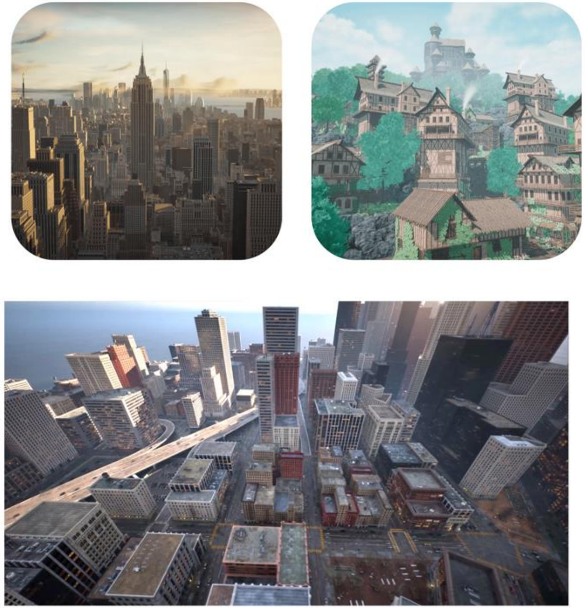
$$\mathbf{x}_t = (x_t^0, x_t^1, \ldots, x_t^S) \sim p_{\theta_2}(\mathbf{x} \mid x_{t-1}^S, a_t),$$

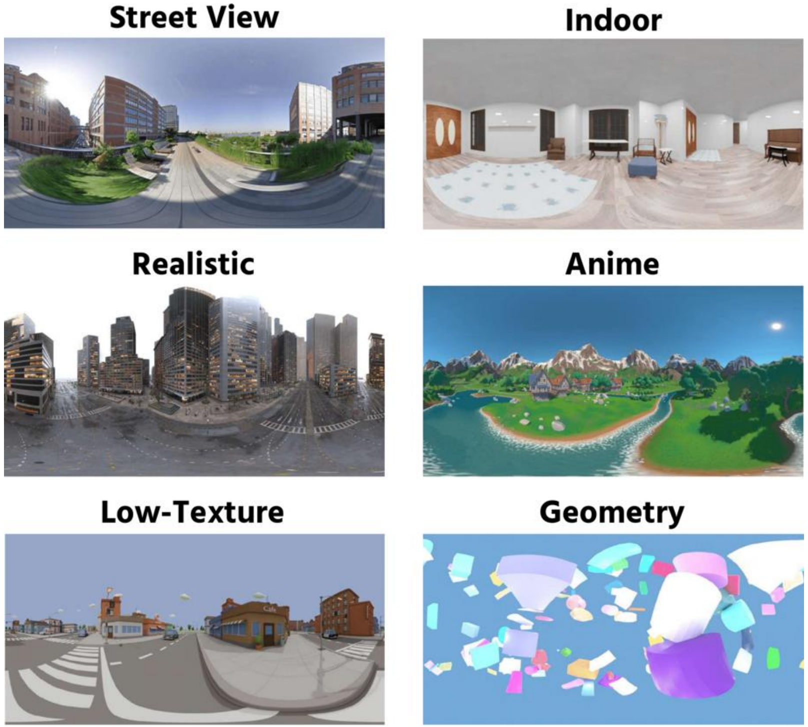
  where θ_2 is a 360° panoramic video generator, $t = 1, \ldots, T$, and $x_0^S := x_0$.

