Where Am I? Exploring the Situational Awareness Capability of Vision-Language Models in Vision-and-Language Navigation

Anonymous ACL submission

Abstract

Intuitively, it is important for humans to localize themselves by understanding their surroundings when navigating to a place, especially when the trajectory is long and complex. Similarly, we believe that this kind of capability, which we call situational awareness, is also crucial for developing better navigation agents. This work aims to explore the situational awareness capability of current popular vision-language model (VLM) based navigation agents in the context of vision-and-language navigation (VLN). We contribute a new dataset, the *Situational Awareness Dataset (SAD)*, comprised of around 100K 360-degree panoramic images and corresponding instructions for this task. We then evaluate multiple prominent VLMs including OpenAI o1, GPT-4o, Gemini 2.0 Flash, and Qwen2.5-VL on this dataset. Our results show that the situational awareness capability of these models is far behind human performance, highlighting substantial opportunities for progress and enhancement in this field. We hope that this work will spark future research to improve navigation agents and VLMs, particularly in their ability to process panoramic image data effectively.

1 Introduction

Situational awareness is a broad concept referring to the capability of perception, comprehension, and projection of the elements in an environment (Endsley, 1995). This capability is crucial for effective decision-making in a variety of tasks, such as aviation and healthcare. Within the realm of vision-and-language navigation (VLN), we simplify this concept to denote an agent's capability to understand its current position based on the observations in the navigation. This understanding is typically the initial step for navigation agents in assessing their progress and making informed decisions. Although fundamental, achieving situational awareness still necessitates intricate spatial reasoning and a nuanced understanding of language.

Recent advancements in large-scale vision-language models (VLMs) have demonstrated great potential across various vision-and-language tasks. Applying these models to the task of vision-and-language navigation in continuous environments (i.e., VLN-CE task; Krantz et al., 2020) using zero-shot learning has been a burgeoning area of research. Despite this interest, the performance of VLMs in this domain still lags far behind the methods that employ supervised learning. For instance, the state-of-the-art VLM-based method, AO-Planner (Chen et al., 2024a), achieves a 22.4% success rate on the RxR-CE dataset (Ku et al., 2020), whereas the popular supervised learning based method ETPNav (An et al., 2024) achieves a 54.8% success rate. Several factors contribute to this performance gap, with the situational awareness capability of these models being a fundamental determinant of their navigation performance. However, research on this capability within the vision-and-language navigation field remains limited. One major obstacle is the scarcity of fine-grained annotated data that aligns navigation instructions with their corresponding observations in the ground-truth trajectories.

To address this issue, we introduce a new dataset, the *Situational Awareness Dataset (SAD)*, which encompasses 100,000 panoramic images paired with corresponding instructions designed to evaluate situational awareness capabilities (see Fig.1). This dataset is constructed utilizing the RxR-CE dataset through the Habitat simulator (Savva et al., 2019; Szot et al., 2021; Puig et al., 2023). The instructions in the dataset are available in three typologically diverse languages–English, Hindi, and Telugu–to facilitate the examination of capabilities within multilingual contexts. Incorporating 360-degree panoramic images, the dataset captures an agent's observational perspective during navigation activities. While this method intuitively enhances

