GenEx: Generating an Explorable World

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

Johns Hopkins University

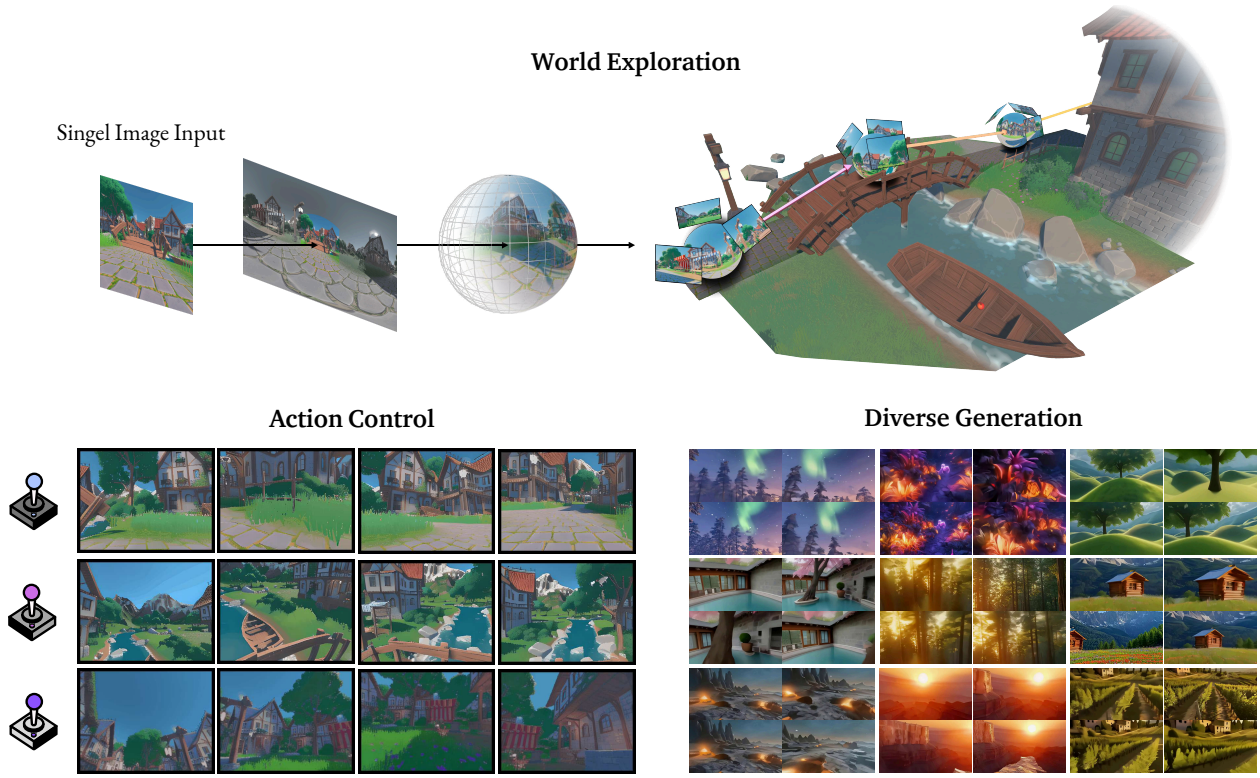

Figure 1 | GenEx explores an imaginative world, created from a single RGB image and brought to life as a generated video. See more examples in our **website (genex.world)**.

Understanding, navigating, and exploring the 3D physical *real* world has long been a central challenge in the development of artificial intelligence. In this work, we take a step toward this goal by introducing GenEx, a system capable of planning complex embodied world exploration, guided by its *generative imagination* that forms priors (expectations) about the surrounding environments.

GenEx generates an entire 3D-consistent imaginative environment from as little as a single RGB image, bringing it to life through panoramic video streams. Leveraging scalable 3D world data curated from Unreal Engine, our generative model is grounded in the physical world. It captures a continuous 360° environment with little effort , offering a boundless landscape for AI agents to explore and interact with. GenEx achieves high-quality world generation and robust loop consistency over long trajectories, and demonstrates strong 3D capabilities such as consistency and active 3D mapping.

Powered by the generative imagination of the world, GPT-assisted agents are equipped to perform complex embodied tasks, including both goal-agnostic exploration and goal-driven navigation. These agents utilize predictive expectations regarding unseen parts of the physical world to refine their beliefs, simulate different outcomes based on potential decisions, and make more informed choices.

In summary, we demonstrate that GenEx provides a transformative platform for advancing embodied AI in imaginative spaces and brings potential for extending these capabilities to real-world exploration.

Keywords: *Generative AI, World Models, Embodied AI, World Explorer*

1. Introduction

Humans explore and interact with the 3D physical world by perceiving their surroundings, taking actions, and engaging with others. Through these interactions, they form mental models that simulate the complexities of their environment. With just a glimpse, humans can construct an internal 3D representation of their surroundings in their minds, enabling reasoning, navigation, and problem-solving. This remarkable ability has long been a central challenge in the development of artificial intelligence.

In this work, we introduce **GenEx**, a platform designed to push this boundary by **Gen**erating an **Ex**plorable world and facilitating explorations in this generated world. GenEx combines two interconnected components: an imaginative world, which dynamically generates 3D environments for exploration, and an embodied agent, which interacts with this environment to refine its understanding and decision-making. Together, these components form a symbiotic system that enables AI to simulate, explore, and learn in ways similar to human cognitive processes.

We begin by constructing an imaginative world that captures a 360°, 3D environment grounded in the physical world, leveraging recent advancements in Generative AI. Starting from a single image, the model generates new environments expansively and dynamically while maintaining coherence and 3D consistency, even during long-distance exploration. This boundless landscape provides endless opportunities for AI agents to explore and interact.

The environment is brought into life in the form of diffusion video generation, conditioned on moving angle, distance, and a single initial view to serve as a starting point. To address field-of-view constraints, we utilize panoramic representations and train our video diffusion models with spherical-consistent learning techniques. This ensures the generated environments maintain coherence and 3D consistency, even during long-distance exploration. To anchor our video generation model in the physical world, we curate training data from physics engines like Unreal Engine, enabling realistic and immersive outputs.

