NEURIPS 2026

Beyond Thinking:
Imagining in 360°
for Humanoid Visual Search

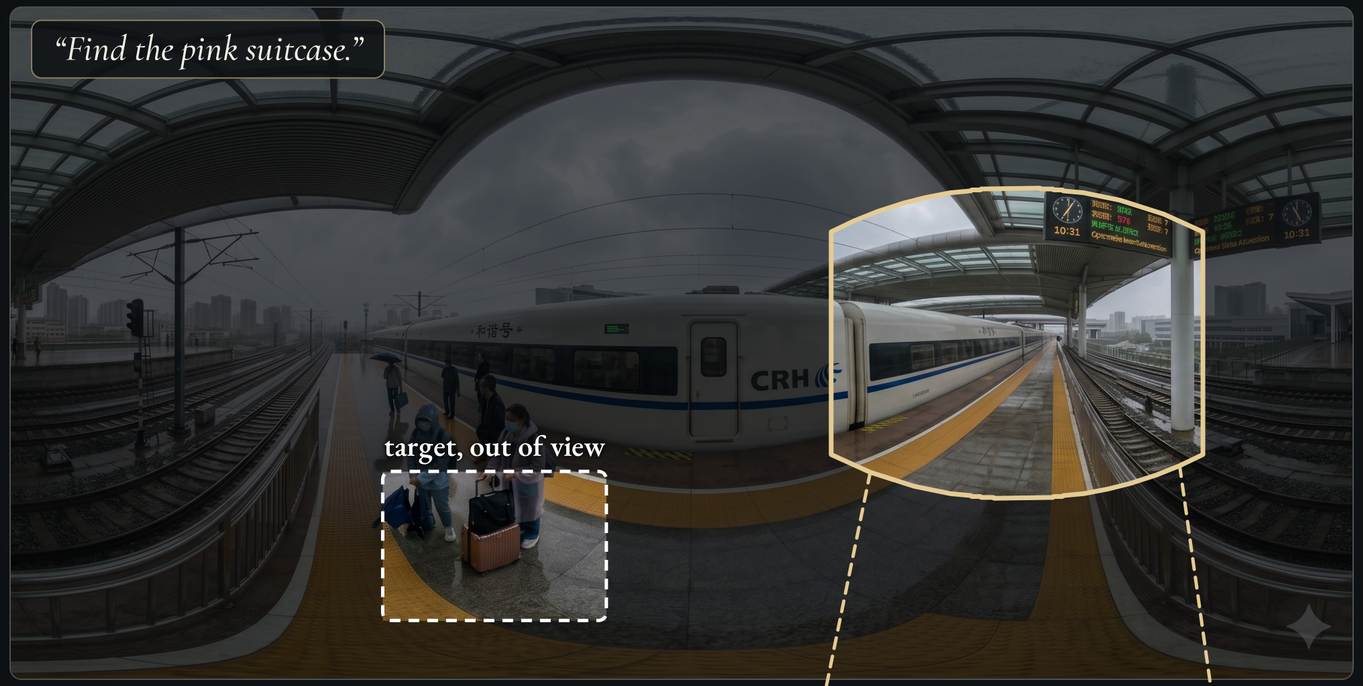
Jingdong Zhang[1] · Yizhou Wang[2] · Zhengzhong Tu[1]
Xin Li[1] · Wenping Wang[1] · Xiaohang Zhan[2]

[1] TEXAS A&M UNIVERSITY [2] Adobe



THE TASK

Humanoid Visual Search





ACTION SPACE

`Rotate(Δφ, Δγ)`	turn the head
`Submit(φ, γ)`	answer with this view

1. **A 360° scene, a 100° view**
   The agent stands inside a 360° scene, yet sees only a narrow field of view.
2. **It turns its head to search**
   At each step it must decide where to look next, within 10 steps.
3. **Two tasks on H*Bench**
   HOS: locate a specific object. **HPS**: find the direction of a navigable path.
4. **Success**
   The submitted view falls within the angular tolerance of the target.

