Generating an Explorable World

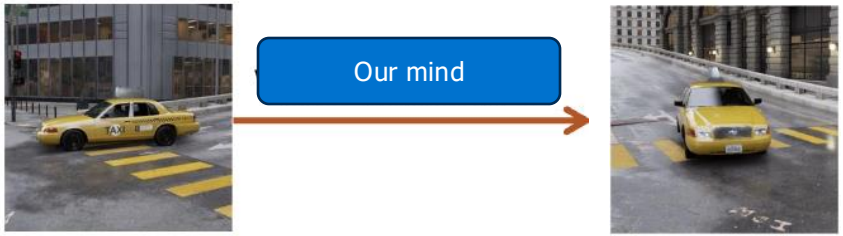
Jieneng Chen
Johns Hopkins University
04/27/2025

Invited Talk at ICLR Workshop on Embodied Intelligence with Large Language Models In Open City Environment

Generating a car

Close your eyes and generate an car in your mind (mental imagery test).

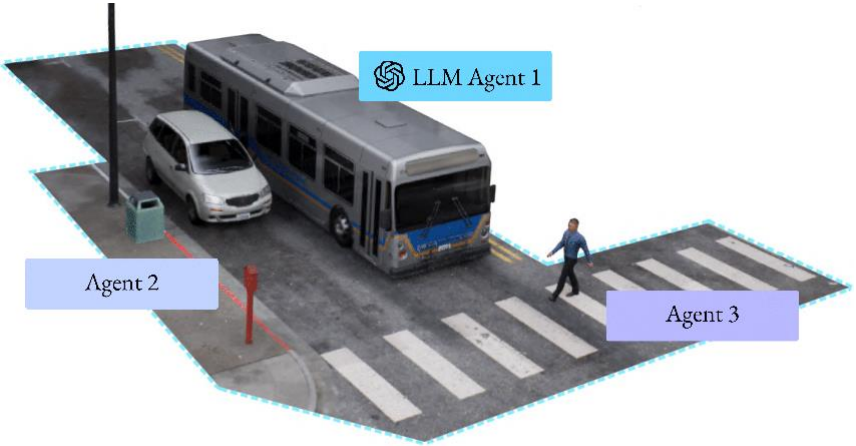
Generating the novel views



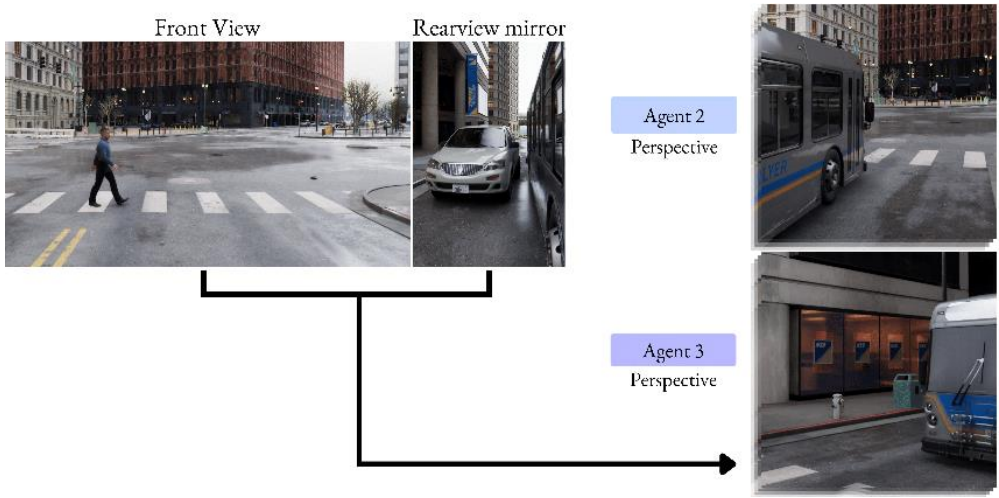
Generating the surroundings

Generating an explorable world

The bus blocks the line of sight between the sedan and the pedestrian.



