E3VS-Bench: A Benchmark for Viewpoint-Dependent Active Perception in 3D Gaussian Splatting Scenes

Koya Sakamoto[1], Taiki Miyanishi[1], Daichi Azuma[1], Shuhei Kurita[2,3,4], Shu Morikuni[1], Naoya Chiba[5], Motoaki Kawanabe[6], Yusuke Iwasawa[1], and Yutaka Matsuo[1]

[1] The University of Tokyo, Japan
[2] National Institute of Informatics, Japan
[3] Institute of Science Tokyo, Japan
[4] NII LLMC, Japan
[5] The University of Osaka, Japan
[6] Advanced Telecommunications Research Institute International, Japan

Abstract. Visual search in 3D environments requires embodied agents to actively explore their surroundings and acquire task-relevant evidence. However, existing visual search and embodied AI benchmarks, including EQA, typically rely on static observations or constrained egocentric motion, and thus do not explicitly evaluate fine-grained viewpoint-dependent phenomena that arise under unrestricted 5-DoF viewpoint control, such as disambiguating object attributes observable only from specific angles. To address this limitation, we introduce E3VS-Bench, a benchmark for embodied 3D visual search where agents must control their viewpoints in 5-DoF to gather viewpoint-dependent evidence for question answering. E3VS-Bench consists of 99 high-fidelity 3D scenes reconstructed using 3D Gaussian Splatting and 2,014 question-driven episodes. 3D Gaussian Splatting enables photorealistic free-viewpoint rendering that preserves fine-grained visual details (e.g., small text and subtle attributes) often degraded in mesh-based simulators, thereby allowing the construction of questions that cannot be answered from a single view and instead require active inspection across viewpoints in 5-DoF. We evaluate multiple state-of-the-art VLMs and compare their performance with humans. Despite strong 2D reasoning ability, all models exhibit a substantial gap from humans, highlighting limitations in active perception and coherent viewpoint planning specifically under full 5-DoF viewpoint changes. The benchmark, code, and dataset are publicly available at `https://k0uya.github.io/e3vs-proj/`.

1 Introduction

Active perception refers to an agent's ability to actively control its sensory apparatus to acquire task-relevant information, rather than passively receiving observations [1]. In modern embodied AI, it enables agents to selectively gather

Fig. 1. Overview of the proposed Embodied 3D Visual Search (E3VS) task. Unlike 2D visual search, E3VS requires an agent to actively control its 5-DoF viewpoint to resolve occlusions and acquire fine-grained visual evidence, such as the production area label on an egg carton.

task-relevant information by controlling their viewpoints and exploring the environment [12]. Such active control is essential for understanding 3D structure and for acquiring information that cannot be obtained from a single static observation. However, existing benchmarks and agents remain largely constrained to two-dimensional behaviors [26,35], and therefore do not fully capture or evaluate these capabilities in real-world 3D environments.

Visual search has long been studied in psychology and neuroscience as the process of locating targets through goal-directed and saliency-driven attention [33, 36, 37]. Inspired by these studies, recent computer vision works have explored visual search using multimodal models [20, 38, 44]. However, their evaluation remains confined to static 2D observations and cannot capture spatial reasoning challenges in real 3D environments, such as occlusions, geometric relations, and depth ambiguity. Within embodied AI, Embodied Question Answering (EQA) requires an agent to navigate a 3D environment to gather visual evidence and answer a question [11, 31]. However, most existing EQA settings treat the agent as a 2D moving camera restricted to planar navigation and limited rotation, and thus do not evaluate the viewpoint control required for realistic 3D perception. In real-world environments, a target may be visible from the initial viewpoint, yet its identifying attributes, such as labels or fine-grained text, remain unreadable without actively adjusting the viewpoint. This type of viewpoint-dependent evidence remains largely unexplored in existing benchmarks.

While recent tasks such as Aerial VLN [19, 22] extend navigation into 3D space, their primary focus is large-scale trajectory following toward distant goals. They do not evaluate the fine-grained viewpoint control required for active inspection. In these situations, the ability to manipulate the 5-DoF viewpoint itself becomes the key to revealing hidden information.

To address this gap, we introduce **E3VS-Bench**, a benchmark for Embodied 3D Visual Search (E3VS), where agents must actively control their viewpoints in 3D space with full 5-DoF to acquire task-relevant visual evidence for question answering. The benchmark contains 99 high-fidelity reconstructed scenes and 2,014 question-driven episodes. E3VS-Bench is built upon publicly available scenes reconstructed using 3D Gaussian Splatting (3DGS) [16]. 3DGS enables photorealistic free-viewpoint rendering that preserves subtle details, such as fine-

grained text and brand logos, often degraded in mesh-based representations. This fidelity makes it possible to define viewpoint-dependent questions that cannot be answered from a single observation and require active viewpoint exploration.

To ensure that benchmark performance reflects genuine exploration ability rather than prior knowledge or favorable initial viewpoints, we introduce a rigorous episode filtering pipeline based on a VLM-as-a-judge framework. This process removes episodes that can be solved without meaningful viewpoint transitions.

Our main contributions are summarized as follows:

- We study **E3VS**, a setting where agents must acquire task-relevant visual evidence through 5-DoF viewpoint control.
- We present **E3VS-Bench**, a benchmark built on 99 photorealistic 3D Gaussian Splatting scenes, with 2,014 human-annotated episodes for evaluating viewpoint-dependent reasoning and active perception in 3D environments.
- We conduct a comprehensive evaluation of state-of-the-art VLMs and reveal that current models struggle with active viewpoint planning and exploration despite strong performance on conventional 2D reasoning tasks.

2 Related Work

Visual Search. Classical visual search studies in psychology and neuroscience [33, 36, 37] investigate how humans locate target objects among distractors through goal-directed and saliency-driven attention. Inspired by these findings, recent computer vision studies have explored visual search using deep and multimodal models [20, 32, 38, 45], employing mechanisms such as curiosity-driven reinforcement learning in pixel space (Pixel Reasoner [32]), dynamic focus (SEAL [38], DyFo [20]), and reinforcement-based fine-tuning (Thyme [45]). To evaluate such capabilities, several benchmarks have been proposed, including Mini-O3 [18], H*Bench [44], and O3-Bench [21]. Mini-O3 collects challenging visual search problems requiring exploratory reasoning, while H*Bench simulates pseudo-actions such as `turn_left` and `turn_right` using panoramic images. O3-Bench requires models to reference multiple image regions during reasoning. These environments, however, remain limited to static 2D imagery and cannot evaluate occlusion reasoning or spatial relationships in 3D environments.