Train on the Data from 3D Synthetic Engines

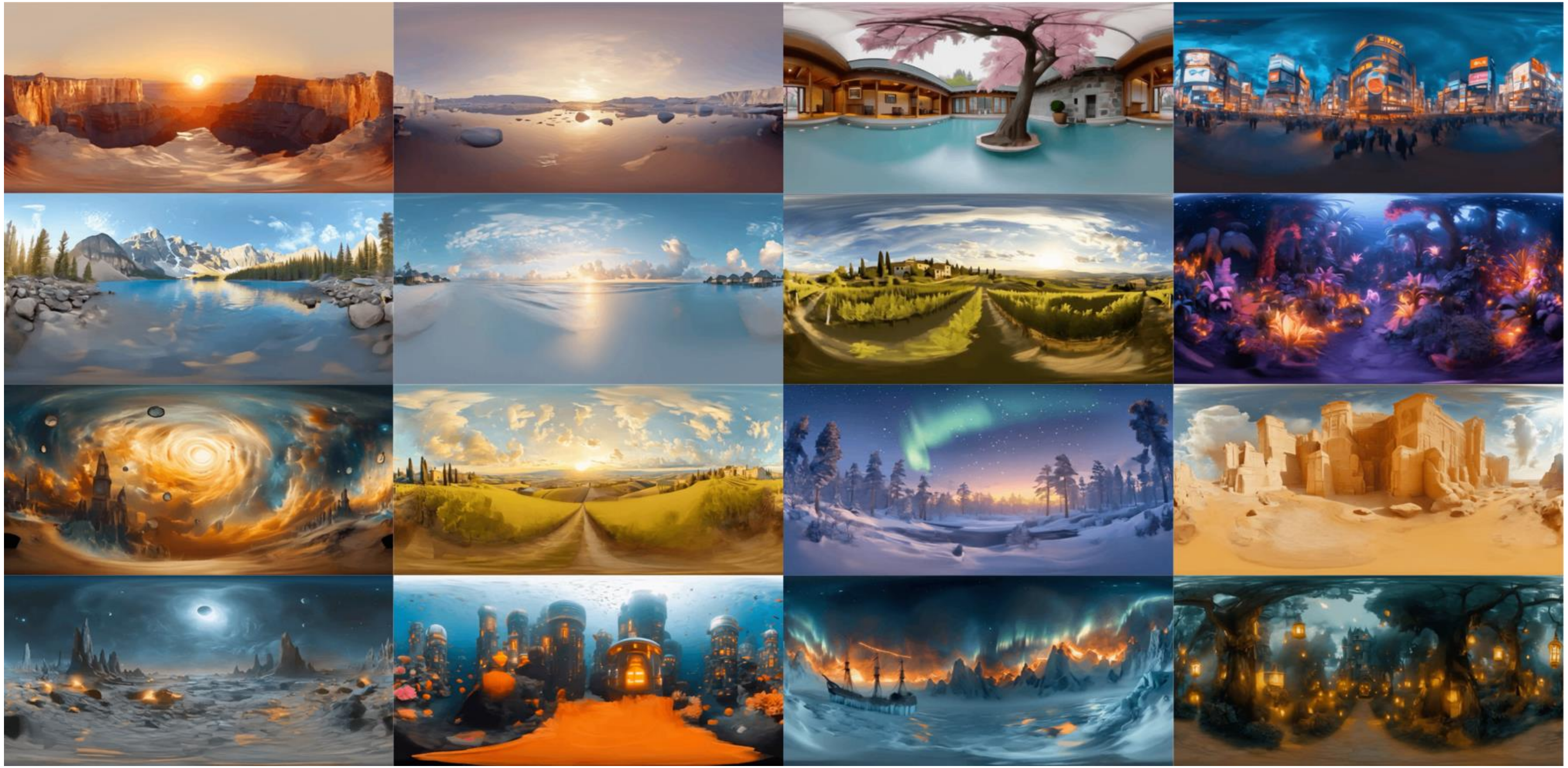
UNREAL ENGINE

Unity





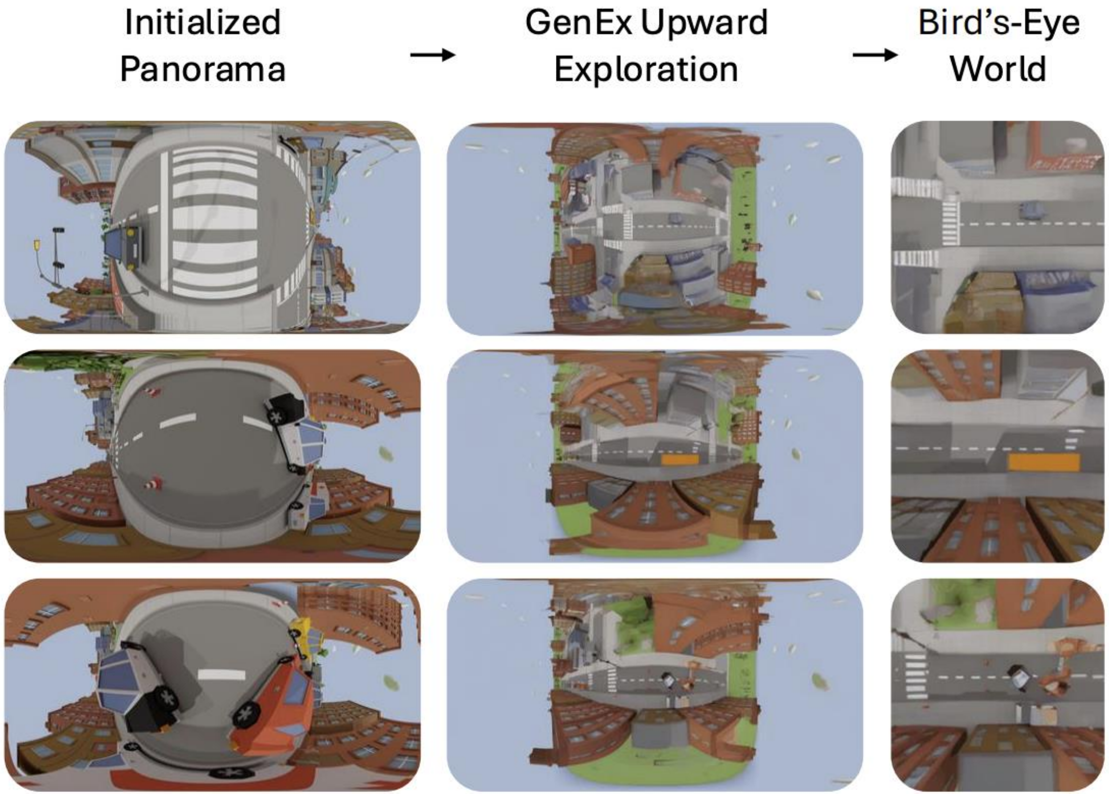