The bus driver can mentally **explore** the viewpoints of other agents.



Are we ready for human-like world generation?

GAN
2014

Diffusion
2020

LDM
2022

Sora
2024

Going from generating one object to producing a realistic video is a decade-long journey.

- ✗ 3D/4D Consistency
- ✗ (Interactive) Action Control
- ✗ Physical Commonsense

Are we ready for human-like world generation?

GAN
2014

Diffusion
2020

LDM
2022

Sora
2024

What is the next decade?

Going from generating one object to producing a realistic video is a decade-long journey.

My bet is on the physical world

GAN
2014

Diffusion
2020

LDM
2022

Sora
2024

Physical World

Going from generating one object to producing a realistic video is a decade-long journey.

Adding 360° physical world to generation

- Scaling data via 3D physical engine.

- High-quality world generator.

Physical World

Adding 360° physical world to generation

- Scaling data via 3D physical engine.
  - World dynamics.
  - Physical exploration and interaction.

[1] Genie 2: A large-scale foundation world model

- High-quality world generator.
  - 360° world exploration.
  - Strong 3D consistency.

[2] Generative object movement with video prior

Adding 360° physical world to generation

- Scaling data via 3D physical engine.
  - World dynamics.
  - Physical exploration and interaction.

High-quality open-source video generation
(e.g., SVD, Cosmos)

- High-quality world generator.
  - 360° world representation.
  - 3D/4D consistency.

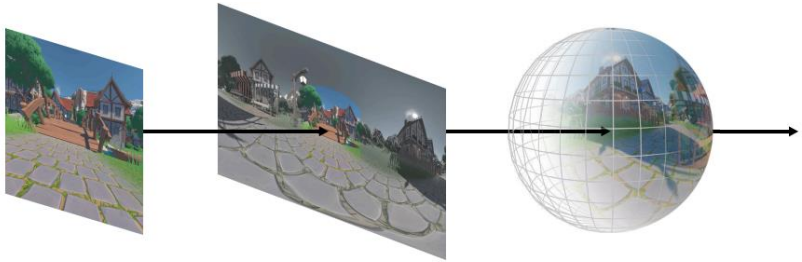
[1] GenEx: generating an explorable world

More than a generated world

GenEx: a type of **world model** that offer a predictive distribution over "changes" in the world**.**

- $\boldsymbol{x_{t-1}}$**:** the past world observation / state
- $\boldsymbol{x_t}$**:** the predicted future world observation / state
- $\boldsymbol{a_t}$**:** the action

$$\boldsymbol{p(x_t|x_{t-1}, a_t)}$$

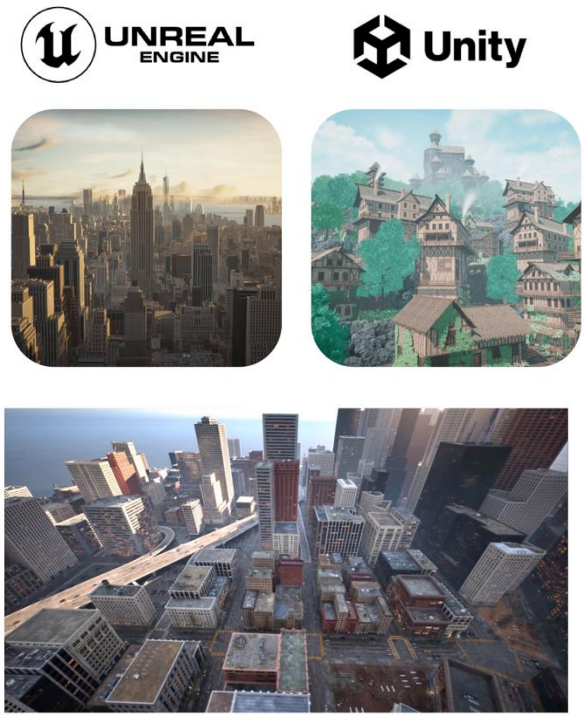


Single image → Image2Pano → Panoramic image as world representation → Spherical-consistent world explorer

