Semantic-GSL: Semantically Guided Gas Source Localization with vision-language models

Hyoungho Park[*,†,1], Hongro Jang[*,1], Seunghwan Kim[1], Junhee Lee[2], Jiwoo Kim[1], and Hyondong Oh[††,2]

***Abstract*— This paper presents Semantic-GSL, the first framework to integrate semantic information into the estimation process for indoor gas source localization (GSL) by leveraging vision-language model (VLM). It addresses a fundamental limitation of conventional GSL approaches, which ignore the crucial real-world prior that emission sources are typically co-located with relevant objects. We employ a two-stage VLM framework to extract semantic information about high-probability source objects from general environmental descriptions without requiring predefined object names. This semantic information is then fused into a semantic-informed particle filter (SIPF), which redistributes particles toward semantically relevant regions, resulting in faster convergence and improved estimation accuracy. Simulation results in complex indoor environments confirm that Semantic-GSL significantly outperforms existing methods.**

I. Introduction

As hazardous gas leaks increasingly threaten ecosystems and human lives [1], gas source localization (GSL) using mobile robots has become a critical task to estimate the source location and release rate [2]. In complex indoor environments, conventional gas-sensor-only GSL methods [3], [4] typically combine Bayesian inference for source parameter (e.g., source locaiton, release rate) estimation with information-theoretic search strategies, but they are often inefficient due to noisy gas measurements and complex airflow-driven gas dispersion.

To overcome this limitation of relying solely on gas measurement, several approaches integrate onboard cameras with object detection models. These models provide the semantic information about potential source objects from a predefined object set and heuristically guide the mobile robot toward them once a certain gas measurement threshold is exceeded [5], [6]. However, by treating detected objects as deterministic search cues, these methods fail to disambiguate the true source when multiple objects are present, leading to inefficient search and poor estimation accuracy.

Motivated by these limitations, we propose Semantic-GSL, the first semantically informed GSL framework: leveraging a two-stage vision-language model (VLM) to extract semantic information (i.e., locations and relevance scores of source-related objects) from general environmental descriptions and using this semantic information within a semantic-informed particle filter (SIPF) to guide particle distribution to semantically informed areas.

The main contributions of this paper are summarized as:

1) We propose the two-stage VLM framework that extracts the semantic information by detecting objects with a high probability of emitting gas using environment descriptions without requiring specific object names (e.g., objects that can leak or store gas).
2) We introduce SIPF, which incorporates semantic information as an informative prior by redistributing particles toward semantically relevant objects, leading to more robust source estimation and improved search efficiency.
3) We validate the effectiveness of the Semantic-GSL through simulations in two realistic indoor environments with different layouts and object settings.

II. Semantic-GSL fraemwork

For robust GSL in the indoor environment, it is essential to incorporate semantic information into the source parameter estimation, enabling fast and robust estimation and efficient search. In this paper, a hazardous gas is assumed to be released at the location $\mathbf{r}_s = [x_s, y_s]^T$ with a release rate Q_s. At each time step k, the mobile robot at location r_k collects a gas measurement z_k and images using its onboard gas sensor and camera. The source parameter set $\theta = [\mathbf{r}_s^T, Q_s]^T$ is estimated using the proposed SIPF, which approximates the posterior distribution $p(\theta_k|z_{1:k})$ with a particle set $\{\theta_k^i, w_k^i\}_{i=1}^N$. Semantic information from the two-stage VLM framework is used to construct a semantically informative prior, biasing particles toward source-relevant objects before updating their weights based on the gas measurement likelihood. Finally, a dual-mode information-theoretic search strategy [7] selects the next locaiton r_{k+1}. The overall structure of the Semantic-GSL can be observed in Fig. 1.

A. Two-Stage VLM Framework

To extract the semantic information about potential gas leak objects, we design a two-stage VLM framework that

∗Equal contributions. †Project lead. ††Corresponding author

[1]Department of Mechanical Engineering, Ulsan National Institute of Science and Technology (UNIST), Republic of Korea.