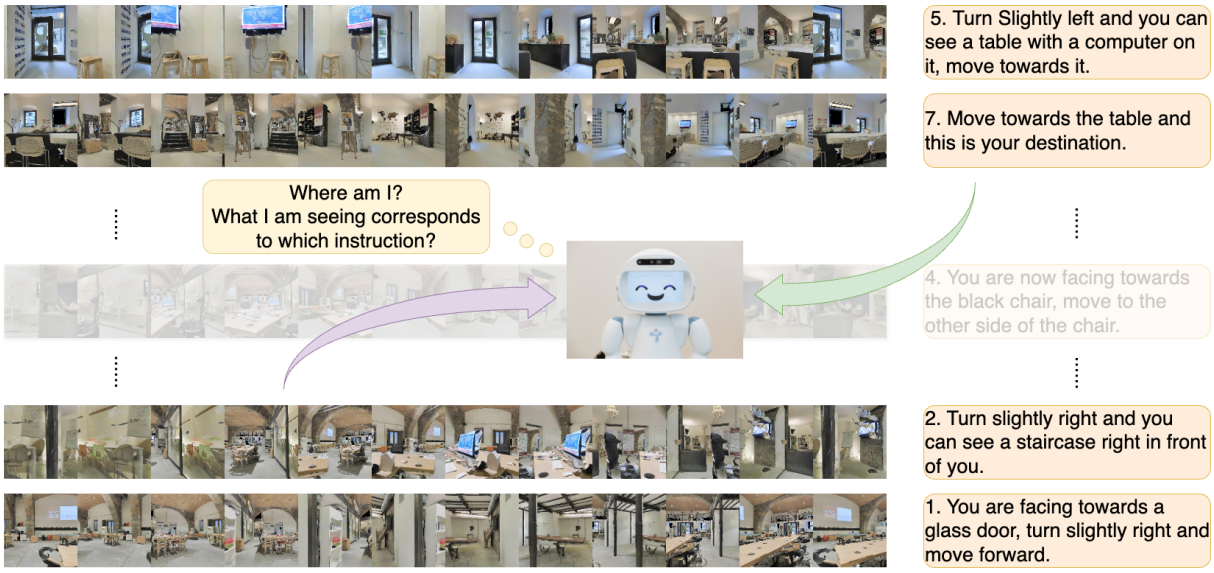

Figure 1: Example context that demonstrates the situational awareness task. The navigation agent takes as input a 360-degree panoramic image and the whole instruction. The agent is required to understand the surrounding observations and language instructions, then predict which sentence in the instruction the current observation corresponds to.

situational representation, it simultaneously introduces unique challenges for models, such as processing the extended field of view and managing significant overlaps within the images. Considering the limited availability of panoramic image datasets, SAD also stands to contribute to research advancements in this domain.

We conducted an evaluation of several prominent commercial and open-source vision-language models to assess their situational awareness capabilities using the SAD dataset. The models tested include OpenAI o1 (OpenAI, 2024b), GPT-4o (OpenAI, 2024a), Gemini 2.0 Flash (DeepMind, 2025), and Qwen2.5-VL-7B/72B-Instruct (QwenTeam, 2025). These models are good representatives of the current state-of-the-art in both commercial and open-source VLM fields. Our findings reveal that even the most advanced model, OpenAI o1, significantly trails human accuracy, achieving a performance of only 33% compared to humans' 87% (see §3.2). This highlights a substantial opportunity for enhancing performance in this area.

To summarize our contributions, we introduce the new task of situational awareness capability evaluation in vision-and-language navigation and contribute a corresponding dataset SAD. We then do a comprehensive evaluation of several most advanced VLMs on the proposed dataset. All datasets and evaluation codes are provided in the supplementary materials and will be made publicly available in the near future. We anticipate that our work will contribute to advancing research focused on improving navigation agents and the development of vision-language models, particularly in their ability to process panoramic images and perform spatial reasoning.

2 Dataset and Evaluation Method

To evaluate an agent's situational awareness capabilities, it is essential to have a dataset that aligns navigation instruction words with observations from corresponding positions. Regrettably, such a dataset does currently not exist. To address this gap, we have developed a dataset named the *Situational Awareness Dataset (SAD)* specifically for this purpose. In order to streamline the evaluation process, we concentrate on the alignment between instructions and observations at the sentence level. This focus means we only assess the correspondence between the conclusion of each instruction sentence and its associated observation.

2.1 Dataset Construction

We develop the Situational Awareness Dataset (SAD) using the Habitat simulator by leveraging the existing RxR-CE dataset. The RxR-CE dataset is a large-scale multilingual vision-and-language navigation resource featuring 126,000 navigation instructions and demonstrations within Matterport3D (Chang et al., 2017) and Habitat en-

Languages	Train	Val	Test
English	10,609	1,210	1,904
Hindi	1,642	202	381
Telugu	10,016	1,141	2,175

Table 1: The number of instances for the three languages in the proposed SAD dataset.

