**Responses to Reviewer 1:**

*Thank you for reviewing our manuscript! Here are our responses to your questions:*

**Question 1 - Related Work:** The related work section is comprehensive and covers ground in both gas source localization (GSL) and large language models (LLMs). However, recent research on anemotactic methods—those that combine chemical sensing with wind flow information—should also be referenced. Anemotactic approaches have shown significant progress in using chemical sensors and anemometers to locate distant odor sources. For example, studies like *Towards Efficient Gas Leak Detection in Built Environments: Data-Driven Plume Modeling for Gas Sensing Robots* (ICRA, 2023) and *Robotic Gas Source Localization with Probabilistic Mapping and Online Dispersion Simulation* (IEEE Transactions on Robotics, 2024) should be discussed.

Additionally, anemotactic approaches are categorized under section 2.1, "olfaction-only," which is inaccurate, as they leverage wind flow data to guide robots, differentiating them from purely olfactory systems.

***Author’s response****:* *Thank you for your feedback!*

*We changed the section 2.1 title from ‘olfaction-only’ to ‘olfactory-based navigation’ algorithm and cited the following anemotactic methods in lines 115-120 of section 2.1.*

*[1] Jin, W., Rahbar, F., Ercolani, C., & Martinoli, A. (2023, May). Towards efficient gas leak detection in built environments: Data-driven plume modeling for gas sensing robots. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 7749-7755). IEEE.*

*[2] Ojeda, P., Monroy, J., & Gonzalez-Jimenez, J. (2024). Robotic gas source localization with probabilistic mapping and online dispersion simulation. IEEE Transactions on Robotics.*

**Question 2: Methodology**

**Sensing Module**: From this section, I see that the authors also rely on wind information. This was not stated before and should be mentioned as it is an important factor (related to the previous comments).

***Author’s response****:* *Thank you for your feedback!*

*In lines 121-122 of section 2.1 we added a statement to explain that our olfaction-based navigation strategy uses wind information.*

**Question 3:**  The **high-level reasoning section**, focuses on using sensor data to direct a multimodal LLM (the core of this work). However, several points require clarification:

* **Prompt Generation**: The authors use a prompt-generation process to translate sensor data into text that the LLM can interpret. However, questions arise about this process:
  + **System Prompt**: More detail on the “system-prompt” would be helpful. The article states it provides “reasoning objective, action selection, constraints, and expected output”. Not only a detailed description of each of these components is needed, but an example would make this easier to understand.