[2]Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea.

gudgh1630@unist.ac.kr, hlomk@unist.ac.kr, kevin6960@unist.ac.kr, ljh0124@kaist.ac.kr, tars0523@unist.ac.kr, h.oh@kaist.ac.kr

This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2023R1A2C2003130), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2020-NR049578) and Unmanned Vehicles Core Technology Research and Development Program through the National Research Foundation of Korea (NRF) and Unmanned Vehicle Advanced Research Center (UVARC) funded by the Ministry of Science and ICT, the Republic of Korea (2020M3C1C1A01082375).

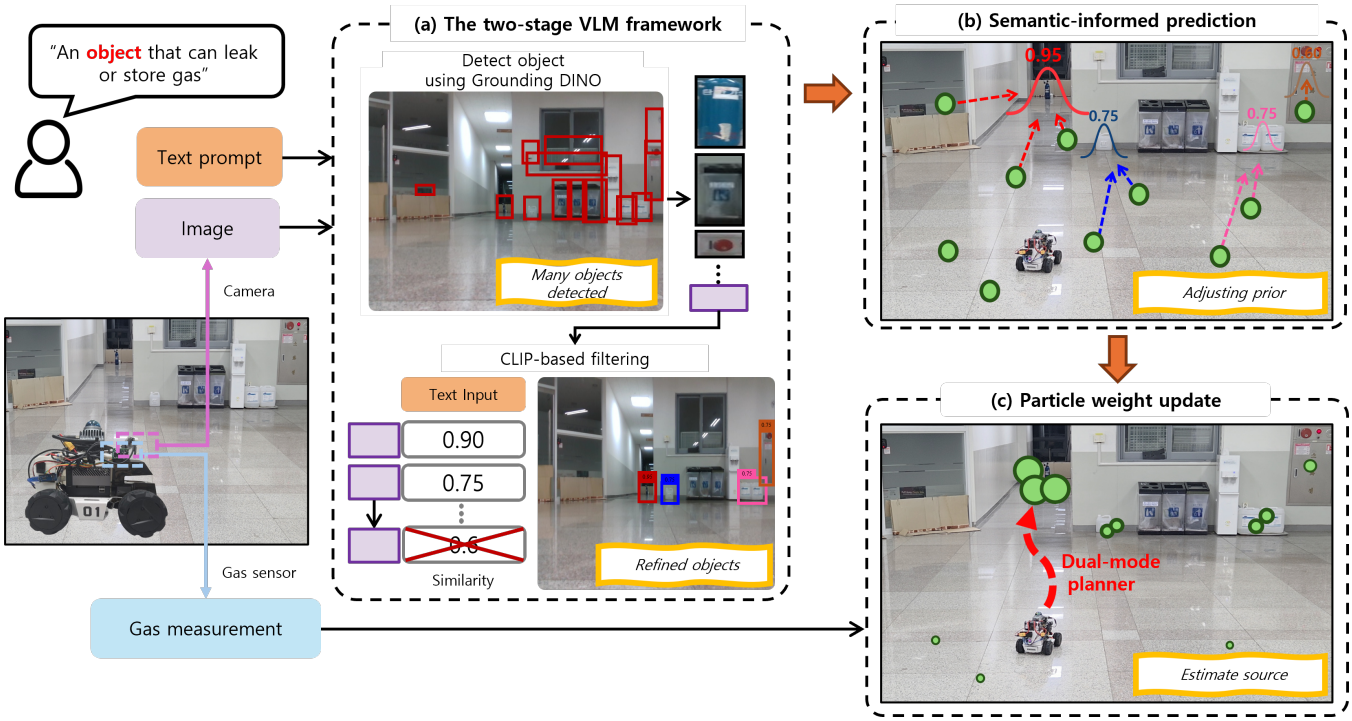

Fig. 1. Overview of the Semantic-GSL framework.

only uses the image and a general environmental text prompt, as shown in Fig. 1(a). At each time step k, after the mobile agent obtains the gas measurement z_k and image, the first stage processes it with Grounding DINO [8]. It identifies all potentially relevant objects from the given environmental descriptive text prompt (e.g., "objects that can leak or store gas"). This yields a set of G_k candidates that is often large and may include false positives or objects only loosely related to the prompt.