THE PROBLEM

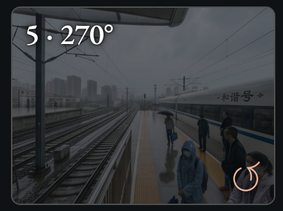
Thinking alone goes in circles











…

Step 1 <think> I notice a train platform with a train on the left and tracks on the right. There is no visible pink suitcase in this view. The platform is wide, and the train is partially visible, suggesting more of the platform might be behind me. Since the goal is to find a pink suitcase, and it's not visible here, rotating 180 degrees will allow me to explore the opposite direction, potentially revealing new areas where the suitcase might be located. Therefore, the next action is rotate(180,0). </think> **rotate(180,0)**

Step 2 <think> I notice a train platform with people walking and a train on the right. There is no visible pink suitcase in this view. The platform is wide, and the train is partially visible, suggesting more of the platform might be behind me. Since the goal is to find a pink suitcase, and it's not visible here, rotating 90 degrees will allow me to explore a new area of the platform, potentially revealing the suitcase. Therefore, the next action is rotate(90,0). </think> **rotate(90,0)**

… steps 3–9: the same reasoning, ~80 words each …

Step 10 <think> … There is no visible pink suitcase in this view. … </think>
rotate(90,0) × not found

HVS-3B alone, real log

Prior work asks one model to carry a long chain of thought, from turn to turn.

1. One model must both **understand the space** and **plan actions**.
2. It has no explicit model of the **unseen** layout.
3. Training needs costly **multi-step CoT trajectories**.

KEY IDEA

Imagine first, then act

Before we even turn our heads, we imagine where it should be.

VIEWS SO FAR



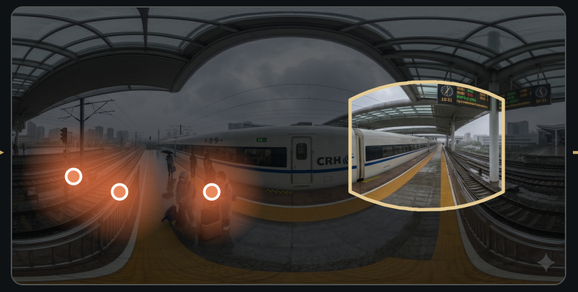
"Find the pink suitcase."

Imaginator

imagines the unseen

ONE STATELESS PASS

SAMPLED HYPOTHESES



RANKED SUGGESTIONS

```
rotate(160,-30)
rotate(130,-20)
rotate(-140,-30)
```

Actor

decides where to look

OFF-THE-SHELF, UNCHANGED

```
Rotate(Δφ, Δγ)
Submit(φ, γ)
```

NEW VIEW

1 Decoupled HVS paradigm

Visual search is decoupled into intuitive spatial imagination and action planning.

2 Probabilistic Imaginator

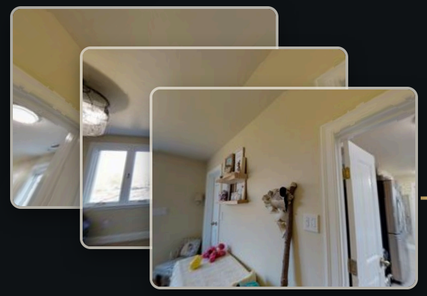
It learns intuitive spatial priors and guides the Actor with probabilistic estimations.

3 Scalable data engine

A fully automated pipeline scales training to 1.92M samples without manual trajectory labels.

THE IMAGINATOR

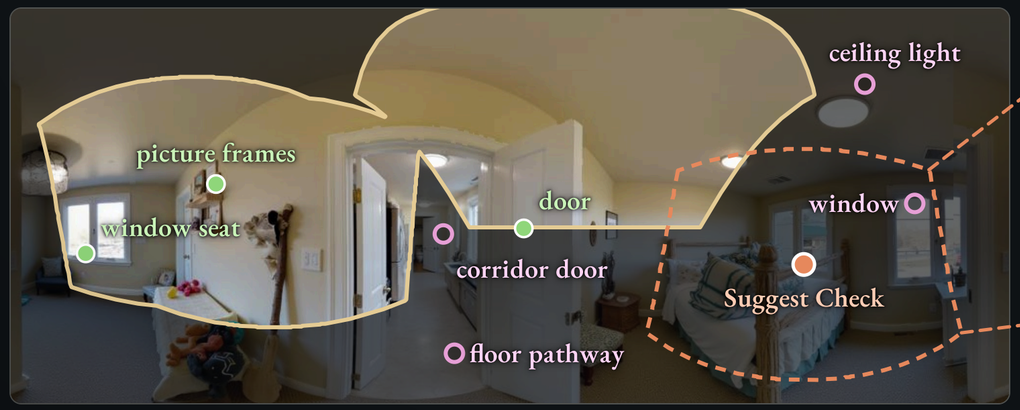
One stateless pass predicts the whole 360º layout



views so far + "Where is the bed with wooden posts?"