Active Perception and Exploration. Building upon the concept of active perception [1, 5], recent embodied AI studies investigate how agents adjust viewpoints to obtain informative observations, through active perception for object understanding and manipulation [17, 40, 41] or exploration-driven strategies that seek novel or semantically meaningful observations [7, 30, 43]. Most of these methods focus on specific objectives such as object detection or navigation. Building upon this line of work, our setting studies how agents actively acquire task-relevant visual evidence through fine-grained 5-DoF viewpoint control for language-guided reasoning.

Question Answering in 3D Scene. Question answering in 3D environments follows two main paradigms: scene-centric 3D-QA and embodied QA. Scene-

Dataset	Active	Env.	Source	DoF	Episodes
ScanQA [2]	✗	Mesh	ScanNet [10]	–	41,363
SQA3D [24]	✗	Mesh	ScanNet [10]	–	33,400
REVERIE [26]	✓	Mesh	Matterport3D [6]	2-DoF (Graph-based)	21,702
EQA [11]	✓	Mesh	House3D [39]	2-DoF	5,281
MP3D-EQA [35]	✓	Mesh	Matterport3D [6]	2-DoF	1,136
OpenEQA [25]	✓	Mesh	Matterport3D [6]	2-DoF	1,600
E3VS-Bench	✓	**3D-GS**	**SceneSplat++** [23]	**5-DoF**	2,014

Table 1. Comparison with embodied QA and 3D visual grounding datasets. E3VS-Bench uses 3DGS scenes with 5-DoF control for occlusion-aware active perception.

centric approaches reason directly over reconstructed 3D representations without requiring physical movement. For example, ScanQA [2] extends the ScanRefer [8] paradigm from 3D object grounding to QA on point-cloud scenes, and SQA3D [24] introduces situated reasoning from a specific pose using pre-defined observations. These approaches assume access to complete scene representations and thus do not evaluate how agents actively acquire missing visual evidence. EQA instead requires an agent to navigate a 3D environment to answer natural-language questions [11, 25, 27, 35]. Answerability Fields [3] estimates spatial answerability distributions, predicting locations where visual evidence required to answer a question is likely to be observable. Subsequent studies enhance relational reasoning through scene graphs and long-horizon planning [9, 13, 29]. Simulation platforms such as Habitat [28] enable benchmarks including MP3D-EQA [35] and EXPRESS-BENCH [15]. These navigation-centric settings, however, often treat the agent as a 2D moving camera and do not evaluate the fine-grained viewpoint control required to reveal task-relevant visual details.

Table 1 summarizes major 3D scene QA and EQA datasets. Although these datasets vary in scale and realism, they typically rely on mesh or point-cloud reconstructions that restrict viewpoint interaction and limit photometric fidelity, and thus do not evaluate how viewpoint-dependent observations influence question answering. In contrast, E3VS-Bench enables agents to actively explore photorealistic 3DGS environments with full 5-DoF viewpoint control, allowing systematic evaluation of viewpoint-dependent reasoning in 3D environments.

3 Benchmark Design and Construction

Unlike existing embodied tasks (e.g., EQA [35]) that rely on limited camera motion, E3VS-Bench requires agents to select viewpoints revealing task-relevant evidence not observable from a single view. By embedding such viewpoint-dependent constraints in photorealistic 3D scenes with unrestricted camera control, it evaluates viewpoint planning and occlusion-aware spatial reasoning.

3.1 Task Formulation

Problem Setup. We formulate E3VS as an active perception task for question answering in photorealistic 3D environments. Unlike conventional 2D visual search [20, 38] or EQA [35], where agents typically move on a ground plane with

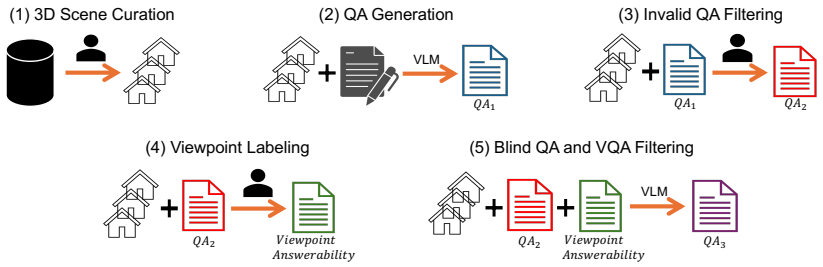

Fig. 2. Dataset construction pipeline for E3VS-Bench. The pipeline consists of five stages: (1) 3D scene curation from SceneSplat++ [23], (2) QA generation using a VLM, (3) invalid QA filtering with human verification, (4) viewpoint labeling to identify answerable viewpoints, and (5) answerability filtering to remove questions solvable without viewpoint transitions.

limited viewpoint control, E3VS requires agents to adjust their viewpoints in 3D space to resolve occlusions, depth ambiguities, and fine-grained visual details.

Formally, an episode is defined by a triplet $(\mathcal{S}, q, v_0)$, where $\mathcal{S}$ denotes a 3D scene reconstructed using 3DGS, q is a natural-language question, and v_0 represents the initial camera viewpoint. At each discrete time step $t \in [0, T]$, the agent receives an egocentric RGB observation $O_t \in \mathbb{R}^{H \times W \times 3}$ rendered from the scene at viewpoint v_t. The initial viewpoint v_0 is centered on the target object; however, the target may be partially or fully occluded by other objects.

Action Space. We represent the agent's state at time t as a viewpoint $v_t = (x_t, y_t, z_t, \theta_t, \phi_t) \in \mathbb{R}^5$, where (x_t, y_t, z_t) denote the 3D Cartesian coordinates and (θ_t, ϕ_t) represent the yaw and pitch angles. The agent interacts with the environment through a discrete 5-DoF action space $\mathcal{A}$, where each action $a_t \in \mathcal{A}$ corresponds to either a translation, a rotation, or termination. The state transition follows the environment dynamics $v_{t+1} = E(v_t, a_t)$, where a collision after executing a_t leaves the state unchanged $v_{t+1} = v_t$.

