# Beyond Thinking: Imagining in 360° for Humanoid Visual Search

**Jingdong Zhang**[1] **Yizhou Wang**[2] **Zhengzhong Tu**[1] **Xin Li**[1]
**Wenping Wang**[1] **Xiaohang Zhan**[2*]

[1]Texas A&M University [2]Adobe

jdzhang@tamu.edu, xiaohangzhan@outlook.com

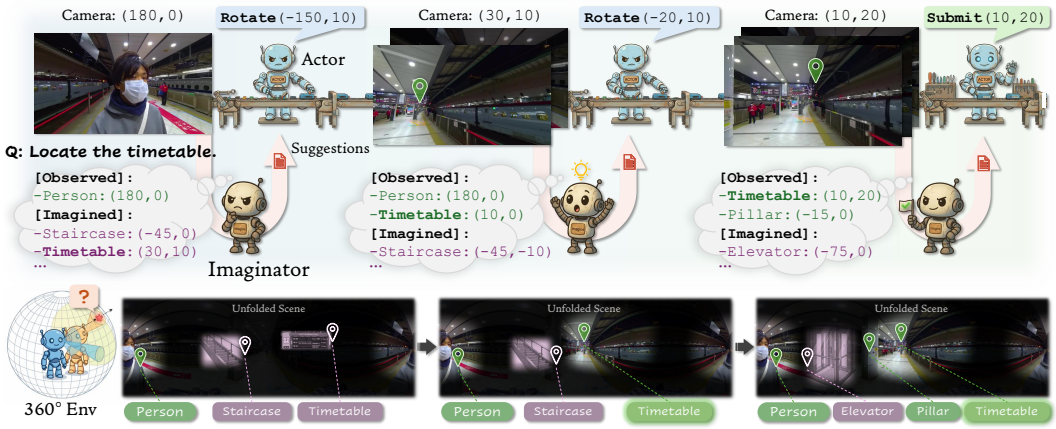



Figure 1: **Imagining in 360°**: We propose a decoupled architecture for Humanoid Visual Search (***HVS***). The **Imaginator** explicitly models the 360° environment by predicting the semantic layout of both *observed* and *unseen* regions, which provides the downstream **Actor** with a sampled distribution of spatial priors as suggestions. This explicit imagination empowers the agent to reorient its head more efficiently, hedging against spatial uncertainty in complex environments.

## Abstract

Humanoid Visual Search (***HVS***) requires agents to actively explore immersive 360° environments. While prior methods treat this as a monolithic task relying on cumulative, multi-turn Chain-of-Thought (CoT) reasoning, they impose heavy cognitive burdens and require expensive trajectory-level annotations. In this paper, we propose **Imagining in 360°**, a novel framework that decouples the exploration process into a specialized **Imaginator** and an **Actor**. The Imaginator functions as a probabilistic predictor of spatial priors; instead of maintaining a cumulative reasoning chain, it infers the semantic layout of both observed and unobserved regions in a single step. By sampling multiple hypotheses within this semantic space, we provide the Actor with a distribution of effective spatial information, offering robust guidance that hedges against uncertainty during active search. This decoupled architecture significantly lowers data engineering costs by eliminating the need for full-trajectory CoT annotations, enabling the generation of approximately 1.92 million curated training samples. Extensive experiments demonstrate that explicitly modeling semantic spatial priors drastically improves search efficiency and success rates on H*Bench and in the wild scenes.

*Corresponding author.
Code: https://github.com/Evergreen0929/Imagining360

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

# 1 Introduction

The human visual system efficiently navigates complex spaces by actively reorienting the head and eyes to deliberately uncover new visual information (Marr, 2010; Stein et al., 2024). Replicating this behavior in artificial agents has driven the development of **Humanoid Visual Search (*HVS*)**, an embodied task where agents must actively explore immersive 360° panoramic environments to locate specific objects (***HOS***) or navigable paths (***HPS***). Unlike passive perception, ***HVS*** requires a closed-loop perception-action cycle (Torralba et al., 2006). To tackle this, a pioneering work, *Thinking in 360°* (Yu et al., 2025a), formulates the search as a sequential decision-making process powered by Multimodal Large Language Models (MLLMs) (Wu and Xie, 2024). Their approach employs a single-model, Chain-of-Thought (CoT) (Wei et al., 2022) framework where the agent iteratively accumulates narrow field-of-view observations, maintains a historical memory of the explored space, and plans the next head-turning action within a unified reasoning chain.

However, this monolithic paradigm introduces significant bottlenecks. First, humanoid visual reasoning in 3D environments remains highly challenging for current MLLMs (Yu et al., 2025a), with the coupling of the challenging spatial understanding with the execution of action planning into a single model. Second, because the model lacks the explicit modeling of spatial relationships, it is forced to implicitly guess the layout of unseen regions, leading to inactive or inaccurate explorations. Finally, teaching the MLLM as a single agent for this process relies heavily on curating high-quality, long-context CoT reasoning trajectories. The prohibitive cost of acquiring such dense, multi-step expert annotations fundamentally limits the ability to scale these systems and hinders their generalization across diverse, in-the-wild scenarios.

In this work, we propose a different way for embodied agents to approach unobserved space. Our core insight is that human-like search is not strictly driven by exhaustive logical chains alone, but is heavily guided by intuitive spatial priors (*e.g.* expecting an open corridor likely leads to an entrance or exit). Thus, we propose **Imagining in 360°**, a framework designed to explicitly decouple probabilistic spatial imagination from rigorous CoT reasoning. By introducing intuitive spatial priors to guide the cumulative reasoning process, we transform the exploration task into a collaboration between probabilistic predicting of spatial priors and rigorous reasoning, which encourages the agent to explore more actively and confidently, grounded in sound spatial assumptions.

To achieve this, we introduce an architecture consisting of a decoupled **Imaginator** and an **Actor**. Rather than relying on computationally heavy pixel-level reconstruction (He et al., 2022) or visual inpainting (Lugmayr et al., 2022) to guess unseen areas, we propose to efficiently model the visual horizon directly within the semantic space of language. Given limited narrow field-of-view (NFoV) observations, the Imaginator predicts the layout of both *observed* and *imagined* coordinates of highly semantic regions. Crucially, the Imaginator functions as a probabilistic predictor of spatial priors, instead of outputting a single deterministic guess, it generates multiple spatial hypotheses per step via temperature sampling to capture the inherent uncertainty of occluded environments. As shown in Fig. 1, these semantic layouts are processed as a ranked list of suggestions and explicitly provided to the downstream Actor's context. By evaluating this probabilistic distribution, the Actor gains comprehensive spatial awareness before executing its next head rotation.

Beyond enhancing spatial reasoning, this decoupling unlocks massive data scalability. Because the Imaginator learns spatial priors directly from single-step semantic annotations rather than cumulative CoT histories, we can introduce a fully automated pseudo-label pipeline. This approach effortlessly transforms unlabeled 360° panoramas into high-quality training data, scaling our dataset to $\sim$1.92 million samples across $\sim$20k scenes without manually annotated CoT trajectories.

In summary, our main contributions are as follows:

- **Decoupled HVS Paradigm:** We decouple visual search into intuitive spatial imagination and action planning, allowing for more active and grounded exploration.
- **Probabilistic Imaginator:** We introduce an Imaginator model that learns intuitive spatial priors and guides the Actor agent with probabilistic estimations.
- **Scalable Data Engine:** Our approach enables a fully automated pipeline that scales training to 1.92M samples without manual trajectory labels.

## 2 Related Works

**Embodied Visual Search.** Traditional visual search relies on visual saliency or contextual cues (Oliva et al., 2003; Torralba et al., 2006; Zelinsky et al., 2005). While Multimodal Large Language Models (MLLMs) like $V^*$ (Wu and Xie, 2024) have advanced these capabilities, they predominantly operate on static, disembodied 2D images, as do recent agents that zoom into image regions, such as DeepEyes (Zheng et al., 2025) and Mini-o3 (Lai et al., 2025). Embodied navigation, including target-driven and object-goal navigation (Zhu et al., 2017; Chaplot et al., 2020b; Majumdar et al., 2022; Chang et al., 2024) and vision-and-language navigation (Anderson et al., 2018; Zhou et al., 2024; Qi et al., 2025), moves an agent through 3D environments, often in simulation (Savva et al., 2019), and many methods build maps over long trajectories (Chaplot et al., 2020a,c). However, real-world search is an active, embodied behavior requiring coordinated head and eye movements to uncover occluded information (Lanman et al., 1978; Barnes, 1979; Solman et al., 2017), which recent robot learning work also models (Kerr et al., 2025; Yu et al., 2025b). Emulating this has driven the emergence of Humanoid Visual Search (***HVS***) (Yu et al., 2025a), where agents actively explore immersive 360° spaces. While pioneering works (Yu et al., 2025a) tackle ***HVS*** via cumulative Chain-of-Thought (CoT) reasoning, this monolithic paradigm imposes a severe cognitive burden and limits data scalability. To overcome this, our framework explicitly decouples spatial imagination from action.

**Agentic Multimodal LLMs.** Multimodal LLMs (MLLMs) represent a significant stride toward unified understanding models among multi-modalities (Alayrac et al., 2022; Liu et al., 2023; Li et al., 2022; Lai et al., 2024; Yang et al., 2025). To transcend their innate perceptual limitations, recent MLLM agents have been empowered with external toolkits (*e.g.*, web browsing, code execution) and refined via multi-turn reinforcement learning (Guo et al., 2025; Jin et al., 2025; Kimi et al., 2025). In the visual domain, agents leverage symbolic tool calls like iterative zooming or region selection to achieve fine-grained understanding (Li et al., 2025; Shen et al., 2024; Zhang et al., 2025c; Wang et al., 2025). Despite these advancements, these tool-use paradigms often occur on a disembodied 2D canvas. While some recent efforts ground MLLMs in embodied reasoning for robotics (Azzolini et al., 2025; Team et al., 2025; Chen et al., 2025b), active visual search driven by a generative spatial prior remains largely underexplored. We introduce an *Imaginator* as a specialized agentic component that provides probabilistic spatial guidance, enabling more efficient and robust search behaviors.

**Spatial Imagination and Scene Extrapolation.** Effective embodied exploration requires anticipating regions beyond the limited narrow field-of-view (NFoV). Prior methods typically rely on pixel-level extrapolation (Shi et al., 2023; Tang et al., 2023; Huang et al., 2025c,b; He et al., 2022; Zhang et al., 2025a), which is computationally expensive and artifact-prone, or implicit latent spaces (LeCun et al., 2022; Assran et al., 2023; Bardes et al., 2024; Huang et al., 2025a), which lack the explicit semantic grounding required for language-driven planning. Meanwhile, out-of-view MLLMs (Chen et al., 2025a) and scene graphs (Gu et al., 2024) often restrict themselves to simple VQA or demand exhaustive multi-view scanning. Departing from these, we frame spatial imagination as generating *probabilistic spatial priors*. Inspired by biological cognitive mapping (Marr, 2010; Bartz, 1966; Tolman, 1948), our *Imaginator* extrapolates the 360° topology entirely in language space. By predicting coordinates for visible and occluded entities, we bypass pixel-level overhead and CoT data curation, offering interpretable and efficient guidance for navigation.

## 3 Method

To overcome the cognitive overload and data-scalability bottlenecks inherent in monolithic CoT search paradigms, we propose a decoupled architecture for Humanoid Visual Search (***HVS***). As illustrated in Fig. 2, our framework separates the active search process into two distinct roles: an **Imaginator** that explicitly models the 360° environment and generates probabilistic spatial priors, and an **Actor** that acts upon these priors to navigate in it.

