Prune-Then-Plan: Step-Level Calibration for Stable Frontier Exploration in Embodied Question Answering

Supplementary Material

1. Overview

We provide the following investigations and results as part of our supplementary materials:

- **SPL Gain Analysis** (Sec 2): An analysis on the SPL gains with respect to question required travel distance and path complexity.
- **Question Category Performance** (Sec 3): Comparison of performance in question categories between different exploration methods.
- **Oscillation Rate Metric** (Sec 4): A metric comparing oscillation behavior between our method and 3D-Mem.
- **Impact of Labeling Noise** (Sec 5): An examination of how annotation noise affects downstream performance of our method.
- **Additional Path Comparisons** (Sec 6): Further visual comparisons of navigation paths generated by our method compared to baselines.

2. SPL Performance Analysis

The SPL gains on EXPRESS-Bench are smaller due to the following reasons: OpenEQA contains *no* questions where the answer is within <10m, whereas ∼80% of questions in EXPRESS-Bench are in the 0–10m range, where our performance is close to 3D-Mem. Our improvements are heavily concentrated in longer-distance episodes (10–20m), where frontier selection and exploration strategies matter more compared to short distance episodes. In addition, our gains are diluted by the large fraction of questions where *both* methods score zero SPL (timeout or wrong answer). To show this, we report the SPL metric separately for different distance to answer ranges in Table 1, computed over the subset where at least one method achieves non-zero SPL; this highlights the substantially larger absolute gains on longer paths.

Dist (m)	#Q (Total)	# A	3D-Mem	Ours
0–5	755	354	**49.2**	48.9
5–10	681	283	**54.0**	52.8
10–20	342	96	46.5	**52.6**

Table 1. We report the average SPL for different distance ranges on ExpressBench, total number of questions per distance range (#Q Total) and questions answered by at least one method (# A).

Performance Overhead Analysis. We additionally measured the wall clock step time for both 3D-Mem and our method (built on top of 3D-Mem). On average, the addition of the pruning increases per-step inference time by 6.6%. Pruning requires only one additional VLM call that reads token logits instead of decoding a full output sequence, making it substantially cheaper than other VLM calls.

	Exploration Strategy		
Question Category	VLM	Closest	Ours
Object	**29.32**	27.23	28.09
Existence	46.41	39.61	**48.01**
Attribute	42.33	41.43	**42.97**
Location	23.51	24.40	**27.17**
State	42.63	**46.62**	45.97
Counting	37.39	35.99	**37.58**
Knowledge	39.66	40.34	**44.83**
Overall	37.82	36.93	**39.54**

Table 2. Visually grounded LLM-Match for each question category in Express-Bench comparing different EQA exploration approaches, VLM-only, closest-frontier, and ours.

3. Question Category Performance

To evaluate the effect of our Prune-Then-Plan based exploration strategy, we run ablations on EXPRESS-Bench using three methods: *(i) a VLM-only planner* that selects each step purely from VLM scores over frontiers, *(ii) a closest-frontier policy* that always moves to the nearest frontier, and *(iii) our approach*, which applies pruning followed by a closest-frontier planner. Our method consistently outperforms the baselines, with notable gains in 'existence', 'location', and 'knowledge' questions seen in Table 2. While a closest-frontier strategy can suffice in small scenes with few options, it can easily trap agents in irrelevant directions in larger environments, reducing path efficiency. In contrast, our approach is both coverage-aware and semantically guided, balancing exploration with efficiency.

4. Oscillation Rate

We compute oscillation rate as an additional metric to measure the improvement of our method over the 3D-Mem baseline in Table 3. Oscillation rate counts the percent of total steps in which the agent chooses among the same set of frontiers. When an agent oscillates it often is not able to reach and update frontiers thus the agent considers the same set of frontiers across many steps. High oscillation rate indicates the agent is making slower progress under the same

step budget and often reconsidering the same frontiers and not reaching any of them. We can only measure this metric for 3D-Mem since Fine-EQA does not directly build a global frontier set to pick from, rather it uses the semantic map and the current view to provide a step specific set of frontier options. Additionally, Fine-EQA forces the agent to reach the selected frontier before determining where to go next, unlike 3D-Mem which chooses every step regardless of whether it has reached the last chosen frontier. The results in Table 3 further highlight behavioral differences between our method and the 3D-Mem baseline. The substantially lower oscillation rate achieved by our agent (**0.28** vs 0.42) indicates that it spends far fewer steps reconsidering the same frontiers across consecutive actions.