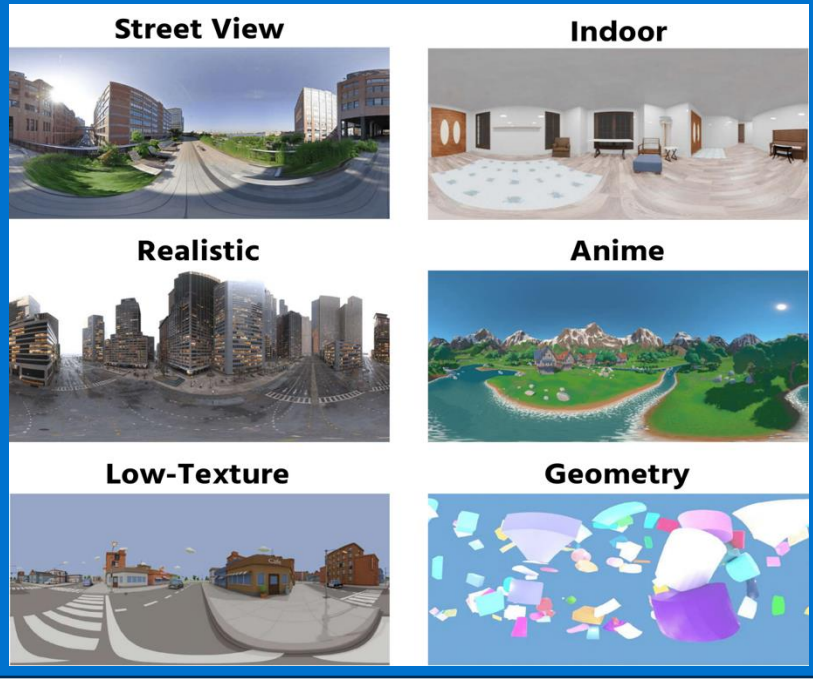
[1] GenEx: generating an explorable world