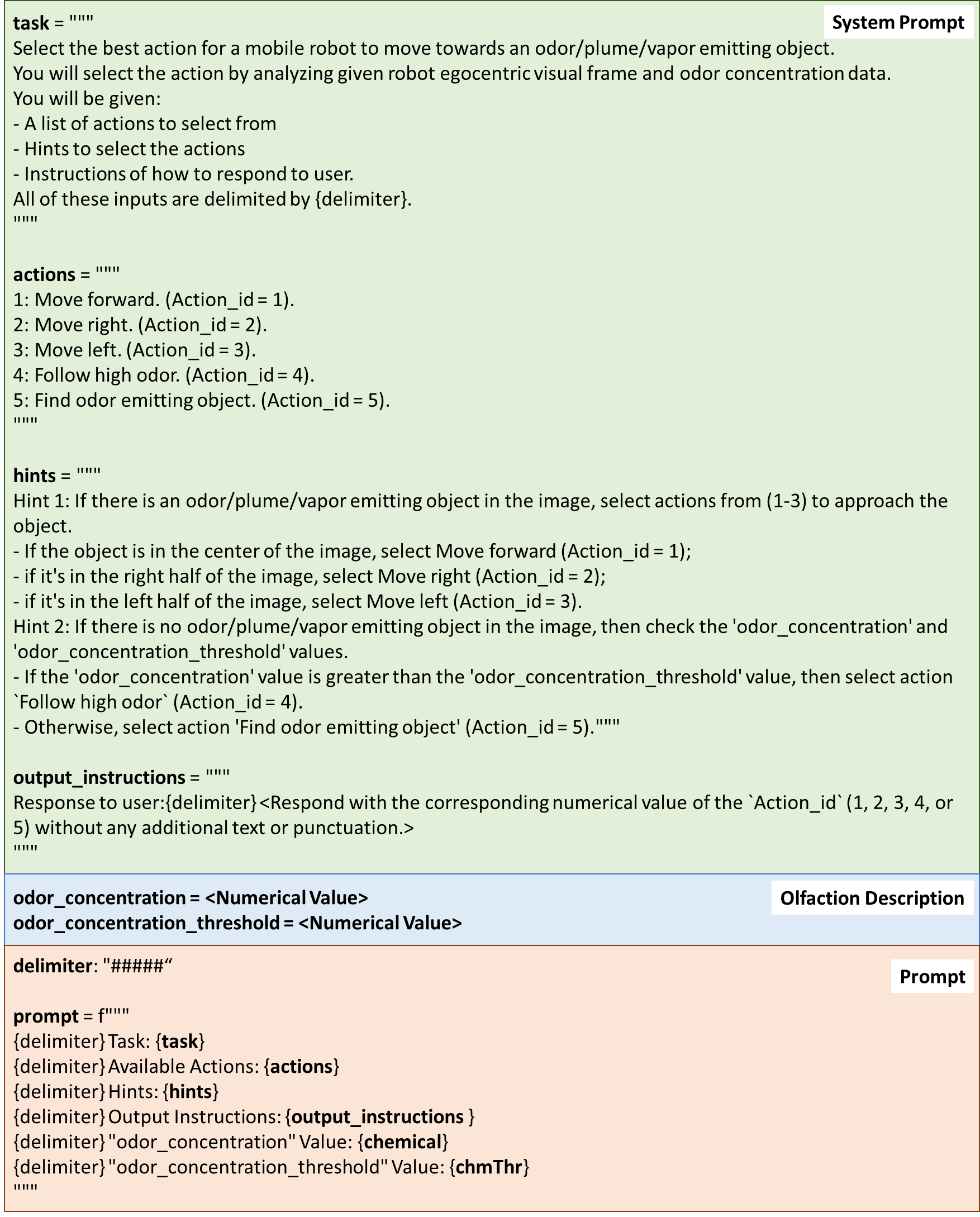
***Author’s response****:* *Thank you for your feedback! In response, we have added a Figure (as shown below) in subsection 3.3 detailing the implementation of the system prompt, olfaction description, and the final prompt.*

*Specifically, the system prompt includes:*

* ***Task****: Describes the objective and process the LLM should follow.*
* ***Actions****: Lists the actions available for the LLM to choose from.*
* ***Hints****: Guides the LLM to select appropriate vision-based or olfaction-based actions based on multi-modal reasoning.*
* ***Output Instruction****: Directs the LLM to generate only the action without additional reasoning.*

*The olfaction description includes the current odor concentration and threshold values. The final prompt integrates all these instructions.*

*We added this Figure as Figure 4 and added the description in lines 267-275 of section 3.3.*

**

**Question 4: Gas Concentration Translation**: How is the gas concentration communicated to the LLM? Is it in PPM values, or more descriptive terms like “high” or “low” concentration? This needs clarification. Again, an example would help.