	# samples	ACC_INSTR	ACC_SENT
Human	200	65.00	87.14

Table 2: Human's performance (%) on our constructed SAD dataset.

vironments. To construct SAD, we utilize both the standard annotation task data and extended pose trace data from the RxR-CE dataset. The annotation task data includes essential components for VLN, such as navigation instructions and reference paths. It also provides a "timed_instruction" field, indicating the start and end times of words or phrases in alignment with the recording. The extended pose trace data offers snapshots detailing the virtual camera parameters and field-of-view from the annotators' perspectives.

We load this dataset into the Habitat simulator and calculate the camera poses and corresponding timestamps based on the supplied camera extrinsic matrix data. By extracting the timestamp of the concluding word in each instruction sentence from the "timed_instruction" data, we align these timestamps with the camera pose data, thereby obtaining the corresponding observations within the Habitat.

For each position's observation, we render a panoramic RGB image composed of 12 RGB sub-images captured from 12 different directions at equally spaced horizontal heading angles: $(0°, 30°, ..., 330°)$. These sub-images are generated in three resolutions: 224×224, 480×480, and 1024×1024. To simplify the task further, we limit our focus to instructions containing a maximum of 10 sentences. More detailed information about the dataset is provided in Table 1.

2.2 Evaluation Method

With the constructed dataset, we evaluate the situational awareness capability of agents through a straightforward question-answering task. Given an instruction and the corresponding panoramic observations, we pose the following question to the agent: "Which sentence in the instruction does this image correspond to the end of?" The agent must predict a list of sentence indices that align with each panoramic observation. Figure 1 provides an example of the task.

We utilize two metrics to assess the agent's performance on this task: (1) Instruction-Level Accuracy (ACC_INSTR): this metric considers a prediction correct if the agent's predicted list exactly matches the ground truth list; (2) Sentence-Level Accuracy (ACC_SENT): this metric evaluates accuracy based on individual sentences in the instruction. Each correct prediction associated with an image contributes to the overall accuracy. These criteria allow us to assess the effectiveness of the agent's situational awareness capabilities based on its ability to align instructions with observations.

2.3 Dataset Quality Evaluation

To ensure the quality of the constructed dataset, we conduct a human evaluation on the English subset. We randomly sample 200 instances from the dataset and have five individuals perform the same task as described in the previous subsection (§2.2). The results indicate an average instruction-level accuracy (ACC_INSTR) of 65% and a sentence-level accuracy (ACC_SENT) of 87%. These findings suggest that the dataset is of high quality and suitable for our proposed task, which involves evaluating the situational awareness capabilities of vision-language model-based agents.

3 Experiments

3.1 Evaluation Settings

Dataset We utilize our constructed Situational Awareness Dataset (SAD) for model evaluation. We test the models across three language splits: English, Hindi, and Telugu. For each example, we limit the number of images to a maximum of 10 and randomly shuffle the input images. Each panorama sub-image is evaluated at a resolution of 224×224. Our preliminary experiments with GPT-4o indicate that higher resolutions do not significantly enhance performance while substantially increasing test time. Further details are provided in Appendix A.1.

Test Models We evaluate the following models on the SAD dataset in a zero-shot setting. We run each model three times and report the average performance in each evaluation setting. All models employ the technique of structured outputs. Specifically, we force the model's output to include the

	English		Hindi		Telugu	
	ACC_INSTR	ACC_SENT	ACC_INSTR	ACC_SENT	ACC_INSTR	ACC_SENT
GPT-4o	6.36	26.74	4.29	25.55	8.15	27.76
OpenAI o1	**11.61**	**32.92**	**17.18**	**37.62**	**15.99**	**37.47**
Gemini 2.0 Flash	6.99	32.13	9.51	35.79	7.71	32.17
Qwen2.5-VL-7B-Instruct	2.84	18.25	4.29	20.94	3.97	21.53
Qwen2.5-VL-72B-Instruct	3.68	20.49	5.52	24.61	5.34	22.58

Table 3: Evaluation results for all the tested models on the SAD dataset. ACC_INSTR denotes the instruction-level accuracy, and ACC_SENT denotes the sentence-level accuracy. All the results are averaged over three runs and reported in percentage.

reasoning steps for each image along with the final answer, formatted in JSON. Further details about the prompts we use are provided in Appendix A.2.