Imaginator
8 B

```
[Observed]
◆ open white bedroom door: (25, 0)
◆ wall-mounted picture frames: (274, 16)
◆ window seat with cushion: (227, -9)
[Imagined]
◇ carpeted floor pathway: (0, -45)
◇ corridor door: (356, -2)
◇ window with outdoor view: (166, 9)
◇ ceiling light fixture: (148, 52)
Suggest Check(126, -13)
```



seen so far · observed · imagined · suggested check



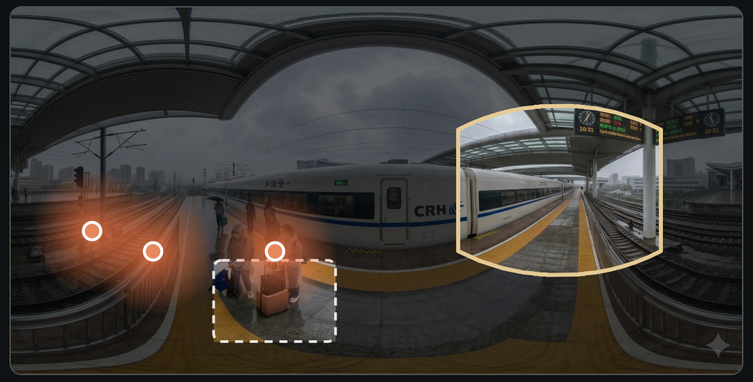
Looking at **(126°, −13°)**: the bed, in a part of the room never seen.

1. **Input**
   All views seen so far, and the instruction.
2. **`[Observed]`**
   Landmarks currently or previously visible, at absolute yaw and pitch.
3. **`[Imagined]`**
   Objects likely existing in regions it has never seen.
4. **`Suggest Check (φ, μ)`**
   A brief recommendation of where to look next.

SEARCH LOOP

The Imaginator suggests where to look; the Actor decides

Step 1 · view (90°, 0°)





SUGGESTED BY IMAGINATOR

```
rotate(160,-30)
rotate(130,-20)
rotate(-140,-30)
```

ACTOR

```
rotate(160,-30)
```

Step 2 · view (250°, −30°)





SUGGESTED BY IMAGINATOR

```
rotate(60,0)
rotate(60,-10)
rotate(50,0)
```

ACTOR

```
rotate(60,0)
```

Step 3 · view (310°, −30°)





SUGGESTED BY IMAGINATOR

```
submit(310,-50)
```

ACTOR

```
submit(310,-50) ✓
```

1 Three hypotheses

One greedy anchor (T = 0), plus two stochastic samples in the first two steps.

2 Ranked suggestions

Absolute predictions become Rotate or Submit, relative to the current view.

3 The Actor decides

Left unchanged, it weighs them against what it sees.

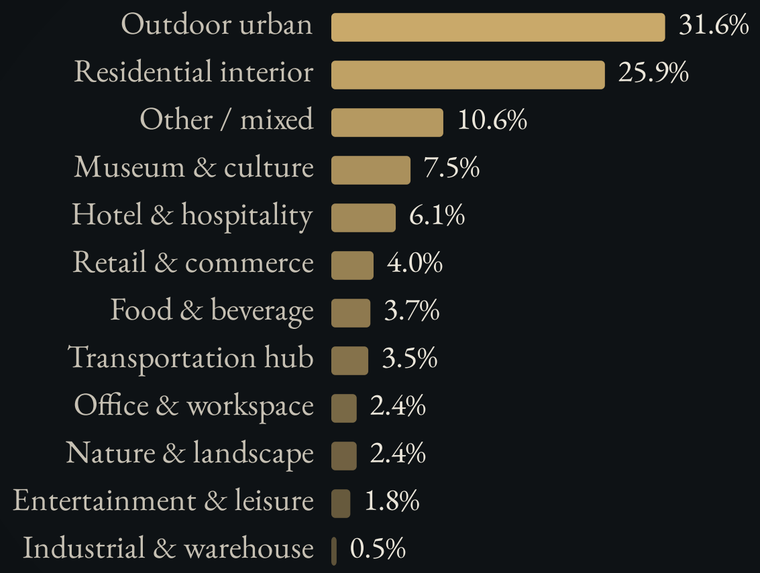
4 Hypotheses narrow

With every new view, until the pink suitcase is found in three steps.