Success Criteria. Each episode is associated with a set of human-annotated answerable viewpoints $\mathcal{V}_{ans}$ and a corresponding goal image O_{goal} rendered from $v \in \mathcal{V}_{ans}$. Upon executing the `stop` action at step T, the agent provides a predicted answer $\hat{y}$ based on the final observation O_T and the question q.

To evaluate the correctness of open-vocabulary responses, we employ a VLM-as-a-judge framework, following LLM-Match in OpenEQA [25]. Specifically, an evaluator model J (e.g., *GPT 5.1*) receives the agent's final observation O_T, the question q, the predicted answer $\hat{y}$, the ground-truth answer y, and the goal image O_{goal} to compute a score $S = J(O_T, q, \hat{y}, y, O_{goal}) \in \{1, \ldots, 5\}$, where 5 is perfectly correct and 1 entirely incorrect. This metric allows for a flexible yet rigorous assessment of the agent's active perception capability.

3.2 Dataset Construction Pipeline

In this section, we present an overview of the question–answer creation pipeline used to construct E3VS-Bench. As shown in Fig. 2, the construction process

consists of five stages: (1) *3D scene curation*, (2) *QA generation*, (3) *invalid QA filtering*, (4) *viewpoint labeling*, and (5) *answerability filtering* , followed by additional data cleaning. We briefly describe each stage below and provide full details in the supplementary.

3D Scene Curation. First, we manually curate 105 high quality reconstructed scenes from SceneSplat++ [23], which represents 3D environments from ScanNet++ [42] using 3DGS [16]. Unlike traditional mesh-based representations that often suffer from texture degradation and geometric oversmoothing, 3DGS preserves the photometric fidelity necessary for reliable viewpoint-dependent reasoning. As shown in Fig. 3, traditional meshes often lack the photorealistic textures needed to capture subtle visual attributes, such as text on a plastic bag or brand logos. The photorealistic rendering of 3DGS enables fine-grained, viewpoint-dependent evaluation of embodied agent perception.

QA Generation. To generate high-quality question–answer pairs at scale, we employ a three-stage pipeline powered by a VLM, specifically *Gemini 2.5 Flash*. (1) We select object categories with few instances per scene. For each instance, we compute a viewing distance so that its projected object bounding box spans roughly three-fifths of the image along either axis. We then uniformly sample viewpoints on a sphere centered at the object and render the corresponding multi-view images. (2) We use a VLM to filter the generated views, discarding images where the object is heavily occluded, outside the reconstruction bounds, or visually ambiguous. (3) Using the remaining valid multi-view images, the VLM synthesizes candidate question-answer pairs that cover a spectrum of visual reasoning, including intrinsic attributes (e.g., color, material, and branding) and extrinsic spatial relations. This pipeline generates 27,877 QAs across 7,578 instances.

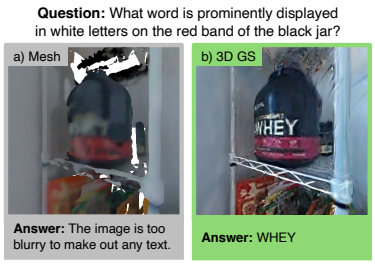

Fig. 3. Comparison of rendering quality between traditional mesh-based (ScanNet++) and 3D Gaussian Splatting (SceneSplat++). 3DGS preserves sharp textures for small text (e.g., "WHEY" label), which is crucial for viewpoint-dependent visual reasoning.

Invalid QA Filtering. We perform human filtering of QA candidates and annotate viewpoint answerability. Annotators verify each candidate against predefined criteria (see Appendix), removing ambiguous, multi-interpretable, or objectively unanswerable questions. Inaccurate answers are corrected accordingly. The overall human annotation process required approximately 1,120 hours.

Viewpoint Labeling. For each retained question, expert annotators select viewpoints that contain sufficient information to answer it. Physically invalid viewpoints, such as those intersecting 3D geometry, are discarded. The annotations are further verified by another annotator to ensure quality and consistency. Finally, unanswerable viewpoints serve as episode starting positions, while answerable viewpoints serve as goal positions.

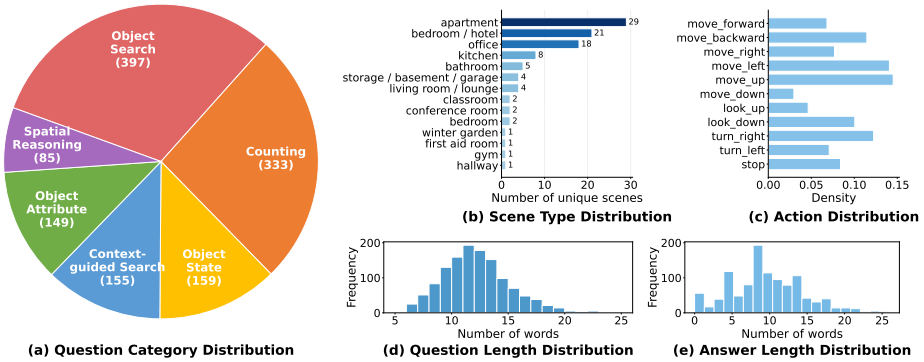

Fig. 4. Dataset Distribution. Each figure represents (a) question category distribution of unique QA pairs classified by *GPT 5.1*. (b) scene type distribution. (c) action distribution. (d, e) the number of words in each question and answer.

Answerability Filtering. To ensure the benchmark properly evaluates active exploration, we remove questions answerable from prior knowledge alone and episodes solvable from the initial viewpoint. For filtering, we use *GPT-5.1*. We perform VQA in both a blind setting (without visual input) and from the designated start viewpoint, evaluating response correctness on a 1–5 scale via a VLM-as-a-Judge protocol (see Section 3.1 for details). QA pairs scoring 3 or above in either setting are excluded, where the threshold is chosen as the midpoint of the 1–5 scale to conservatively filter out cases that show any indication of being answerable without active exploration. The remaining episodes cannot be solved from prior knowledge or the initial observation alone, and require active 5-DoF viewpoint transitions and information gathering for accurate answers.