### 3.1 Problem Formulation

**Objective.** We formulate the environment as a continuous 360° panorama. Because the embodied agent possesses a limited narrow field-of-view (NFoV), it must actively rotate its head to explore the surrounding space. At any step, the agent receives a NFoV observation $o_{\phi,\gamma} \in \mathcal{O}$, defined by its current azimuth ($\phi$) and polar angle ($\gamma$). Given a natural language instruction $x$, the overarching goal of ***HVS*** is to iteratively gather visual evidence and ultimately predict a final viewing direction $(\hat{\phi}, \hat{\gamma})$

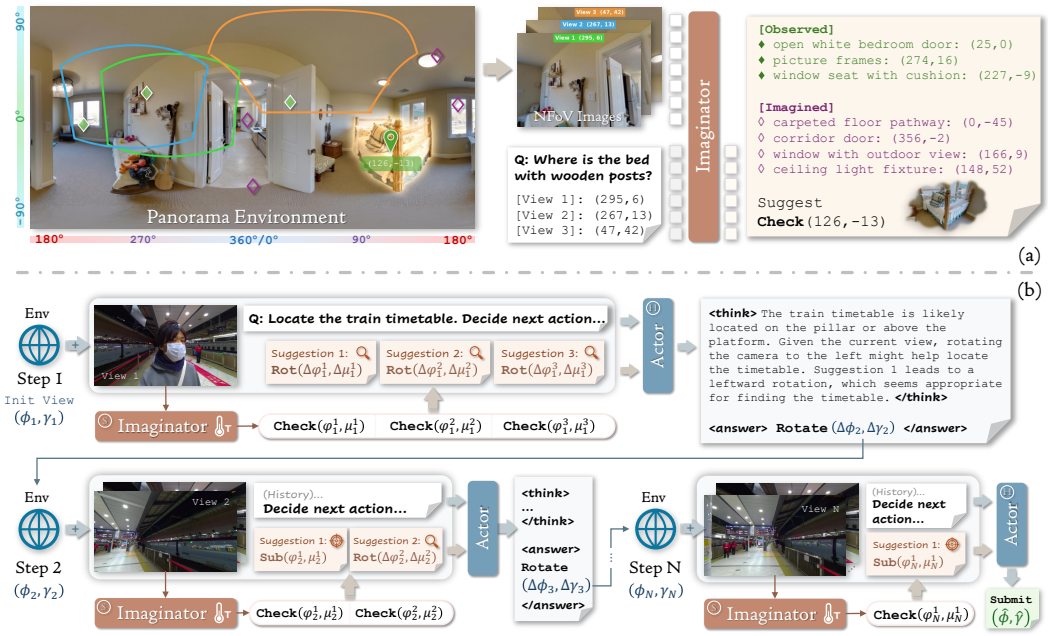



Figure 2: **Framework Overview.** (a) **360° Env Imagination**: The Imaginator performs a single-step estimation of the global layout by predicting $(\varphi, \mu)$ coordinates for both observed landmarks and imagined entities to capture high-level spatial relationships. (b) **Collaborative HVS Pipeline**: The Actor integrates the sampled distribution of probabilistic spatial suggestions into its reasoning chain, allowing the agent to hedge against environmental uncertainty and guide active search efficiently. `Rot`/`Sub` in (b) abbreviate `Rotate`/`Submit`.

that maximizes the probability of task success $r_s$:

$$(\hat{\phi}, \hat{\gamma}) = \arg\max_{\phi,\gamma} P(r_s \mid o_{\phi,\gamma}, x). \tag{1}$$

Throughout, $(\phi, \gamma)$ denotes the viewing direction of the Actor, and $(\varphi, \mu)$ denotes the absolute direction predicted by the Imaginator. Following the task definition in Yu et al. (2025a), **Humanoid Object Search (HOS)** requires the agent to locate a specific target and submit a final view $(\hat{\phi}, \hat{\gamma})$ that locates the object, while **Humanoid Path Search (HPS)** involves identifying a final orientation $\hat{\phi}$ aligned with a navigable path.

### 3.2 The Imaginator Model

Rather than forcing a single agent to implicitly guess the layout of unobserved areas, we introduce the **Imaginator**. Its sole purpose is to construct a semantic layout map of both observed and unobserved regions based on the current visual history, as shown in Fig. 2(a), acting as a generative spatial prior.

**Task Setup and Model Execution.** Spatial object layouts in the physical world strictly adhere to inherent geometric and semantic rules (*e.g.*, a chandelier typically hangs from the ceiling rather than sitting on the floor; a teacup rests on a table rather than a wall). Internalizing these structural priors is crucial for accurate ***HVS***, as they drastically narrow down redundant explorations and facilitate fast convergence. Therefore, our primary objective is to enable the model to learn these spatial priors from the data distribution to robustly guide the search process.

A naive approach might treat this as a direct target detection task, simply mapping the instruction $x$ and historical observations to a target coordinate $(\varphi, \mu)$. However, due to the severely limited field-of-view, this direct mapping often causes the solution space to collapse into a sharply peaked distribution. Deprived of sufficient visual context, the model is highly susceptible to generating high-confidence but erroneous predictions.

To foster reasonable exploratory behaviors akin to human intuition, we design the Imaginator to perform a comprehensive layout estimation rather than a solitary point prediction. At step $t$, receiving

all visible FoV images $\{o_{\phi_1,\gamma_1}, \ldots, o_{\phi_t,\gamma_t}\}$ and the instruction $x$, the Imaginator executes a stateless, single-step forward pass. It outputs a structured block that explicitly lists semantic entities and their predicted absolute coordinates $(\varphi, \mu)$ in the equirectangular space:

- `[Observed]`: An estimation of landmarks currently or previously visible in the NFoVs (*e.g.*, "*picture frames: (274, 16)*").
- `[Imagined]`: A probabilistic prediction of objects likely existing in *unseen* regions (*e.g.*, "*window with outdoor view: (166, 9)*").

By enforcing this joint prediction of both observed and imagined entities, the Imaginator explicitly models the spatial relationships of key visual anchors. This structural grounding makes it significantly easier to infer the target's location, naturally producing a smoother distribution supporting the sampling of spatial priors. The final output is a brief recommendation `Suggest Check` $(\varphi, \mu)$.

**Automated Scale-Up Data Engine.** Training the Imaginator is inherently more scalable than prior CoT-based methods (Yu et al., 2025a). Instead of relying on teacher models for long, multi-step trajectories that require costly manual curation, our decoupled architecture frames spatial learning simply as a *masked spatial modeling* task on panoramas. Specifically, we utilize a powerful teacher model to perform accurate and comprehensive labeling across the complete 360° environment. From these fully annotated panoramas, we synthetically sample trajectories and reveal only 1 to 8 NFoV observations per training instance (with 50% targets kept out of views to force spatial layout imagination). By feeding the Imaginator these limited views and tasking it to predict the entire layout (following the `[Observed]` and `[Imagined]` format discussed above), we actively force the model to comprehend the global environment and internalize spatial relationships. This paradigm completely bypasses the prohibitive cost of trajectory-level CoT annotations, enabling a fully automated, low-cost pseudo-labeling pipeline. As summarized in Fig. 3, we scale our training set using approximately 20k diverse panoramic environments filtered from Zhang et al. (2026), including real-world and synthetic subsets. Through deterministic trajectory expansion, this efficiently yields ∼1.92 million high-quality, single-step spatial reasoning samples.

**Training Objective.** Formally, given the instruction $x$ and visual history $\mathcal{O}_t = \{o_{\phi_1,\gamma_1}, \ldots, o_{\phi_t,\gamma_t}\}$, the Imaginator generates the target sequence $y = \{y_1, \ldots, y_L\}$, comprising the structured layouts and spatial suggestions. We optimize the model parameter $\theta$ using standard auto-regressive cross-entropy loss applied only to the completion tokens:

$$\mathcal{L} = -\sum_{i=1}^{L} \log P_\theta(y_i \mid x, \mathcal{O}_t, y_{<i}) \tag{2}$$

To efficiently internalize 3D spatial priors, the model is first pretrained on the large-scale synthetic data, and subsequently finetuned on the processed ***HVS*** training split in Yu et al. (2025a) with the same format for precise task alignment. Both stages use the objective above: the ∼1.92M pseudo labels are abundant and diverse but noisy, while the small, human-annotated split recovers precision and refines the task output format.

### 3.3 The Actor Model

The **Actor** is the agent responsible for executing the physical search within the environment. Our framework is agnostic to the specific architecture of the Actor; it can be any off-the-shelf MLLM or a fine-tuned search model, *e.g.*, the HVS-3B model proposed in Yu et al. (2025a). Following their standard ***HVS*** formulation, the **Actor** is a tool-augmented model responsible for executing the physical search. At step $t$, the Actor receives the target instruction $x$, the current observation $o_{\phi_t,\gamma_t}$, its historical memory, and crucially, the spatial suggestions $\mathcal{S}_t$ provided by the Imaginator.

Empowered by these probabilistic spatial priors, the Actor generates a textual reasoning chain (the `<think>` block) to assess the plausibility of the suggested layouts against its own visual memory. It concludes by outputting an `<answer>` block specifying the next action from two primitives:

- `Rotate` $(\Delta\phi_{t+1}, \Delta\gamma_{t+1})$: Adjusts the viewing direction relative to the current camera pose.
- `Submit` $(\hat{\phi}, \hat{\gamma})$: Commits the current view as the final target estimate and terminates the search.

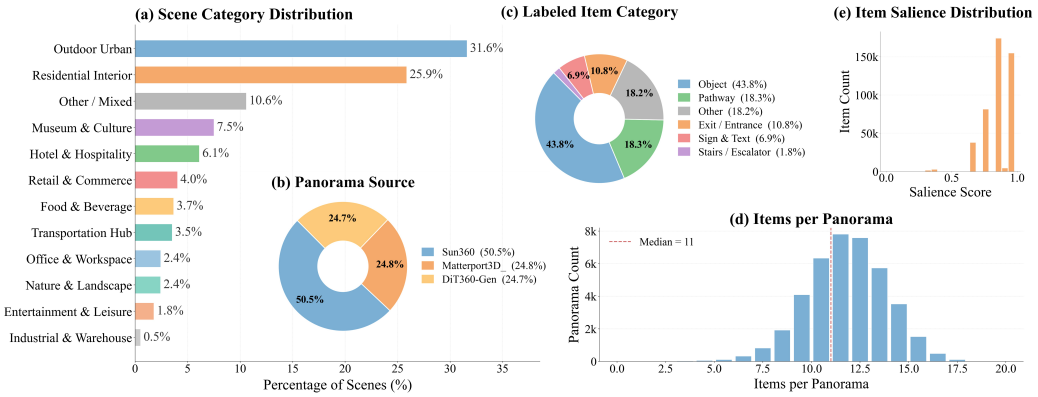



Figure 3: **Overview of the Curated Training Dataset.** (a) Distribution of scene categories, highlighting a diverse mix dominated by most commonly encountered outdoor urban (31.6%) and residential interior (25.9%) environments. (b) The composition of unlabeled panorama sources. (c) The distribution of labeled semantic item categories. (d) The histogram of labeled items per panorama, with a median of 11 items, demonstrates dense semantic coverage for spatial modeling. (e) Salience distribution of the labeled items in the environment.

### 3.4 Imagination-Guided HVS Pipeline

While the Imaginator provides scalable spatial priors, its stochastic predictions are unsuitable for final navigation decisions. Conversely, a standalone Actor lacks explicit spatial modeling and often hallucinates in unseen regions. We thus propose a complementary pipeline: the Imaginator generates probabilistic priors, which the Actor grounds via rigorous sequential reasoning.

As shown in Fig. 2 (b), to fully exploit the Imaginator's spatial knowledge without over-relying on a single potentially flawed layout, we formulate its output as a probabilistic distribution. At each step $t$, we sample $K$ candidate coordinates, $\mathcal{C}_t = \{(\varphi_t^1, \mu_t^1), \ldots, (\varphi_t^K, \mu_t^K)\}$, extracted from its `Suggest Check` $(\varphi, \mu)$ outputs. The first candidate is generated via greedy decoding ($T = 0$) to anchor the most confident prediction. During the first steps of the search, the remaining $K - 1$ candidates are stochastically sampled ($T > 0$) to capture diverse spatial hypotheses. The temperature $T$ follows a step-wise decay schedule over the search steps: as the agent accumulates more visual clues, the environmental uncertainty naturally decreases, transitioning the sampling strategy from broad exploration to focused prediction.

