INDIAN CONTROL CONFERENCE
18-20 Dec 2025 • IISc Bengaluru

OpenVLM-Nav: Training-Free Zero Shot Object-Goal Navigation via Vision-Language Guidance

Athira Krishnan R.   Swapnil Bag   Sumohana S. Channappayya

Indian Institute of Technology Hyderabad

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Abstract

We propose **OpenVLM-Nav**, a **training-free framework for zero-shot object-goal navigation** using **open-source vision–language models**. Using , **BLIP**, and **Qwen3-VL-2B**, the agent interprets object descriptions directly from images without task-specific training. Qwen3-VL-2B performs the best, and we further study two extensions: a **history module** for temporal context and a **depth module** for geometric cues. Depth provides the most significant gain, improving the **Success Rate (SR) from 0.08 to 0.14** and reducing the **Distance-to-Goal (DTG) from 7.824 to 7.567**. History gives minor, but consistent improvements. These results demonstrate that simple, training-free VLM-based navigation can be enhanced through the incorporation of temporal reasoning and depth information.

Problem Setting

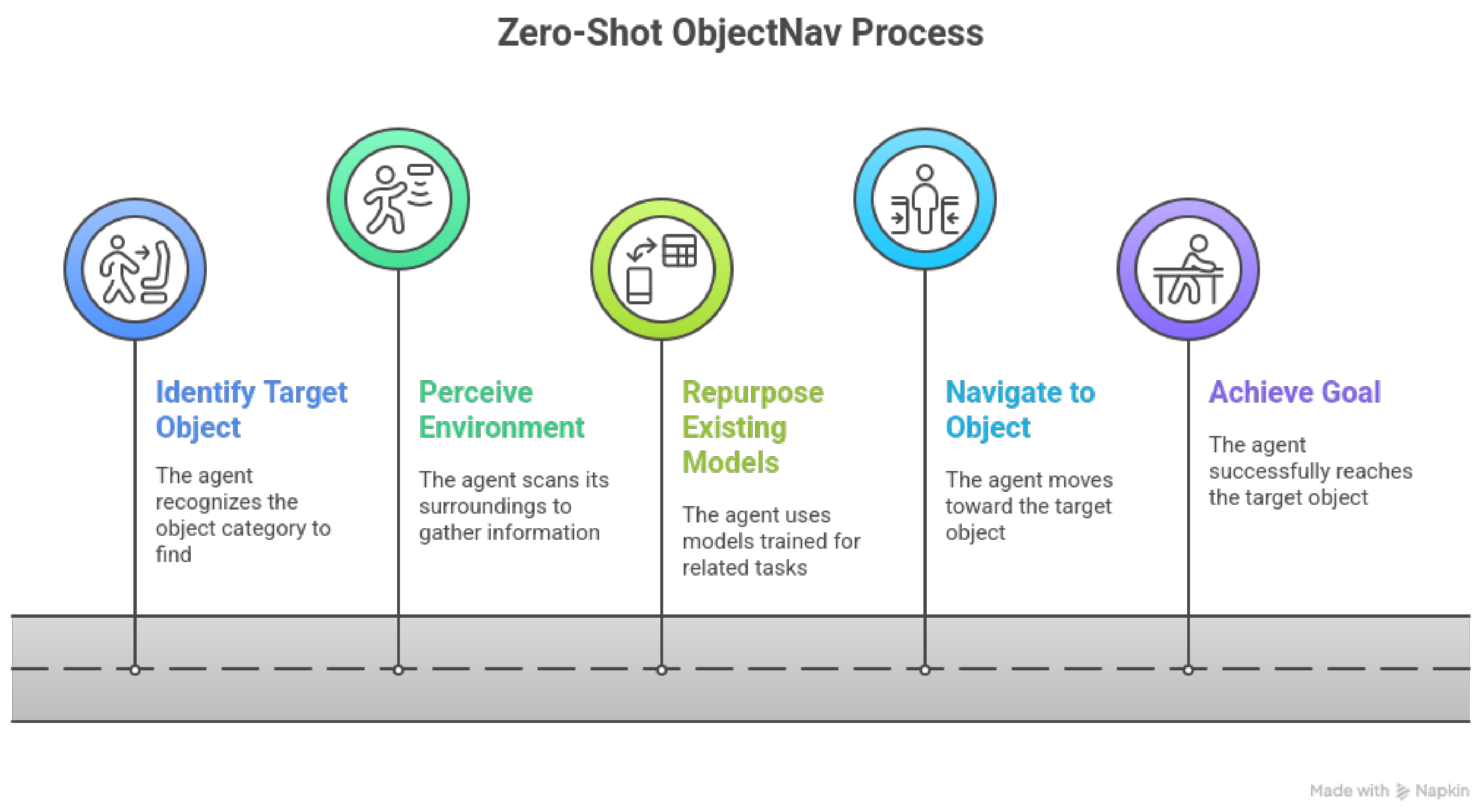

Zero-Shot ObjectNav Process
Identify Target Object
The agent recognizes the object category to find
Perceive Environment
The agent scans its surroundings to gather information
Repurpose Existing Models
The agent uses models trained for related tasks
Navigate to Object
The agent moves toward the target object
Achieve Goal
The agent successfully reaches the target object
Made with Napkin

Figure 1. Steps in Zero-shot object goal navigation. Here, the agent must perceive its environment and move toward a target object category, such as "find a chair."

Our Training-Free Open-Source Approach: We explore whether open-source VLMs can replace proprietary models for scene understanding and goal reasoning. Our experiments on the MP3D dataset within the Habitat simulator evaluate these models in a fully training-free setting, and study how history and depth cues influence decision-making in embodied navigation.