Generating future observations

Training recipe: purely synthetic physical engine



One million meters of data

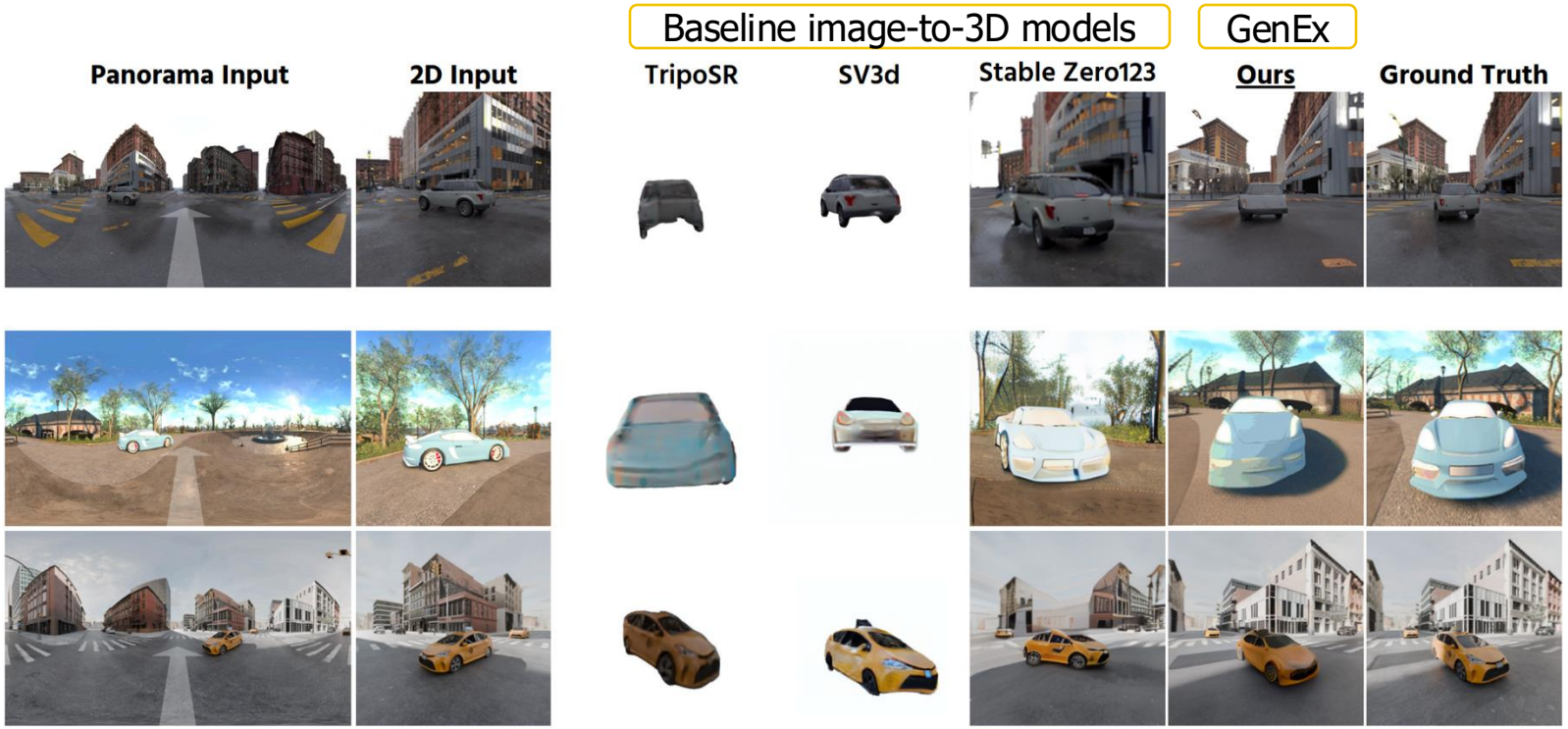


Inference on unseen diverse scene

Inference on unseen real world (JHU campus in left)

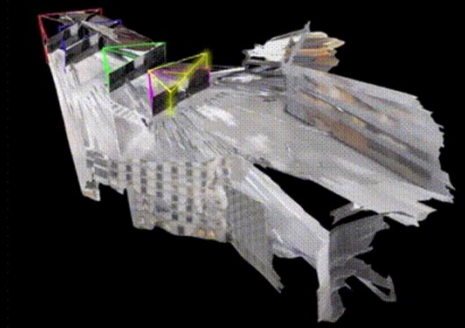
Different to prior world models focusing on AI gaming, this is one of the first to show the real-world generalizability.