Within this imaginative landscape, embodied agents play a crucial role. Enhanced by GPTs, these agents can explore unseen parts of the physical world with imagined observations to refine their understanding of surroundings, simulate different outcomes based on potential decisions, and make more informed choices. Furthermore, GenEx supports multi-agent scenarios, allowing agents to mentally navigate others' positions, share imagined beliefs, and collaboratively refine their strategies.

In summary, GenEx represents a transformative step forward in the development of AI, offering a platform that bridges the generative and physically grounded world. By enabling AI to explore, learn, and interact in boundless, dynamically generated environments, GenEx opens the door to applications ranging from real-world navigation, interactive gaming, and VR/AR to embodied AI.

2. Generating an Explorable World

We define the explorable generative world and the problem in § 2.1, present the world initialization in § 2.2 and world transition in § 2.3.

2.1. Problem Formulation

Defining an explorable generative world. We define an explorable generative world as an AI-generated virtual environment, constrained to the agent's immediate surroundings. The generative world is both physically plausible and visually coherent. This environment is represented by the agent's egocentric panoramic observations, denoted as $\mathbf{x}$. While $\mathbf{x}$ is synthesized, it remains grounded in intuitive physical principles and realistic appearance, akin to a high-fidelity, physically realistic video game environment.

Crucially, the explorable nature of our generative world ensures the agent's experience is not limited to a static scene. Instead, the environment dynamically evolves in response to the agent's movements and actions, simulating continuous and coherent exploration. Formally, let a_t be the agent's action at step t, encompass-

ing both view rotation α and forward distance d. Let $\mathbf{x}_t = (x_t^0, x_t^1, \ldots, x_t^S)$ represent the sequence of panoramic observations encountered as the agent moves according to a_t, where S corresponds to sequence length in $\mathbf{x}_t$, or the traveled distance. Each x_t^s in $\mathbf{x}_t$ is generated to reflect the environment's currently perceivable state, ensuring that the agent's evolving viewpoint remains coherent and physically meaningful.

We train our models using data harvested from a controlled, simulated setting. By employing a physics-based data engine (§2.2), we ensure realistic and diverse training scenarios that capture the intricate variations encountered in complex, virtual landscapes.

Task formulation: We reformulate the task of "exploring a generative world" as the problem of generating an initial panoramic world view x_0 and a sequence of world views represented by panoramic videos $\mathbf{x}_{1:T}$, together represented as $\mathbf{x}_{0:T}$, given a single initial image i_0, a description l_0, and action a_t at each step t, where $t = 1, \ldots, T$. Formally, we have

$$p(\mathbf{x}_{0:T} \mid i_0, l_0) = \underbrace{p_{\theta_1}(x \mid i_0, l_0)}_{\text{world initialization}} \underbrace{\prod_{t=1}^{T} p_{\theta_2}(\mathbf{x}_t \mid x_{t-1}^S, a_t)}_{\text{world transition}}$$

In this unified form, the core terms are:

- **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:

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

  where θ_1 is an image-to-panorama generator.

- **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 $x_0^S := x_0$.

Algorithm 1 Generating an Explorable World $p(\mathbf{x}_{0:T} \mid i_0, l_0)$

Require:
- A initial single-view image i_0.
- A language description l_0 specifying the desired panoramic world initialization.
- A conditional distribution $p_{\theta_1}(x \mid i_0, l_0)$, parameterized by an image-to-panorama generation model θ_1 to initialize the 360° world.
- Action space $\mathcal{A}$ defined in the physical engine, from which an action is sampled: $a_t \sim \mathcal{A}$.
- A conditional distribution $p_{\theta_2}(\mathbf{x} \mid x_{t-1}^S, a_t)$, parameterized by a panoramic video generation model θ_2.

Notation: Let $\mathbf{x}_t = (x_t^0, x_t^1, \ldots, x_t^S)$ denote the generated panoramic video at exploration step t. Here, x_t^S is the latest explored panoramic view.

World initialization: Initialize a 360° panoramic world from a single image:

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

for $t = 1$ to T **do**

  World transition at step t: Given $a_t \sim \mathcal{A}$ and the latest explored world x_{t-1}^S where $x_0^S := x_0$, we sample the new panoramic video $\mathbf{x}_t$:

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

end for

return The initial 360° panoramic world view x_0 and a sequence of generated panoramic states $\mathbf{x}_{1:T}$, which together represent one explorable generative world, denoted as $\mathbf{x}_{0:T}$.

2.2. World Initialization

Preliminary: data and representation. Collecting diverse world exploration data in the real world is challenging due to resource constraints and environmental variability. Thus, we utilize physics engines such as Unreal Engine 5 and Unity in Figure 2 for data curation. These engines allow for the creation of rich, diverse virtual environments where we can simulate exploration trajectories and collect corresponding data efficiently.

We represent the 360° world using the panoramic view of the agent. Panoramic images capture a complete 360° × 180° view of a scene from a fixed viewpoint. One common panoramic representation is the *cubemap*, which projects a 360° view onto the six faces of a cube. Each face

Figure 2 | Our data curation leverages physical engines, utilizing realistic city assets from UE5 and animated world assets from Unity.

captures a 90° field of view, resulting in six perspective images that can be seamlessly stitched together. Due to its simplicity and compatibility with rendering engines, we directly collect cubemaps in the physics engine to represent the egocentric world. Notably, cubemaps, equirectangular panorama, and sphere are three representations of 360° panoramic world. The curated cubemaps will be projected to equirectangular panoramas for video generation in the world exploration stage, and projected to spherical space when changing the exploring angle.

Given predefined exploration trajectories, we collect sequences of cubemaps to represent different exploration outcomes in the virtual world. By sampling a large number of exploration directions uniformly, we curate an extensive dataset of world exploration scenarios, which serves as the training data for our models.