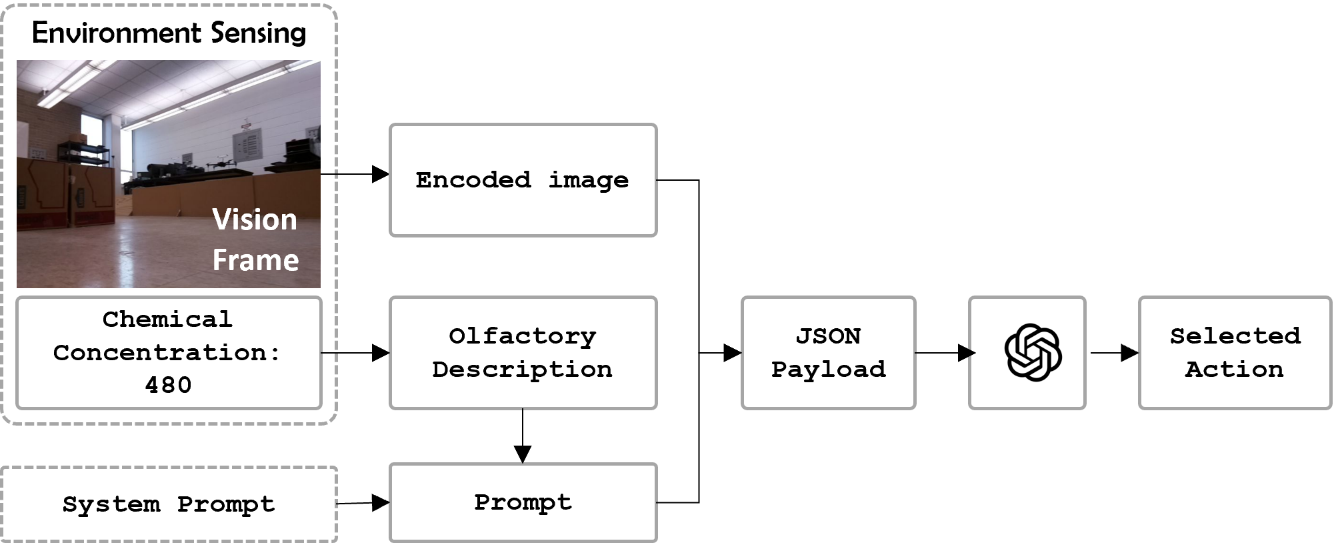
***Author’s response***: *Thank you for your feedback! The measured gas concentration is passed to the prompt as a numerical value. We added a sentence in lines 274-275 of section 3.3 to explain this question.*

**Question 5: Binary-to-Text Translation**: Since images are converted to text, the LLM does not directly analyze images, raising questions about the rationale behind not performing object detection or semantic analysis, which could enhance the model’s understanding. This important point must be clarified.

***Author’s response****:* *Thank you for your feedback!*

*The LLM analyzes the input image to determine the robot's next direction toward the odor source. In this work, we employed GPT-4 as the multi-modal LLM. By default, requests are sent to GPT-4 as a JSON payload, where the image input is encoded to text string using the default 'base64' function. Upon receiving the payload, GPT-4 then decodes this string back into an image format for processing. All these processes are encapsulated inside the GPT-4. That means that the GPT-4 is able to analyze the graphic information.*

*The Figure below shows the process of sending the request to the GPT-4 model.*

**

*To clarify this question, we added this Figure as Figure 5 and the added description in lines 276-281 of section 3.3.*

**Question 6: LLM Reasoning Stage**: The second stage appears to make a binary choice between vision-based and olfaction-based navigation, being the second one only chosen when there are no vision clues. The latter rises some questions:

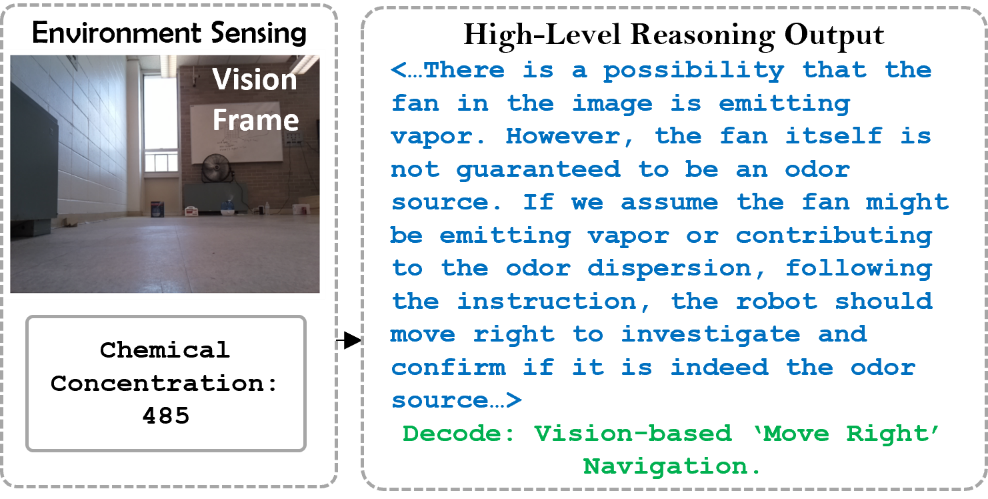
* There is no "fusion" of sensing modalities. It only selects vision based navigation, and in the absence of visual clues, then odor based navigation.

***Author’s response****:* *Thank you for your feedback! The proposed navigation strategy was developed based on human odor-search behaviors, where vision and olfaction are used* ***sequentially*** *rather than* ***simultaneously****. Upon detecting an odor, vision is used first to locate the odor source. If visual reasoning fails to identify the source, the olfaction-based approach is employed to guide the robot toward the odor source. To answer this question, we added a statement in lines 292-296 of section 3.3.*

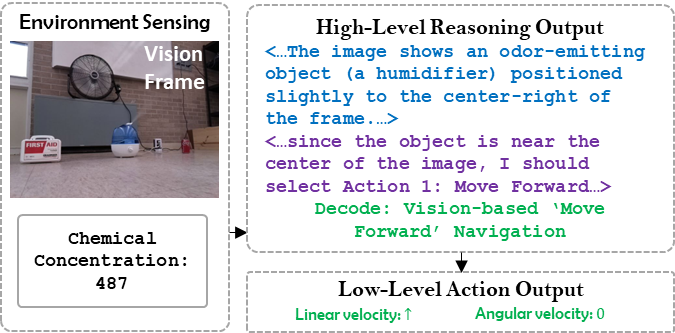
**Question 7:** It’s unclear if the LLM infers likely odor sources among visible objects, as suggested, or if it only tracks visible plumes, creating a discrepancy in the article's description.

***Author’s response****:* *Thank you for your feedback!*

*The LLM infers potential odor sources by reasoning over both object semantic definitions and visual cues. For instance, the image below shows that even when visible plumes are not detected, the LLM may deduce that approaching a fan could lead the robot closer to the odor source.*

**

*In other cases, like in the Figure below, when the LLM identifies a plume-emitting object (e.g., a humidifier), it directs the robot to approach it, as visible plumes are strong indicators of the odor source.*

**

*To explain this question, we added new frames Figure 11 and new sentences in lines 455-458 of section 4.4.*

**Question 8:** The LLM’s ability to infer navigation direction from a single image is also not well-explained. Information on the LLM’s training, how it evaluates multiple potential sources, and whether it provides confidence scores are necessary aspects to be commented.

How does the proposed method behave when one image displays multiple "potential source candidates"?

***Author’s response***: *Thank you for your feedback! In this project, the LLM is instructed to navigate towards a single odor source in zero-shot manner, no model training is involved in this process. The model does not provide a prioritized list of potential odor sources or any confidence scores. This is because in this work, the LLM is not commanded to choose which object is odor source, but to choose which action the robot should select to approach the odor source. To clarify the issue, we added lines 443-448 in section 4.4.*

**Question 9:** Additionally, section 3.4 implies the movement strategy is fully reactive, lacking mapping or memory. While reactive strategies may work in simple environments (as the one presented in the experimental section), they are likely to struggle in complex settings like houses, hospitals, or factories, which often contain multiple rooms and obstacles.