Before injecting the hypotheses into the Actor's context, they are translated into actionable primitives relative to the Actor's current camera pose $(\phi_t, \gamma_t)$. We deploy a suggestion converter to process each absolute predicted coordinate $(\varphi_t^i, \mu_t^i)$. First, it calculates the required angular displacement:

$$\Delta\varphi_t^i = (\varphi_t^i - \phi_t) \bmod 360^\circ, \quad \Delta\mu_t^i = \mu_t^i - \gamma_t \tag{3}$$

Next, the converter evaluates whether the imagined target location already falls within the boundaries of the current NFoV, denoted by angular bounds $(F_\phi, F_\gamma)$. If the coordinate is within the current view, further exploration is unnecessary, and the suggestion is mapped to a termination action. Otherwise, it is mapped to a rotation action directing the agent toward the unobserved region. Formally, the mapped suggestion $s_t^i$ is defined as:

$$s_t^i = \begin{cases} \texttt{Submit}(\varphi_t^i, \mu_t^i), & \text{if } |\Delta\varphi_t^i| \le \frac{F_\phi}{2} \text{ and } |\Delta\mu_t^i| \le \frac{F_\gamma}{2} \\ \texttt{Rotate}(\Delta\varphi_t^i, \Delta\mu_t^i), & \text{otherwise} \end{cases} \tag{4}$$

This mapping yields a final, ranked set of actionable suggestions $\mathcal{S}_t = \{s_t^1, s_t^2, \ldots, s_t^K\}$. These suggestions are seamlessly injected into the Actor's prompt under the specific tag `[Spatial Imagination Suggestions]`. By explicitly evaluating this distribution of spatial priors, the Actor obtains awareness of the spatial layout, significantly suppressing overconfident exploration errors and accelerating convergence in complex environments.

Table 1: **Results on H*Bench.** We report the Success Rate (%) across various Vision-Language Model families. '×' denotes the baseline Actor relying solely on its internal CoT. '✓' indicates the Actor plugged in with our Imaginator's spatial priors.

| Actor | | Imaginator | Humanoid Object Search | | | | Humanoid Path Search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Family | Params | | **Overall** | Easy | Medium | Hard | **Overall** | Easy | Medium | Hard | Extreme |
| *Open-Source Models* | | | | | | | | | | | |
| Qwen2.5-VL | 3B | × | 21.17 | 18.54 | 16.48 | 22.68 | 7.69 | 9.20 | 9.65 | 3.50 | 8.33 |
| | | ✓ | 49.54 | 59.19 | 51.70 | 45.62 | 25.94 | 34.80 | 22.15 | 24.77 | 15.74 |
| | 7B | × | 12.79 | 31.22 | 7.39 | 6.34 | 9.88 | 12.20 | 11.18 | 7.24 | 6.94 |
| | | ✓ | 57.42 | 70.73 | 51.70 | 52.95 | 28.38 | 35.20 | 24.34 | 29.44 | 18.98 |
| | 72B | × | 29.54 | 36.91 | 21.59 | 27.59 | 12.94 | 17.40 | 9.21 | 15.65 | 5.09 |
| | | ✓ | 62.25 | 70.08 | 60.80 | 59.42 | 38.19 | 45.00 | 36.18 | 39.25 | 24.54 |
| Qwen3-VL | 32B | × | 34.38 | 66.83 | 36.36 | 21.75 | 23.19 | 26.00 | 24.78 | 22.43 | 14.81 |
| | | ✓ | 57.46 | 73.98 | 51.70 | 51.77 | 31.31 | 37.60 | 30.04 | 30.14 | 21.76 |
| | 235B | × | 23.62 | 57.24 | 26.70 | 10.44 | 21.44 | 26.80 | 19.96 | 23.36 | 8.33 |
| | | ✓ | 65.04 | 81.46 | 60.23 | 59.29 | 38.25 | 46.60 | 35.31 | 37.15 | 27.31 |
| Gemma-3 | 4B | × | 18.17 | 41.79 | 25.57 | 8.33 | 14.50 | 17.80 | 16.67 | 10.98 | 9.26 |
| | | ✓ | 57.75 | 72.03 | 60.23 | 52.02 | 35.06 | 43.80 | 27.41 | 36.45 | 28.24 |
| | 12B | × | 22.88 | 54.96 | 32.39 | 9.57 | 15.31 | 16.80 | 16.23 | 14.49 | 11.57 |
| | | ✓ | 54.62 | 63.90 | 48.86 | 51.71 | 33.56 | 43.00 | 30.04 | 32.24 | 21.76 |
| Kimi-VL | A3B | × | 4.33 | 12.36 | 0.57 | 1.68 | 5.62 | 7.20 | 6.80 | 2.34 | 6.02 |
| | | ✓ | 32.46 | 40.00 | 27.84 | 30.08 | 21.00 | 28.20 | 16.89 | 21.96 | 11.11 |
| *Proprietary Models* | | | | | | | | | | | |
| Gemini-2.5 | Flash | × | 43.70 | 79.50 | 59.70 | 28.30 | 27.70 | 35.80 | 22.60 | 31.30 | 12.50 |
| | | ✓ | 63.20 | 86.30 | 69.30 | 53.70 | 37.70 | 50.20 | 30.30 | 42.30 | 15.30 |
| | Pro | × | 55.00 | 88.00 | 62.50 | 41.60 | 37.90 | 43.80 | 36.40 | 40.70 | 21.80 |
| | | ✓ | 66.10 | 89.60 | 68.80 | 56.80 | 43.10 | 51.00 | 39.90 | 48.80 | 20.40 |
| *Finetuned Models* | | | | | | | | | | | |
| HVS | 3B | × | 48.04 | 61.79 | 26.14 | 45.18 | 24.12 | 36.60 | 17.76 | 22.66 | 11.57 |
| | | ✓ | 62.75 | 75.77 | 64.20 | 57.61 | 39.38 | 49.40 | 32.89 | 41.12 | 26.39 |

# 4 Experiment

## 4.1 Implementation Details

**Data Curation.** To construct the pseudo-pretrained dataset, we employ Qwen3-VL-235B as the teacher model and to filter ∼20k valid indoor/outdoor scenes from Sun360 (Xiao et al., 2012), Matterport3D (Chang et al., 2017), and synthesized by DiT360 (Feng et al., 2025). Then we localize 5-15 salient semantic targets per panorama. For each scene, we synthetically render 24 NFoV trajectories, in which 50% of the targets are deliberately withheld from the field-of-view. Deterministically expanding these trajectories based on FoV coverage yields ∼1.92M single-step `[Observed]/[Imagined]` samples. We also curate 2,100 ***HOS*** and 2,148 ***HPS*** imagine data with the same format on H*Bench (Yu et al., 2025a) for task alignment.

**Imaginator Training.** The Imaginator is initialized from Qwen3-VL-8B-Instruct (Yang et al., 2025). We first train the full model on 1.92M pseudo dataset for 1 epoch with a global batch size of 192 to generate structured imagination formats, instilling broad imagination priors across diverse scene types. Subsequently, we perform full-parameter supervised finetuning on the processed HVS training split to focus strictly on task alignment. This SFT runs for 2 epochs with a global batch size of 8 and a learning rate of $1 \times 10^{-5}$. The experiments are conducted on 8 NVIDIA H200/A100 GPUs.

**Actor and Inference.** Our framework is architecture-agnostic. We validate generalization across off-the-shelf models including *open-source*: Qwen2.5-VL (Xu et al., 2025), Qwen3-VL (Yang et al., 2025), Gemma-3 (Gemma Team et al., 2025), and Kimi-VL (Team, 2025); *proprietary models*: Gemini 2.5 (Sadigh et al., 2025); *finetuned models:* HVS-3B (Yu et al., 2025a). During inference, the Imaginator generates three spatial hypotheses per step: one greedy anchor ($T = 0$) and two stochastic samples ($k = 50$, initial $T_1 = 0.7$ with a 0.85 step-wise decay) in the first two steps. These absolute predictions are converted into ego-centric relative actions and appended to the Actor's context.

**Evaluation Protocol.** Following Yu et al. (2025a), search episodes are within 10 steps. NFoV observations are rendered with a 100° horizontal FoV at $1920 \times 1080$ during evaluation, while training views are $960 \times 720$, the resolution of the H*Bench SFT data. Success requires the final

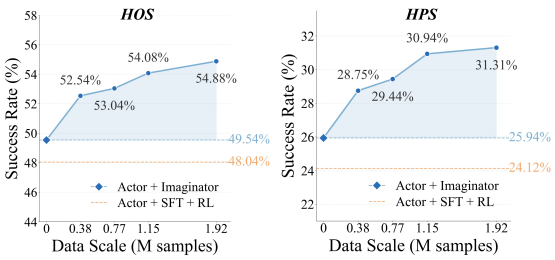



Figure 4: **Scaling Laws of Imagination.** We use Qwen2.5-VL-3B as the Actor model. The dashed line (Actor + SFT + RL) is HVS-3B.

Table 2: **Ablation on Guidance Type.** Evaluated on (***HOS***) with HVS-3B as Actor Model, where *layout* represents the whole-scene layout modeling during imagination and $T$=0 means without probabilistic sampling.

| **Imagination Guidance** | **Overall** | Easy | Medium | Hard |
|---|---|---|---|---|
| Actor | 48.04 | 61.79 | 26.14 | 45.18 |
| Actor + Random Suggestions | 20.83 | 23.09 | 19.32 | 20.14 |
| Actor + Pixel Imagination | 46.17 | 61.46 | 27.27 | 42.39 |
| Actor + Imaginator (*w.o. layout*) | 41.00 | 59.84 | 34.09 | 34.56 |
| Actor + Imaginator ($T$=0) | 61.54 | 74.47 | 58.52 | 56.93 |
| Actor + Imaginator (*full*) | **62.75** | **75.77** | **64.20** | **57.61** |

submitted view to fall within the bounding-box-centered angular tolerance: $\tau_\phi = 30°$ and $\tau_\gamma = 20°$ for ***HOS***, and $\tau_\phi = 10°$ for ***HPS***.

**Train/Test Separation.** We adopt the H*Bench split of Yu et al. (2025a) unchanged: 1,000 instances (600 ***HOS***, 400 ***HPS***) form the evaluation set, and our stage-2 data is drawn only from its SFT set of 250 ***HOS*** and 250 ***HPS*** instances. The pseudo-pretraining data does not overlap with the evaluation panorama environments (App. A.3).

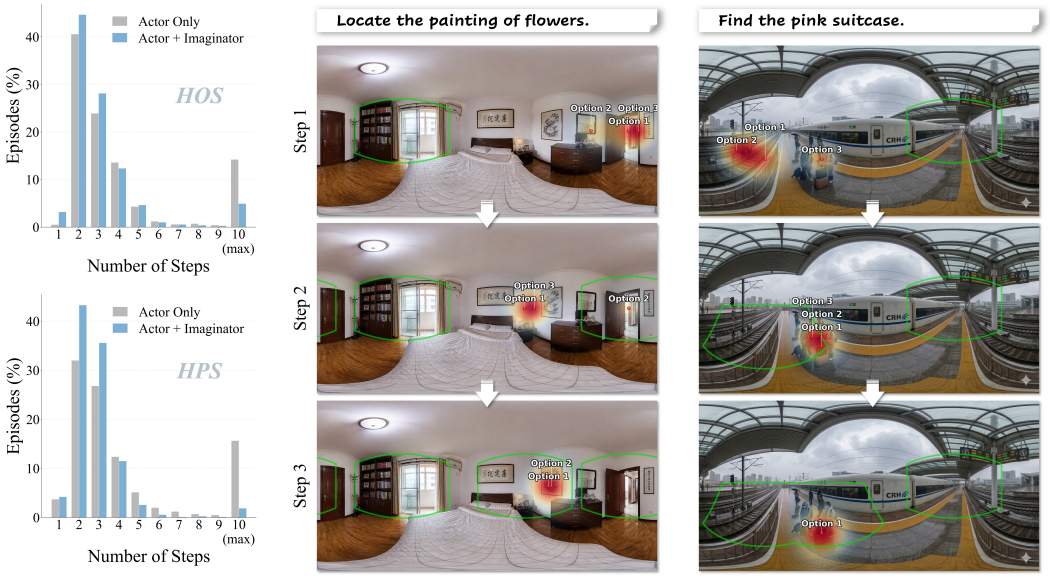
### 4.2 H*Bench Evaluation

Table 1 presents the comprehensive evaluation of our framework across a diverse collection of *Open-Source*, *Proprietary*, and domain-specific *Finetuned* MLLMs. For off-the-shelf open-source models, the monolithic CoT approach (Actor only) generally struggles with the spatial reasoning, often resulting in overall success rates below 35% on ***HOS*** and 25% on ***HPS***. However, when augmented with the Imaginator, models like Qwen3-VL (235B) surge from 23.62% to 65.04% on ***HOS***, and from 21.44% to 38.25% on ***HPS***. Remarkably, this plug-and-play compatibility extends to powerful proprietary models; Gemini-2.5 Flash/Pro receives notable gain on both splits. Most importantly, our decoupled pipeline significantly outperforms the previous state-of-the-art established by RL-finetuned models. The baseline HVS-3B model achieves 48.04% on ***HOS*** and 24.12% on ***HPS***. When integrated into our joint pipeline as the Actor, its performance jumps to 62.75% and 39.38%, respectively. This validates the limitation, the lack of explicit, probabilistic spatial modeling, in current single-agent architectures.