Test and Explore in Diverse Scenes



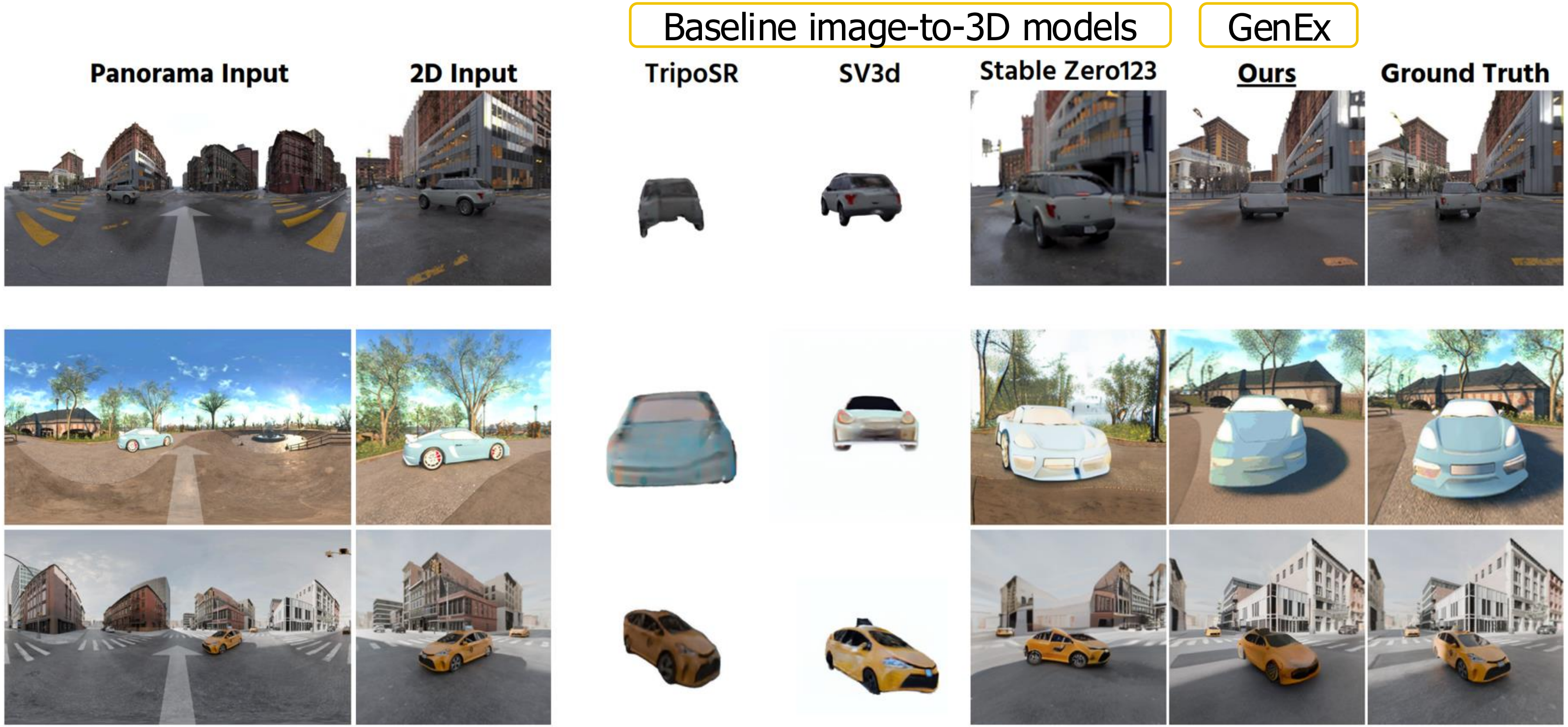
Step into the picture. Imagine the world within.