Metric	3D-Mem	Ours
Oscillation rate ↓	0.42	**0.28**

Table 3. Additional metrics for describing the agents exploration behavior. Oscillation rate is number of steps where the agent reconsiders the same set of frontiers. We report on Express-Bench.

5. Annotation Noise

Our calibration procedure depends on annotator labels on the data, to understand the impact of this label quality on the final downstream performance, we perform experiments with varying label noise. In total we have approximately 5,500 labeled frontiers of which roughly 3,600 are considered bad. To inject noise, we invert the labels for a random x% of the 5,500 labeled frontiers. The noisy data is used to construct empirical distribution functions which are then used to run evaluation on OpenEQA to understand the impact of label quality on downstream performance. We evaluate under varying levels of noise and report the results below.

Method	LLM-Match/EAC ↑
3D-Mem	29.2
Ours 0% noise	**34.3**
Ours 5% noise	**34.3**
Ours 10% noise	32.3

Table 4. Comparison of our method to 3D-Mem on OpenEQA under varying levels of noise in calibration data labeling.

The results in Table 4 highlight an important practical observation: our calibration procedure does not require perfectly clean annotations to remain effective. Even when we intentionally introduce noise - by inverting 5% of all frontier labels - the overall LLM-Match/EAC performance remains unchanged at **34.3**, the same as the noise-free setting. This shows our method is robust even with small errors in the data labeling. When the noise level is doubled to 10% do we observe a modest drop in accuracy (to 32.3), yet the performance still remains higher than the 3D-Mem baseline (29.2).

6. Additional Path Comparisons

We show additional comparisons of agent exploration paths between our method, 3D-Mem, and Fine-EQA in Figs. 1,2,3,4. We observe that although the methods may end up with similar final snapshots and answers, our method can significantly improve the exploration efficiency. We also note that due to its exploration policy, Fine-EQA paths can appear to warp through objects as the agent completely reaches the frontier before considering its next choice regardless of distance. This is in contrast with 3D-Mem which logs the position at each step which is a fixed distance in meters, allowing for a smoother and more realistic looking path visual.

Figure 1. Comparison of our method with 3D-Mem and Fine-EQA across three embodied question-answering scenarios. In the top row, both our method and 3D-Mem answer correctly by locating the correct viewpoint, but our agent reaches the target chair more efficiently, while Fine-EQA explores locally and produces an incorrect, ungrounded response. In the middle row, the question concerns the number of plants in the entryway. Again, both our method and 3D-Mem identify the single visible plant and answer correctly, but 3D-Mem takes a substantially longer and less direct route. Fine-EQA, due to highly inefficient exploration behavior, overestimates the number of plants and fails to visually ground its answer. In the bottom row, the agent must determine whether the door to the living room is closed. Our method efficiently reaches a clear vantage point and answers correctly. In contrast, 3D-Mem ends at an incorrect pose and misclassifies the door state, while Fine-EQA produces an ungrounded and incorrect answer after extensive, inefficient wandering. Across all three cases, the efficiency of our exploration directly contributes to both accurate and visually grounded answers, whereas baseline methods suffer from inefficient trajectories that lead to incorrect or ungrounded predictions.