### 4.3 Ablation Studies

**Necessity of Structured Layouts and Probabilistic Sampling.** Table 2 dissects the specific components of our imagination guidance using the HVS-3B Actor. First, injecting random suggestions severely degrades performance (20.83%), confirming the Actor's reliance on external priors. To evaluate alternative representation spaces, we implement a *pixel imagination* baseline by fine-tuning FLUX.1-Fill (Labs et al., 2025) on ∼1.4M panoramas. At each step, FLUX out-paints the unobserved regions conditioned on the accumulated NFoVs, and the generated panorama is fed into the Actor. However, modeling low-level pixel noise proves sub-optimal for high-level reasoning, yielding inferior navigation signals (46.17%). Furthermore, ablating the structured `[Observed]` and `[Imagined]` context (*w.o. layout*) drops the success rate to 41.00%, lagging even behind the unaugmented baseline Actor (48.04%). This confirms that explicitly modeling global spatial relationships is essential to prevent hallucinated priors. Finally, our probabilistic sampling (*full*, 62.75%) consistently outperforms deterministic greedy decoding ($T$=0, 61.54%).

**Controls for Context, Capacity and Distillation.** Table 5 (App. A.5) replaces the Imaginator's output with other content at the same position and format in the Actor prompt: a matched-token generic hint, uniformly spaced directions, a scene-category prior and a retrieval prior all fall below the Actor alone. Using the Qwen3-VL-235B teacher directly gives 40.17%, and spending the same 3B+8B budget on a second, 8B Actor gives 50.12%, against 62.75% with our Imaginator. Letting one model both imagine and act, with the layout inside its own reasoning chain, gives 47.04% on ***HOS*** and 24.08% on ***HPS***, level with the Actor alone.




(a) Step Distribution (b) Imagination Distribution

Figure 5: **Distribution Analysis.** (a) Histograms of the step distribution, illustrating the percentage of search episodes that terminate at each step for both the baseline (Actor Only) and our joint pipeline across the ***HOS*** and ***HPS*** tasks. (b) Step-by-step heatmap visualizations of the Imaginator's probabilistic spatial suggestions. The heat areas indicate the probability distribution of the imagined target coordinates, which are projected onto the complete 360° environment.

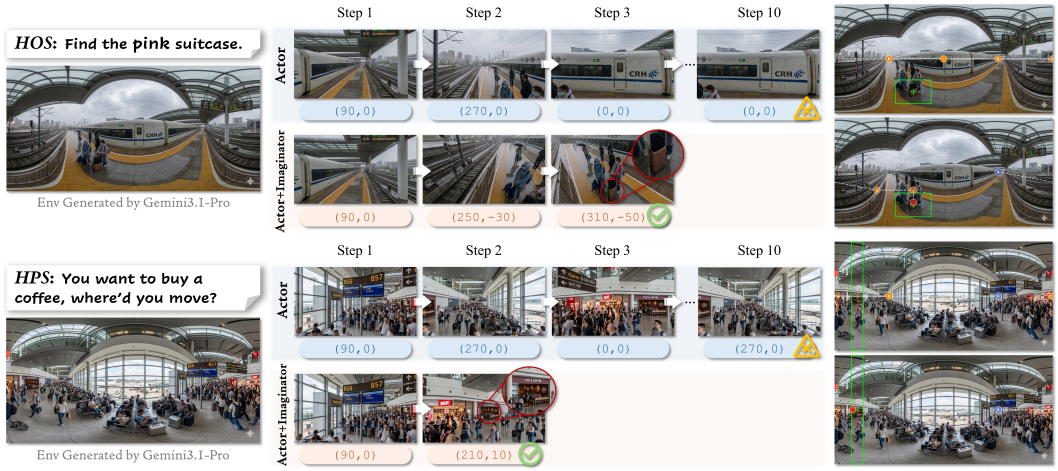



Figure 6: **Qualitative Search Trajectories in the wild.** Evaluated in complex environments generated by Gemini 3.1-Pro. The Actor loses context and falls into an infinite dead loop, whereas our decoupled framework leverages layout imagination to locate the target within 2-3 efficient steps.

**Data Scaling Laws.** Fig. 4 demonstrates the scaling dynamics of the **Imaginator**. Increasing its pseudo-pretrained data from 0 to 1.92M samples yields monotonic improvements across both ***HOS*** and ***HPS*** tasks. Notably, equipping the Actor with an Imaginator trained on merely 0.38M samples is already sufficient to surpass the costly RL-finetuned baseline (orange dashed line). At zero pseudo-labeled samples, the Imaginator is trained only on the H*Bench SFT split, the same annotations HVS-3B is trained on, and the pipeline reaches 49.54% on ***HOS*** and 25.94% on ***HPS***, against 48.04% and 24.12% for HVS-3B. We also discover evaluating the Imaginator in isolation yields modest results (***HOS*** 37.62%, ***HPS*** 25.31%), as its single-step sampling lacks sequential reasoning. Yet, integrating it with the Actor unlocks massive gains, proving our framework is not a trivial ensemble, but a complementary architecture that grounds generative spatial priors with logical action planning.

**Direct Evaluation of the Imaginator.** Scored directly against the targets from the initial view, the Imaginator's greedy suggestion hits 39.04% (***HOS***) and 26.12% (***HPS***), and the deployed $K$=3 hypotheses 51.72% and 42.29%. Entity localization and uncertainty analyses are in App. A.6.

**Search Efficiency and Distribution Analysis.** Beyond success rates, spatial priors dramatically improve search efficiency. As shown in Fig. 5(a), the baseline Actor frequently tails toward the 10-step failure cutoff due to redundant blind rotations. Conversely, our joint pipeline shifts the distribution sharply leftward, with most searches rapidly converging within 1-3 steps. Over the 2,400 ***HOS*** episodes, 53.5% of the 8,080 issued suggestions fall inside the current view and are presented as `Submit`. Terminating remains the Actor's decision: 94.1% of episodes end with the Actor's own `Submit`, and the remaining 5.9% exhaust the 10-step budget. Compared to pixel-level imagination (e.g., FLUX.1), which incurs a prohibitive latency of 20-30s per step, a single Imaginator forward pass requires only $\sim$2.5s. Even when sampling multiple spatial suggestions, the total latency remains constrained to 5-7s, drastically reducing inference costs. Fig. 5(b) visualizes this underlying mechanism. Initially, given a limited field-of-view, the Imaginator generates a broad but logical distribution of spatial hypotheses to guide early exploration. Crucially, as the agent rotates and accumulates visual evidence, the environmental uncertainty rapidly decreases. The predictive heatmap progressively narrows step-by-step, seamlessly converging onto the exact target location and completely eliminating the need for aimless 360° sweeping.

### 4.4 Qualitative Results in the Wild

To validate real-world robustness, we evaluate the framework on complex, out-of-distribution panoramas generated by Gemini 3.1-Pro (Fig. 6). In the ***HOS*** scenario (train platform), the baseline Actor is trapped in an oscillating loop and fails. Conversely, the Imaginator instantly models the longitudinal layout, guiding the Actor to foveate the target by Step 3. Similarly, in the ***HPS*** scenario (airport terminal), the joint pipeline imagines occluded storefronts behind the crowd to locate the coffee shop pathway in just 2 steps. Additional qualitative results are provided in the supplementary material.

**Held-out Benchmark.** We further evaluate on a held-out benchmark of 15 panoramas (8 Google Street View, 7 generated), disjoint from all training data and H*Bench, with 47 targets annotated by human experts. The Imaginator raises the success rate from 47.87% to 58.51% for HVS-3B and from 17.55% to 46.28% for Qwen2.5-VL-3B (Table 6, App. A.7).

## 5 Conclusion and Limitations

In this paper, we introduced *Imagining in 360°*, a decoupled framework for Humanoid Visual Search (***HVS***) that separates exploration into a generative Imaginator and an executive Actor. This mitigates the cognitive and data-scalability bottlenecks of monolithic models. By leveraging masked world modeling, the Imaginator provides probabilistic semantic layouts that enable the Actor to hedge against uncertainty and conduct efficient, directed searches. Experiments validate that the Imaginator universally boosts foundational MLLMs on ***HVS*** tasks. Despite its success, our framework relies on an open-loop execution paradigm where the Actor cannot provide feedback on the Imaginator's generated priors. Future work will explore incorporating a *Critic* module to dynamically evaluate and correct hypotheses, enabling a fully closed-loop perception-action-evaluation workflow. In addition, the deployment of active 360° perception systems may raise privacy concerns regarding data collection in residential or public spaces, and there is a risk of misuse in unauthorized surveillance applications.

## Acknowledgments and Disclosure of Funding

The authors would like to thank Zekai Shao for his constructive suggestions. This work was partially supported by NSF grants 2615437 and 2115405.

## References


Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Graham R Barnes. Vestibulo-ocular function during co-ordinated head and eye movements to acquire visual targets. *The Journal of Physiology*, 287(1):127–147, 1979.

Albert E Bartz. Eye and head movements in peripheral vision: nature of compensatory eye movements. *Science*, 152(3729):1644–1645, 1966.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Matthew Chang, Théophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, et al. Goat: Go to any thing. In *Robotics: Science and Systems*, 2024.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *8th International Conference on Learning Representations*, 2020a.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020b.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12875–12884, 2020c.

Qixiang Chen, Cheng Zhang, Chi-Wing Fu, Jingwen Ye, and Jianfei Cai. Openview: Empowering mllms with out-of-view vqa. *arXiv preprint arXiv:2512.18563*, 2025a.

William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025b.

Zhe Chen, Bin Fu, Junjun He, Yu Qiao, Dahua Lin, Limin Wang, Xizhou Zhu, Weijie Su, Jifeng Dai, Yinan He, and et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025c.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. Dit360: High-fidelity panoramic image generation via hybrid training. *arXiv preprint arXiv:2510.11712*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and et al. Gemma 3 technical report, 2025.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Hai Huang, Yann LeCun, and Randall Balestriero. Llm-jepa: Large language models meet joint embedding predictive architectures. *arXiv preprint arXiv:2509.14252*, 2025a.

Yanjia Huang, Xianshun Jiang, Xiangbo Gao, Mingyang Wu, and Zhengzhong Tu. Vistav2: World imagination for indoor vision-and-language navigation. *arXiv preprint arXiv:2512.00041*, 2025b.

Yanjia Huang, Mingyang Wu, Renjie Li, and Zhengzhong Tu. Vista: Generative visual imagination for vision-and-language navigation. *arXiv preprint arXiv:2505.07868*, 2025c.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, tyler bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a bc-rl perception-action loop. In *Proceedings of The 9th Conference on Robot Learning*, pages 3647–3664. PMLR, 2025.

Team Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

Jeremy Lanman, Emilio Bizzi, and John Allum. The coordination of eye and head movement during smooth pursuit. *Brain research*, 153(1):39–53, 1978.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9098–9108, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352, 2022.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, pages I–253. IEEE, 2003.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*, 2025.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pages 1–16. IEEE, 2020.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Dorsa Sadigh, Arvind Neelakantan, Pranav Shyam, Noam Shazeer, Archit Sharma, Naman Goyal, Yossi Matias, Lawrence Chan, Zhe Chen, and et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. *arXiv preprint arXiv:2411.16044*, 2024.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Grayden JF Solman, Tom Foulsham, and Alan Kingstone. Eye and head movements are complementary in visual selection. *Royal Society Open Science*, 4(1):160569, 2017.

Niklas Stein, Tamara Watson, Markus Lappe, Maren Westendorf, and Szonya Durant. Eye and head movements in visual search in the extended field of view. *Scientific Reports*, 14(1):8907, 2024.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Kimi Team. Kimi-vl technical report, 2025.

Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2695–2702. IEEE, 2012.

Yiheng Xu, Peng Wang, Hang Zhang, Pengfei Wang, Shuai Bai, Shijie Wang, Junyang Lin, Tianbao Xie, Yuanzhi Zhu, Zhibo Yang, and et al. Qwen2.5-vl technical report, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Heyang Yu, Yinan Han, Xiangyu Zhang, Baiqiao Yin, Bowen Chang, Xiangyu Han, Xinhao Liu, Jing Zhang, Marco Pavone, Chen Feng, et al. Thinking in 360 {\deg}: Humanoid visual search in the wild. *arXiv preprint arXiv:2511.20351*, 2025a.