Proposed Methodology

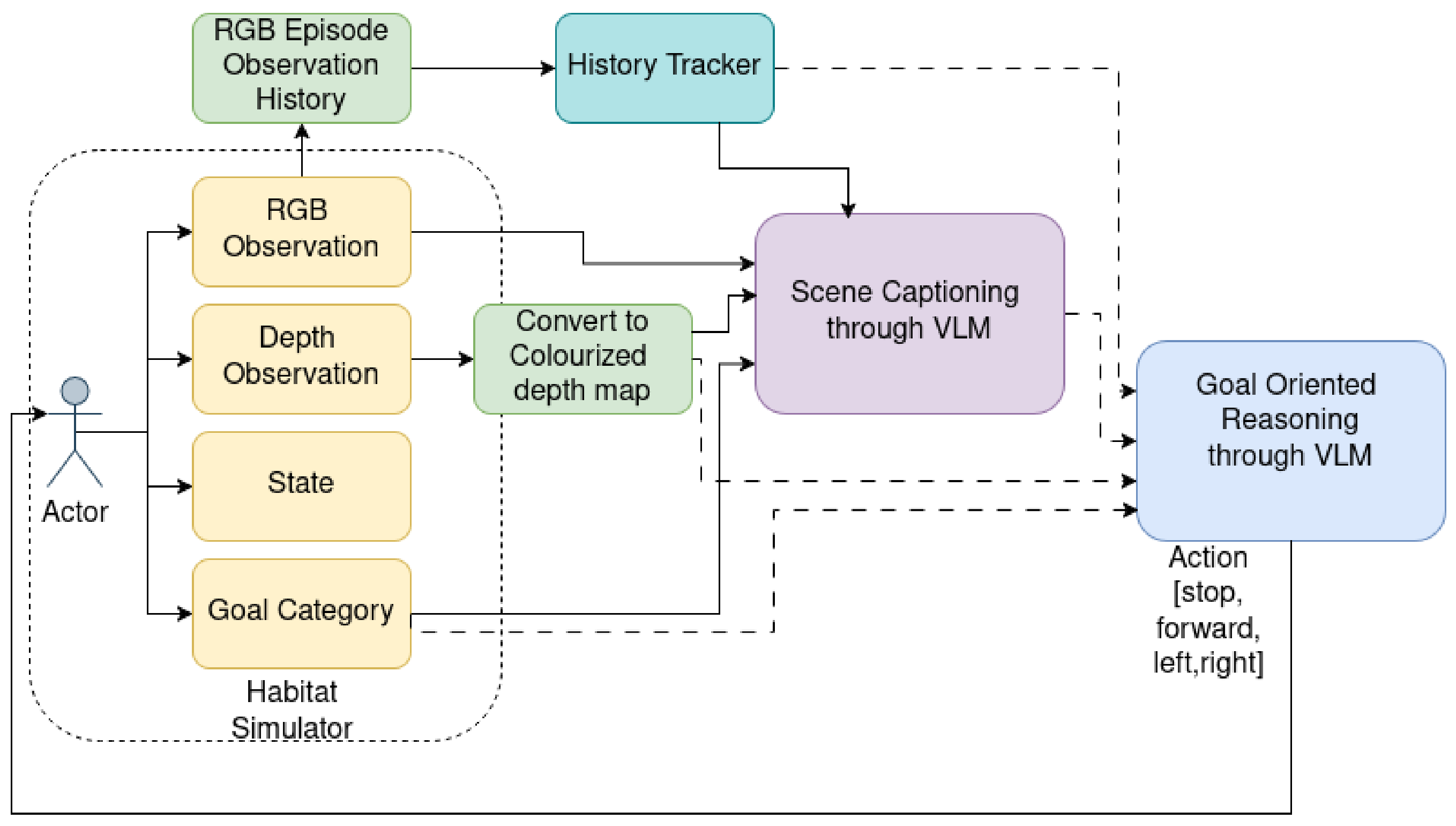