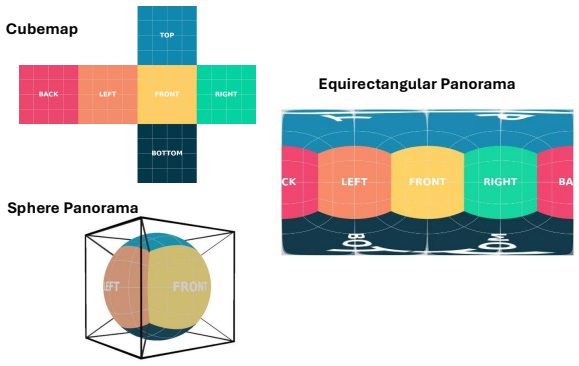

Figure 3 | Three panorama representations that can be transformed into one another.

World initialization model. Starting from a single input image i_0, we aim to construct a full 360° panoramic representation x_0 of the agent's environment. To achieve this, we condition a pretrained text-to-image diffusion model on both the input image i_0 and a text description l_0 of the desired 3D world, yielding a high-dynamic-range panorama. Thus, x_0 is drawn from the conditional distribution $p(x \mid i_0, l_0)$.

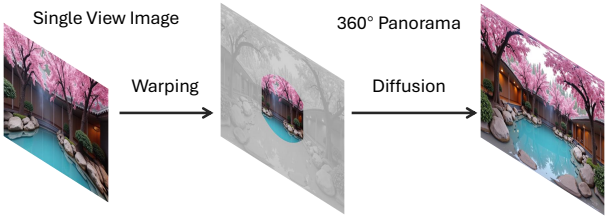

Figure 4 | From single view to 360° panorama.

Our world initialization model is built up on a state-of-the-art text-to-panorama model (Bilcke, 2024) tuned from the state-of-the-art text-to-image model FLUX.1 (Labs, 2024). The text-to-panorama model (Bilcke, 2024) generates a panorama from a text description l_0:

$$x_0 \sim p_{\text{flux}}(x \mid l_0)\,.$$

However, without being conditioned on the single image, this approach cannot guarantee the coherence of generated panorama x_0 with the provided reference image i_0.

We extend the model to condition on both textual input and a single image. This adaptation allows the model to produce a full 360-degree environment that aligns with the provided image:

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

Although this yields a coherent, image-consistent panorama, the scene remains static and does not permit dynamic movement or exploration. To enable deeper interaction within the generative world, we introduce the world transition.

2.3. World Transition

When the agent moves within the imaginative environment, its egocentric 360° view changes, prompting a *world transition*. We model this transition as an action-driven panoramic video generation process, transforming the previously observed panorama into a new, forward-looking view as the agent progresses.

Transition objective. The goal is to sample $\mathbf{x}_t = (x_t^0, x_t^1, ..., x_t^S)$, a newly explored panoramic video, conditioned on the previous panorama x_{t-1}^S and the action $a_t = (\alpha_t, d_t)$. Here, α_t is the moving angle and d_t is the distance. Formally, we have the transition objective:

$$\mathbf{x}_t \sim p(\mathbf{x} \mid x_{t-1}^S, a_t)\,.$$

The transition procedure has core modules:

- **Action sampling**: Consider an action sequence $a_{1:T}$ drawn from an infinitely large action set in the Unreal Engine and Unity. We can denote the action space as: $\mathcal{A}$, where $|\mathcal{A}| = \infty$. Each element of the sequence for $t = 1, \ldots, T$ is sampled from $\mathcal{A}$:

  $$a_t \sim \mathcal{A}, \quad t = 1, \ldots, T,$$

  As a result, the entire action sequence $a_{1:T} = (a_1, \ldots, a_T)$ lies in $\mathcal{A}^T$.
- **Sphere rotation**: The action a_t determines a rotation angle α_t, which we apply to the spherical representation of the equirectangular panorama x_{t-1}^S. This yields a rotated equirectangular panorama ${x_{t-1}^S}'$:

  $${x_{t-1}^S}' = \mathcal{T}(x_{t-1}^S, \alpha_t)\,,$$

  where $\mathcal{T}$ is a known rotation geometric transformation defined to Equation 3 in Appendix.
- **Panoramic video generation**: We next generate videos to travel in the imaginative space by distance d_t. Our video generator is adapted from a video diffusion model conditioned on the latest explored view ${x_{t-1}^S}'$ and randomly sampled noise $\epsilon_t \sim \mathcal{N}(0, I)$:

  $$\mathbf{x}_t \sim p_{\theta_2}(\mathbf{x} \mid {x_{t-1}^S}', \epsilon_t)\,.$$

  This approach ensures that each generated panoramic video remains consistent with the prior view, while incorporating stochastic variations to represent an explorable world.

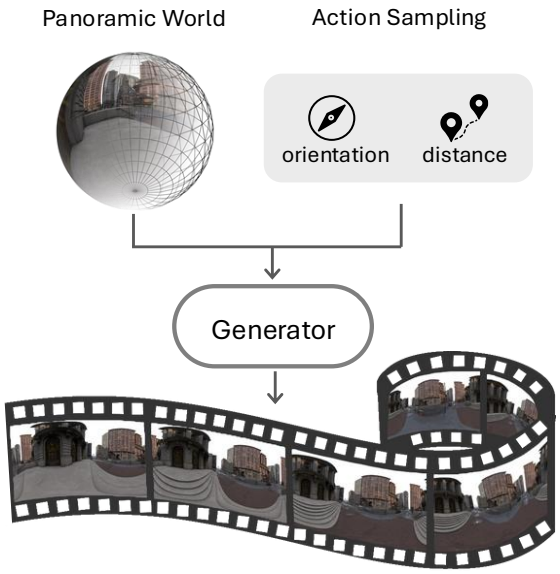

Figure 5 | We model the world transition as a panoramic video generation process. Given the last explored 360° panorama and an action that rotates the viewing sphere, the model produces a sequence of newly generated panoramic views

We aim to learn to produce panoramic videos that remain visually coherent on a spherical surface. Without additional constraints, training on equirectangular panorama alone can result in discontinuities at the panorama edges. To address this, we adopt spherical-consistency learning, or SCL, detailed in (Lu et al., 2024), which promotes smooth and continuous imagery across all viewing directions on the sphere.