Justin Yu, Yide Shentu, Di Wu, Pieter Abbeel, Ken Goldberg, and Philipp Wu. Egomi: Learning active vision and whole-body manipulation from egocentric human demonstrations. *arXiv preprint arXiv:2511.00153*, 2025b.

Gregory Zelinsky, Wei Zhang, Bing Yu, Xin Chen, and Dimitris Samaras. The role of top-down and bottom-up processes in guiding eye movements during visual search. In *Advances in neural information processing systems*, 2005.

Jingdong Zhang, Weikai Chen, Yuan Liu, Jionghao Wang, Zhengming Yu, Zhuowen Shen, Bo Yang, Wenping Wang, and Xin Li. Spgen: Spherical projection as consistent and flexible representation for single image 3d shape generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025a.

Jingdong Zhang, Lingzhi Zhang, Qing Liu, Mang Tik Chiu, Connelly Barnes, Yizhou Wang, Haoran You, Xiaoyang Liu, Yuqian Zhou, Zhe Lin, et al. Uniser: A foundation model for unified soft effects removal. *arXiv preprint arXiv:2511.14183*, 2025b.

Jingdong Zhang, Xiaohang Zhan, Lingzhi Zhang, Yizhou Wang, Zhengming Yu, Jionghao Wang, Wenping Wang, and Xin Li. Mtpano: Multi-task panoramic scene understanding via label-free integration of dense prediction priors. *arXiv preprint arXiv:2602.05330*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025c.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278, 2024.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.

# A Technical Appendices and Supplementary Material

## A.1 More Details on Data Curation

**Dataset Distributions.** To complement the main text, we provide a comprehensive breakdown of our automated data curation pipeline, the cleaning process for supervised fine-tuning (SFT), and the statistical distribution of the curated dataset.

**Automated Pseudo-Label Pipeline.** As discussed in the main paper, our massive training corpus of $\sim$1.92M samples is constructed without any manual trajectory annotations. The automated pipeline consists of three stages:

1. **Scene Filter**: We utilize Qwen3-VL-235B (Yang et al., 2025) to classify and filter a mixed pool of unlabeled panoramas (including Sun360 (Xiao et al., 2012), Matterport3D (Chang et al., 2017), DiT360 generated (Feng et al., 2025)). Approximately 20k high-quality panoramas are retained across indoor, transit, and urban categories.
2. **Detection Labels**: The teacher model proposes $\sim$20 semantically salient targets per scene, localizing each to an $(x, y)$ position in equirectangular coordinates. After token-level perplexity filtering and spatial deduplication, we obtain 5-15 high-confidence semantic anchors per panorama.
3. **Trajectory Expansion**: For each labeled panorama, we select 3 salient targets and synthetically generate 24 narrow field-of-view (NFoV) trajectories (rendered at $960 \times 720$, FOV=$100^\circ$). Crucially, half of these (12 trajectories) are "target-avoiding" paths specifically designed so that all views remain distant from the target. This forces the model to imagine target locations from non-overlapping views. This deterministic expansion yields the final $\sim$1.92M single-step `[Observed]/[Imagined]` SFT samples.

**SFT Data Cleaning.** For the final task alignment stage, we aggressively filter the raw H*Bench (Yu et al., 2025a) SFT annotations to ensure high-quality reasoning. We enforce a strict structural requirement: a training sample is retained if and only if its response contains both an `[Observed]` and an `[Imagined]` block, since some of the cases lack raw panoramas, so the intermediate structured layouts are also absent and replies degenerate into a simple spatial guess (e.g., just outputting `suggest check(yaw, pitch)`), as they would teach the model to guess rather than explicitly model the space. This rigorous cleaning process retains 2,100 HOS samples (73.8% retention) and 2,148 HPS samples (81.0% retention), forming a highly curated alignment corpus of 4,248 samples.

Figure 7 illustrates the diversity and semantic richness of our curated dataset. As shown in Fig. 7(a) and (b), the dataset covers a highly diverse distribution of environments ranging from indoor domains (e.g., modern living rooms, luxury bedroom suites) to outdoor areas (e.g., city square plazas), accommodating a wide array of corresponding navigation instructions. Furthermore, Fig. 7(c-e) detail the semantic composition of the spatial layouts. Notably, "Architectural Elements" (26.2%) and "Doors & Entrances" (16.2%) dominate the `[Imagined]` categories (Fig. 7c). This inherently aligns with the cognitive requirements of our task: architectural features and doors serve as critical structural anchors for extrapolating the layout of unseen regions. Similarly, "Doors & Entrances" and "Objects" constitute large portions (29.1% and 21.5%) of the search targets (Fig. 7e), accurately reflecting the object/path-finding and navigational nature of practical ***HVS*** applications.

## A.2 More Details on Training and Inference

In this section, we provide extended implementation details regarding the two-stage training of the Imaginator, the pixel imagination baseline model, and the inference-time sampling strategy.

**Imaginator Optimization.** The Imaginator is initialized from Qwen3-VL-8B-Instruct (Yang et al., 2025) and optimized via full-parameter supervised fine-tuning (SFT). We utilize the AdamW optimizer with a learning rate of $1 \times 10^{-5}$, a cosine decay schedule, and a 3% linear warmup. To ensure training stability and memory efficiency, we employ DeepSpeed (Rasley et al., 2020) ZeRO (Rajbhandari et al., 2020) Stage 2 with gradient sharding and BF16 mixed precision.

- **Stage 1 (Pseudo-pretraining)**: The model is trained on the $\sim$1.92M synthetic sample corpus for 1 epoch. We use a global batch size of 192 (8 GPUs $\times$ 6 per-device $\times$ 4 gradient accumulation).

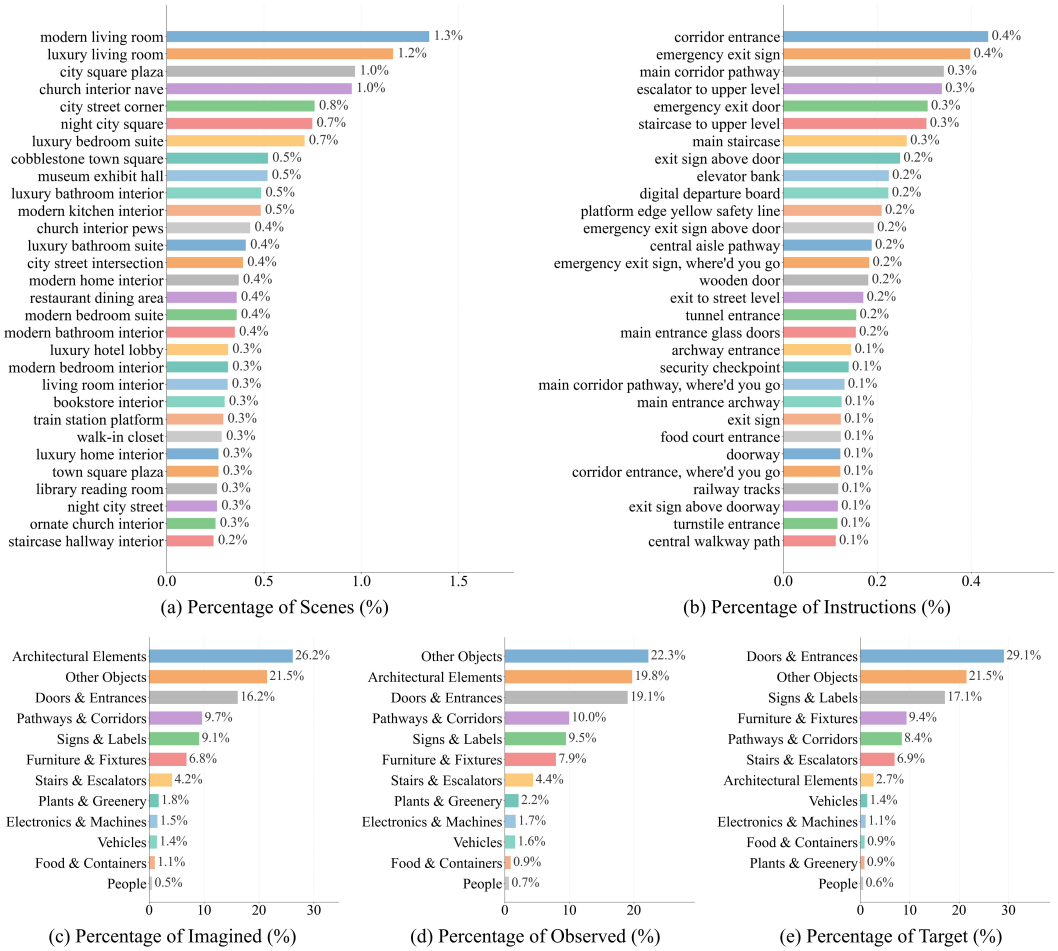



Figure 7: **Overview of the Curated Training Dataset.** (a) Fine-grained scene category distribution. (b) Distribution of target-search instructions. (c-e) Semantic category distributions for Imagined entities, Observed entities, and Search Targets, respectively.

- **Stage 2 (Clean-data SFT)**: The pseudo-pretrained checkpoints are fine-tuned on the curated H*Bench set (∼4,248 samples) for 2 epochs to compensate for the smaller corpus size. The global batch size is adjusted to 8.

Input images are resized to a maximum of 921,600 pixels (1280×720), and sequences are truncated at 8,192 tokens, which covers the 99th-percentile length of our structured layouts.

**Pixel Imagination Baseline.** With the rapid development of diffusion-based generative models (Lugmayr et al., 2022; Rombach et al., 2022; Zhang et al., 2023; Podell et al., 2023; Zhang et al., 2025b), effective generative priors are gained by pretraining on billions of images representing the world's knowledge. To evaluate alternative representation spaces in our ablation study, we implemented a pixel-level imagination baseline using a panoramic outpainting model. Built upon `FLUX.1-Fill-dev` (Labs et al., 2025), the model is adapted via LoRA fine-tuning (rank 64, $\alpha = 64$) and initialized from the DiT360 (Feng et al., 2025) generation checkpoint, which provides strong equirectangular geometry priors. At each step, this model out-paints the unobserved regions conditioned on the accumulated NFoV observations to generate a complete panorama. It is optimized using a masked flow-matching MSE loss to ensure gradients flow exclusively through the generated unknown regions:

$$\mathcal{L} = \frac{\sum[w(\sigma) \cdot \|v_{\text{pred}} - v_{\text{target}}\|^2 \cdot M_{\text{unknown}}]}{|M_{\text{unknown}}| \cdot C} \tag{5}$$

Table 3: **Full Results on H*Bench.** We report the Success Rate (%) across various Vision-Language Model families. '×' denotes the baseline Actor relying solely on its internal CoT. '✓' indicates the Actor plugged in with our Imaginator's spatial priors.