3.3 Dataset Statistics and Analysis

After meticulous filtering, our dataset contains 99 scenes, 2,014 episodes derived from 1290 unique question-answer pairs. Fig. 4 shows that the dataset covers diverse scene categories and exhibits a wide distribution of question and answer lengths. The final benchmark focuses on six question categories that broadly evaluate an agent's capability in active perception and spatial reasoning, illustrated in Fig. 5. These categories test an agent's ability to (1) perform Object Search (OS) by locating items explicitly mentioned in the question, (2) recognize Object States (OST) such as open or closed, which requires navigating to a viewpoint where the state is observable, (3) identify Object Attributes (OA) such as physical characteristics and features, (4) conduct Context-guided Search (CGS) to find objects by functional or situational context when the object name is not given, (5) perform Spatial Reasoning (SR) over relative positions and sizes, and (6) perform Counting (CNT) over instances of an object class.

In addition, the action distribution in Fig. 4(c) shows that all action types are well utilized, including those beyond planar navigation, suggesting the need for 5-DoF active viewpoint control. Notably, *move_backward* is more frequent

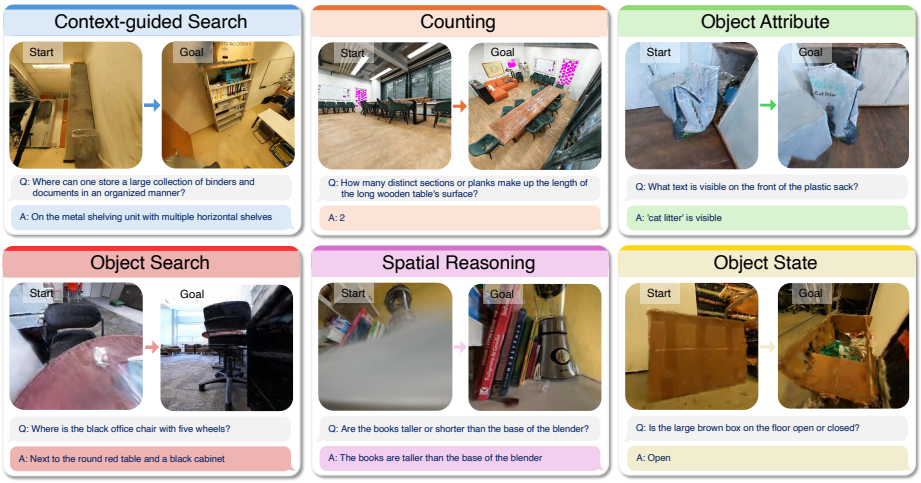

Fig. 5. Examples of E3VS task defined in our dataset. Each example illustrates a distinct reasoning type that requires viewpoint control in reconstructed 3D environments.

than *move_forward*, indicating that gaining context by moving away from the target is often necessary. Moreover, *move_up* is the most frequent action, and *look_down* appears more often than *look_up*, suggesting that relevant visual evidence is often obtained by observing target objects from lateral or slightly elevated viewpoints rather than from below.

4 Evaluation Protocol and Baselines

We design evaluation settings to assess whether VLMs can actively acquire visual evidence through viewpoint control, resolving spatial ambiguities such as occlusions and depth uncertainty by selecting informative viewpoints.

4.1 E3VS Framework for VLMs

To evaluate VLMs on the E3VS task, we implement a closed-loop perception–action framework in which the model iteratively selects actions based on egocentric observations. At each time step t, the agent receives an observation O_t rendered from the 3DGS scene G at its viewpoint $v_t = (x, y, z, \theta, \phi)$, and the VLM processes O_t with the question to select an action $a_t \in \mathcal{A}$. This continues until the agent executes `stop`, after which it generates the final answer.

To interface VLMs with the environment, we construct a structured prompt that conditions the model on both the environment configuration and the current interaction state. The prompt consists of two components:

– **System Prompt:** Defines the world coordinate system (Z-axis as up) and the constraints of the action space, including fixed translation (0.25 m) and rotation (30°) intervals.

– **User Prompt:** Provides task input and dynamic state information, including the question, the current observation image, step count, 3D coordinates, the previous action, and feedback such as collision detection.

Unless otherwise specified, our experiments use a single-frame observation setting without additional reasoning modules to establish a clear baseline. Full prompt templates are provided in the Appendix.

4.2 Baseline Models

To examine different levels of perceptual access and spatial reasoning, we define five categories of baseline agents: (1) blind VLMs, (2) VQA at Start, (3) VQA at Birdview, (4) VQA at Goal, and (5) 2D Visual Search at Start. These baselines form a progressive spectrum of perceptual conditions, from no visual input to privileged viewpoints, providing a reference for interpreting embodied agent performance on E3VS-Bench.

(1) Blind VLMs. In this setting, models are asked to answer questions using only textual input without any visual observation. This baseline measures how well models can infer answers from linguistic cues and prior knowledge alone.

(2–4) VQA from Fixed Viewpoints. In these settings, VLMs perform visual question answering from a single static image captured at a predefined viewpoint, without any viewpoint movement.

VQA at Start uses the image from the initial viewpoint and evaluates whether the question can be answered without exploration.

VQA at Goal uses an image from human-annotated answerable viewpoints, representing an upper bound where sufficient visual evidence is available.

VQA at Birdview provides a bird's-eye image from directly above the scene, evaluating whether a global overview alone suffices. Many questions in E3VS-Bench still require fine-grained, viewpoint-dependent evidence (e.g., text, object states, occluded regions) that cannot be resolved without moving to appropriate viewpoints.

(5) 2D Visual Search at Start. This baseline examines whether questions can be solved by image-space exploration alone, without 3D viewpoint transitions. We adopt representative methods such as SEAL [38] and DyFo [20] to show the limitations of image-based search relative to embodied 3D visual search.

5 Experiments

5.1 Experimental Setting

Dataset Split. The dataset is split into train, validation, and test sets at the scene level. The train set contains 1,406 episodes from 900 question–answer pairs across 68 scenes, the validation set contains 231 episodes from 132 pairs across 10 scenes for model selection and hyperparameter tuning, and the test set contains 377 episodes from 258 pairs across 21 scenes for final evaluation. All splits are

going to be publicly released. In this work, we focus on evaluating zero-shot VLM-based frameworks.