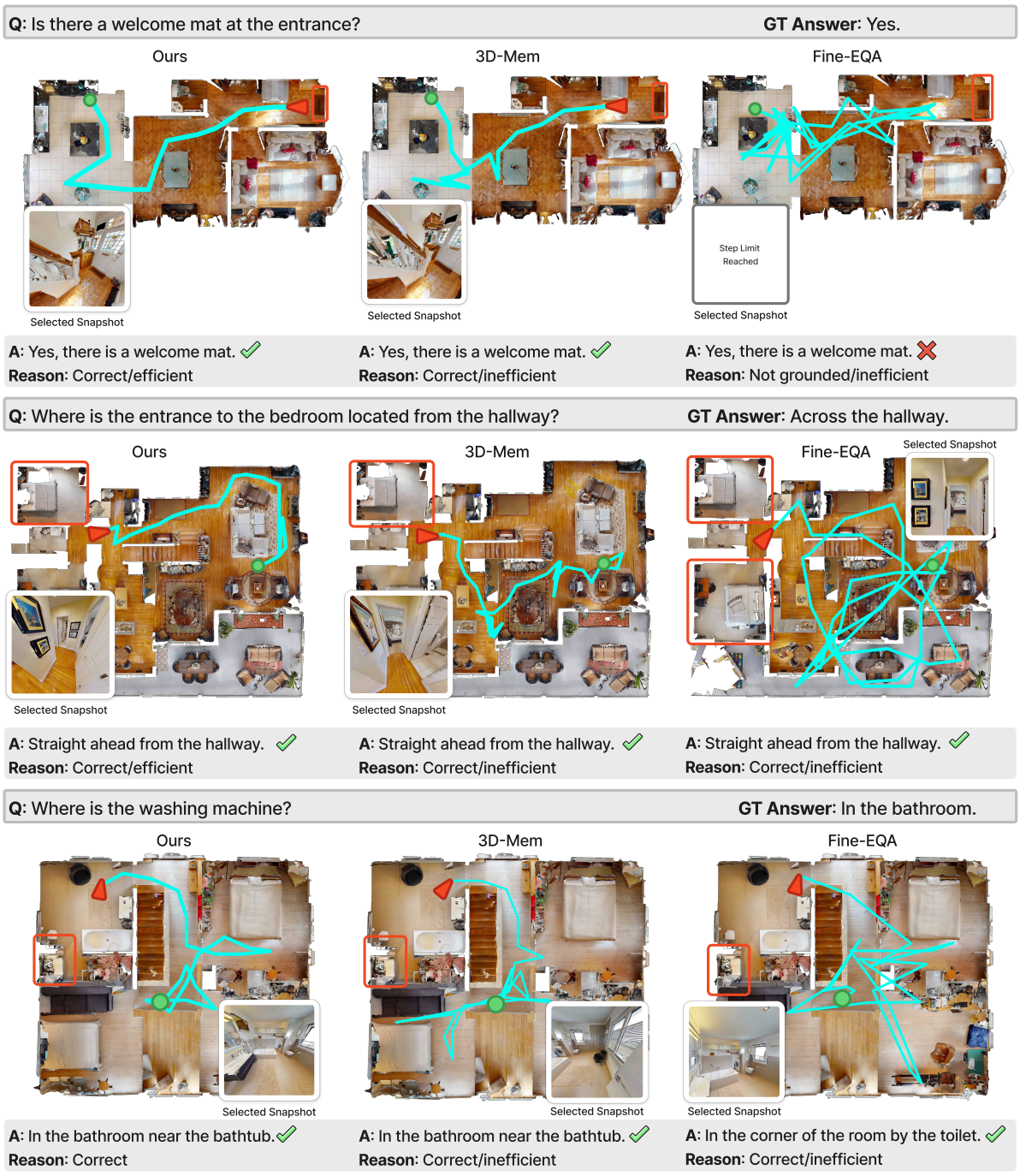

Figure 2. Evaluation of our method, 3D-Mem, and Fine-EQA on three additional navigation-and-grounding queries. In the top row, the agent must verify the presence of a welcome mat at the entrance. All methods ultimately respond "yes," but Fine-EQA's scattered trajectory prevents it from grounding its answer in the selected snapshot, whereas our method identifies the mat through a concise and direct route. In the middle row, the question asks for the location of the bedroom entrance relative to the hallway. All three methods reach the correct interpretation, but 3D-Mem and Fine-EQA require significantly more movement through the scene, reflecting weaker spatial efficiency in locating simple relational cues. In the bottom row, the agent searches for the washing machine. While all approaches arrive at the correct answer, 3D-Mem again follows a lengthier, less targeted path, and Fine-EQA briefly diverges into irrelevant regions before finding the correct viewpoint. These examples highlight how even in cases where baselines eventually answer correctly, inefficient scene traversal leads to longer trajectories and less reliable grounding.