Result 1: 3D Consistency



Result 2: Reconstructing 3D World

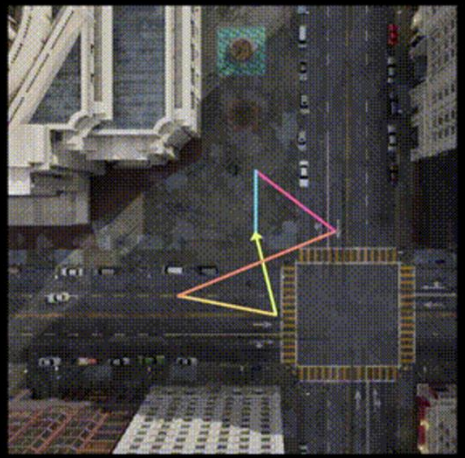
Generating the World



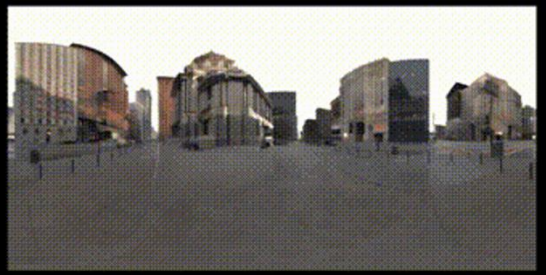
Reconstructing the World

Result 3: Loop Closure in the Generated World

Initial



Path

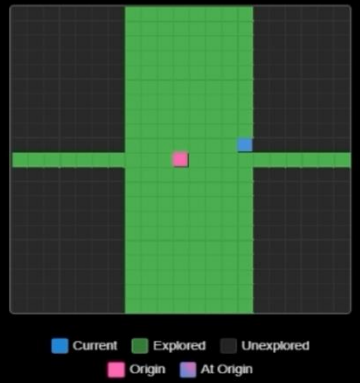


Navigated

GenEx: Generating an Explorable World

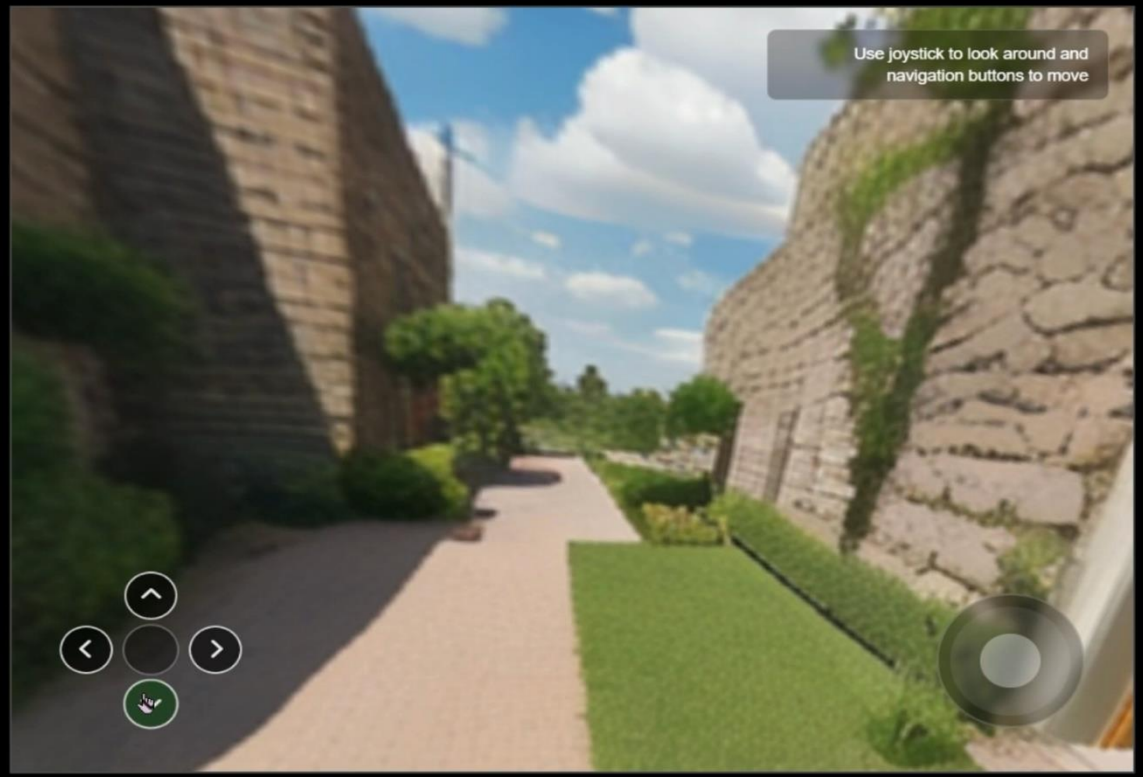
A fantasy town

Start Explorer



(4, 0) | 181 explored

[12:09:52] Loading panorama...
[12:09:52] Exploring: 368 areas generated, 4 remaining. You can move to any green area.
[12:09:52] Loading panorama...
[12:09:51] Loading panorama...
[12:09:51] Exploring: 367 areas generated, 4 remaining. You can move to any green area.
[12:09:51] Loading panorama...
[12:09:50] Loading panorama...
[12:09:49] Exploring: 366 areas generated, 4 remaining. You can move to any green area.
[12:09:49] This area is not yet generated (not green on the map).