Summary. In essence, the world transition step updates the agent's observed 360° panorama into a newly explored view sequence. Through action-driven rotation, spherical adjustments, and a diffusion-based video model, we achieve seamless transitions and maintain coherent, panoramic representations as the agent navigates the generative environment.

3. Exploration in the Generative World

After generating the explorable world, human or embodied agents can explore the virtual world with an exploration policy, defined in §3.1. We then introduce three exploration modes in §3.2.

3.1. Exploration Policy

The exploration action a_t is decided by a policy:

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

where $\mathcal{I}$ is the instruction that specifies the exploration mode to be either human interaction or assisted by a GPT, detailed in §3.2. Note that x^S_{t-1} denotes the latest explored view from the previous step $t-1$. At $t=1$, it corresponds to the initial panorama x_0. The action $a_t = (\alpha_t, d_t)$ defines how the agent rotates its field of view with the rotation angle α_t and moves forward with d_t distance, shaping the direction and extent of exploration.

3.2. Exploration Modes

The GenEx framework enables agents to explore within an imaginative world by streaming video generation, based on current single view image i_0 and the given exploration action a.

We support three modes for generative world exploration, including (a) interactive exploration, (b) GPT-assisted free exploration, and (c) goal-driven navigation, illustrated in Figure 6.

Interactive exploration. GenEx enables the agent to freely explore the synthetic world with an unlimited range of orientations, enhancing its understanding of the surrounding environment. Users can control the agent's movement directions and distances, allowing for continuous exploration of the virtual world.

GPT-assisted free exploration. However, human-provided commands can sometimes lead the model to collapse. For example, if users instruct the agent to move excessively close to a wall, the resulting viewpoint may reduce the quality of subsequent generated video frames.

To mitigate this, we employ a GPT-4o (Achiam et al., 2023) as a "pilot" to determine exploration configurations, encompassing full 360° explorable directions and distances. Given that generation quality can compoundingly degrade over time, GPT-4o acts as a policy that selects actions to maximize the fidelity of generative worlds and avoid model collapsing.

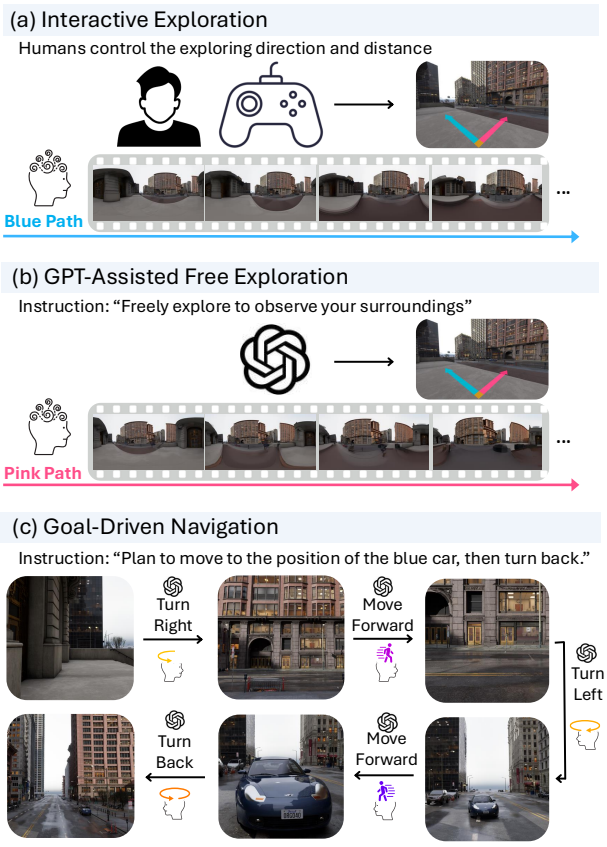

Figure 6 | Three exploration modes — interactive, GPT-assisted, and goal-driven — each defined by distinct exploration instructions.

Goal-driven navigation. The agent receives a goal with navigation instruction $\mathcal{I}$, such as, "Move to the blue car's position and orientation." GPT performs high-level planning based on the instruction and initial image, generating low-level exploration configurations in an iterative manner. GenEx then processes these configurations step-by-step, updating images progressively throughout the imaginative exploration. This allows for greater control and targeted exploration.

4. Advancing Embodied AI

In our generative world, we can explore previously unobserved regions of the physical environment, gather more comprehensive information, and refine our beliefs for more informed decision-making. We frame this process in a form of human-like decision-making—an "imagination-augmented policy"—that could play a crucial role in shaping the future of embodied AI.

Preliminary. We first denote a common embodied policy as $\pi_\theta(A|o, g)$ where θ is a GPT-based planner, o is the agent's observation, g is the goal to answer questions such as "Danger ahead. Stop or go ahead?". Here, A denotes higher-level embodied actions (*e.g.,* answering the questions or generating navigation plans), which differ from the exploration actions a introduced earlier. However, if the observation is limited to a single initial image i_0, then executing $\arg\max_A \pi_\theta(A|o = i_0, g)$ may fail because it provides no visibility into unseen parts of the environment.

The decision can become more informed if the agent gains a clearer understanding of its surroundings (Fan et al., 2024). By navigating through the physical space, the agent gathers additional information about its environment ("Physical" path in the cyan color in Figure 7), enabling more accurate assessments and better choices moving forward.

Nevertheless, physically traversing the space is inefficient, expensive, and even impossible in dangerous scenarios. To streamline this process, we use imagination as a pathway for the agent to simulate outcomes without physically traversing ("Imaginative" path in purple color in Figure 7).