SCALABLE DATA ENGINE

Training data, no CoT labels

Scene categories



Panorama sources



Labeled item categories



~20k
unlabeled panoramas

→

5–15
items per panorama

→

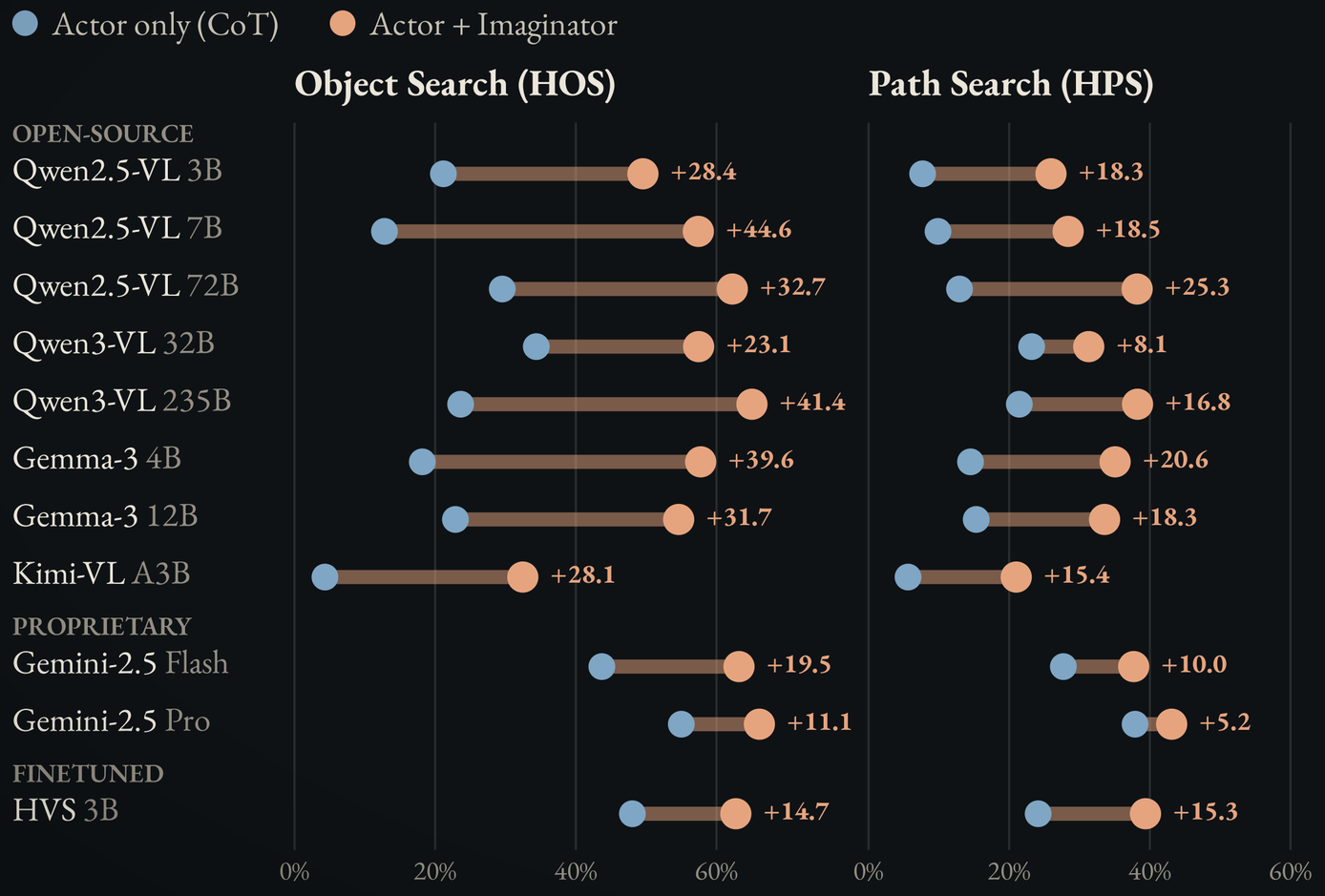
24
NFoV trajectories per scene, 50% of the targets withheld