Evaluation Metrics. We evaluate models from three aspects: answer correctness, exploration efficiency and navigation safety. Regarding answer correctness, we employ a VLM-as-a-judge framework in accordance with OpenEQA [25], using *GPT 5.1* as the evaluator. The judge VLM receives the predicted response and ground-truth answer, along with the end and goal images, and outputs a score of 5 for correct predictions and 1 for incorrect ones. This metric shows a Spearman correlation of $\rho = 0.54$ with human evaluation, indicating reasonable agreement. Exploration efficiency is quantified by the average number of steps, and navigation safety by Collision Rate, a binary indicator that is 1 if any collision occurs within an episode and 0 otherwise.

Baseline Settings. Across the baseline settings, we evaluate a diverse set of proprietary and open-source models, including *Gemini 2.5 Pro/Flash*, *Gemini 3.0 Pro/Flash*, *GPT 5.1*, and open-source models such as *Qwen3-VL 8B/30B*, *InternVL3.5-8B*, and *Step3-VL-10B*. These models are evaluated under embodied E3VS agent settings, while a subset of models is used for static baselines to enable controlled comparison. For static baseline settings, we evaluate representative models including *Gemini 2.5 Flash*, *GPT 5.1*, and *Qwen3-VL-8B*. To avoid evaluation bias, *GPT 5.1* is excluded from the blind and start-view settings, since it was used during dataset filtering. For the same reason, *GPT 5.1* is used for answer generation in all E3VS agent conditions except Human: because episodes it could answer from the start viewpoint were already removed (Section 3.2), using a different generator would introduce model-specific bias.

Implementation Details. We use images with a resolution of 512×512 pixels and a field of view (FOV) of $90°$. Each episode is limited to a maximum of 25 steps, providing a sufficient exploration budget while preventing excessively long trajectories. For locomotion actions, the agent moves 0.25 m per step, while rotation actions rotate the camera by $30°$ in place. The maximum output token length is 128 when reasoning is disabled and 256 or more when enabled, depending on the model. If no valid action is produced within this limit, the agent defaults to `move_forward`.

5.2 Main Results

We evaluate a wide range of VLMs on the E3VS-Bench, categorized into static baselines and E3VS agents. The results are summarized in Table 2, 3.

Performance Across Model Families on E3VS. The evaluation on E3VS shows that *Gemini 3.0 Flash* substantially outperforms all other models across nearly all categories. In contrast, *GPT 5.1* performs comparably to open-source models such as *Qwen3-VL-8B* and *Step3-VL-10B*, and overall these models achieve scores only slightly above a Random Action baseline. This suggests that while current VLMs possess strong 2D image recognition capabilities, they show limitations in 3D visual search settings that require viewpoint changes. A substantial gap also remains between current models and human performance.

Model	OS	OST	OA	CGS	SR	CNT	Avg.	Steps	Coll.
Random Action	2.38	2.18	2.60	2.20	2.50	2.00	2.28	10.15	0.39
Proprietary Models									
Gemini 2.5 Pro	3.14	2.36	2.45	2.60	2.25	**2.04**	2.54	7.80	0.31
Gemini 3.0 Flash	**3.21**	**2.82**	**3.18**	**3.53**	2.75	1.88	**2.79**	11.29	0.43
Gemini 3.0 Pro	3.07	2.55	**3.18**	3.40	**3.00**	1.96	2.75	10.17	0.34
GPT 5.1	2.90	2.18	2.53	2.73	2.25	1.88	2.42	15.04	0.49
Open Source Models									
Qwen3-VL-8B [4]	2.86	2.36	2.60	3.13	1.88	1.96	2.46	18.73	0.30
Qwen3-VL-30B [4]	2.67	2.09	2.60	2.47	2.25	1.84	2.32	14.18	0.50
InternVL3.5-8B [34]	2.66	2.18	2.24	2.60	2.12	1.60	2.21	10.85	0.65
Step3-VL-10B [14]	2.38	2.27	2.38	3.13	2.50	1.64	2.24	11.16	0.42
Human	3.12	3.59	4.06	4.06	3.35	3.59	3.53	11.21	-

Table 2. Main results on E3VS-Bench. We evaluate various baseline models across six fine-grained question types. All values except Steps and Coll. are VLM judge scores on a scale of 1 to 5 where a higher score indicates better performance. Steps and Coll. denote the average number of Navigation Steps and collision rate respectively.

Comparison of OS and CGS. Models often perform better on CGS than on explicit OS (Table 2). CGS queries can often be solved by recognizing relevant objects or functional zones (e.g., "where to store binders"). In contrast, OS requires not only recognizing the target object but also understanding its spatial relations and placement in the scene. This additional spatial reasoning likely increases task complexity and leads to lower performance.

Challenges in CNT. In CNT, models already struggle in the 2D single-view setting and degrade further in the full E3VS setting (Tables 3, 2), whereas humans score 3.82 given the goal viewpoint. This reveals a fundamental limitation: counting requires accumulating evidence across a 5-DoF trajectory, which single-observation EQA approaches cannot handle.

Emerging Viewpoint Capability for SR. SR questions require understanding geometric relationships such as relative positions, distances, and size comparisons between objects, and the quality of such geometric evidence is highly viewpoint-dependent; comparing object heights, for example, is far easier from a lateral viewpoint than from an oblique or top-down view. In this context, *Gemini 3.0 Pro* achieves E3VS performance (Table 2) comparable to the VQA at Goal upper bound (Table 3), indicating that it can not only recognize spatial relationships but also actively select viewpoints that improve their observability, effectively integrating viewpoint selection into SR.