| Actor | | Imaginator | Humanoid Object Search | | | | Humanoid Path Search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Family | Params | | **Overall** | Easy | Medium | Hard | **Overall** | Easy | Medium | Hard | Extreme |
| *Open-Source Models* | | | | | | | | | | | |
| Qwen2.5-VL | 3B | × | 21.17 | 18.54 | 16.48 | 22.68 | 7.69 | 9.20 | 9.65 | 3.50 | 8.33 |
| | | ✓ | 49.54 | 59.19 | 51.70 | 45.62 | 25.94 | 34.80 | 22.15 | 24.77 | 15.74 |
| | 7B | × | 12.79 | 31.22 | 7.39 | 6.34 | 9.88 | 12.20 | 11.18 | 7.24 | 6.94 |
| | | ✓ | 57.42 | 70.73 | 51.70 | 52.95 | 28.38 | 35.20 | 24.34 | 29.44 | 18.98 |
| | 32B | × | 32.08 | 50.57 | 27.84 | 25.48 | 17.56 | 21.80 | 15.35 | 19.39 | 8.80 |
| | | ✓ | 41.12 | 66.34 | 36.36 | 32.01 | 21.12 | 25.60 | 19.74 | 21.26 | 13.43 |
| | 72B | × | 29.54 | 36.91 | 21.59 | 27.59 | 12.94 | 17.40 | 9.21 | 15.65 | 5.09 |
| | | ✓ | 62.25 | 70.08 | 60.80 | 59.42 | 38.19 | 45.00 | 36.18 | 39.25 | 24.54 |
| Qwen3-VL | 30B-A3B | × | 24.96 | 54.63 | 30.11 | 13.05 | 8.44 | 11.40 | 7.24 | 8.88 | 3.24 |
| | | ✓ | 56.67 | 74.15 | 55.68 | 50.09 | 30.94 | 41.80 | 26.32 | 30.14 | 17.13 |
| | 32B | × | 34.38 | 66.83 | 36.36 | 21.75 | 23.19 | 26.00 | 24.78 | 22.43 | 14.81 |
| | | ✓ | 57.46 | 73.98 | 51.70 | 51.77 | 31.31 | 37.60 | 30.04 | 30.14 | 21.76 |
| | 235B | × | 23.62 | 57.24 | 26.70 | 10.44 | 21.44 | 26.80 | 19.96 | 23.36 | 8.33 |
| | | ✓ | 65.04 | 81.46 | 60.23 | 59.29 | 38.25 | 46.60 | 35.31 | 37.15 | 27.31 |
| Gemma-3 | 4B | × | 18.17 | 41.79 | 25.57 | 8.33 | 14.50 | 17.80 | 16.67 | 10.98 | 9.26 |
| | | ✓ | 57.75 | 72.03 | 60.23 | 52.02 | 35.06 | 43.80 | 27.41 | 36.45 | 28.24 |
| | 12B | × | 22.88 | 54.96 | 32.39 | 9.57 | 15.31 | 16.80 | 16.23 | 14.49 | 11.57 |
| | | ✓ | 54.62 | 63.90 | 48.86 | 51.71 | 33.56 | 43.00 | 30.04 | 32.24 | 21.76 |
| Kimi-VL | A3B | × | 4.33 | 12.36 | 0.57 | 1.68 | 5.62 | 7.20 | 6.80 | 2.34 | 6.02 |
| | | ✓ | 32.46 | 40.00 | 27.84 | 30.08 | 21.00 | 28.20 | 16.89 | 21.96 | 11.11 |
| InternVL3.5 | 4B | × | 10.00 | 14.80 | 13.07 | 7.83 | 8.62 | 12.80 | 9.43 | 6.31 | 1.85 |
| | | ✓ | 62.42 | 69.59 | 62.50 | 59.66 | 34.75 | 44.80 | 28.95 | 35.51 | 22.22 |
| | 8B | × | 11.04 | 23.74 | 6.25 | 6.71 | 7.06 | 9.60 | 6.80 | 6.31 | 3.24 |
| | | ✓ | 66.33 | 76.10 | 68.18 | 62.40 | 38.38 | 50.60 | 28.73 | 39.95 | 27.31 |
| *Proprietary Models* | | | | | | | | | | | |
| Gemini-2.5 | Flash | × | 43.70 | 79.50 | 59.70 | 28.30 | 27.70 | 35.80 | 22.60 | 31.30 | 12.50 |
| | | ✓ | 63.20 | 86.30 | 69.30 | 53.70 | 37.70 | 50.20 | 30.30 | 42.30 | 15.30 |
| | Pro | × | 55.00 | 88.00 | 62.50 | 41.60 | 37.90 | 43.80 | 36.40 | 40.70 | 21.80 |
| | | ✓ | 66.10 | 89.60 | 68.80 | 56.80 | 43.10 | 51.00 | 39.90 | 48.80 | 20.40 |
| *Finetuned Models* | | | | | | | | | | | |
| HVS | 3B | × | 48.04 | 61.79 | 26.14 | 45.18 | 24.12 | 36.60 | 17.76 | 22.66 | 11.57 |
| | | ✓ | 62.75 | 75.77 | 64.20 | 57.61 | 39.38 | 49.40 | 32.89 | 41.12 | 26.39 |

where $M_{\text{unknown}}$ represents the mask of the unobserved regions. Training is conducted for 10,000 steps with an AdamW optimizer ($lr = 2 \times 10^{-5}$) on a multi-source panorama corpus from Zhang et al. (2026).

In the inference cycle, once the agent receives a partial panoramic canvas built from current NFoV views (containing both explored and masked unknown regions), the pixel imagination model executes a 30-step diffusion process via the finetuned DiT360 to outpaint the unobserved areas, which are then blended with the original known regions to ensure visual consistency. The HVS agent subsequently perceives both the original partial observation and the completed reference panorama to guide its next action.

**Inference-time Sampling Strategy.** During the Humanoid Visual Search process, the Imaginator captures spatial uncertainty through a temperature-decay sampling schedule. At each step $t$, we sample $N$ spatial hypotheses where the temperature $T_t$ follows:

$$T_t = T_1 \cdot \gamma^{(t-1)} \tag{6}$$

We set the initial temperature $T_1 = 0.7$ and the decay factor $\gamma = 0.85$. This decay schedule ensures that the agent explores diverse spatial possibilities in the early stages of search (high $T$) and transitions to focused exploitation (low $T$) as visual evidence accumulates. All suggestions are filtered via top-$k$ logit sampling ($k = 50$) to maintain semantic coherence. We use this temperature sampling strategy in the first two steps ($T_1 = 0.7, T_2 = 0.595$) and provide only the greedy prediction afterwards, since uncertain search mostly happens at the beginning stage, with the observation gaining, the uncertainty quickly drops and converges to the target.

**Imaginator Prompt and Inference Settings.** The Imaginator's input and output follow the structure below (the ***HPS*** variant replaces the object-search wording with navigation wording); its suggestions are converted (Sec. 3.4) and appended to the Actor's turn. The Actor uses the prompts and greedy decoding of Yu et al. (2025a). Each episode allows at most 10 steps and starts at pitch 0° from one of four initial headings, and the Imaginator generates at most 4,096 new tokens per call.

```
[System] You are a spatial reasoning assistant for a visual search agent. ... Each image is annotated with
      its viewing direction as (yaw,pitch).
  1. In <think>, list landmarks under [Observed] and predict objects in unseen areas under [Imagined], each
       with absolute (yaw,pitch).
  2. Reason about where the target is most likely located.
  3. In <answer>, output suggest check(yaw,pitch).
[User] [View 1] viewing direction: (yaw_1,pitch_1) <image> ... [View n] viewing direction: (yaw_n,pitch_n)
      <image>
        Human Instruction: {instruction} Decide your next action.
[Output] <think>[Observed] - picture frames: (274,16) ... [Imagined] - window with outdoor view: (166,9)
     ... (reasoning)</think>
        <answer>suggest check(yaw,pitch)</answer>
[Injected into the Actor's turn]
        [Spatial Imagination Suggestions]: Option1: suggest rotate(dyaw,dpitch) | submit(yaw,pitch)
             Option2: ... Option3: ...
```

### A.3 Train/Test Separation

**Split construction.** Yu et al. (2025a) annotate approximately 3,000 task instances. 1,000 instances (600 ***HOS*** and 400 ***HPS***) are reserved as the H*Bench evaluation set; from the remainder, 250 ***HOS*** and 250 ***HPS*** instances are randomly sampled to form the SFT set, and all leftover instances are used only for RL training. We adopt this partition unchanged. Our stage-2 alignment data is drawn from the SFT portion only, and no evaluation instance enters any training stage. The 1,000 evaluation instances rest on 857 distinct panoramas (457 ***HOS***, 400 ***HPS***), since several instances may pose different targets within one scene; each instance is expanded into four episodes by its initial heading, which yields the 2,400 ***HOS*** and 1,600 ***HPS*** episodes reported throughout.

**Provenance.** The pseudo-pretraining panoramas and the evaluation panoramas come from different collections. The teacher model (Qwen3-VL-235B) generates pseudo layout labels only for the pseudo-pretraining panoramas and is never run on any evaluation panorama.

- **SUN360** (Xiao et al., 2012) (pseudo-pretraining): high-resolution still panoramas downloaded from the Internet and grouped into 80 place categories.
- **Matterport3D** (Chang et al., 2017) (pseudo-pretraining): indoor panoramas of 90 building-scale properties, scanned with an RGB-D camera.
- **DiT360** (Feng et al., 2025) (pseudo-pretraining): panoramas synthesized by DiT360, a specialist panorama generative model tuned from FLUX, with no real scene behind them.
- **H*Bench** (Yu et al., 2025a) (evaluation, and the source of our stage-2 data): frames of 360° panoramic videos, self-collected in metropolitan areas and taken from open platforms across 12 countries.

**Deduplication audit.** Each of the 857 evaluation panoramas was compared with every panorama of the pseudo-pretraining corpus, without pre-filtering. Because equirectangular panoramas are cyclic in yaw, a yaw-rotated copy of the same place counts as a duplicate; every pair was therefore scored at the best of 36 yaw rolls (10° steps). *MAE*: both panoramas are converted to 8-bit greyscale and resized to a common resolution, and we take the mean absolute pixel difference, minimized over the rolls; a value below 2.0 (0.8% of the range) indicates the same photograph. *Pearson* $r$: on the same aligned pair, $r = \mathrm{cov}(A, B)/(\sigma_A \sigma_B)$, which is invariant to any global brightness or contrast change and therefore detects the same viewpoint under different exposure or colour grading; the threshold is $r \geq 0.90$. Table 4 reports the distributions. No pair meets either criterion; the closest pair has an MAE of 31.25. The same check on the stage-2 alignment data also returns no match.

### A.4 Pseudo-Label Quality, Bias and Noise

We stratified-sampled 250 panoramas from the curated corpus, covering 743 search targets, and human experts judged every target label against its panorama: 85.20% are correct. Two automatic judges, GroundingDINO (Liu et al., 2024) and GPT-5.6, applied to the same 743 targets give 87.17%

Table 4: **Deduplication statistics** between the 857 H*Bench evaluation panoramas and the pseudo-pretraining corpus. Top: nearest-neighbour Pearson $r$ per evaluation panorama (median 0.517, mean 0.525). Bottom: MAE over all pairs (median 52.10, mean 54.05).

| Pearson $r$ | 0–0.3 | 0.3–0.6 | 0.6–0.9 | $\geq$0.9 (match) | |
|---|---|---|---|---|---|
| Evaluation panoramas | 0.2% | 80.4% | 19.4% | 0.0% | |
| MAE | $<$2.0 (match) | 2.0–25 | 25–50 | 50–75 | $>$75 |
| Pairs | 0.0% | 0.0% | 45.5% | 45.0% | 9.5% |

on average, and 88.79% on the whole corpus. Label deviations are small: the median absolute error is 5.0° in yaw and 9.6° in pitch.

**Bias and noise.** Fig. 3 characterizes corpus-level bias in scene categories, item categories and salience. A spatial-position bias also exists: object categories occur with very unequal frequency in the world, and 360° panoramas are captured under strong conventions on camera placement and height, so certain directions and elevations are over-represented. The labels are generated automatically and contain noise, and the images carry their own capture or synthesis noise.

### A.5 Controls for the Source of the Gain

All controls follow the protocol of Table 2: they are evaluated on the 2,400 ***HOS*** episodes with HVS-3B as the Actor, and their results are given in Table 5. The first five controls keep the suggestion block in the Actor prompt at the same position and in the same format, and change only its content; the last two change how the models are used.

- **Generic hint**: the block contains task-neutral text with the same number of tokens as a real suggestion.
- **Uniform directions**: the block suggests evenly spaced yaw angles, $t \cdot 360°/N$ at step $t$, so that the suggestions sweep the full 360° in $N$ steps.
- **Scene-category prior**: the block suggests directions drawn from the training-data statistics of the current scene's category, swept over the episode in the same way as the uniform directions.
- **Retrieval prior**: the block suggests the direction in which the target lies in the most similar training view, applied relative to the current view.
- **Teacher suggestions**: the Qwen3-VL-235B teacher produces the suggestions under the Imaginator's prompt.
- **Ensemble**: the same 3B+8B budget is spent on a second Actor instead of the Imaginator. An 8B Actor trained on the same CoT trajectories as HVS-3B runs independently on every episode, and the answers of the two Actors are combined per episode by confidence.
- **One model imagines and acts**: the `[Observed]/[Imagined]` layout is placed inside the Actor's own reasoning chain. This setting reaches 47.04% on ***HOS*** and 24.08% on ***HPS***.

The direct target-coordinate predictor is the *w.o. layout* row of Table 2: the Imaginator is trained on the same data to output only the target direction, without the `[Observed]/[Imagined]` layout.