→

1.92M
pseudo-labeled samples

1. **Masked spatial modeling**
   The Imaginator sees limited views and predicts the entire layout.
2. **Fully automated**
   A teacher model labels the 360° layout of each panorama; no trajectory-level CoT labels.
3. **Diverse scenes, dense coverage**
   Indoor and outdoor panoramas from three sources, with a median of 11 labeled items each.
4. **Two training stages**
   Pretraining on the pseudo labels, then finetuning on the H*Bench SFT split for task alignment.

A plug-and-play gain for every Actor on H*Bench



Success rate (%) within 10 steps.

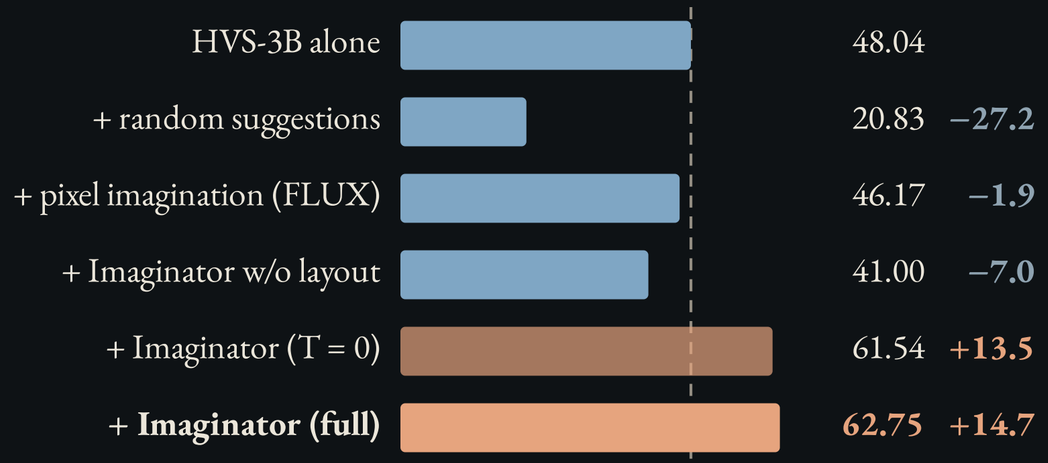
1. **Plug-and-play**
   The same Imaginator-8B guides every Actor; none is fine-tuned on its output.
2. **HVS-3B**
   48.04 → 62.75 on object search, **24.12 → 39.38** on path search.
3. **Qwen3-VL-235B**
   23.62 → 65.04 on object search: more than doubles.
4. **Gemini-2.5 Pro**
   55.00 → 66.10: proprietary models gain too.

ABLATION

Which guidance helps?

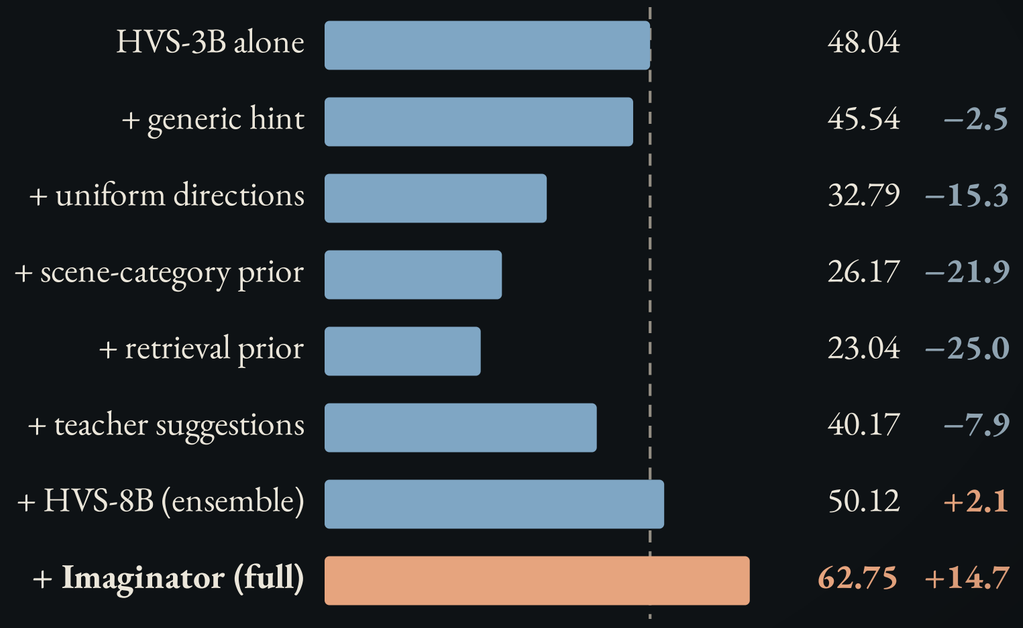
HOS success rate (%), HVS-3B as the Actor