To filter and rank this set of G_k candidates, the second stage employs CLIP [9] for semantic refinement. For each candidate $g_k \in \{1, \ldots, G_k\}$, the corresponding image patch is compared against the text prompt to compute their semantic similarity score $s_{g_k} \in [0, 1]$. This contrastive scoring enables pruning of irrelevant objects by applying a similarity threshold, τ_s, retaining only those with $s_{g_k} > \tau_s$ (set to 0.6 in our framework) The resulting refined set of D_k high-confidence objects ($D_k \leq G_k$) constitutes the semantic information. Each object d is then represented by a semantic vector $\mathbf{S}_{d_k}$ that combines its bounding box center location $\mathbf{c}_d \in \mathbb{R}^2$ and its semantic similarity score s_{d_k}:

$$\mathbf{S}_{d_k} = [\mathbf{c}_{d_k}^T, s_{d_k}]^T \in \mathbb{R}^3, \quad d_k \in \{1, \ldots, D_k\}, \tag{1}$$

B. Semantic-Informed Particle Filter

1) Semantically Informed Prediction: Since the gas source is assumed static, the prediciton step simply propagates each particle as $\theta_{k-1}^i \leftarrow \theta_k^i$ forming the prior $p(\theta_k | z_{1:k-1})$ with the particle set $\{\theta_k^i, w_{k-1}^i\}$. To enrich this prior with semantic information $\{\mathbf{S}_{d_k}\}_{d_k=1}^{D_k}$, we apply an independence Metropolis-Hastings (MH) move to each particle. Each particle first samples a candidate object o_k^i from the normalized semantic object distribution:

$$p_{\text{obj}}(d_k) = \frac{s_{d_k}}{\sum_{d_k=1}^{D_k} s_{d_k}}. \quad d_k \in \{1, \ldots, D_k\}. \tag{2}$$

The proposal state $\tilde{\theta}_k^i$ is then drawn from a Gaussian distribution centered at the selected object location $\mathbf{c}_{o_k^i}$:

$$\tilde{\theta}_k^i \sim \mathcal{N}(\mathbf{c}_{o_k^i}, \sigma_q^2 I), \tag{3}$$

where σ_q controls the spread of the proposal and I is the identity matrix. The proposal is accepted with the MH acception probability α:

$$\alpha(\theta_k^i, \tilde{\theta}_k^i) = \min\left\{1, \exp\left(-\frac{\|\tilde{\theta}_k^i - \theta_k^i\|^2}{2\sigma_p^2}\right)\right\}, \tag{4}$$

where σ_p regulates sensitivity to the move distance. Accepted proposals replace the current particles, redistributing the prior particle set toward semantically relevant regions before the weight update step, as illustrated in Fig. 1(b).

2) Particle Weight Update: For every particle $i \in \{1, \ldots, N\}$, the particle weight w_{k-1}^i is updated using the gas measurement z_k as:

$$\bar{w}_k^i = p(z_k | \theta_k^i) \cdot w_{k-1}^i, \tag{5}$$

where $p(z_k | \theta_k^i)$ denotes the likelihood of the gas measurement z_k. To mitigate particle degeneracy, we apply systematic resampling [12] whenever the effective sample size falls below a threshold. With updated particle set $\{\theta_k^i, w_k^i\}$, the dual-mode information-theoretic search strategy [7] determines the next location r_{k+1} for the mobile robot, as illustrated in Fig. 1(c).