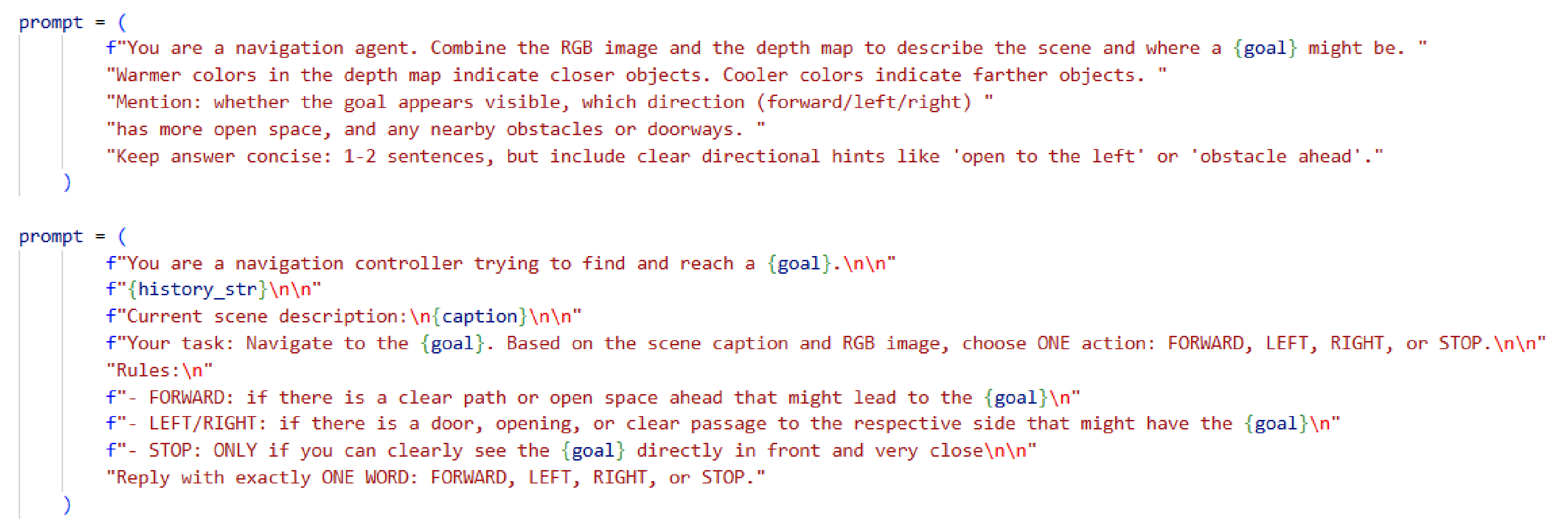