Guidance type



Random suggestions severely degrade performance. **Pixel imagination** and an Imaginator **without the layout** both fall below the Actor alone. **Probabilistic sampling** outperforms greedy decoding (T = 0).

Controls for the source of the gain

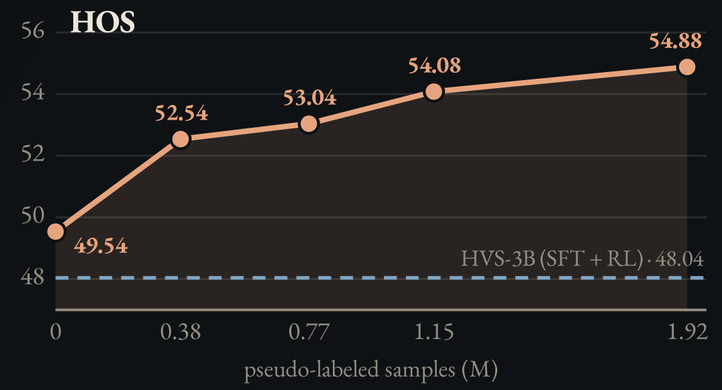


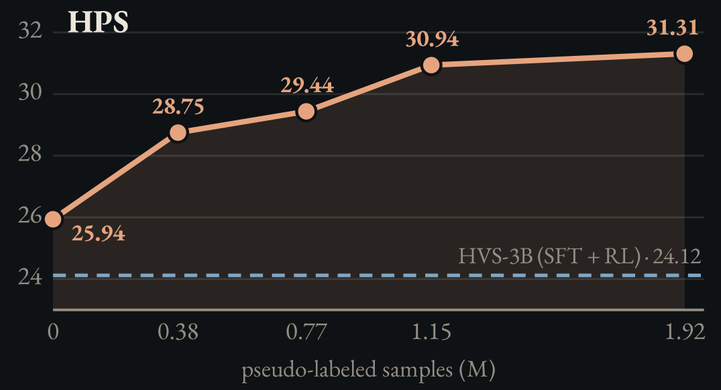
Other content at the same position in the prompt falls below the Actor alone. **The teacher** used directly gives 40.17; **a second, 8B Actor** with the same 3B+8B budget gives 50.12.