Scalability

Scalability

Aim: turn **any** single image/text into a fully explorable world.

More than a generated world

A step further for human-like world exploration

GAN
2014

Diffusion
2020

LDM
2022

Sora
2024

Embodied
Physical World

Embodied intelligence amplification with generated world

$$A = \arg\max_{A} \pi_\theta(A \mid \text{instruct}, \text{observation}, \text{goal})$$

Intelligence amplification

$$A = \arg\max_{A} \pi_\theta(A \mid \text{instruct}, \text{observation in generated world}, \text{goal})$$

Embodied question answering benchmark

Single-agent

200 scenarios for EQAs

Multi-agent

Scene: I arrive at an intersection and want to turn left. The front path is clear, there is no car. ... I see another car at the intersection, on the left view moving slowly.

Question: What should I make the turn?

Choices:

(A) Stop in place to wait for the car to make the turn first.
(B) Honk to warn other cars to avoid collision.
(C) Pull over and wait for traffic to clear.
(D) Carefully continue the turn to avoid traffic congestion.

Scene: I arrive at an intersection and want to drive forward. ... I see the car opposite to myself suddenly stop. Also, I hear what seems to be an alarm, possibly from an emergency vehicle."

Question: What should I do?

Choices:

(A) Change lanes to bypass the car carefully.
(B) Stop passing the intersection and move a little bit left to clear the way.
(C) Stop in place to observe the environment.
(D) Continue to proceed through the intersection since the traffic light is green.

Scene: I am driving down a street. Ahead, there is a car stopped in my lane. I can't see what is in front of this car because it is blocking my view. The traffic is light, ...

Question: How should I proceed?

Choices:

(A) Change lanes to pass the stopped car quickly, since there is no visible obstruction.
(B) Honk to signal the stopped car to move.
(C) Slow down and keep to my lane, proceeding with caution.
(D) Wait for the car ahead to start moving.

Scene: I am approaching an intersection with a "Do Not Enter" sign. ... Ahead, there is a police car in view, but it is unclear whether the police car is waiting or needs to move.

Question: How should I respond to this situation?

Choices:

(A) Wait at the intersection for the police car to move first.
(B) Change lanes to pass through.
(C) Honk to signal the police car to move.
(D) Slow down and proceed cautiously, assuming the police car will stay in place.

Scene: I arrive at an intersection to proceed forward. The intersection does not have a traffic light and is busy. There is a pedestrian on my right side crossing the road fast ...

Question: What should I do now?

Choices:

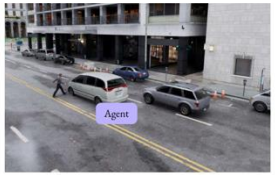
(A) Drive forward as normal.
(B) Block the pedestrian for a few seconds to avoid hitting by other cars.
(C) Accelerate to avoid collision with other cars.
(D) Pull over and wait for traffic to clear.

Scene: I'm at an intersection with a red light, where right turns are allowed. ... A fast car is approaching to turn right, and a pedestrian is crossing in front of me.

Question: What do I need to do?

Choices:

(A) Signal the car to stop for the pedestrian.
(B) Stay in place and wait for the green light.
(C) Honk to alert the pedestrian of the approaching car.
(D) Proceed cautiously while monitoring both the car and pedestrian.



Scene: I'm driving on a street. The front path is clear. ... I see a car in my back try to bypass me. There is also a pedestrain crossing the street on my left side.

Question: What would I do?

Choices:

(A) Move a little bit to the left to allow the other car to pass.
(B) Continue drive forward fast.
(C) Slow down to avoid the car bypass now to protect the pedestrain.
(D) Suddenly stop in place to block the back car.