Explore

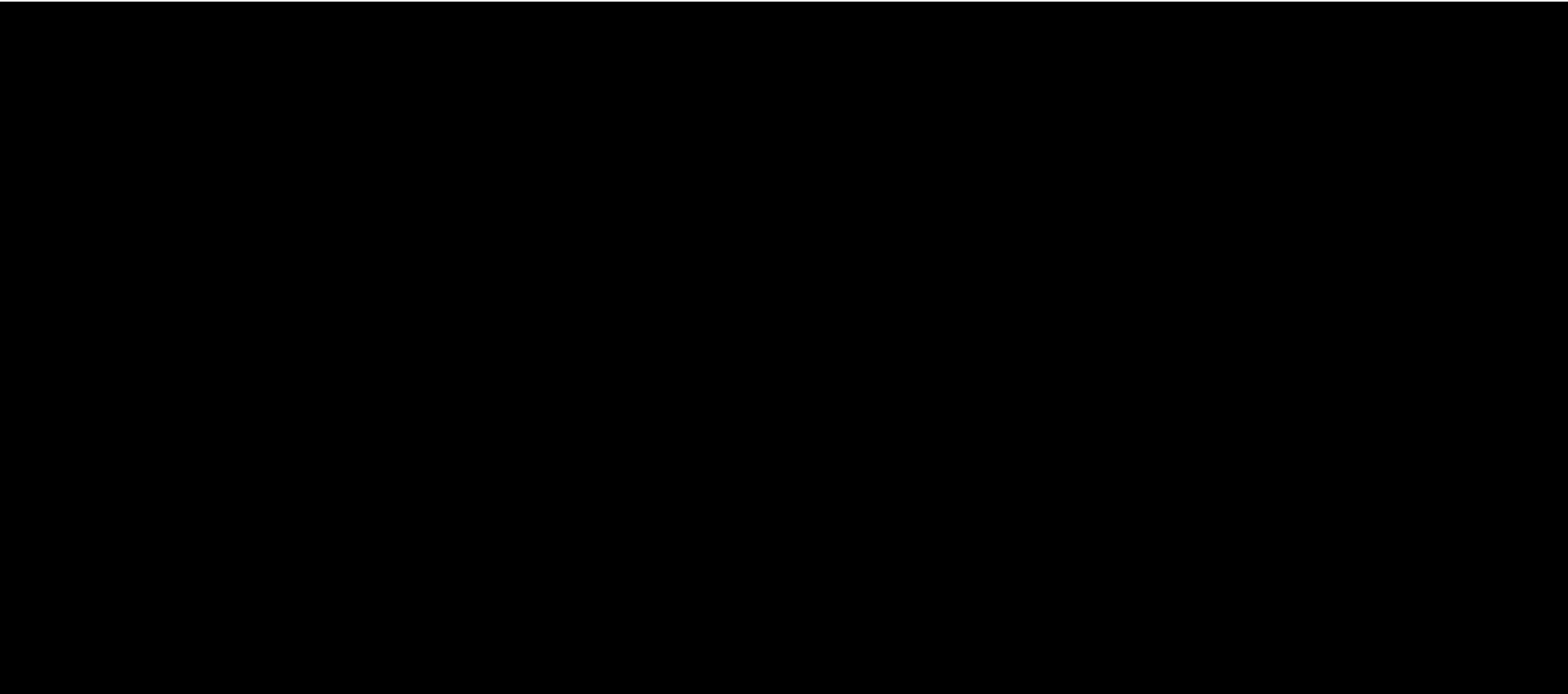
Generating Bird's-Eye Worlds



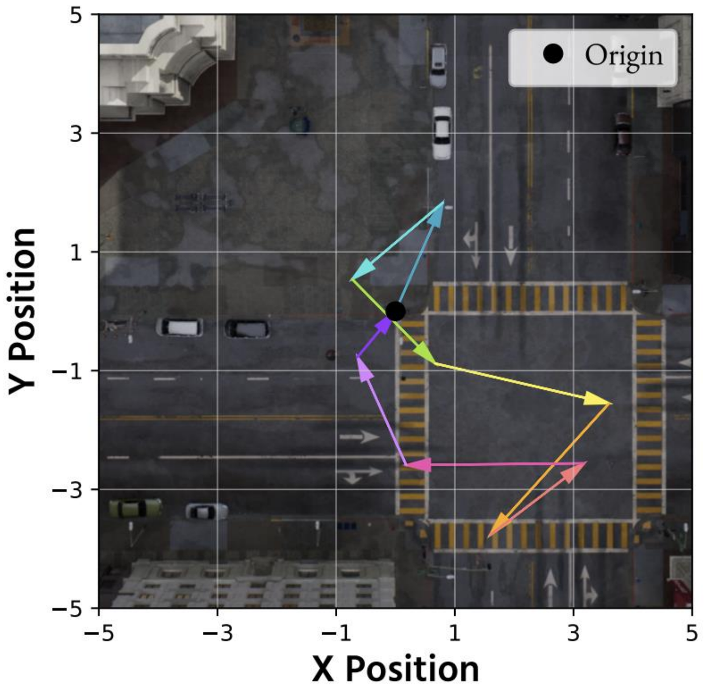
3D Consistency

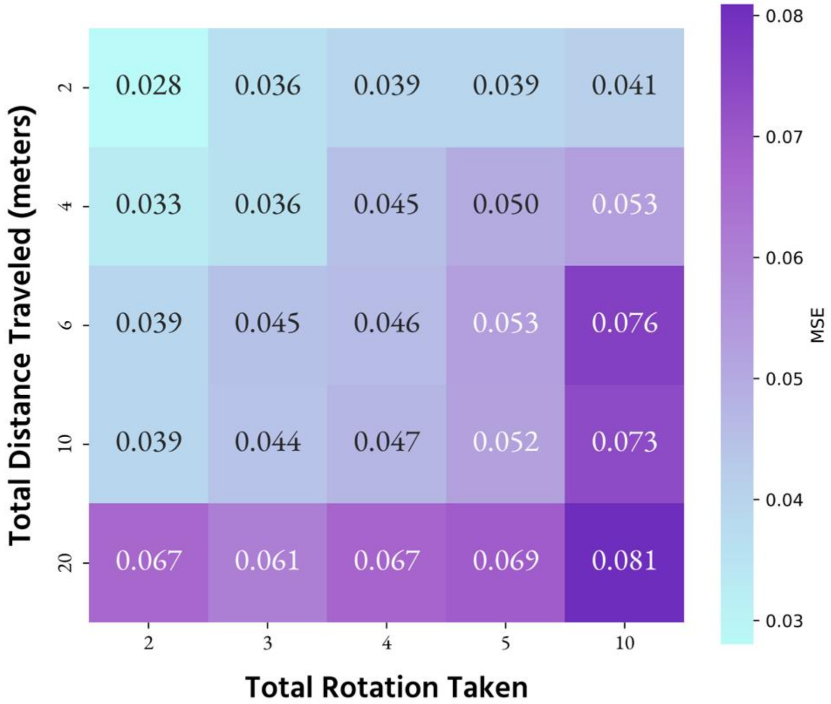


Loop Consistency when Navigating in the city

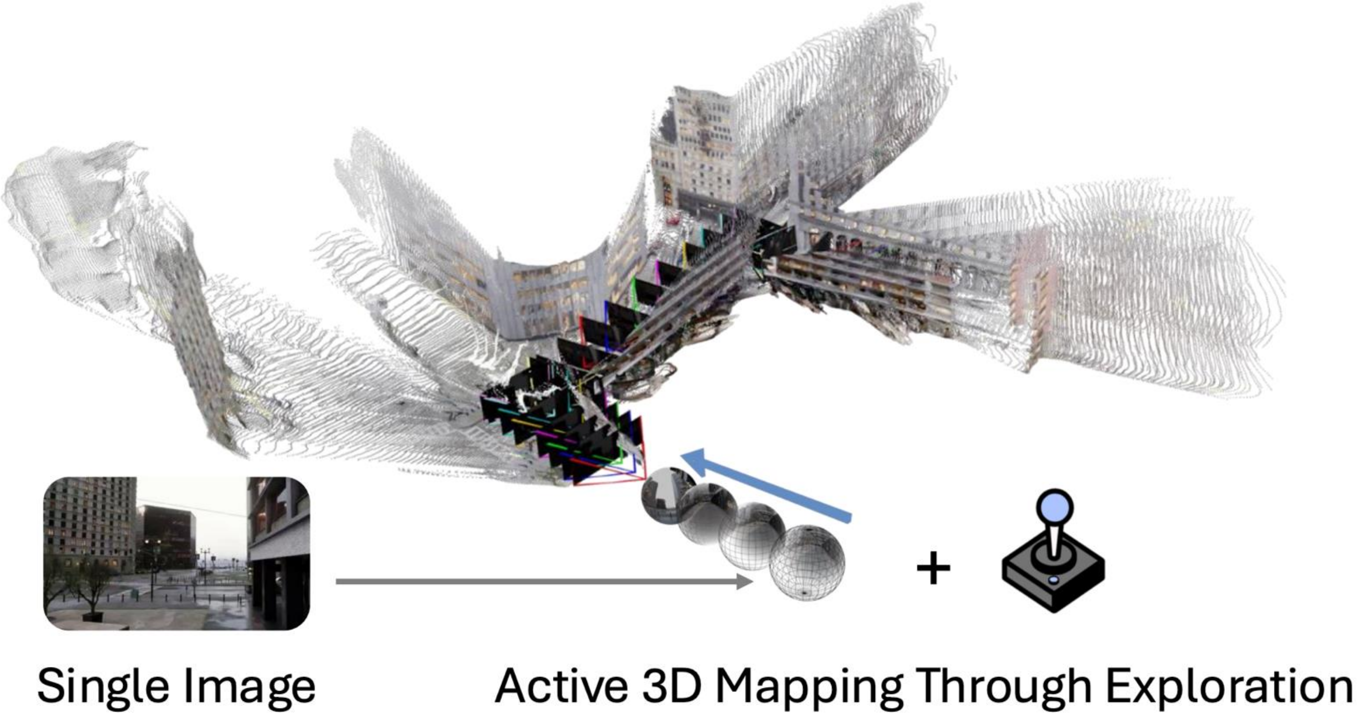


Loop Consistency