ANALYSIS

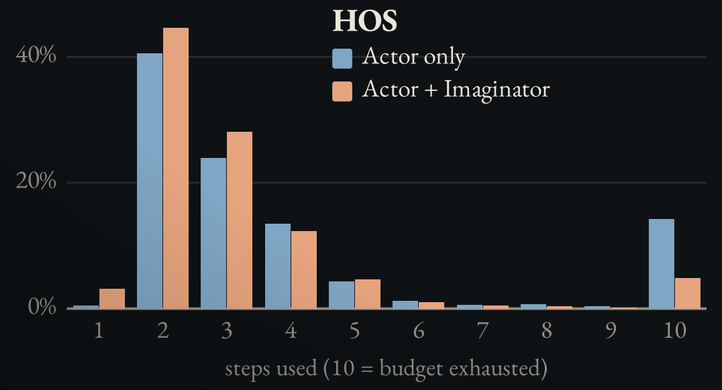
Data scaling and search efficiency

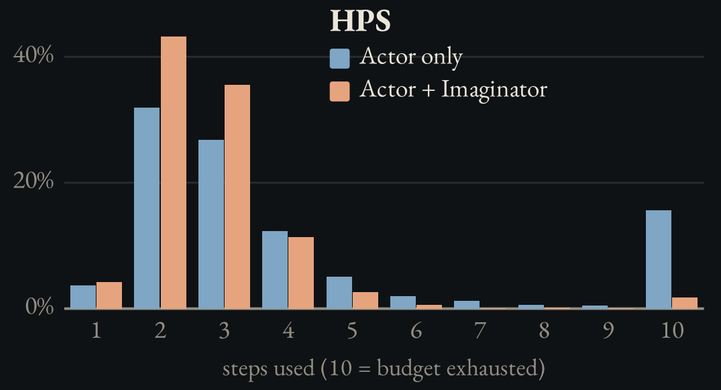
Imagination is scalable *Qwen2.5-VL-3B as the Actor*





Fewer steps to finish *share of episodes ending at each step*





1. **More data keeps helping**
   From 0 to 1.92M pseudo-labeled samples, both tasks improve monotonically.

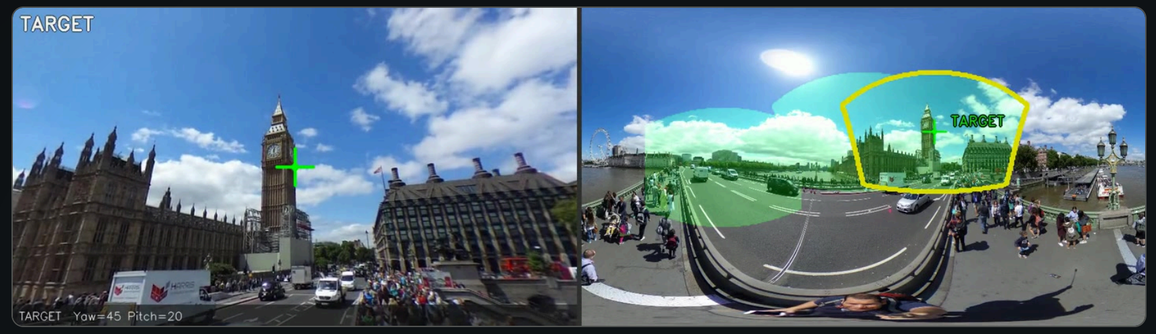
2. **0.38M samples already suffice**
   to beat HVS-3B, the Actor fine-tuned with SFT and RL (dashed line).

3. **Faster convergence**
   The Actor alone often runs into the 10-step cutoff; with the Imaginator most searches converge within 1–3 steps.

4. **Low latency**
   ~2.5 s per Imaginator pass, 5–7 s with sampling; pixel-level imagination takes 20–30 s per step.

IN THE WILD · ZERO-SHOT

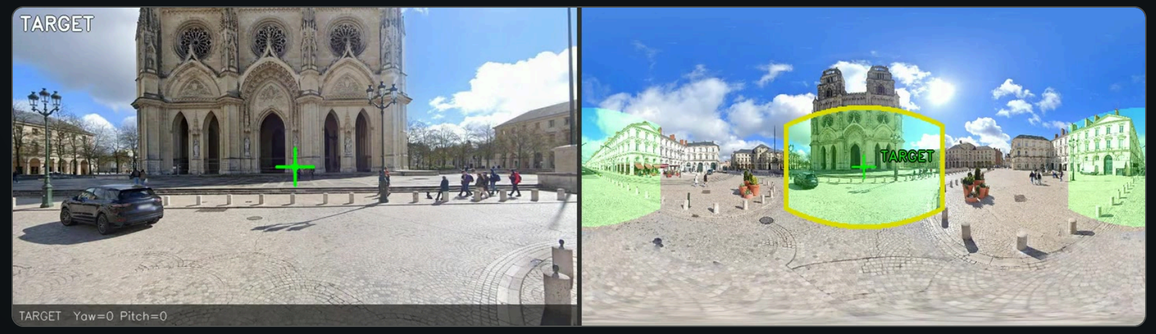
Real streets, from Google Street View



“Locate the Big Ben Tower.” **found in 3 steps**



“If you want to enter the red house, where should you move?” **found in 4 steps**