TABLE I

SIMULATION RESULTS IN ENVIRONMENT 1 AND ENVIRONMENT 2

Env	Method	Success Rate (%)	Search Time	Travel Distance (m)	Estimation Error (m)	Travel Time [s]
1	Dual-mode planner	86	47.73	76.20	2.47	301.71
	Mode change	85	41.01	68.77	2.58	286.92
	Grounding DINO + SIPF	83	43.11	70.90	1.70	292.38
	Semantic-GSL (Ours)	**90**	**38.41**	**64.78**	**0.94**	**250.83**
2	Dual-mode planner	75	51.32	77.13	1.98	359.76
	Mode change	61	45.30	70.88	2.33	321.69
	Semantic-GSL (Only Grounding DINO)	63	46.87	72.20	1.40	391.77
	Semantic-GSL (Ours)	**87**	**38.24**	**62.08**	**1.34**	**257.80**

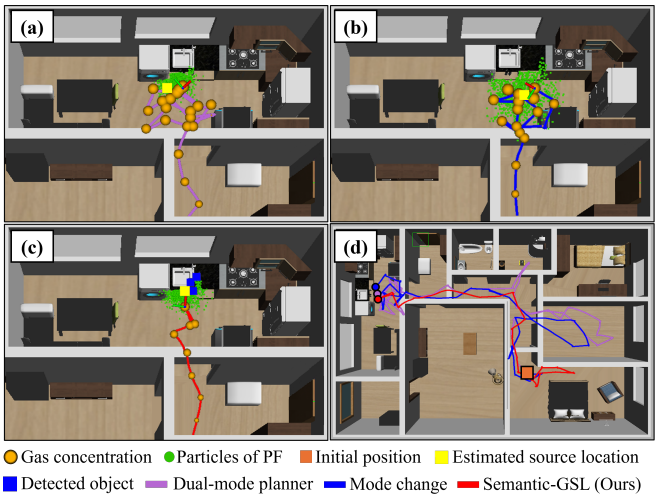

Fig. 2. (a), (b), and (c) show the mobile robot path and the gas measurement locations for the dual-mode, mode change, and Semantic-GSL in environment 1. Each case completed the search in timestep $k = 49$, $k = 43$, and $k = 25$. The total path is shown in (d).

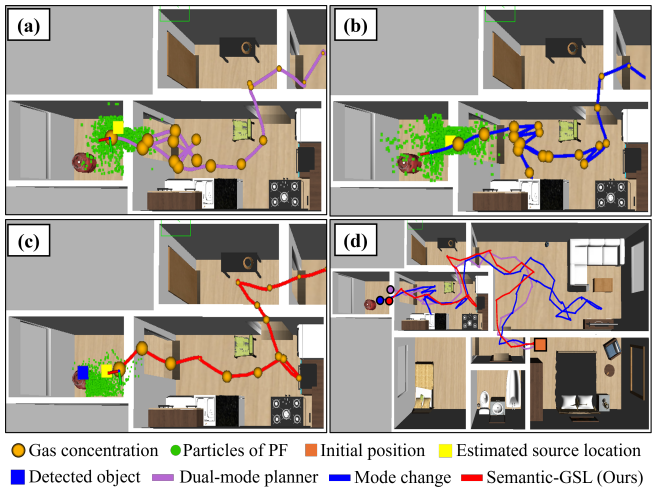

Fig. 3. (a), (b), and (c) show the mobile robot path and the gas measurement locations for the dual-mode, mode change, and Semantic-GSL in environment 2. Each case completed the search in timestep $k = 36$, $k = 46$, $k = 31$. The total path is shown in (d).

III. SIMULATION RESULTS

The proposed Semantic-GSL is compared with three baselines: the dual-mode planner [7], which relies solely on gas measurements; the mode-change strategy, which uses semantic information only for search by guiding the robot toward detected objects once a gas threshold is exceeded [5]; and a variant using Grounding DINO + SIPF to evaluate the benefit of the two-stage VLM. All methods employ the dual-mode planner for trajectory generation, and 100 Monte Carlo simulations are conducted in two distinct indoor environments. Table I summarizes the results, and detailed performance comparisons are shown in Fig. 2 and Fig. 3.