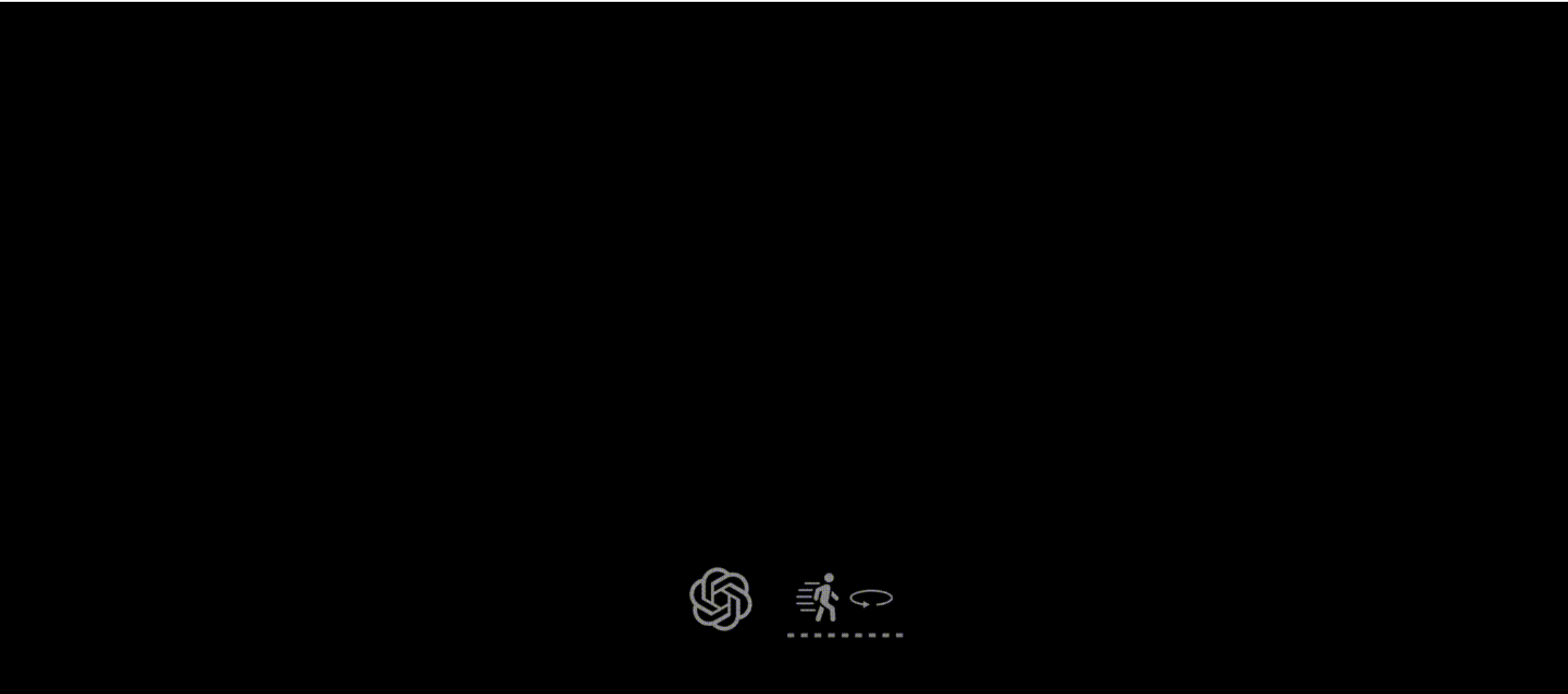




Active 3D mapping through exploration

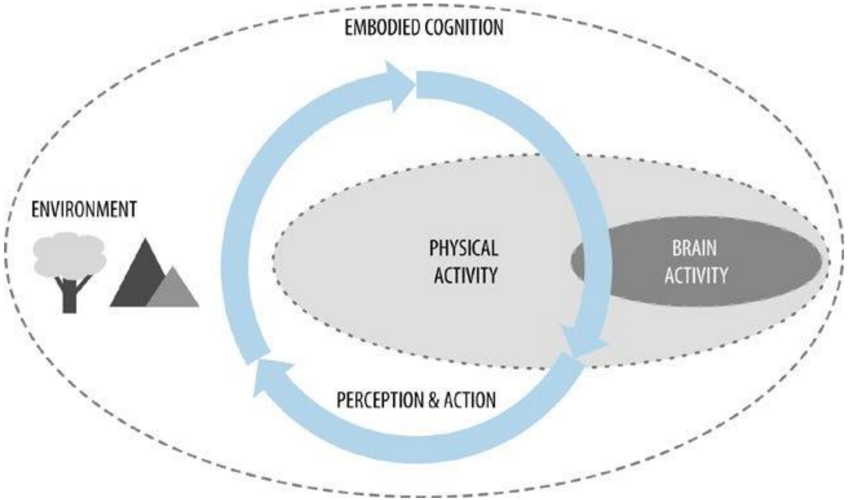


Active 3D mapping through exploration



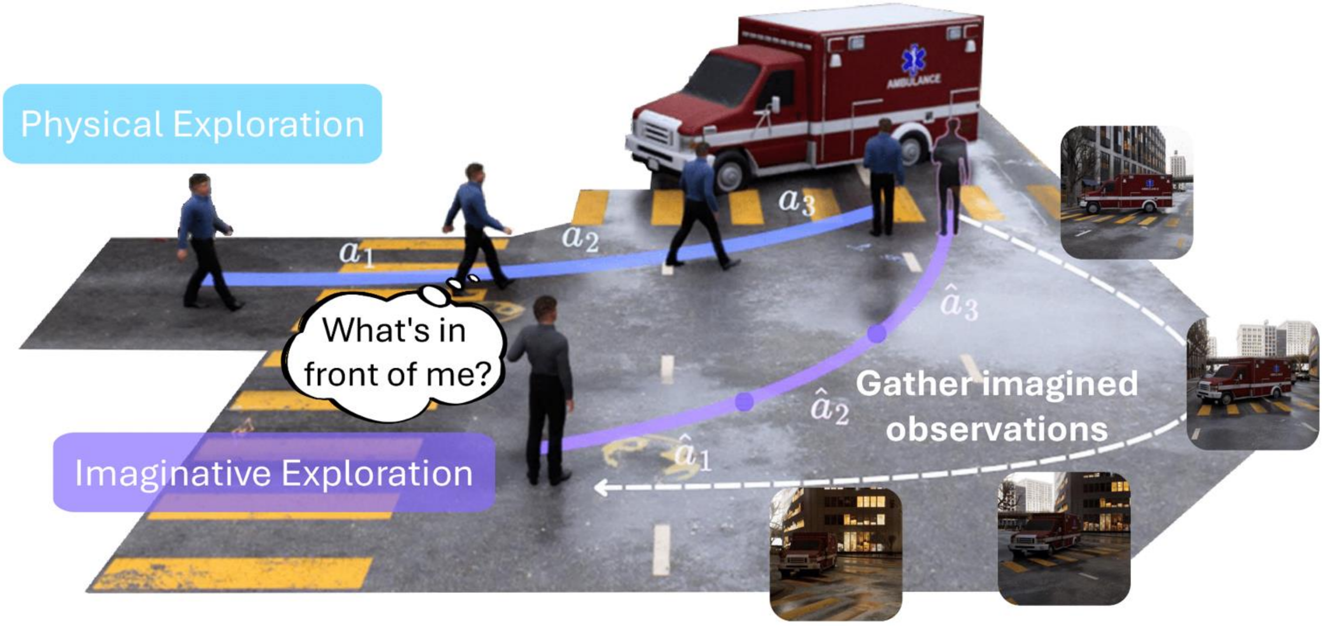
Connecting GenEx to Embodied AI

Embodied AI



- Definition of embodied AI:
  - The **embodiment hypothesis**, also known as embodied cognition, is the idea that intelligence is a result of how an agent **interacts** with its environment.
- Connect GenEx to Embodied AI:
  - Predict the change of environment after interaction (agent exploration).