Limited Gain from Viewpoint Selection in OST. OST questions typically involve binary decisions (e.g., open vs. closed), which can often be predicted from prior knowledge or dataset bias even without sufficient visual evidence. Importantly, during dataset construction, episodes that *GPT 5.1* could answer correctly from the initial viewpoint were filtered out. However, similar filtering was not applied to other models such as Qwen or Gemini. As a result, VQA at Start performance for these models still reflects the influence of bias. Empirically,

Model	OS	OST	OA	CGS	SR	CNT	Avg.
Static Baselines (No Agent Movement)							
Blind VLM							
Gemini 2.5 Flash	1.72	2.00	2.09	1.67	1.88	1.72	1.82
Qwen3-VL-8B	1.38	2.64	1.58	2.20	1.50	1.52	1.67
2D Visual Search at Start							
SEAL [38]	1.62	2.45	1.44	1.53	2.50	1.32	1.68
DyFO [20]	1.90	2.45	1.44	1.67	2.00	1.80	1.86
VQA at Start							
Gemini 2.5 Flash	2.21	2.64	1.73	2.33	2.12	2.00	2.14
Qwen3-VL-8B	2.10	2.82	1.87	2.20	2.12	1.56	2.02
VQA at Birdview							
Gemini 2.5 Flash	1.31	1.64	1.44	1.67	1.75	1.48	1.48
GPT 5.1	1.48	2.09	1.58	2.20	1.62	1.16	1.55
Qwen3-VL-8B	1.55	2.18	2.09	1.80	1.38	1.12	1.59
VQA at Goal							
Gemini 2.5 Flash	3.79	3.27	3.84	4.20	3.50	3.16	3.58
GPT 5.1	4.17	3.36	3.91	4.20	3.00	2.88	3.60
Qwen3-VL-8B	3.07	3.55	3.40	4.20	3.00	2.20	3.03
Human	4.25	3.82	4.47	4.25	4.14	3.82	4.12

Table 3. Comparison of static baselines on E3VS-Bench. The large performance gap between 'VQA at Start' and 'VQA at Goal' indicates that initial viewpoints are often insufficient, necessitating active exploration.

we observe that the VQA at Start performance of *Qwen3-VL-8B* is comparable to the E3VS performance of *Gemini 3.0 Flash* evaluated with *GPT 5.1* as the judge. This suggests that active viewpoint selection yields only marginal improvements beyond compensating for such biases. Furthermore, the viewpoint required to resolve the state is relatively predictable (e.g., Fig. 5), reducing OST to viewpoint selection and path planning rather than exploration. Despite this, gains remain limited, indicating that models still struggle to reach the appropriate viewpoint even when it is relatively clear.

High Viewpoint Sensitivity in OA. OA questions require identifying fine-grained, instance-specific properties of a target object, such as text, material, color, or subtle visual features. While the target is explicitly specified (as in OS), it is often unclear from the initial viewpoint where the relevant evidence resides on the object. Unlike OST, whose observation region is predictable (e.g., a door for open/closed), OA thus requires actively exploring the object to discover informative evidence for each instance. Empirically, while *Gemini 3.0* models achieve relatively high performance on OA, other models perform at or below the Random Action baseline. This highlights the difficulty of acquiring fine-grained visual evidence through 5-DoF active viewpoint control, suggesting that OA is highly sensitive to viewpoint selection and requires both exploratory viewpoint discovery and detailed visual recognition.

Model	OS	OST	OA	CGS	SR	CNT	Avg. Steps	Coll.
Gemini 3.0 Flash	3.21	2.82	3.18	3.53	2.75	1.88	2.79	11.29	0.43
Gemini 3.0 Flash+Th.	3.21	2.82	2.96	3.27	3.00	2.00	2.79	19.80	0.41
GPT 5.1	2.90	2.18	2.53	2.73	2.25	1.88	2.42	15.04	0.49
GPT 5.1+Th.	2.97	2.64	3.25	2.87	2.38	2.16	2.70	13.78	0.38

Table 4. Ablation study on the effect of the thinking process on E3VS-Bench. Th. indicates whether the model performs the Thinking process.

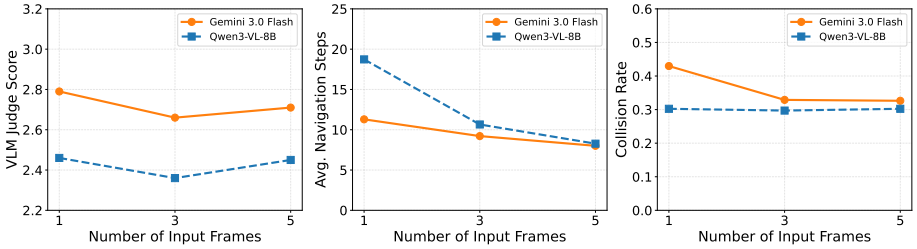

Fig. 6. Effect of the number of input frames on E3VS performance. While more frames do not significantly impact the VLM judge score, they consistently lead to more efficient navigation (fewer steps) and safer trajectories (lower collision rates).

5.3 Ablation Study

Impact of Internal Reasoning. We analyze the effect of thinking-related parameters on E3VS performance without using explicit reasoning prompts (Table 4). For *GPT 5.1*, increasing the thinking budget consistently improves performance, suggesting that additional internal computation benefits viewpoint planning and decision making. In contrast, *Gemini 3.0 Flash* shows little sensitivity to the reasoning budget. Overall, the effectiveness of increased reasoning appears to be model-dependent, though the cause remains unclear and may relate to differences in architecture or training data.

Effect of Memory. We study the effect of temporal information by varying the number of input frames (1, 3, and 5) for *Gemini 3.0 Flash* and *Qwen3-VL-8B*. As shown in Fig. 6, multi-frame inputs do not significantly improve VLM Judge scores, but consistently reduce navigation steps and collision rates. Step reduction is likely due to fewer action "deadlocks," where single-frame agents repeat the same actions. Collision reduction appears to result from improved action–observation understanding, enabling better anticipation of viewpoint changes and more effective collision avoidance.