The largest performance gap appears in estimation error, clearly separating methods that use SIPF from those that do not. This confirms that a semantic information-guided redistribution of particles effectively serves as an informative prior, concentrating particles near likely source objects and improving estimation accuracy.

However, the relatively long travel times and lower success rates of Grounding DINO + SIPF highlight the importance of reliable semantic information. Without the refinement stage, many irrelevant objects are detected, producing nearly uniform priors that misguide the search.

Similar trends are observed with the mode-change method. Although it avoids irrelevant detections by using a predefined object set by using the two-stage VLM, its heuristic search strategy toward detected objects causes frequent detours to non-source objects whenever gas concentration is below the threshold, leading to inefficient exploration.

Finally, the dual-mode planner records the longest travel distance and time due to its lack of prior knowledge, but achieves relatively high success rates, indicating that poor or misleading semantic information can be more harmful than having none.

Overall, Semantic-GSL achieves the best results across all metrics, demonstrating that combining a two-stage VLM with SIPF leads to faster, more accurate, and robust GSL.

IV. CONCLUSION

This work introduced Semantic-GSL, the first semantically informed framework for STE that integrates a two-stage VLM with a SIPF. Extracting semantic information from the two-stage VLM and by guiding SIPF with an informative, semantics-driven prior, the proposed Semantic-GSL achieves fast and robust GSL. Simulation results also verify that Semantic-GSL consistently achieves the best performance across all evaluation metrics. Future work will focus on real-world deployment and extending evaluation to more diverse and dynamic environments.

REFERENCES

[1] W. Tsujita, A. Yoshino, H. Ishida, and T. Moriizumi, "Gas sensor network for air-pollution monitoring," *Sensors and Actuators B: Chemical*, vol. 110, no. 2, pp. 304–311, 2005.

[2] M. Hutchinson, H. Oh, and W.-H. Chen, "A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors," *Information Fusion*, vol. 36, pp. 130–148, 2017.

[3] M. Vergassola, E. Villermaux, and B. I. Shraiman, "'Infotaxis' as a strategy for searching without gradients," *Nature*, vol. 445, no. 7126, pp. 406–409, 2007.

[4] S. An, M. Park, and H. Oh, "Receding-horizon RRT-Infotaxis for autonomous source search in urban environments," *Aerospace Science and Technology*, vol. 120, p. 107276, 2022.

[5] S. Ma, J. Yuan, Z. Guo, and Q. Wu, "Autonomous plume near-source search assisted by intermittent visible plume information using finite state machine and YOLOv3-tiny," *Expert Systems with Applications*, vol. 228, p. 120350, 2023.

[6] J. Monroy, J.-R. Ruiz-Sarmiento, F.-A. Moreno, F. Melendez-Fernandez, C. Galindo, and J. Gonzalez-Jimenez, "A semantic-based gas source localization with a mobile robot combining vision and chemical sensing," *Sensors*, vol. 18, no. 12, p. 4174, 2018.

[7] S. Kim, J. Seo, H. Jang, C. Kim, M. Kim, J. Pyo, and H. Oh, "Gas source localization in unknown indoor environments using dual-mode information-theoretic search," *IEEE Robotics and Automation Letters*, 2024.

[8] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 38–55, 2024.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 8748–8763, 2021.

[10] S. Sun, R. Li, P. Torr, X. Gu, and S. Li, "CLIP as RNN: Segment countless visual concepts without training endeavor," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 13171–13182, 2024.

[11] X. Wu, F. Zhu, R. Zhao, and H. Li, "CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 7031–7040, 2023.

[12] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Boston, MA: Artech House, 2003.