The key question is:

*How can an agent make more informed decisions through **exploration** in a generative 360° world?*

4.1. Imagination-Augmented Policy

We propose a new policy based on imagined observations in the generative world, described in Algorithm 2. The Imagination-Augmented Policy consists of the following two steps:

- **Step 1**: Gather imagined observations sampled from GenEx (Algorithm 1):

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

- **Step 2**: Select an action conditioned on the imagined observations to maximize the policy:

$$A = \arg\max_A \pi_{\theta_3}(A \mid i_0, \mathbf{x}_{0:T}, g) .$$

In our work, we apply GenEx for imaginative exploration and an LMM as the policy model π_{θ_3}, with examples in Figure 8.

Compared to $\arg\max_a \pi_{\theta_3}(A \mid i_0, g)$ the common policy which selects the action based solely on real observations i_0, the Imagination-Augmented Policy selects actions using both actual and imagined observations $(i_0, \mathbf{x}_{0:T})$, potentially leading to more informed decisions.

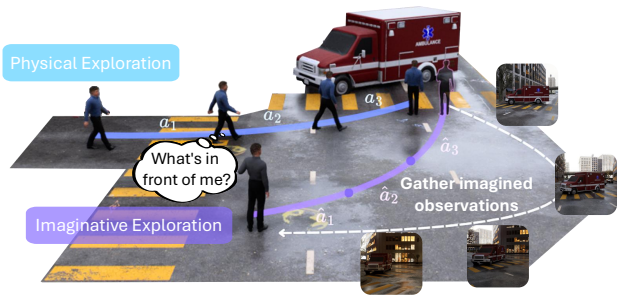

Figure 7 | GenEx-driven imaginative exploration can gather observations that are just as informed as those obtained through physical exploration.

Algorithm 2 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:

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

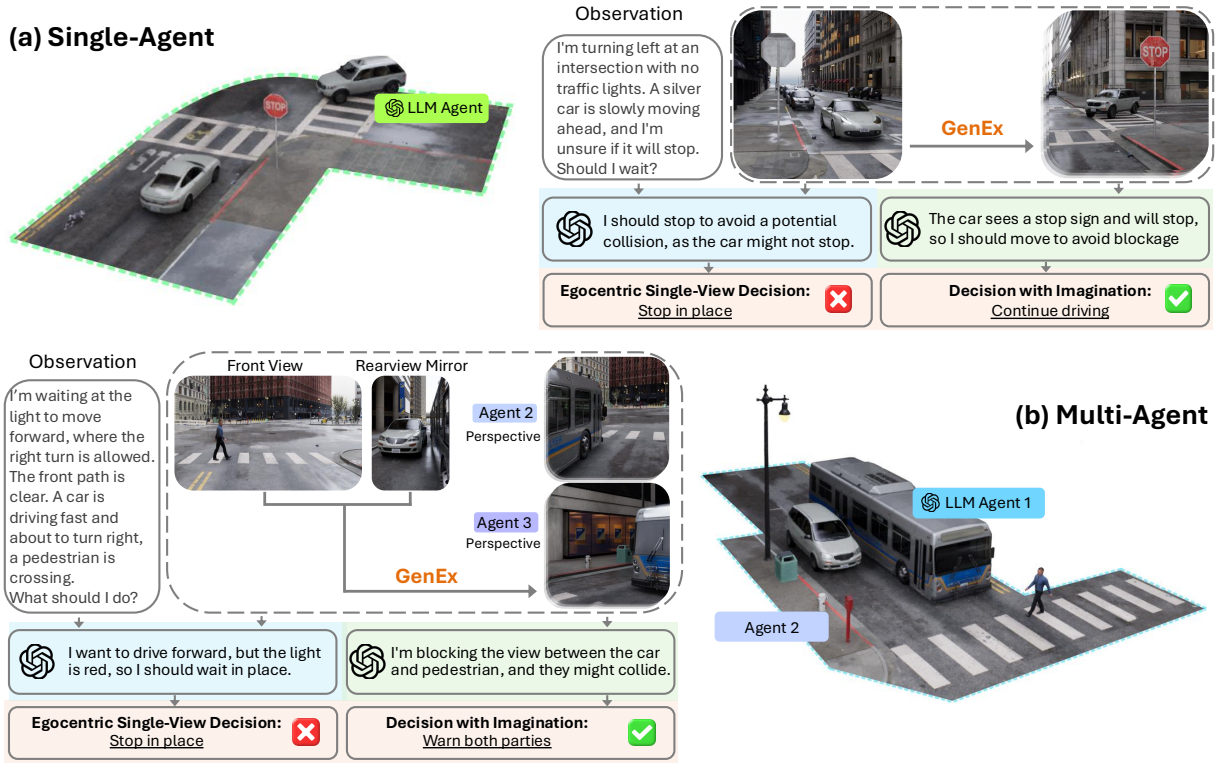

Figure 8 | Single agent reasoning with imagination and multi-agent reasoning and planning with imagination. (a) The single agent can imagine previously unobserved views to better understand the environment. (b) In the multi-agent scenario, the agent infers the perspective of others to make decisions based on a more complete understanding of the situation. Input and generated images are panoramic; cubes are extracted for visualization.

4.2. Multi-Agent Imagination-Augmented Policy

Our Imagination-Augmented Policy can be generalized to the multi-agent scenario. An agent can explore the position of other agents. This predicts other agents' observations and infers their understanding of the surrounding environments.

Technically, we can create multiple exploration paths by providing instructions like "navigate to the position of agent-k". The agent can then explore the generated 360° environment to reach agent-k's location.

By extending Algorithm 2, the Multi-Agent Imagination-Augmented Policy has three steps:

- **Step 1**: Gather imagined observations by exploring the position to agent-k using Algorithm 1, with instruction $\mathcal{I}_k$ "navigate to the position of 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:

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

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

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

When exploring another agent's surrounding environment, we can predict what that agent sees, understands, and might do next, which in turn helps us adjust our own actions with more complete information.

5. Applications

5.1. Generation Quality

We evaluate the **video generation quality** using FVD (Unterthiner et al., 2019), SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and PSNR (Horé and Ziou, 2010). Table 1 shows our earlier GenEx version (Lu et al., 2024) has high video quality in all metrics.