Decoupling Recognition and Exploration via Observable Evidence. We investigate whether VLMs can correctly recognize visual evidence and terminate the episode when the required information is already visible from the current viewpoint. To this end, we construct a setting where the initial view contains sufficient visual evidence to answer the question, removing the need for further exploration. As shown in Table 5, both models achieve strong performance under this condition, approaching the VQA at Goal performance of *GPT 5.1* reported

Model	Init. at Goal	OS	OST	OA	CGS	SR	CNT	Avg. Steps	Coll.
Qwen3-VL-8B	✗	2.86	2.36	2.60	3.13	1.88	1.96	2.46	18.73	0.30
Qwen3-VL-8B	✓	3.90	3.36	3.84	3.93	2.75	2.56	3.38	12.06	0.23
Gemini 3.0 Flash	✗	3.21	2.82	3.18	3.53	2.75	1.88	2.79	11.29	0.43
Gemini 3.0 Flash	✓	4.21	3.55	3.33	3.93	3.25	2.38	3.41	6.60	0.29

Table 5. Ablation study with goal-initialized viewpoints, where the initial viewpoint is set to the goal viewpoint containing sufficient visual evidence. This removes the need for exploration and isolates the model's ability to recognize visual evidence and make appropriate stopping decisions.

in Table 3. However, their stopping behaviors differ notably. *Gemini 3.0 Flash* terminates efficiently, requiring only 6.60 steps on average, whereas *Qwen3-VL-8B* takes nearly twice as many steps (12.06), suggesting that *Qwen3-VL-8B* tends to continue exploring even when sufficient visual evidence is already available. Across most question types, *Gemini 3.0 Flash* outperforms *Qwen3-VL-8B*, indicating stronger overall evidence utilization. In contrast, for OA questions, *Qwen3-VL-8B* achieves substantially higher scores, implying a stronger ability to recognize fine-grained object properties and make stopping decisions based on such cues. Interestingly, this trend reverses in the standard E3VS setting with unanswerable initial viewpoints, where Gemini consistently outperforms *Qwen3-VL-8B*. This suggests that the two models exhibit different strengths: Gemini is more effective at actively exploring to acquire missing information, whereas *Qwen3-VL-8B* shows stronger capability in recognizing fine-grained visual attributes when evidence is present, but does not consistently translate this recognition into appropriate stopping behavior.

5.4 Qualitative Results

Fig. 7 compares the viewpoints selected by humans and *Gemini 3.0 Flash*. Human-selected viewpoints are typically intuitive, deliberately centering the target object or revealing the discriminative features required to answer the question. In contrast, the "answerable" viewpoints reached by agents are frequently sub-optimal, as they do not always capture sufficient visual evidence even when the target object is partially visible. Since the VLM-as-a-judge penalizes answers whose supporting evidence is absent from the final observation, such stops receive low scores (e.g., Fig. 7, first row, second column, where *Gemini 3.0 Flash* halts before the task-relevant features are visible). This highlights the importance of actively selecting evidence-revealing viewpoints, which remains challenging for current VLM-based agents.

6 Conclusion

In this paper, we introduced **E3VS-Bench**, a benchmark for viewpoint-dependent active perception in photorealistic 3D Gaussian Splatting environments. E3VS requires agents to manipulate viewpoints in 5-DoF to acquire task-critical evidence. The benchmark consists of 99 high-fidelity scenes and 2,014 episodes

Fig. 7. Qualitative Results. The orange bars represent the predicted answers after visual search by *Gemini 3.0 Flash*, while the blue bars indicate the human performance.

covering diverse reasoning types, including object search, spatial reasoning, attribute recognition, and counting. We evaluated a range of state-of-the-art proprietary and open-source VLMs under both static and active settings. While several models demonstrate strong 2D visual reasoning ability, all evaluated systems exhibit a substantial performance gap compared to human performance on E3VS. These results suggest that current VLM-based agents still lack robust viewpoint planning in 3D environments.

Limitations. We adopt a VLM-as-a-judge framework to evaluate open-vocabulary answers. However, we observe a relatively weak correlation between human evaluation and VLM-based judging. In some cases, the agent's final observation differs substantially from the human-annotated goal viewpoint while receiving a similar judge score, indicating that current automated evaluation may insufficiently capture viewpoint adequacy.

Future Work. This work focuses on zero-shot evaluation without exploiting the train/validation splits for optimization. We release each split to facilitate future learning-based agents.

Acknowledgments. This work was supported by JST PRESTO (Grant Number JPMJPR22P8), JST CRONOS (Grant Number JPMJCS24K6), JST SPRING (Grant Number JPMJSP2108), JSPS KAKENHI (Grant Number 25K03177), and JST K Program (Grant Number JPMJKP25V2), Japan.

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. International Journal of Computer Vision **1**, 333–356 (1988)
2. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19129–19139 (June 2022)
3. Azuma, D., Miyanishi, T., Kurita, S., Sakamoto, K., Kawanabe, M.: Answerability fields: Answerable location estimation via diffusion models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2024)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Bajcsy, R.: Active perception. Proceedings of the IEEE **76**(8), 966–1005 (1988)
6. Chang, A., et al.: Matterport3d: Learning from rgb-d data in indoor environments. In: International Conference on 3D Vision (3DV) (2017)
7. Chaplot, D.S., Jiang, H., Gupta, S., Gupta, A.: Semantic curiosity for active visual learning. In: Proc. European Conference on Computer Vision (ECCV) (2020)
8. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: Proc. European Conference on Computer Vision (ECCV) (2020)
9. Cheng, K., Li, Z., Sun, X., Min, B.C., Bedi, A.S., Bera, A.: Efficienteqa: An efficient approach to open-vocabulary embodied question answering. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2025)
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5829–5839 (2017)
11. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied Question Answering. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
12. Feng, T., Wang, X., Jiang, Y.G., Zhu, W.: Embodied ai: From llms to world models. arXiv preprint arXiv:2509.20021 (2025)
13. Ginting, M.F., Kim, D.K., Meng, X., Reinke, A.M., Krishna, B.J., Kayhani, N., Peltzer, O., Fan, D., Shaban, A., Kim, S.K., Kochenderfer, M., Agha-mohammadi, A.a., Omidshafiei, S.: Enter the mind palace: Reasoning and planning for long-term active embodied question answering. In: Proc. Conference on Robot Learning (CoRL) (2025)
14. Huang, A., Yao, C., Han, C., Wan, F., Guo, H., Lv, H., Zhou, H., Wang, J., Zhou, J., Sun, J., Hu, J., Lin, K., Zhao, L., Huang, M., Yuan, S., Qu, W., Wang, X., Lai, Y., Zhao, Y., Zhang, Y., Shi, Y., Chen, Y., Weng, Z., Meng, Z., Li, A., Kong, A., Dong, B., Wan, C., Wang, D., Qi, D., Li, D., Yu, E., Li, G., Yin, H., Zhou, H., Zhang, H., Yan, H., Zhou, H., Peng, H., Zhang, J., Lv, J., Fu, J., Cheng, J., Zhou, J., Yin, J., Xie, J., Wu, J., Zhang, J., Liu, J., Tan, K., Yan, K., Chen, L., Chen, L., Li, M., Zhao, Q., Sun, Q., Pang, S., Fan, S., Shang, S., Zhang, S., You, T., Ji, W.,