RGB Episode Observation History
History Tracker
RGB Observation
Depth Observation
Convert to Colourized depth map
Scene Captioning through VLM
Goal Oriented Reasoning through VLM
State
Goal Category
Actor
Habitat Simulator
Action [stop, forward, left,right]

Figure 2. The proposed methodology is depicted here. The Observations from the actor in the habitat simulator are preprocessed and passed to the Scene Captioning module and the Goal Reasoning module, along with the history and goal category as prompts. The directions returned by the agent are passed to the agent to act. State is used to compute the metrics.

We employ a two-stage prompting pipeline for zero-shot object navigation using a VLM. First, the model generates a structured spatial caption that describes objects, free space, and navigable cues (left/center/right). This caption, together with the raw image and goal category, is used in a second prompt that constrains the model to select a discrete action (FORWARD, LEFT, RIGHT, STOP).



```
prompt = (
        f"You are a navigation agent. Combine the RGB image and the depth map to describe the scene and where a {goal} might be. "
        "Warmer colors in the depth map indicate closer objects. Cooler colors indicate farther objects. "
        "Mention: whether the goal appears visible, which direction (forward/left/right) "
        "has more open space, and any nearby obstacles or doorways. "
        "Keep answer concise: 1-2 sentences, but include clear directional hints like 'open to the left' or 'obstacle ahead'."
    )

prompt = (
        f"You are a navigation controller trying to find and reach a {goal}.\n\n"
        f"{history_str}\n\n"
        f"Current scene description:\n{caption}\n\n"
        f"Your task: Navigate to the {goal}. Based on the scene caption and RGB image, choose ONE action: FORWARD, LEFT, RIGHT, or STOP.\n\n"
        "Rules:\n"
        f"- FORWARD: if there is a clear path or open space ahead that might lead to the {goal}\n"
        f"- LEFT/RIGHT: if there is a door, opening, or clear passage to the respective side that might have the {goal}\n"
        f"- STOP: ONLY if you can clearly see the {goal} directly in front and very close\n\n"
        "Reply with exactly ONE WORD: FORWARD, LEFT, RIGHT, or STOP."
    )
```

Figure 3. Prompt templates for the VLM: (upper prompt is used to generate scene captions, while the lower prompt is used to predict navigation actions)

- **Depth**: We provide a depth heatmap where warm colors indicate proximity and cool colors indicate free space, enabling the VLM to understand spatial layout, avoid obstacles, and choose safer navigation actions.
- **History**:We feed recent actions and scene captions back to the VLM, giving it short-term temporal context so it can reason about how it has been moving, not just what it sees in the current frame.

Results

Model Type	Model	SR (↑)	SPL (↑)	DTG (↓)
Models trained on embodied navigation task	ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings [5]	0.1467	0.0637	4.68
	RGB PixNav with BLIP [10]	0.0526	0.0312	5.03
	Depth PixNav with BLIP [10]	0.0421	0.0059	4.84
	Ensemble (RGB + Depth) PixNav with BLIP [10]	0.0600	0.00275	7.03
Open-source VLMs (no domain knowledge)	CLIP + Phi-3	0	0	9.484
	BLIP	0	0	7.990
	Qwen3-VL-2B	0.08	0.061	7.824
Effect of depth and history on VLMs	Qwen3-VL-2B with history	0.10	**0.0867**	7.728
	Qwen3-VL-2B with depth	**0.14**	0.0755	**7.567**
	Qwen3-VL-2B with history + depth	0.14	0.0556	7.569