(1) GPT-4o-2024-08-06 (OpenAI, 2024a), OpenAI's versatile flagship model that accepts input any combination of text, audio, image, and video.

(2) OpenAI o1-2024-12-17 (OpenAI, 2024b), OpenAI's reasoning model, trained with reinforcement learning and employing chain-of-thought to excel at complex reasoning tasks.

(3) Gemini 2.0 Flash (DeepMind, 2025), DeepMind's latest large language model, offering a 1 million token context window and built for the era of Agents.

(4) Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-72B-Instruct (QwenTeam, 2025), Qwen's latest open-source flagship vision-language models, capable of functioning as a visual agent and understanding long videos.

3.2 Evaluation Results

Table 3 presents the evaluation results of the tested models on the SAD dataset. The approximate accuracy estimates for random guesses are 0.02% and 14.29%, respectively.[1] In terms of exact match instruction-level accuracy (ACC_INSTR), all models underperform. Among them, OpenAI o1 emerges as the leader, outperforming others by approximately 50levels, while the open-sourced Qwen2.5-VL-7B/72B-Instruct models perform the poorest. This suggests that the OpenAI o1 model demonstrates a superior comprehensive reasoning capability in understanding complete trajectories compared to the other models. For sentence-level accuracy (ACC_SENT), OpenAI o1 once again achieves the highest performance, though Gemini 2.0 Flash closely follows. The Qwen2.5-VL-7B/72B-Instruct models still lag significantly behind, showing a marked gap with the other models. Furthermore, the evaluation across different language splits reveals no substantial performance differences, suggesting consistent model capabilities across various languages.

[1]These values are calculated as $1/7! \times 100\% \approx 0.02\%$ and $1/7 \times 100\% \approx 14.29\%$, where 7 represents the average number of images per example.

4 Related Work

Situational Awareness The concept of situational awareness is extensively studied in the field of cognitive science, psychology, human factors, aviation, healthcare, and more (Munir et al., 2022; Endsley, 2021; Stanton et al., 2001). Recently, Berglund et al. (2023) studies the emergence of situational awareness in large language models (LLMs). We further specify this concept in the context of VLN task in this work.

VLN with LLMs and VLMs The VLN task is a representative research topic in the field of embodied AI, and how to make use of LLMs and VLMs to solve this task has attracted much attention (Zhou et al., 2024; Chen et al., 2024b; Long et al., 2024; Zhang et al., 2024; Lin et al., 2024; Chen et al., 2023; Cai et al., 2024; Chen et al., 2024a). However, little work studies the fundamental situational awareness capability of these models. This work aims to explore this subject.

5 Conclusion

In this work, we examine the essential capability of situational awareness in VLM-based navigation agents within the VLN task. We introduce the SAD dataset and evaluate five leading VLMs using this dataset. Our findings indicate that the situational awareness capability of these models remains limited, potentially affecting their effectiveness in navigation tasks. We hope that our dataset and evaluation results will encourage future research aimed at developing improved navigation agents and VLMs.

6 Limitations

Our work has two main limitations. First, the format of the evaluation is a straightforward question-answering task, which is not able to be directly applied to evaluate the agents trained with supervised learning. Second, we did not check whether a VLM-based navigation agent's performance in the VLN-CE task can be improved by enhancing the situational awareness capability. We will add this experiment in the future.

Use of AI Assistance We used AI assistance tools (ChatGPT and GitHub Copilot) to aid in rewriting code and text. All AI-generated content was thoroughly reviewed and verified by the authors. AI was not used to generate new research ideas or original findings; rather, it served as a support tool to improve clarity, efficiency, and organization. In accordance with ACL guidelines, our use of AI aligns with permitted assistance categories, and we have transparently reported all relevant usage in this paper. While AI contributed to enhancing the quality of the work, no direct research outputs are the result of AI assistance.

References

Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. 2024. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. 2024. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5228–5234. IEEE.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*.

Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. 2024a. Affordances-oriented planning using foundation models for continuous vision-language navigation. *arXiv preprint arXiv:2407.05890*.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. 2024b. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9796–9810.

Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. 2023. A$\hat{2}$ nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. *arXiv preprint arXiv:2308.07997*.

DeepMind. 2025. Gemini 2.0 flash.

Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64.