***Author’s response***: *Thank you for your feedback!*

*We used a reactive method due to its simplicity and lower computational cost compared to engineering-based methods (such as those involving mapping or memory). Adding mapping and memory would require more computational resources and could slow down the decision-making process. However, it should be noted that engineering-based methods can be implemented in the proposed ‘Low-level Action’ module without requiring any changes to the existing ‘High-level Reasoning’ module.*

*To answer this question, we added lines 344-349 of section 3.4.*

**Question 10 – Experiments:** The experimental setup is somewhat simplistic, involving a central obstacle with aligned “candidate sources” all within the robot’s field of view. While acceptable as a proof of concept, this setup limits the validity of the approach in more realistic, multi-room settings. Overall, to properly evaluate the impact of the proposal, more experiments are necessary.

***Author’s response***: *Thank you for your feedback!*

*The main purpose of the experiment was to demonstrate the concept of using LLM for robotic odor source localization tasks. In our experiments, we included obstacles to simulate complex indoor environments. In future work, we plan to conduct experiments in actual environments, such as office or household environments.*

*To explain this question, we added a sentence in the Section 5 at lines 524-525.*

**Question 11 – Declaration Stage:** The goal of reaching the object within 0.8 meters could pose issues when multiple potential sources exist. A more refined method for distinguishing between sources is necessary. For example, by scoring candidates in the visual frame, and validating them with the gas sensing once the robot reaches them.

***Author’s response****:* *Thank you for your feedback!*

*It is important to note that this work primarily focuses on the design of the navigation algorithm. The key research question we aim to address is how to process multi-modal sensory inputs to compute robot actions that guide the robot toward the odor source. Source declaration in this work is defined by a distance threshold, meaning that reaching the threshold is considered as detecting the source. Designing a more sophisticated source declaration algorithm that can rank potential odor source objects is part of our future work.*

*To clarify this point, we have added the relevant information in lines 367-372 of section 4.1 and 535-536 of the section 5.*

**Question 12 - Fan 2’s role:** It’s unclear what “fan2” contributes to the setup—perhaps it prevents gas dispersal, but this should be clarified.

***Author’s response****:* *Thank you for your feedback! The role of the ‘Fan 2’ is to create turbulent airflows. In our experiments, as shown in Figure 8, ‘Fan2’ is place perpendicular to the main airflow direction, resulting in a turbulent airflow field. To answer this question, we added lines 362-364 of section 4.1.*

**Question 13 - Comparative Analysis**: The comparisons focus on simple models (like surge-cast) and the authors’ previous work. Including recent approaches in GSL, especially the anemotactic techniques previously mentioned, would strengthen this analysis.

***Author’s response***: *Thank you for your feedback!*

*The primary goal of our experiments is to demonstrate that the proposed multi-modal OSL navigation algorithm outperforms single-modal algorithms, which include both olfaction-based and vision-based approaches. Most recent advancements in GSL are single-modal algorithms, primarily olfaction-based. For this study, we selected a representative olfaction-based algorithm, the moth-inspired method, due to the availability of its control code, which allowed us to implement it in our robotic agent. In future work, we plan to compare our proposed method with additional recent olfaction-based navigation algorithms.*

*To better answer this question, we added lines 533-536 of section 5.*

**Question 14 - Gas Concentration Threshold**: The term "threshold" is used in Fig. 9 for gas concentration, yet it’s not clear how this threshold is set or used. Clarification on this point and on the success/failure rate of the LLM’s movement prediction would improve understanding of the model’s effectiveness.

***Author’s response***: *Thank you for your feedback!*

*The threshold is set to the background concentration, which represents the concentration when no alcohol plumes are present in the environment. It is important to note that the navigation problem is a long-horizon task, where the success or failure of each step is less significant than the final result. In other words, the primary concern is whether the robot can reach the odor source location; the intermediate decision-making steps are less important.*

*To clarify this, we have added details in lines 364-366 of section 4.1.*