Table 5: **Controls** on ***HOS*** with HVS-3B as the Actor. The middle block replaces the Imaginator's suggestions with other content or, for the ensemble, with a second Actor.

| **Imagination Guidance** | **Overall** | Easy | Medium | Hard |
|---|---|---|---|---|
| Actor | 48.04 | 61.79 | 26.14 | 45.18 |
| Actor + Generic Hint | 45.54 | 52.20 | 23.86 | 45.37 |
| Actor + Uniform Directions | 32.79 | 38.21 | 17.61 | 32.38 |
| Actor + Scene-Category Prior | 26.17 | 30.08 | 14.20 | 25.98 |
| Actor + Retrieval Prior | 23.04 | 33.17 | 19.89 | 19.52 |
| Actor + Teacher Suggestions | 40.17 | 47.64 | 28.98 | 38.53 |
| Actor + HVS-8B (Ensemble) | 50.12 | 61.14 | 34.09 | 47.67 |
| Actor + Imaginator (*full*) | **62.75** | **75.77** | **64.20** | **57.61** |

### A.6 Direct Evaluation of the Imaginator

**Protocol.** For each H*Bench episode, the Imaginator receives the initial view and produces one greedy prediction (top-1) and $K$=8 temperature-sampled predictions. A prediction hits when it falls

within the task tolerance of the annotated target. Best-of-$K$ hit rates average over all subsets of size $K$ of the eight samples rather than taking the first $K$ (Fig. 8a). The deployed $K$=3 already recovers $+12.67$ points on ***HOS*** and $+16.16$ on ***HPS*** over greedy decoding. The median angular error falls from 45.9° (top-1) to 21.6° (best of $K$=8) on ***HOS*** and from 60.6° to 14.7° on ***HPS***. By target visibility at the first step, the ***HOS*** top-1 / $K$=8 hit rates are 82.84 / 86.19 when the target is fully visible, 54.35 / 70.71 when partially visible and 22.05 / 56.48 when fully out of view.

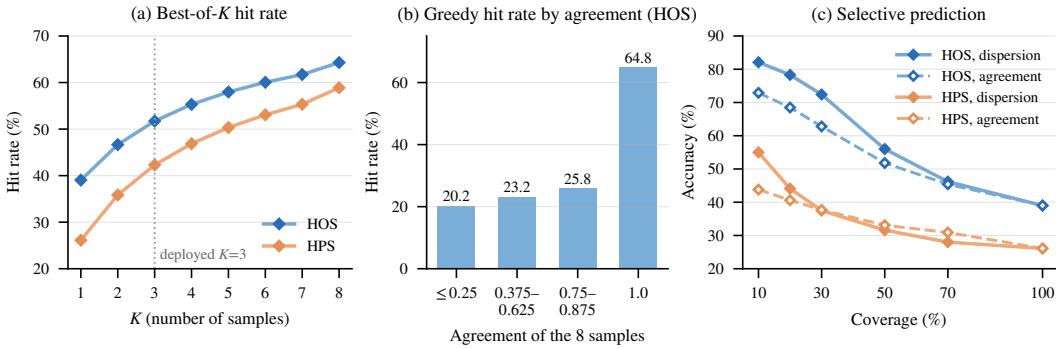



Figure 8: **Direct evaluation of the Imaginator** from the initial view. (a) Best-of-$K$ hit rate under the task tolerance. (b) ***HOS*** hit rate of the greedy prediction, grouped by the agreement of the eight samples. (c) Accuracy of the greedy prediction on the most confident fraction of episodes (coverage), with episodes ranked by dispersion (solid) or agreement (dashed).

**Entity localization.** For every entity in the predicted layout we render the view at its claimed direction, and GPT-5.6 and GroundingDINO (Liu et al., 2024) judge whether the named object is present; we report their mean. The evaluator sees the full panorama in both cases, so `[Observed]` (84.8%) and `[Imagined]` (45.3%) differ only in whether the agent had seen that direction.

**Certainty of the samples.** The eight samples also show how certain the Imaginator is about a prediction. We summarize them in two ways: *agreement*, the fraction of the eight samples that fall within the ***HOS*** yaw tolerance of the greedy prediction, and *dispersion*, the mean pairwise angular distance among the eight samples. High agreement and low dispersion mean that the samples point in the same direction. When all eight samples agree, the greedy prediction is correct 64.8% of the time, against 20–26% in the lower agreement bins (Fig. 8b), and the dispersion is 22.7° for correct predictions against 44.6° for incorrect ones. Used as scores for predicting whether the greedy prediction is correct, agreement and dispersion reach an AUROC of 0.727 and 0.745 on ***HOS*** and 0.653 and 0.608 on ***HPS***. Dispersion also grows as the target leaves the initial view: 11.8° when it is fully visible, 22.0° when partially visible, and 45.3° when fully out of view.

**Selective prediction.** This certainty tells when a suggestion can be trusted. We sort the episodes from most to least confident, by dispersion or by agreement, and measure the accuracy of the greedy prediction on the most confident fraction of them, called the coverage (Fig. 8c). Accuracy is highest at low coverage and decreases as less confident episodes are added: on ***HOS***, the 10% most confident episodes by dispersion are correct 82.1% of the time, against 39.0% over all episodes.

**Failure modes.** We call an incorrect suggestion plausible when it points within 15° of an entity the Imaginator itself listed. Two such modes account for 14.2% of all episodes: the correct object named but misplaced (7.0%), or a semantically adjacent object substituted (7.2%). A further 16.1% fail with no listed entity nearby, and 7.0% fail despite a correct suggestion.

### A.7 Held-out Benchmark

**Sources.** The fifteen 360° panoramas were assembled for this evaluation and are disjoint from every source used in training and in H*Bench (Table 7). *Real-captured*: eight Google Street View panoramas, selected at random subject only to covering distinct environment types (dense urban architecture, road traffic, institutional space and natural outdoor terrain), then cleaned: we discarded panoramas with stitching failures, heavy motion blur or visible capture apparatus, and verified that

Table 6: **Results on the held-out benchmark.** Success rate (%) and average number of steps.

| Actor | **Overall** | Easy | Hard | Steps |
|---|---|---|---|---|
| HVS-3B (Yu et al., 2025a) | 47.87 | 44.44 | 51.02 | 3.98 |
| + Imaginator | **58.51** | **54.44** | **62.24** | **3.75** |
| Qwen2.5-VL-3B | 17.55 | 11.11 | 23.47 | 6.17 |
| + Imaginator | **46.28** | **36.67** | **55.10** | **3.57** |

Table 7: **Composition of the held-out benchmark.**

| Source | Scenes | Targets | Episodes |
|---|---|---|---|
| Real (Google Street View) | 8 | 20 | 80 |
| Synthesized (GPT-Image) | 4 | 15 | 60 |
| Synthesized (Gemini 3.1-Pro) | 3 | 12 | 48 |
| Total | 15 | 47 | 188 |

each remaining one is a complete equirectangular sphere. *Synthesized*: seven panoramas produced by GPT-Image and Gemini 3.1-Pro from prompts covering environment types under-represented in the captured set, including indoor spaces and transport hubs. A person inspected each one and kept it only if the scene is physically plausible, the wrap-around seam is continuous, and no object appears in an implausible configuration; panoramas failing any check were discarded rather than repaired.

**Tasks and annotation.** The 47 targets follow the task definitions of H*Bench. An object-search instruction names the target directly (*Find the tram.*) or by the text it carries (*Find 'Gate B56'.*), and a path-search instruction asks in which direction to move to reach a goal (*You want to buy a coffee, where to move?*). All instructions and targets were annotated by human experts, and each target is marked by its yaw and pitch interval (e.g., yaw $[227.5°, 287.5°]$, pitch $[-20.4°, 19.6°]$). Each target is evaluated from initial headings of 0°, 90°, 180° and 270°, giving 188 episodes, and scored under the H*Bench tolerances. An episode is Easy when the target interval fully overlaps the initial field of view and Hard when it does not overlap it, which gives 90 Easy and 98 Hard episodes.

**Statistical reliability.** The held-out benchmark is small, so we test whether the gain of Qwen2.5-VL-3B from the Imaginator could arise by chance (Table 8). The 95% confidence interval of the gain comes from a paired bootstrap over episodes (10,000 resamples), and the $p$-value from an exact two-sided McNemar test on the paired episode outcomes; H*Bench ***HOS*** is listed for reference. Because the four initial headings of a target, and the targets of one panorama, are not independent, we repeat the bootstrap by resampling whole targets or whole panoramas instead of episodes, which gives intervals of $[+20.74, +36.70]$ and $[+20.56, +36.41]$. Per panorama, 13 of the 15 improve, 1 is unchanged and 1 degrades.

Table 8: **Significance of the gain** of Qwen2.5-VL-3B from the Imaginator. Success rate (%) without and with the Imaginator, the gain, its 95% bootstrap confidence interval and the McNemar $p$-value.

| Benchmark | w/o Imaginator | w/ Imaginator | Gain | 95% CI of gain | McNemar $p$ |
|---|---|---|---|---|---|
| Held-out | 17.55 | 46.28 | +28.72 | $[+20.21, +36.70]$ | $6.4 \times 10^{-10}$ |
| H*Bench ***HOS*** | 21.17 | 49.54 | +28.37 | $[+22.66, +33.01]$ | $4.6 \times 10^{-23}$ |

### A.8 More Quantitative Comparisons

To provide a comprehensive evaluation of our proposed framework, Table 3 details the full quantitative results on the H*Bench (Yu et al., 2025a) dataset, expanding upon the main paper by including a wider array of model families (such as InternVL3.5 (Chen et al., 2025c) and Qwen3-VL-30B-A3B) and breaking down the success rates across all difficulty tiers (Easy, Medium, Hard, and Extreme).

**Universal Boosting Across Model Scales.** The most prominent takeaway from Table 3 is the universal applicability of the Imaginator's spatial priors. As a plug-and-play context provider, it drastically enhances the search capabilities of open-source models regardless of their parameter size. For instance, the InternVL3.5 family exhibits staggering improvements: the 8B model jumps from a mere 11.04% to 66.33% on Humanoid Object Search (HOS), and from 7.06% to 38.38% on Humanoid Path Search (HPS). Similar transformative gains are observed across the Qwen2.5-VL, Qwen3-VL, and Gemma-3 families. This proves that explicit 3D spatial modeling effectively bridges the capability gap for smaller or less embodied-aligned models.

**Efficacy on State-of-the-Art Proprietary Models.** While cutting-edge proprietary models like Gemini-2.5 Pro exhibit strong baseline performance relying on their massive internal world knowledge

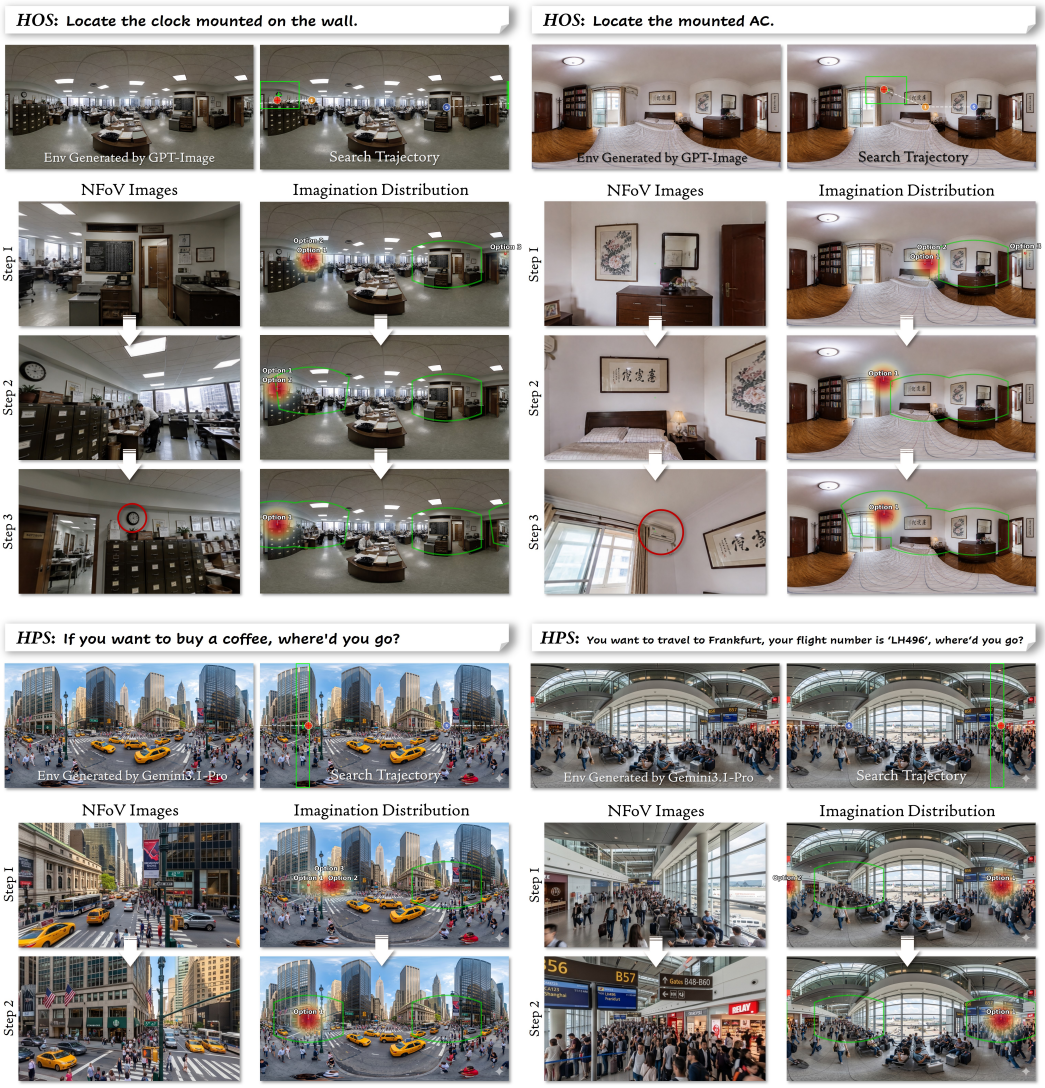



Figure 9: **Qualitative search trajectories in synthetic panoramas.** Evaluated in complex indoor and outdoor environments generated by GPT-Image and Gemini 3.1-Pro. The step-by-step imagination distributions (heatmaps) demonstrate how the framework progressively refines its spatial hypotheses as new visual evidence is gathered, effectively guiding both HOS and HPS tasks.