Mica R Endsley. 2021. Situation awareness. *Handbook of human factors and ergonomics*, pages 434–455.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. *Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments*, page 104–120. Springer International Publishing.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. 2024. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *arXiv preprint arXiv:2403.07376*.

Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. 2024. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387. IEEE.

Arslan Munir, Alexander Aved, and Erik Blasch. 2022. Situational awareness: techniques, challenges, and prospects. *AI*, 3(1):55–77.

OpenAI. 2024a. Hello gpt-4o.

OpenAI. 2024b. Learning to reason with llms.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain,

Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots. *Preprint*, arXiv:2310.13724.

QwenTeam. 2025. Qwen2.5-vl.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Neville A Stanton, Peter RG Chambers, and John Piggott. 2001. Situational awareness and safety. *Safety science*, 39(3):189–204.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. 2024. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

A Experiments

A.1 Effects of Different Image Resolutions

We study the effects of different image resolutions on the performance of GPT-4o on our proposed SAD dataset. We evaluate the model on three different image resolutions: 224×224, 480×480, and 1024×1024. The results are shown in Table 4. We find that the higher resolutions do not bring significant improvement in the performance while significantly increasing the test time. Therefore, we use the image resolution of 224×224 for evaluation in the main experiments.

Image Resolution	ACC_Instr	ACC_Sent	Inf. Time
224×224	6.36	26.74	30min
480×480	7.36	26.78	52min
1024×1024	6.93	26.85	20.5h

Table 4: Effects of image resolutions on the performance of GPT-4o on our proposed SAD dataset.

A.2 Prompts

We present the prompts we use for GPT-4o in the following code snippet (see Listing 1). It contains the system prompt and the user prompt. We also use the technique of structured outputs to force the model to output the reasoning steps and answers in a json format. We use the same prompts for all the models we evaluate in this work.

```python
class SingleImageStep(pydantic.BaseModel):
    explanation: str
    answer: int


class SituationalAwarenessOutput(pydantic.BaseModel):
    number_of_input_images: int
    reasoning_steps: list[SingleImageStep]
    answer: list[int]

SYSTEM_PROMPT = inspect.cleandoc(
    """You are an agent navigating through a virtual environment according to
    the given instruction. But now your task is not to navigate, but to predict
    the positions of the given observation images in the corresponding
    instruction.  You would be given a set of images and an corresponding
    instruction.  The given images are the RGB {image_type} observation of your
    current position. Each panoramic image is comprised of 12
    sub-egocentric-images, where each sub-image corresponds to a different
    direction.  You need to think of where the position is in the instruction.
    The entire instruction is comprised of multiple sub-instructions.  Each
    sub-instruction starts with '#' followed by a number, which is the index of
    the sub-instruction.  Each position is the end of each sub-instruction.  So
    your task is to predict at the end of which sub-instruction you could see
    the current given image.  Note that the number of input images are strictly
    equal to the number of sub-instructions. Moreover, There will not be two
    images corresponding to the same position.  Your final answer should be a
    list of integers, where each integer represents that image's positions in
    the instruction.   For example, "[2, 3, 1, 4]" means you would observe the
    first input image at the end of the second sub-instruction, the second
    input image corresponds to the end of the third sub-instruction, the third
    input image corresponds to the end of the first sub-instruction, and the
    fourth input image corresponds to the end of the fourth sub-instruction.
    """
).replace("\n", " ")

USER_PROMPT = inspect.cleandoc(
    """Given the following {num_input_images} images, please predict their
    observation positions in the instruction.  The instruction is:
    {instruction_with_index}"""
).replace("\n", " ")


response = client.beta.chat.completions.parse(
    model=test_model,
    messages=[
        {
            "role": "system",
            "content": [
                {
                    "type": "text",
                    "text": SYSTEM_PROMPT.format(image_type=image_type),
                }
            ],
        },
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": USER_PROMPT.format(
                        num_input_images=len(multiple_images_input),
                        instruction_with_index=instruction_with_index,
                    ),
                }
            ]
            + multiple_images_input,
        },
    ],
    response_format=SituationalAwarenessOutput,
```

```
)
```

Listing 1: Prompts for GPT-4o and o1 models.