Figure 4. ZSON performance on MP3D Dataset

- Results from a few models trained on embodied navigation, which were extended to zero-shot object goal navigation, are compared.From the numbers, it is evident that Qwen3 achieves a competitive zero-shot baseline on ZSON.
- Across settings, adding depth and/or history improves navigation performance. Depth provides the highest performance gains in terms of SR, whereas adding history alone yields moderate improvements, and combining both maintains the high SR. Depth forces the VLM to make an informed trade-off between describing the scene and avoiding obstacles.
- The low SPL in the case of added depth and history modality is mostly due to confusion when making an informed choice with a lot of prior information.

Sample Navigation Trajectory

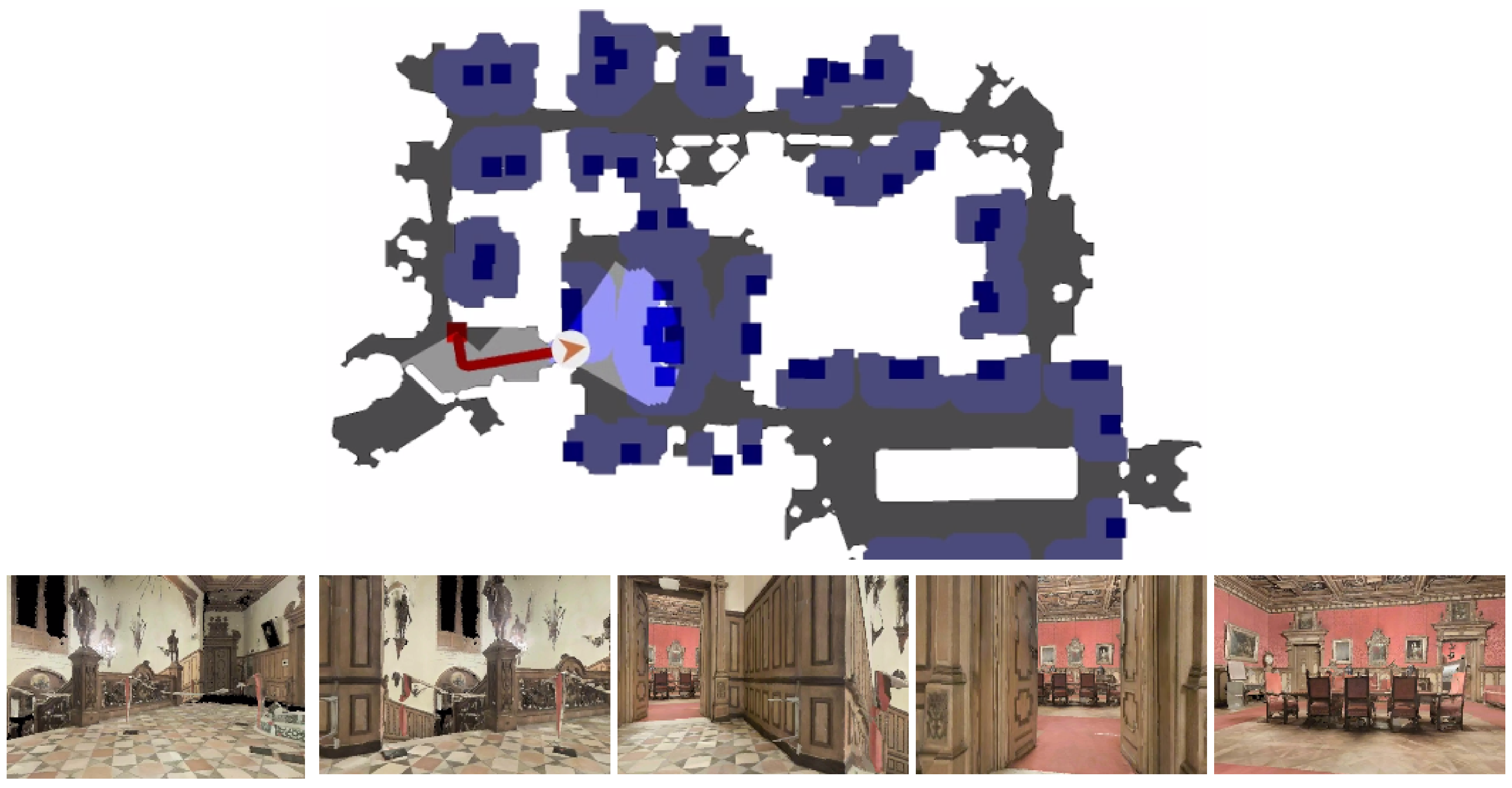


