Reviewer 1:

Thank you for reviewing our manuscript! Here are our responses:

**Feedback 1:** 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!*

*Response.*

*According to your feedback, we have added lines 0-0 of Section 5.*

**Feedback 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).

LX W Comment: Add a sentence in related works where you mention the Anemotactic approaches. This sentence mentions that our proposed method includes wind measurements.

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.

LX W Comment: Add system prompt in pic or quote the system prompt sentences in the main article. Try both, see which one gives you better presentation.

* + **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.

LX W Comment: Give an example of how Gas Concentration is presented.

* + **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.

LX W Comment: Explain that Multi-Modal LLM can reason on images. Give examples and Multi-Modal LLM response.

* **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.

LX W Comment: Emphasize that the navigation strategy was developed based on human odor search behaviors. Explain human search behaviors and emphasize that vision and olfaction are used sequentially, not simultaneously.

* + 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.

LX W Comment: Explain how LLM infers odor source objects. Visible plumes are strong indicator of odor source, but LLM can also infer odor source locations based on semantic definition of objects. Give an example where LLM say fans will transport odor plumes so the robot is moved towards fans.

* + 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.

LX W Comment: Mention that LLM is commanded to choose only one odor source. Explain that in the vision branch, the robot commands are generated based on this odor source detection, such as the robot turns left if the object is on left and turns right if the object is on right. Mention that the LLM does not list potential odor sources in a priority list. It gives the most likely odor source directly, and there is only one odor source.

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

LX W Comment: Similar to the previous question. Exmplain that the LLM only output one source.

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.

LX W Comment: Mention that our experimental field includes obstacles to mimic complex indoor environments. Mention that memory-based or mapping-based method requires additional computational cost that will slow down the decision-making process. Mention that we choose reactive olfactory-based navigation strategy due to its effectiveness, simplicity, and easy to implement on robotic agents.

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

*Response.*

*According to your feedback, we have added lines 0-0 of Section 5.*

**Feedback 2:**

**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.

* **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.

LX W Comment: In this work, declaration is simply defined as a distance threshold, and if the robot reaches this threshold, the robot is considered as detecting the source. Mention that this work primarily focuses on the navigation algorithm design instead of the source declaration algorithm design, which involves a sophisticated design. Mention that in future work we will add source declarations.

* **Fan2's Role**: It’s unclear what “fan2” contributes to the setup—perhaps it prevents gas dispersal, but this should be clarified.

LX W Comment: Explain that the role of fan2 is to make turbulent flows, and show that if the robot relying only on olfactory sensor and anemometer cannot correctly find the odor source (use Figure 11 as the evidence).

* **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.

LX W Comment: Explain that we have the code for these navigation algorithms. For some anemotactic methods, we cannot access the control code (I am not sure if this is correct, verify this sentence. If the code is open access, mention that we will add comparison in future works. Meanwhile, mention that the purpose of experiment design is to show that multi-modal navigation is better than single-modal navigation)

* **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.

LX W Comment: mention the threshold value here. Explain that there is no success/failure rate to evaluate LLM’s movement prediction since there is no unique movement combination to solve the OSL problem. Mention that this navigation problem is a long-horizon problem where the success/failure is measured by whether or not the robot can reach the correct odor source location, the middle decision-making is less important.

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

*Response.*

*According to your feedback, we have added lines 0-0 of Section 5.*