Replacing Physical Exploration



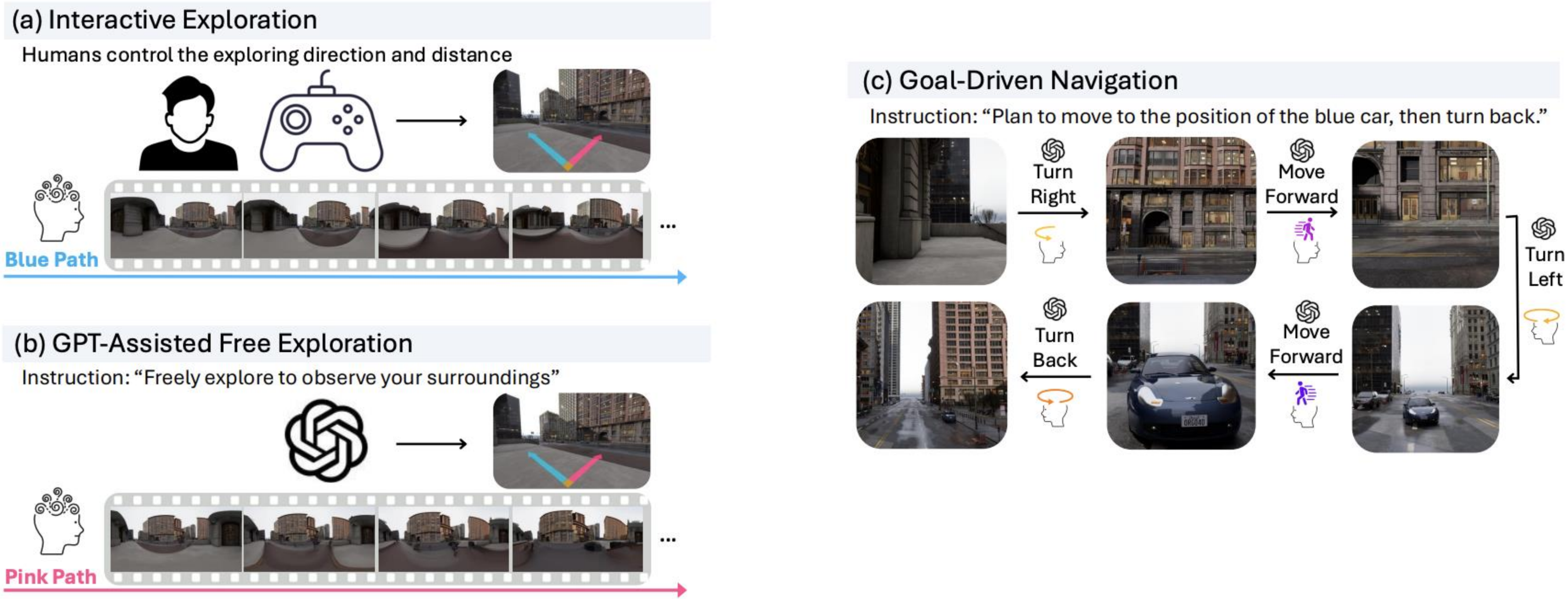
Exploration Policy

- The exploration action is decided by a policy:

$$a_t = \arg\max_a \pi_{explore}(a|x_{t-1}^S, \mathcal{I})$$

  - I is the instruction that specifies the exploration mode to be either human interaction or assisted by a GPT
  - x_{t-1}^S denotes the latest explored view from the previous step $t-1$.
  - action $a_t = (\alpha_t, d_t)$ defines how the agent rotates its field of view with the rotation angle αt and moves forward with dt distance

Embodied Exploration: Three Modes



Imagination-Augmented Policy

Require: • Initial observation i_0 and world initialization description l_0

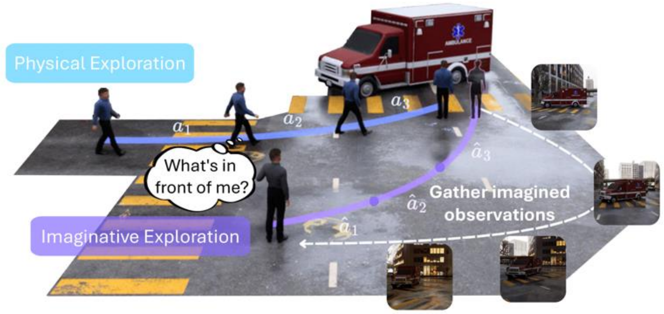
- A goal g to answer embodied questions. *E.g*, "Danger ahead—stop or go ahead?"
- A navigation instruction $\mathcal{I}$. *E.g*, "Navigate to the unseen parts of the environment."
- GenEx $p(\mathbf{x}_{0:T}|i_0, l_0, \mathcal{I})$ defined in § 2.1 and Algorithm 1.
- An embodied policy $\pi_{\theta_3}(A|o, g)$ conditioned on observation variable o and goal g.

1: **Gather imagined observations** with GenEx:

$$\mathbf{x}_{0:T} \sim p(\mathbf{x}_{0:T} \mid i_0, l_0, \mathcal{I})$$

2: **Select an action with imagined observations** to maximize the policy:

$$A = \arg\max_{A} \pi_{\theta}(A \mid i_0, \mathbf{x}_{0:T}, g)$$



Multi-Agent Imagination-Augmented Policy

- Step 1 : Gather imagined observations by exploring the position to agent-k

$$\mathbf{x}_{0:T}^{(k)} \sim p(\mathbf{x}_{0:T} \mid i_0, l_0, \mathcal{I}_k)$$

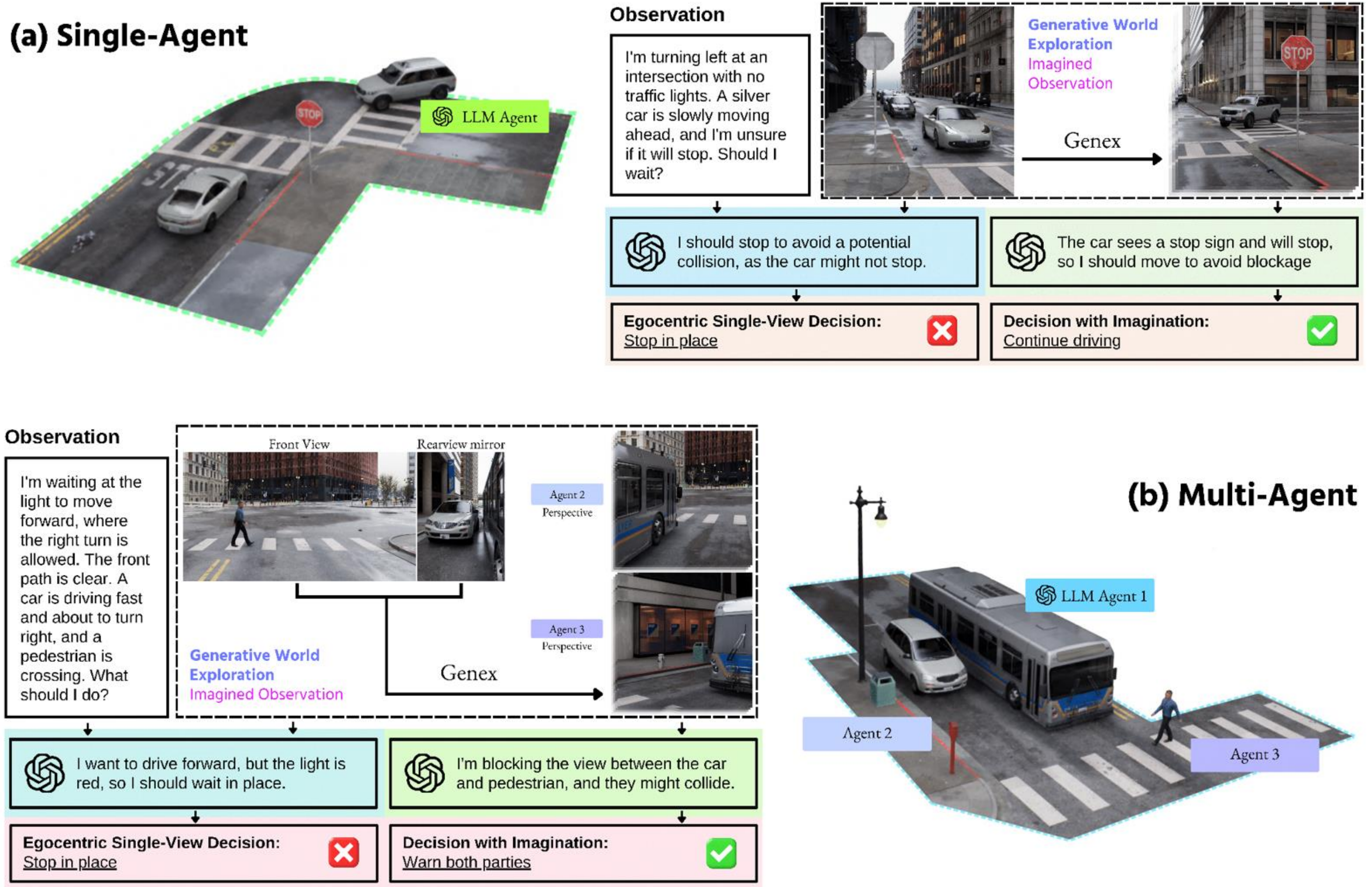
- Step 2: Repeat Step 1 a total of K times, then imaginatively explore the resulting positions of all K agents in our generated explorable world

$$\{\mathbf{x}_{1:T}^{(k)}\}_{k=1}^{K} = (\mathbf{x}_{1:T}^{(1)}, \mathbf{x}_{1:T}^{(2)}, \ldots, \mathbf{x}_{1:T}^{(K)})$$

- Step 3: Select an embodied action A with imagined observations to maximize the policy

$$A = \arg\max_{A} \pi_{\theta_3}(A \mid i_0, \{\mathbf{x}_{1:T}^{(k)}\}_{k=1}^{K}, g)$$

Embodied Decision Making



Multi-Agent Imagination-Augmented Policy

Enrich real observation with imaginative observation

Method	Acc. (%)	Confidence (%)	Logic Acc. (%)
Random	25.00	25.00	-
Human Text-only	21.21	11.56	13.50
Human with Image	55.24	58.67	46.49
Human with **GenEx**	**77.41**	**71.54**	**72.73**
Unimodal Gemini-1.5	26.04	24.37	5.56
Unimodal GPT-4o	25.88	26.99	5.00
Multimodal Gemini-1.5	11.54	15.35	0.0
Multimodal GPT-4o	21.88	21.16	6.25
GPT4-o with GenEx	**94.87**	**69.21**	**72.11**

- Augment human decision making
- Augment GPT decision making

Imagination-Augmented Policy

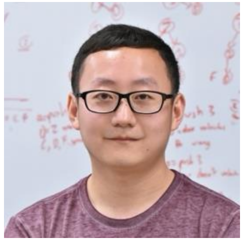
Enrich real observation with imaginative observation

Method	Acc. (%)	Confidence (%)	Logic Acc. (%)
Random	25.00	25.00	-
Human Text-only	44.82	52.19	46.82
Human with Image	91.50	80.22	70.93
Human with **GenEx**	**94.00**	**90.77**	**86.19**
Unimodal Gemini-1.5	30.56	29.46	13.89
Unimodal GPT-4o	27.71	26.38	20.22
Multimodal Gemini-1.5	46.73	36.70	0.0
Multimodal GPT-4o	46.10	44.10	12.51
GPT4-o with GenEx	**85.22**	**77.68**	**83.88**

- Augment human decision making
- Augment GPT decision making

Thank you! Question?





Acknowledgement:

Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan L. Yuille, and Jieneng Chen