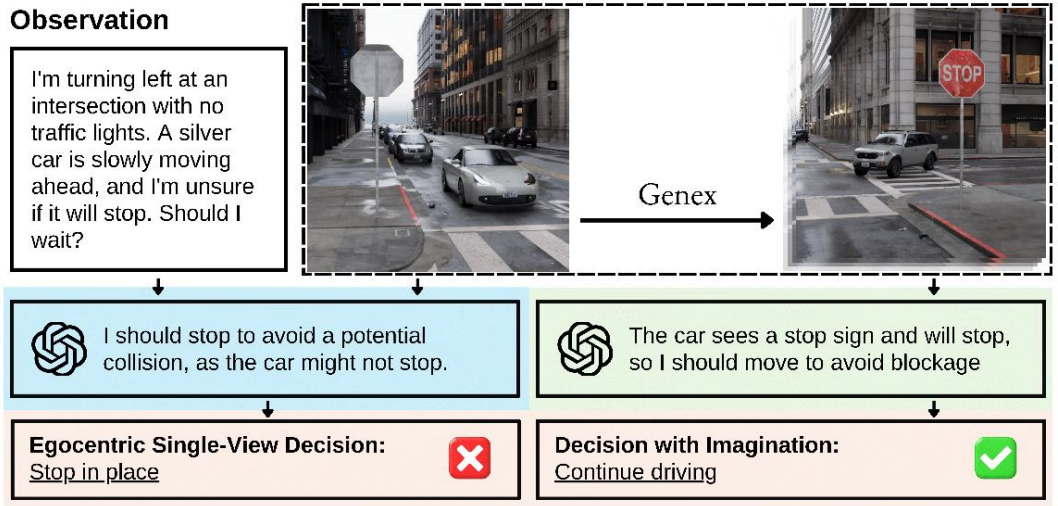
Scene: I'm driving on the right lane on a street. On the other lane, there is a car approaching fast. ... I can also see a pedestrain on the left side trying to cross the street.

Question: What to do now?

Choices:

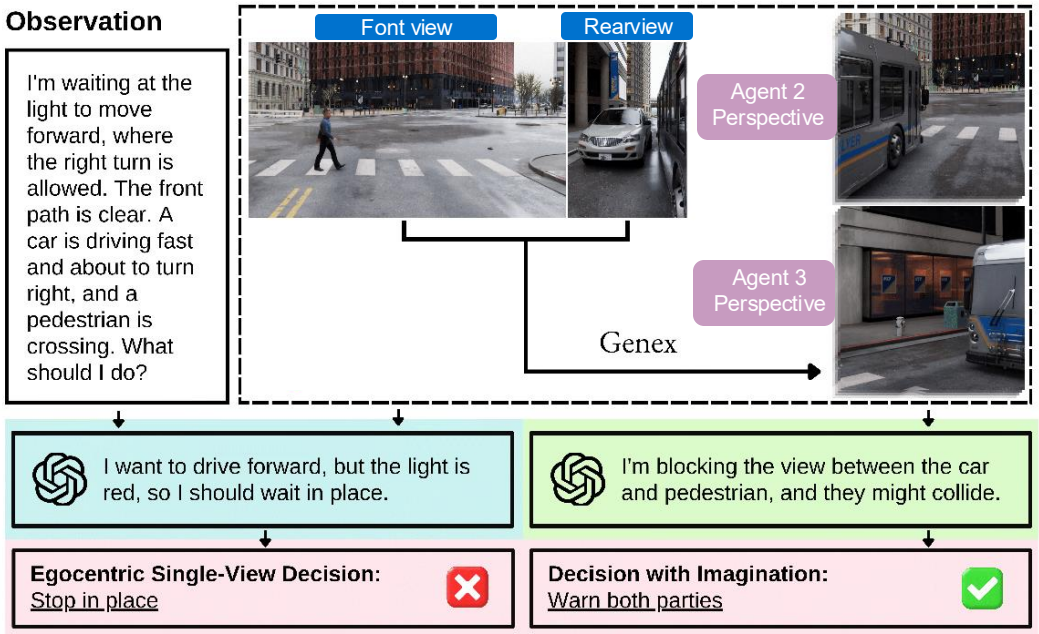
(A) Continue forward as the path is clear.
(B) Honk to signal the front car to avoid collision with me.
(C) Pull over to the right.
(D) Warn both pedestrain and the car for a potentail collision.