Figure 5. Top down map view (top) and the observations seen by the agent (bottom) during a successful episode of locating a chair

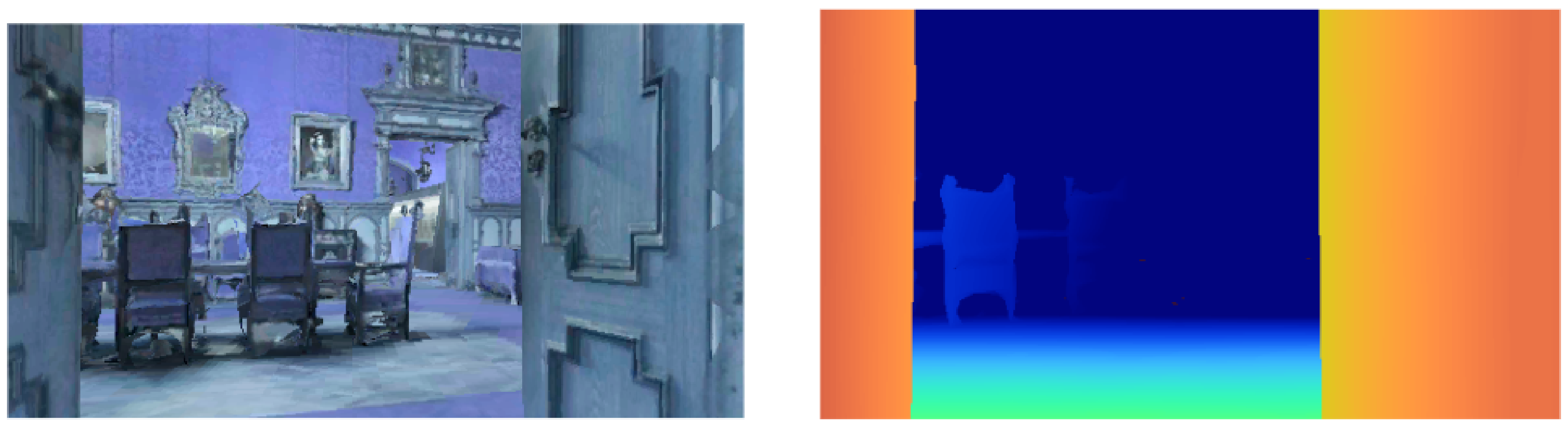


Figure 6. RGB observation and corresponding RGB depth map

Key Takeaways

Our experiments demonstrate that open-source VLMs trained without domain knowledge can serve as viable agents for zero-shot object goal navigation. With Qwen3 and depth, we got an **inference time of ≈200s** for an episode with around 80 steps. This shows that the approach is **low-cost and lightweight**, making it feasible for **local inference and edge-computing deployments**. Minor modifications to the prompts lead to substantial changes in performance. This highlights the importance of prompt design and suggests a clear direction for future work on **prompt tuning**.