Model	Representation	FVD ↓	MSE ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Baseline	6-view cubemaps	196.7	0.10	0.09	26.1	0.88
GenEx w/o SCL	panorama	81.9	0.05	0.05	29.4	0.91
GenEx	panorama	**69.5**	**0.04**	**0.03**	**30.2**	**0.94**

Table 1 | GenEx with high generation quality.

5.2. Exploration Loop Consistency

We propose **Imaginative Exploration Loop Consistency** (IELC) to measure long-range exploration fidelity. For each randomly sampled closed-loop path, we compute the latent MSE between the initial real image and the final generated image, and then average these values over 1000 loops with varying rotations and distances, discarding blocked paths. As shown in Figure 9, the IELC remains high even for 20m loops and multiple consecutive videos, maintaining latent MSE below 0.1 and thus indicating minimal drift. This robustness stems from preserving spherical consistency, ensuring that rotations do not compromise image quality.

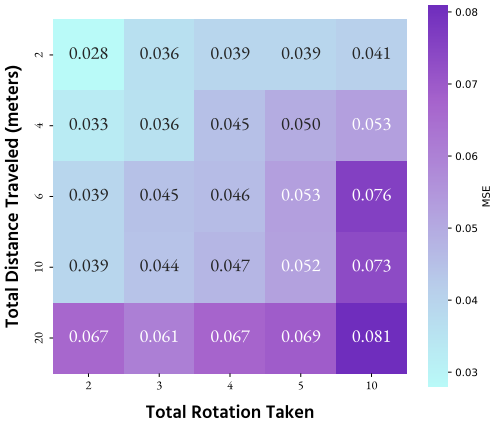

Figure 9 | Imaginative Exploration Loop Consistency (IELC) varying distance and rotations.

5.3. Generating Bird's-Eye Worlds

By exploring upward along the z-axis, our method generates top-down (bird's-eye view) maps directly from a single panoramic image. As shown in Figure 10, these overhead layouts give the agent an objective, third-person understanding of the scene, thereby improving reasoning.

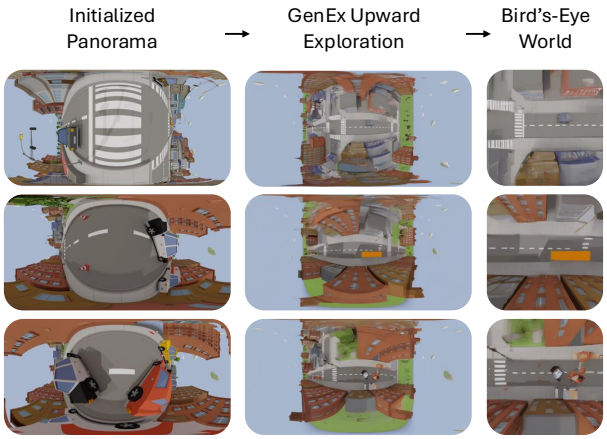

Figure 10 | Through generative exploration in *z-axis*, we are able to generate the 2D bird-eye world view of the current scene.

5.4. 3D Consistency

Our method enables the generation of multi-view videos of an object through imaginative exploration with a path circling around it. Our model demonstrates superior performance compared with the SOTA open-source models. Importantly, it maintains near-perfect background consistency and effectively simulates scene lighting, object orientation, and 3D relationships as in Figure 11.

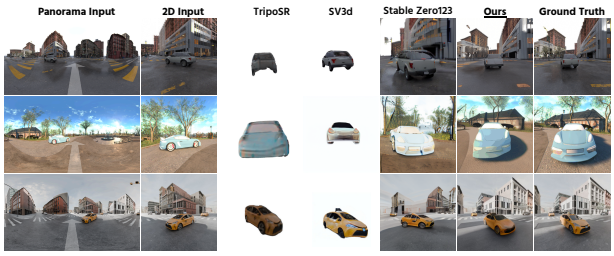

Figure 11 | Through exploration, our model achieves higher quality in novel view synthesis for objects and better consistency in background synthesis, compared to SOTA 3D reconstruction models (StabilityAI, 2023; Tochilkin et al., 2024; Voleti et al., 2024).

5.5. Active 3D Mapping in Generated Worlds

When the agent actively explores the generative world, it continuously gathers observations that can be leveraged to reconstruct a 3D map using DUSt3R (Wang et al., 2024b), shown in Figure 12.

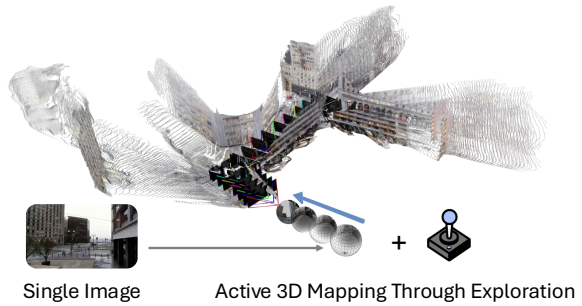

Figure 12 | Active 3D mapping from a single image.

5.6. Embodied Decision Making

We next evaluate the Imagination-Augmented Policy proposed in §4 and share two key findings.

Evaluation. We evaluate our Imagination-Augmented Policy (§4.1) in Table 2. We extend the Genex-EQA in (Lu et al., 2024) with a controlled counterpart for each scenario. We use *Unimodal* to refer to agents receiving only text context, while *Multimodal* reasoning demonstrates LLM decision when prompted along with an egocentric visual view. GenEx shows the performance of models equipped as agents with a generative world explorer. We evaluate our Multi-Agent Imagination-Augmented Policy (§4.2) in Table 3.

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**

Table 2 | Eval of Imagination-Augmented Policy.

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**

Table 3 | Evaluation of Multi-Agent Imagination-Augmented Policy.

Findings. We identified two findings based on the results from human policy (grey row) and GenEx-enhanced GPT policy (blue row).