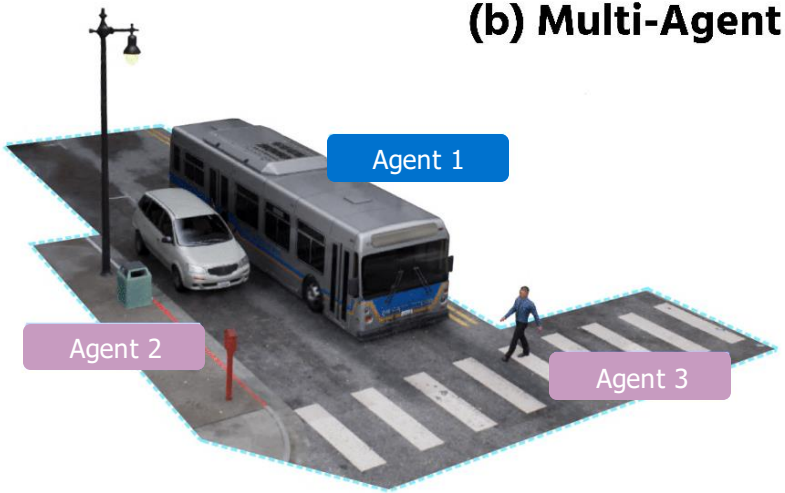
Embodied decision making



[1] GenEx: generating an explorable world

Embodied decision making



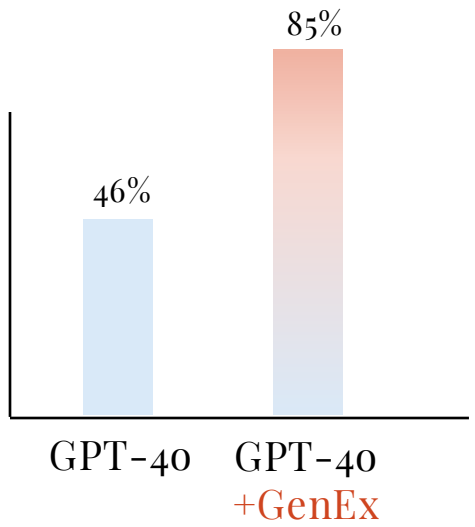


[1] GenEx: generating an explorable world

Collaborative / decentralized world models

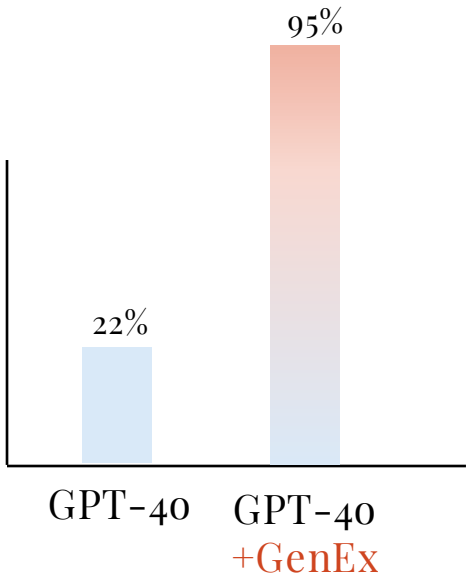
Boost GPT agent decision making

+ 39% accuracy



Single-Agent Setting

+ 73% accuracy



Multi-Agent Setting

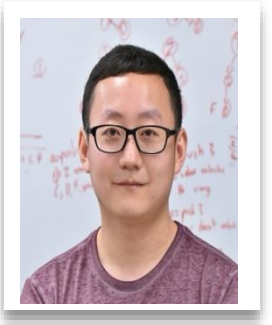
[1] GenEx: generating an explorable world

Summary

- One of the first models to generate explorable (real) world from a single image.
- Intelligence amplification in the embodied generated world.

Thanks for wonderful collaborators on these projects!!




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

Thank you! Question?