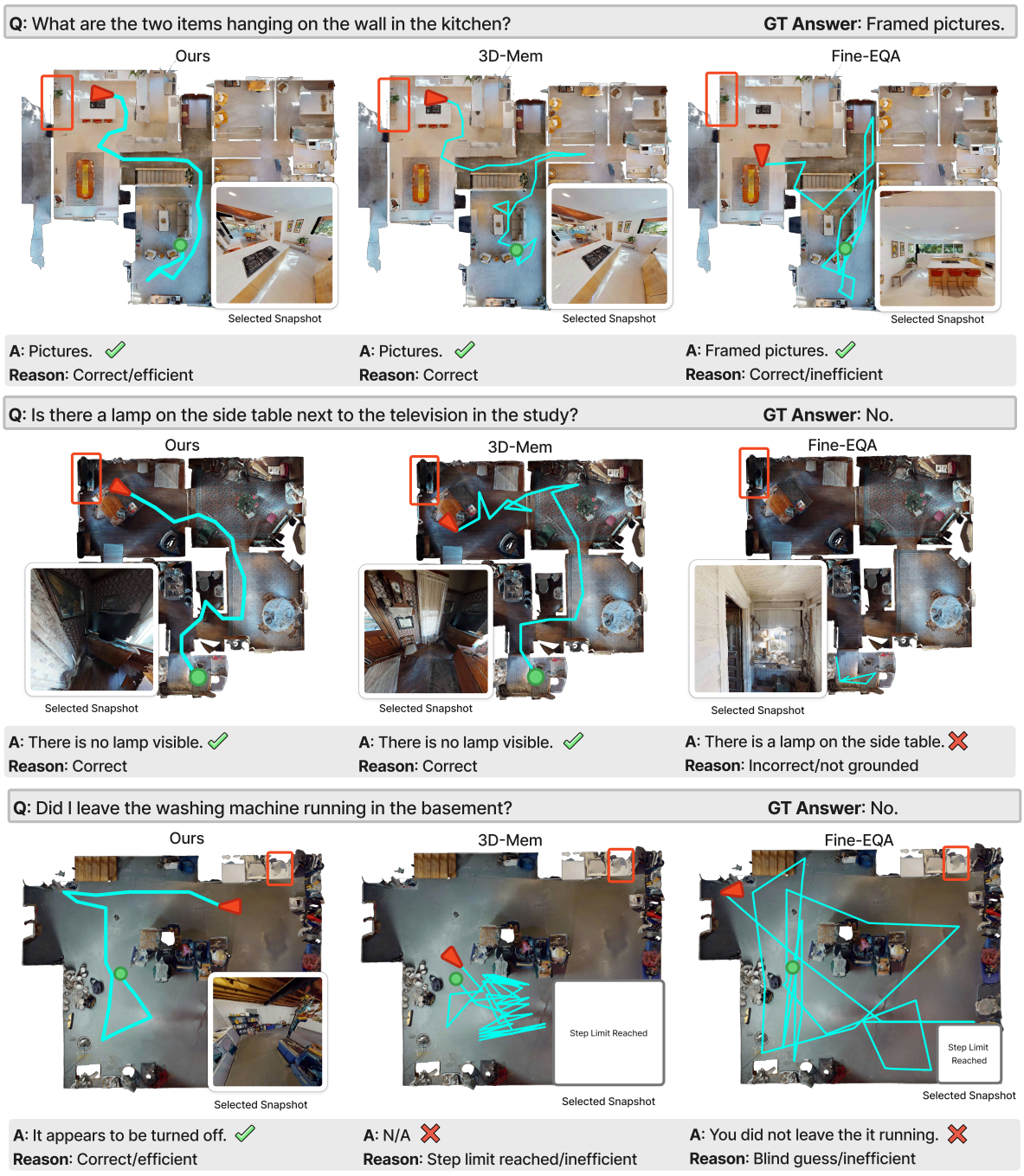

Figure 3. Additional qualitative comparisons illustrating how differing exploration behaviors affect correctness and grounding across three scenarios. In the top row, the agent must identify the two items hanging on the kitchen wall. All methods answer correctly, but Fine-EQA's wandering trajectory again yields a less grounded snapshot, whereas our method isolates the framed pictures with a concise, targeted sweep. In the middle row, the question concerns the presence of a lamp on the side table in the study. Both our method and 3D-Mem correctly determine that no lamp is present, while Fine-EQA incorrectly reports one due to its failure to reach a vantage point covering the side table. In the bottom row, the agent must verify whether the washing machine in the basement was left running. Our method navigates directly to the appliance and confirms that it is off. By contrast, 3D-Mem hits its step limit before reaching a meaningful viewpoint, and Fine-EQA produces an ungrounded guess after an erratic trajectory. These examples further emphasize how precise, question-aware exploration yields reliable and visually supported answers, whereas inefficient motion patterns frequently lead baseline methods to incomplete or incorrect conclusions.

Figure 4. Qualitative comparison on two additional queries highlighting how exploration reliability impacts correctness in lower-visibility scenarios. In the top row, the agent must determine whether any sections of the bathroom wall differ in texture. Our method successfully inspects the relevant surfaces and confirms consistent texture, while 3D-Mem fails to reach the bathroom before exhausting its step budget, and Fine-EQA incorrectly reports texture differences after a diffuse, poorly targeted trajectory. In the bottom row, the agent is asked what to do to grab a snack. Our method efficiently reaches the kitchen area and produces the correct, grounded answer. By contrast, both baselines terminate prematurely due to inefficient motion; 3D-Mem never reaches the kitchen region, and Fine-EQA issues an ungrounded guess after wandering extensively. Together, these examples illustrate how question-guided, deliberate navigation remains crucial for answering even simple queries when visual evidence lies behind occlusions or requires scene-level context.