“If you want to enter the church, where’d you move?” **found in 2 steps**



“Locate the SLOW banner.” **found in 2 steps**

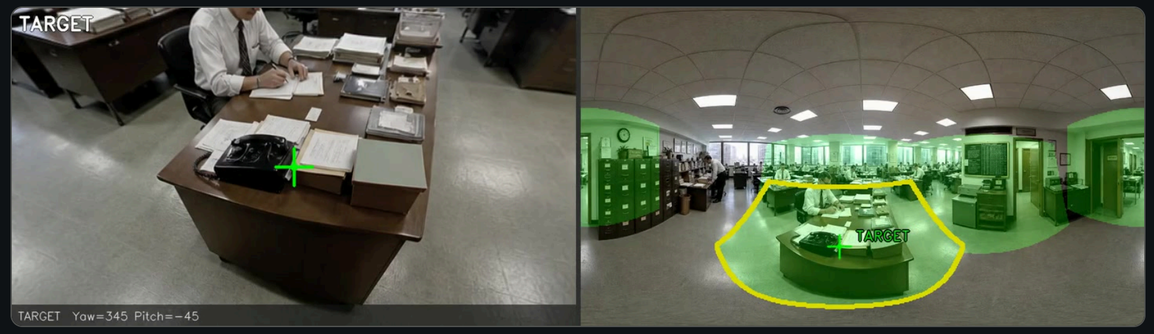
1. **Zero-shot**
   Unstructured, real-world panoramas sampled directly from Google Street View.
2. **Early cues**
   The Imaginator leverages early architectural cues to hypothesize the target’s location.
3. **2 to 4 steps**
   The agent locks onto the target without sweeping the full 360°.

IN THE WILD · ZERO-SHOT

Scenes generated by GPT-Image and Gemini



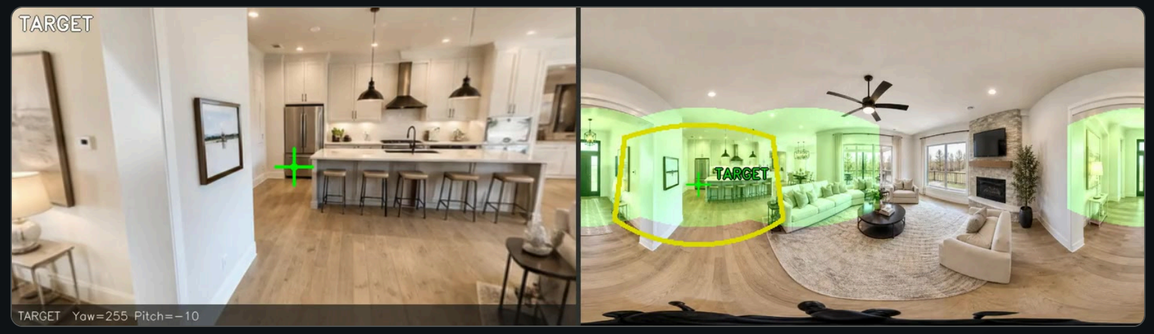
"Find the sign '5th Ave'." GEMINI 3.1-PRO **found in 2 steps**



"Locate the black telephone." GPT-IMAGE **found in 4 steps**



"Locate the TV mounted on the wall." GPT-IMAGE **found in 3 steps**



"Find the built-in refrigerator." GPT-IMAGE **found in 3 steps**

Held-out benchmark

15 panoramas (8 Google Street View, 7 generated) with 47 targets annotated by human experts, disjoint from all training data and H*Bench.

● Actor only → ● + Imaginator · success rate (%)

HVS-3B	47.87 → 58.51
Qwen2.5-VL-3B	17.55 → 46.28

CONCLUSION

Imagining in 360°

We imagine before we look.

Now, our agents can too.

1. **A decoupled framework**
   Exploration is separated into a generative Imaginator and an executive Actor.

2. **Probabilistic semantic layouts**
   They let the Actor hedge against uncertainty and search efficiently; the Imaginator boosts every Actor we tested.

3. **Limitation**
   Execution is open-loop: the Actor cannot give feedback on the priors. A Critic module is future work.

PROJECT PAGE

CODE

evergreen0929.github.io/projects/imagining360.html

github.com/Evergreen0929/Imagining360

arXiv:2605.09146