Xie, W., Yang, X., Hou, X., Jiao, X., Ren, X., Kong, X., Huang, X., Wu, X., Chen, X., Wang, X., Zhang, X., Wei, Y., Li, Y., Xu, Y., Shen, Y., Peng, Y., Peng, Y., Zhou, Y., Li, Y., Yang, Y., Zhang, Y., Xie, Z., Huang, Z., Lu, Z., Fan, Z., Cheng, Z., Jiang, D., Han, Q., Zhang, X., Zhu, Y., Ge, Z.: Step3-vl-10b technical report. arXiv preprint arXiv:2601.09668 (2026)
15. Jiang, K., Liu, Y., Chen, W., Luo, J., Chen, Z., Pan, L., Li, G., Lin, L.: Beyond the destination: A novel benchmark for exploration-aware embodied question answering. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9091–9101 (October 2025)
16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) **42**(4), 1–14 (July 2023)
17. Kerr, J., Hari, K., Weber, E., Kim, C.M., Yi, B., Bonnen, T., Goldberg, K., Kanazawa, A.: Eye, robot: Learning to look to act with a bc-rl perception-action loop. In: Proc. Conference on Robot Learning (CoRL) (2025)
18. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In: Proc. International Conference on Learning Representations (ICLR) (2026)
19. Lee, J., Miyanishi, T., Kurita, S., Sakamoto, K., Azuma, D., Matsuo, Y., Inoue, N.: Citynav: Language-goal aerial navigation dataset with geographic information. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5912–5922 (October 2025)
20. Li, G., Xu, J., Zhao, Y., Peng, Y.: Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9098–9108 (June 2025)
21. Li, K., Yao, L., Wu, J., Yu, T., Chen, J., Bai, H., Hou, L., Hong, L., Zhang, W., Zhang, N.L.: Insight-o3: Empowering multimodal foundation models with generalized visual search. In: Proc. International Conference on Learning Representations (ICLR) (2026)
22. Liu, S., Zhang, H., Qi, Y., Wang, P., Zhang, Y., Wu, Q.: Aerialvln: Vision-and-language navigation for uavs. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15384–15394 (October 2023)
23. Ma, M., Ma, Q., Li, Y., Cheng, J., Yang, R., Ren, B., Popovic, N., Wei, M., Sebe, N., Van Gool, L., et al.: Scenesplat++: A large dataset and comprehensive benchmark for language gaussian splatting. In: Proc. Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
24. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. In: Proc. International Conference on Learning Representations (ICLR) (2023)
25. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., Yadav, K., Li, Q., Newman, B., Sharma, M., Berges, V., Zhang, S., Agrawal, P., Bisk, Y., Batra, D., Kalakrishnan, M., Meier, F., Paxton, C., Sax, S., Rajeswaran, A.: Openeqa: Embodied question answering in the era of foundation models. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
26. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., van den Hengel, A.: Reverie: Remote embodied visual referring expression in real indoor environments. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

27. Sakamoto, K., Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Map-based modular approach for zero-shot embodied question answering. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2024)
28. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A platform for embodied ai research. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
29. Saxena, S., Buchanan, B., Paxton, C., Liu, P., Chen, B., Vaskevicius, N., Palmieri, L., Francis, J., Kroemer, O.: Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. In: Proc. Conference on Robot Learning (CoRL) (2025)
30. Scarpellini, G., Rosa, S., Morerio, P., Natale, L., Bue, A.D.: Look around and learn: self-improving object detection by exploration. In: Proc. European Conference on Computer Vision (ECCV) (2024)
31. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
32. Su, A., Wang, H., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. In: Proc. Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
33. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes. Psychological Review (2006)
34. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
35. Wijmans, E., Datta, S., Maksymets, O., Das, A., Gkioxari, G., Lee, S., Essa, I., Parikh, D., Batra, D.: Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6659–6668 (2019)
36. Wolfe, J.M.: Visual search: How do we find what we are looking for? Annual Review of Vision Science **6**, 539–562 (Sep 2020)
37. Wolfe, J.M., Vo, M.L.H., Evans, K.K., Greene, M.R.: Visual search in scenes involves selective and nonselective pathways. Trends in Cognitive Sciences (2011)
38. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13084–13094 (June 2024)
39. Wu, Y., Wu, Y., Gkioxari, G., Tian, Y.: Building generalizable agents with a realistic and rich 3d environment. arXiv preprint arXiv:1801.02209 (2018)
40. Xiong, H., Xu, X., Wu, J., Hou, Y., Bohg, J., Song, S.: Vision in action: Learning active perception from human demonstrations. In: Proc. Conference on Robot Learning (CoRL) (2025)
41. Yang, J., Ren, Z., Xu, M., Chen, X., Crandall, D.J., Parikh, D., Batra, D.: Embodied amodal recognition: Learning to move to perceive objects. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2040–2050 (2019)
42. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12–22 (2023)
43. Yin, T., Mei, Z., Sun, T., Zha, L., Zhou, E., Bao, J., Yamane, M., Shorinwa, O., Majumdar, A.: Womap: World models for embodied open-vocabulary object localization. In: Proc. Conference on Robot Learning (CoRL) (2025)

44. Yu, H., Han, Y., Zhang, X., Yin, B., Chang, B., Han, X., Liu, X., Zhang, J., Pavone, M., Feng, C., Xie, S., Li, Y.: Thinking in 360°: Humanoid visual search in the wild. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2026)
45. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., et al.: Thyme: Think beyond images. arXiv preprint arXiv:2508.11630 (2025)