(e.g., 55.00% on HOS), they are not immune to the cognitive burdens of monolithic 360° reasoning. By integrating our Imaginator, Gemini-2.5 Pro reaches 66.10% on HOS and 43.10% on HPS. Furthermore, Gemini-2.5 Flash sees a massive +19.5% absolute improvement on HOS (43.70% $\rightarrow$ 63.20%), highlighting that explicit spatial hypotheses are crucial even for highly advanced multimodal systems.

**Surpassing Task-Specific RL Models.** Our framework's superiority is further solidified when evaluated alongside HVS-3B, a model specifically fine-tuned with Reinforcement Learning (RL) for this exact task. A standalone HVS-3B achieves 48.04% on HOS. However, when deployed as the Actor within our decoupled pipeline, its performance surges to 62.75%. This indicates that forcing a single model to simultaneously imagine layouts and plan actions imposes a fundamental ceiling—one that can be broken by explicitly decoupling the probabilistic spatial priors.

**Robustness in Extreme Environments.** The difficulty breakdown in Table 3 reveals that the baseline Actor models degrade rapidly as the environments become more complex (Hard and Extreme

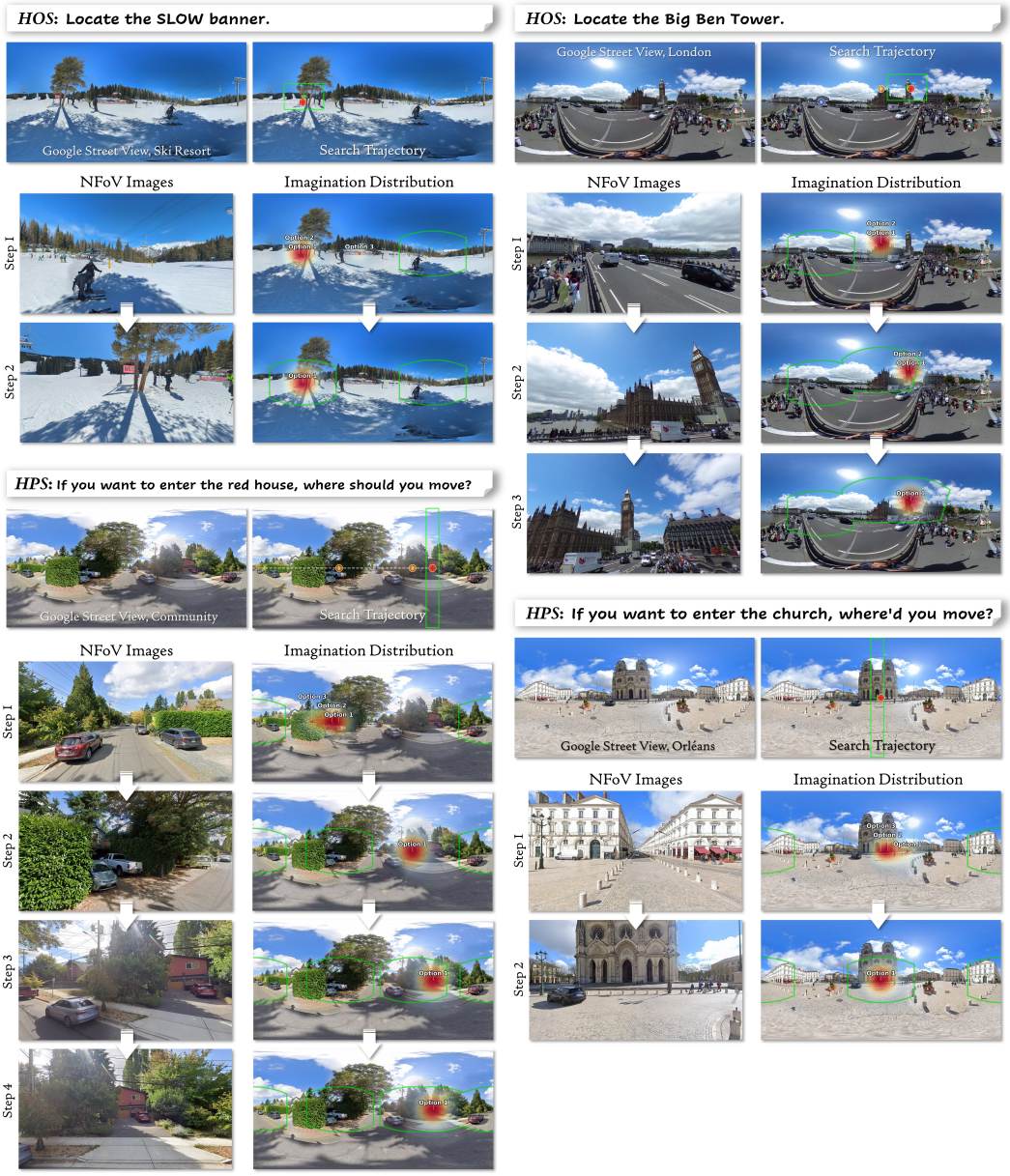



Figure 10: **Qualitative search trajectories in real-world panoramas.** Evaluated in diverse, in-the-wild Google Street View environments. The decoupled pipeline demonstrates strong zero-shot robustness in unstructured outdoor scenes (e.g., ski resorts, bustling city streets, and suburban neighborhoods), efficiently narrowing down target locations.

subsets). Without the Imaginator, most open-source models fall below 10% success rate in Extreme HPS scenarios, often succumbing to blind 360° sweeping. In contrast, the joint pipeline maintains remarkable robustness, frequently tripling or quadrupling the success rates in these highly occluded and challenging conditions.

### A.9 More Qualitative Visualizations

To further demonstrate the robustness and generalizability of our decoupled framework, we provide additional step-by-step search trajectories in highly complex, out-of-distribution (OOD) environments across both Humanoid Object Search (HOS) and Humanoid Path Search (HPS) tasks.

**Search in Synthetic Environments.** Figure 9 showcases the agent's performance in panoramic environments synthesized by state-of-the-art generative models, including GPT-Image and Gemini 3.1-Pro. These scenes feature densely packed indoor rooms (e.g., offices with cluttered desks) and busy outdoor hubs (e.g., city intersections, airport terminals). Despite the unconventional layouts or minor artifacts inherent to AI-generated images, the Imaginator successfully captures the structural semantics of the unobserved regions. As the agent accumulates Narrow Field-of-View (NFoV) observations step-by-step, the imagination distribution (shown as predictive heatmaps) progressively narrows, eliminating spatial uncertainty and efficiently guiding the Actor toward the targets (e.g., locating a wall-mounted clock or the pathway to flight LH496).

**Search in Real-World Environments.** Figure 10 illustrates search trajectories in unstructured, real-world panoramas sampled directly from Google Street View. These environments present significant navigational challenges due to their vast scales, dynamic elements (e.g., pedestrians, skiers), and extreme topological diversity, ranging from snowy mountain resorts to historic European plazas. Nevertheless, the Imaginator consistently generates logical spatial priors. For instance, when tasked with locating the "Big Ben Tower" or determining the path to "enter the church," the model effectively leverages early architectural cues to hypothesize the target's location. This allows the agent to lock onto the correct trajectory within just 2 to 4 steps, completely avoiding redundant 360° sweeping. The model also handles partial occlusion: in *Locate the SLOW banner*, the banner is seen from behind and partly hidden by a tree trunk, yet the model places its hypothesis in the correct direction.

### A.10 Social Impacts

The proposed framework for Humanoid Visual Search (***HVS***) has significant potential for positive social impact, particularly in enhancing the autonomy of service robots and search-and-rescue agents in complex, unstructured environments. By improving search efficiency and robustness, our approach can facilitate more reliable assistance in household chores, elderly care, and emergency response, ultimately improving quality of life and operational safety.

However, we also acknowledge potential ethical considerations. The deployment of active 360° perception systems may raise privacy concerns regarding data collection in residential or public spaces. Furthermore, like any advanced vision-action system, there is a risk of misuse in unauthorized surveillance applications. We encourage the research community to prioritize privacy-preserving mechanisms and adhere to strict ethical guidelines in future embodied AI deployments.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state our claims regarding the decoupled Imaginator–Actor architecture and data scalability in the Abstract and Introduction, and support them with evaluations on H*Bench and a held-out benchmark, controls and direct evaluations in Section 4.1–4.4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our open-loop execution paradigm and the absence of a Critic module in Section 5 (Conclusion and Limitations), together with privacy and surveillance risks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper proposes a novel empirical framework and architecture for Humanoid Visual Search; it focuses on empirical evaluations and does not introduce new theoretical bounds or formal mathematical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 and Appendices A.1, A.2 and A.3 give the data pipeline, the train/test split, training hyperparameters, prompts, decoding configuration and per-model inference settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code: `https://github.com/Evergreen0929/Imagining360`; Imaginator weights: `https://huggingface.co/jdzhang0929/Imaginator`; stage-2 alignment data: `https://huggingface.co/datasets/jdzhang0929/Imagine-in-360-Dataset`. The repository README gives instructions to reproduce training and evaluation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details are in Section 4.1 and Appendix A.2; the exact train/test split and its construction are in Section 4.1 (Train/Test Separation) and Appendix A.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report paired bootstrap 95% confidence intervals (10,000 resamples) and exact McNemar tests for the gain of Qwen2.5-VL-3B from the Imaginator on H*Bench ***HOS*** and on the held-out benchmark, together with target- and panorama-clustered intervals (Appendix A.7).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify the hardware infrastructure used for our experiments (8 NVIDIA H200/A100 GPUs) in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research relies on existing public datasets or synthesized panoramas and strictly adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 discusses privacy and surveillance risks of active 360° perception; Appendix A.10 discusses positive impacts and mitigation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our framework is designed for embodied visual search within simulated or pre-captured panoramic environments and poses no direct high risk for malicious misuse requiring specific safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the datasets (SUN360, Matterport3D, DiT360, H*Bench) and models (Qwen2.5-VL, Qwen3-VL, Gemma-3, Kimi-VL, InternVL3.5, Gemini 2.5, FLUX.1-Fill-dev, GroundingDINO) we use and follow their terms of use; Google Street View imagery and panoramas generated with GPT-Image and Gemini 3.1-Pro are used only for non-commercial research evaluation.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released Imaginator weights and stage-2 alignment data are documented in the repository README and dataset card; the data pipeline is described in Section 4.1 and Appendix A.1.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing or research with human subjects; the pseudo-label evaluation and the held-out benchmark were annotated by human experts (Appendices A.4 and A.7).

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable as our research does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Qwen3-VL-235B is the teacher for pseudo-labels (Section 4.1) and a control (Section 4.3); GPT-5.6 together with GroundingDINO serves as an automatic judge of label and layout accuracy (Appendices A.4 and A.6); GPT-Image and Gemini 3.1-Pro synthesize part of the held-out panoramas (Appendix A.7).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.