- *Vision without imagination can be misleading for GPTs*. Interestingly, a unimodal response that relies solely on the environment's text description often outperforms its multimodal counterparts, which incorporate both text and egocentric visual inputs. This suggests that vision without imagination can be misleading, as it may lead to incorrect inferences due to the lack of spatial context and relying only on language-based commonsense reasoning. This highlights the importance of integrating visual imagination to enhance the accuracy and reliability of the agent's decision-making processes.
- *GenEx has the potential to enhance cognitive abilities for humans*. Human performance results reveal several key insights. First, individuals using both visual and textual information achieve significantly higher decision accuracy compared to those relying solely on text. This indicates that multimodal inputs enhance reasoning. Secondly, when provided with imagined videos generated by GenEx, humans make even more accurate and informed decisions than in the conventional image-only setting, especially in multi-agent scenarios that require advanced spatial reasoning. These findings demonstrate GenEx's potential to enhance cognitive abilities for effective social collaboration and situational awareness.

6. Discussion

Related works. Advances in single-image 3D modeling (Tewari et al., 2023; Yu et al., 2024) enable novel view synthesis but are limited by render distances or fields of view, relying heavily on depth estimator. Meanwhile, video generation methods (Blattmann et al., 2023; Kondratyuk et al., 2024; OpenAI, 2024) excel at producing diverse videos but often lack physical grounding, reducing their utility for exploration. Video generation models (Bu et al., 2024; Du et al., 2024a,b; Wang et al., 2024a; Yang et al., 2024) are capable of directly synthesizing visual plans for decision-making, but world exploration for imagined observations remains unexamined. Our approach unites these domains by drawing on physically grounded data to generate 3D-consistent, explorable worlds and advance embodied AI.

Extension to our earlier work. Our earlier work (Lu et al., 2024), published on arXiv in November 2024, conceptualized world transitions, exploration, and applications in embodied AI, but it did not address the crucial aspect of world initialization from a single image.

Relation to concurrent industrial progress. WorldLabs (WorldLabs, 2024) recently released demos of anime-world generation from a single image. DeepMind (DeepMind, 2024) released a blog on interactive world models. Our work complements these ongoing industrial efforts, jointly contributing toward a shared vision: creating rich, interactive, 3D-consistent generative worlds. Importantly, we offer our technical details. Beyond this, we also introduce the concept of an Imagination-Augmented Policy by exploring the generative world, further expanding the frontiers of embodied AI.

Challenges. Bridging imaginative and real-world environments remains a core challenge in AI. Current approaches rely on physical engines. Future work must address several key limitations, including sim-to-real adaptation, real sensor integration, dynamic conditions, and ethical safeguards, to ultimately enable reliable deployment of embodied AI in diverse physical settings.

7. Conclusion

We introduce **GenEx**, a platform that **Gen**erates an **Ex**plorable world and enables agents, either instructed by human users or a GPT, to freely explore in this imaginative panoramic world. By generating 3D-consistent environments from a single image, our approach enables the creation of immersive and interactive worlds offering a boundless landscape, grounded in the physical world, and explored by agents. We demonstrate diverse applications of GenEx, showing that this generative explorable world technique can create diverse and consistent 3D environments, build active 3D mappings, and advance embodied decision-making by allowing agents to create more informed and effective plans. Furthermore, GenEx's framework supports multi-agent interactions, paving the way for more advanced and cooperative AI systems. This work marks an advancement toward real-world navigation, interactive gaming, and achieving human-like intelligence in embodied AI.

Author Contributions

We list author contributions here alphabetically by last name. Please direct all correspondence to the project lead **Jieneng Chen** (jchen293@jh.edu).

Core Contributors

- **Taiming Lu**: project leadership, data engine, model research and pipeline, infrastructure
- **Tianmin Shu**: embodied policy research, writing, revising, technical advice
- **Junfei Xiao**: image-to-panorama data and model research, writing, editing

Contributors and Advisors

- **Rama Chellappa**: device support, advice
- **Daniel Khashabi**: writing, technical advice
- **Cheng Peng**: data support, editing
- **Jiahao Wang**: math, postprocessing, editing
- **Chen Wei**: revising, editing, writing advice
- **Luoxin Ye**: model, postprocessing
- **Alan L. Yuille**: math revising, funding, editing, writing advice, technical advice

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

J. Bilcke. Flux.1-[dev] panorama lora (v2), 2024. URL https://huggingface.co/jbilcke-hf/flux-dev-panorama-lora-2. Accessed: 2024-12-05.

A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Q. Bu, J. Zeng, L. Chen, Y. Yang, G. Zhou, J. Yan, P. Luo, H. Cui, Y. Ma, and H. Li. Closed-loop visuomotor control with generative expectation for robotic manipulation. *arXiv preprint arXiv:2409.09016*, 2024.

DeepMind. Genie 2: A large-scale foundation world model, 2024. URL deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/. Accessed: 2024-12-10.

Y. Du, M. Yang, P. Florence, F. Xia, A. Wahid, B. Ichter, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum, et al. Video language planning. *ICLR*, 2024a.

Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. In *NeurIPS*, 2024b.

L. Fan, M. Liang, Y. Li, G. Hua, and Y. Wu. Evidential active recognition: Intelligent and prudent open-world embodied perception. In *CVPR*, 2024.

A. Horé and D. Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010.

D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung, H. Adam, H. Akbari, Y. Alon, V. Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *ICML*, 2024.

B. F. Labs. Flux.1 [dev], 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Accessed: 2024-12-05.

T. Lu, T. Shu, A. Yuille, D. Khashabi, and J. Chen. Generative world explorer. *arXiv preprint arXiv:2411.11844*, 2024.

OpenAI. Video generation models as world simulators, 2024.

StabilityAI. Stable zero123, 2023.

A. Tewari, T. Yin, G. Cazenavette, S. Rezchikov, J. Tenenbaum, F. Durand, B. Freeman, and V. Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *NeurIPS*, 2023.

D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric and challenges, 2019. URL https://arxiv.org/abs/1812.01717.

V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

B. Wang, N. Sridhar, C. Feng, M. Van der Merwe, A. Fishman, N. Fazeli, and J. J. Park. This&that: Language-gesture controlled video generation for robot planning. *arXiv preprint arXiv:2407.05530*, 2024a.

S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024b.

Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

WorldLabs. Generating worlds, 2024. URL https://www.worldlabs.ai/blog. Accessed: 2024-12-10.

S. Yang, J. Walker, J. Parker-Holder, Y. Du, J. Bruce, A. Barreto, P. Abbeel, and D. Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.

H.-X. Yu, H. Duan, C. Herrmann, W. T. Freeman, and J. Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

Appendix

A.1. Preliminary: Equirectangular Panorama Images

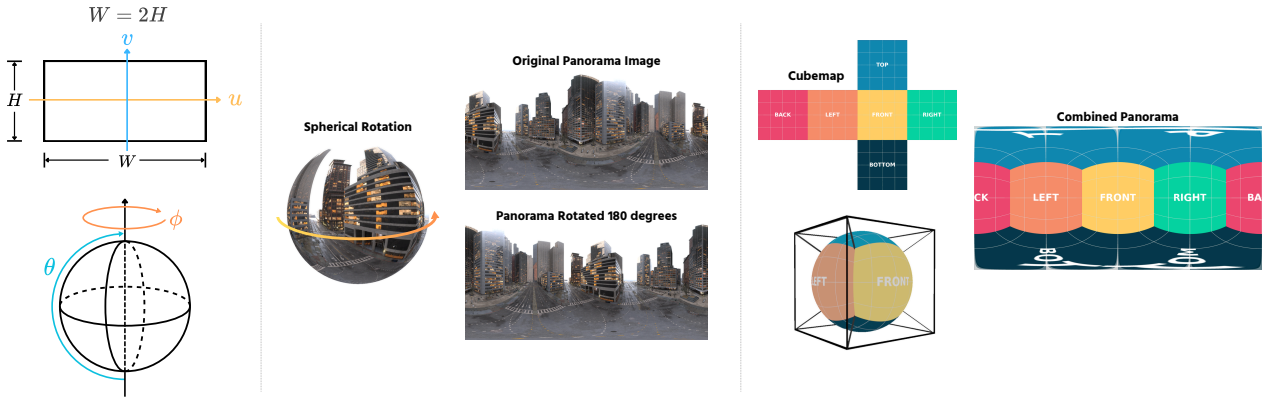

Figure 13 | Left: Pixel Grid coordinate and Spherical Polar coordinate systems; Middle: rotation in Spherical coordinates corresponds to rotation in 2D image; Right: expansion from panorama to cubemap or composition in reverse.

A.1.1. Coordinate Systems

An *Equirectangular Panorama Image* captures all perspectives from an egocentric viewpoint into a 2D image. Essentially, it represents a spherical coordinate system on a 2D grid.

Definition D.1 (Spherical polar coordinate system)**.** $\mathcal{S}$: Taking the origin as the central point, a point in this system is represented by coordinates $(\phi, \theta, r) \in \mathcal{S}$, where ϕ denotes the longitude, θ the latitude, and r the radial distance from the origin. The ranges for these coordinates are $\phi \in [-\pi, \pi)$, $\theta \in [-\pi/2, \pi/2]$, and $r > 0$.

Definition D.2 (Cartesian coordinate system for panoramic image)**.** $\mathcal{P}$: In this system, a pixel is identified by the coordinates $(u, v) \in \mathcal{P}$, where u and v correspond to the column and row positions on the 2D panoramic image plane, respectively. Here, u ranges from 0 to $W - 1$ and v ranges from 0 to $H - 1$.

Definition D.3 (Sphere-to-Cartesian Coordinate Transformation)**.** The transformation between the spherical polar coordinates and the panoramic pixel grid coordinates can be defined by the following functions:

$$f_{\mathcal{S}\to\mathcal{P}}(\phi, \theta) = \left(\frac{W}{2\pi}(\phi + \pi), \frac{H}{\pi}\left(\frac{\pi}{2} - \theta\right)\right) \tag{1}$$

$$f_{\mathcal{P}\to\mathcal{S}}(u, v) = \left(\frac{2\pi u}{W} - \pi, \frac{\pi}{2} - \frac{\pi v}{H}\right) \tag{2}$$

Here, the function $f_{\mathcal{S}\to\mathcal{P}}$ maps the spherical coordinates (ϕ, θ) to the pixel coordinates (u, v), and the inverse function $f_{\mathcal{P}\to\mathcal{S}}$ maps the pixel coordinates (u, v) back to the spherical coordinates (ϕ, θ). This transformation ensures that the entire spherical surface is represented on the 2D panoramic image.

Panorama effectively stores every perspective of the world from a single location. In our work, due to the nature of panoramic images, we are able to preserve the global context during spatial navigation. This allows us to maintain consistency in world information from the conditional image, ensuring that the generated content aligns coherently with the surrounding environment.

A.1.2. Panorama Image transformations

The spherical format allows various image processing tasks. For example, the image can be rotated by an arbitrary angle without any loss of information due to the spherical representation. Additionally, it can be broken down into cubemaps for 2D visualization, as shown in Figure 13.

Definition D.4 (Rotation Transformation in Spherical Polar Coordinate System)**.** Since a panorama image is in a spherical format, we can rotate the image to face a different angle while preserving the original image quality. The rotation can be performed using the following formula:

$$\mathcal{T}(u, v, \Delta\phi, \Delta\theta) = f_{\mathcal{S}\to\mathcal{P}}\left(\mathcal{R}\left(f_{\mathcal{P}\to\mathcal{S}}(u, v), \Delta\phi, \Delta\theta\right)\right) \tag{3}$$

Where the rotation function $\mathcal{R}$ is defined as:

$$\mathcal{R}(\phi, \theta, \Delta\phi, \Delta\theta) = (\phi + \Delta\phi \pmod{2\pi}, \theta + \Delta\theta \pmod{\pi}) \tag{4}$$

If there is no explicit input, both $\Delta\phi$ and $\Delta\theta$ can be set to 0.