Parse, Search, and Confirmation: Training-Free Aerial Vision-and-Dialog Navigation with Chain-of-Thought Reasoning and Structured Spatial Memory

Yu Qi[1] Hongyu Li[4] Shaofei Huang[5] Tianrui Hui[1,2,3*]
Yaxiong Wang[1] Lechao Cheng[1] Zhun Zhong[1*] Si Liu[4] Meng Wang[1]
[1]School of Computer Science and Information Engineering, Hefei University of Technology
[2]Jianghuai Advance Technology Center [3]Anhui Provincial Key Laboratory of Humanoid Robots
[4]School of Artificial Intelligence, Beihang University [5]University of Macau

*Corresponding authors.

Parse, Search, and Confirmation: Training-Free Aerial Vision-and-Dialog Navigation with Chain-of-Thought Reasoning and Structured Spatial Memory

Supplementary Material

1. More Dataset Details

To provide a more comprehensive background on the ANDH [2] and ANDH-Full [2] datasets used in this experiment, we elaborate on their data composition and collection design. Both datasets are derived from the AVDN base dataset [2], which is constructed using a customized UAV simulator that generates an original georeferenced map by cropping X-View satellite imagery. Each map corresponds to a fixed real-world area, accurately preserving realistic aerial visual characteristics. Specifically, the dataset contains $3,064$ full navigation trajectories, each associated with multiple rounds of natural language dialogue. These trajectories are further segmented into $6,269$ sub-trajectories, each corresponding to a single dialogue cycle. The average length of a full trajectory is 287 meters, while sub-trajectories range from 142 to 148 meters on average.

2. More Implementation Details

To map any pixel coordinate (u, v) from the multi-scale visual observation of our SSM module back to its corresponding location on the original georeferenced map, we employ an inverse homography transformation.

First, we compute the inverse homography matrix H_{inv} based on the correspondence between the four corners of the visual observation, denoted as $(0,0), (S,0), (0,S), (S,S)$, and their corresponding points (X_i, Y_i) on the original georeferenced map. Let $H_{\text{inv}} = [h_{ij}]_{3\times3}$. The relationship between the visual observation coordinates and the homogeneous georeferenced map coordinates is given by:

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \propto H_{\text{inv}} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{1}$$

Specifically, the Cartesian pixel coordinates (X, Y) on the georeferenced map are obtained via:

$$X = \frac{h_{00}u + h_{01}v + h_{02}}{h_{20}u + h_{21}v + h_{22}}, \tag{2}$$

$$Y = \frac{h_{10}u + h_{11}v + h_{12}}{h_{20}u + h_{21}v + h_{22}}. \tag{3}$$

Finally, we convert (X, Y) to geographic coordinates (lng, lat) using linear interpolation, where lng and lat denotes longitude and latitude, respectively. The georeferenced map follows a standard image coordinate system with the origin at the top-left corner, whereas the geographic coordinate system is defined from the bottom-left. Let the georeferenced map have a resolution of $W_{\text{map}} \times H_{\text{map}}$, and its geospatial bounds is defined by the lower-left coordinate $(\text{lng}_{\min}, \text{lat}_{\min})$ and the upper-right coordinate $(\text{lng}_{\max}, \text{lat}_{\max})$. The conversion is defined as:

$$\text{lng} = \text{lng}_{\min} + \frac{X}{W_{\text{map}}} (\text{lng}_{\max} - \text{lng}_{\min}), \tag{4}$$

$$\text{lat} = \text{lat}_{\max} - \frac{Y}{H_{\text{map}}} (\text{lat}_{\max} - \text{lat}_{\min}). \tag{5}$$

3. More Ablation Studies

Image resolution	SPL	SR	GP
512×512	15.1	19.7	38.0
768×768	**17.8**	**22.6**	**39.2**
1024×1024	16.7	20.9	36.3

Table 1. Ablation results of different image resolutions of the main view.

As shown in Table 1, we conduct an ablation study on the cropping resolution of the main view by comparing three input sizes: 512×512, 768×768, and 1024×1024. The results indicate that the 768×768 configuration achieves the best overall performance across all three evaluation metrics. This suggests that lower resolutions tend to lose critical structural details, thereby hindering accurate target localization. In contrast, excessively high resolutions introduce more irrelevant background and noise, and further exacerbate the instability of Qwen's [1] visual encoding and geometric mapping, which ultimately degrades performance. The 768×768 setting strikes a favorable balance between fine-grained detail preservation and sufficient global context, and is therefore adopted as the default field of view size in all subsequent experiments.

Steps	SPL	SR	GP
2	**17.9**	20.2	38.4
3	17.8	**22.6**	**39.2**
4	15.7	20.0	37.2
5	15.6	19.7	33.7

Table 2. Ablation results of different maximum search steps.

As reported in Table 2, we further conduct an ablation study on the maximum number of search steps by increasing the limit from 2 to 5 steps. The results show that using 3 steps yields the best overall performance and clearly

Method	ANDH								
	Seen Val.			Unseen Val.			Unseen Test		
	SPL	**SR**	**GP**	**SPL**	**SR**	**GP**	**SPL**	**SR**	**GP**
GeoGround [7]	2.8	3.0	-2.5	5.6	5.6	-3.8	2.8	2.8	-2.3
RemoteSAM [6]	5.3	6.2	6.6	6.1	6.3	6.8	4.8	5.0	6.1
PSC-AVDN (Ours)	**16.3**	**18.6**	**37.4**	**17.8**	**22.6**	**39.2**	**13.5**	**16.4**	**28.2**

Table 3. Comparison with remote sensing large models on the ANDH dataset for training-free AVDN task.

surpasses both the 2 step setting and larger step counts. Although the 2 step configuration achieves a slightly higher SPL, its SR and GP scores are substantially lower, indicating that insufficient search often forces the agent to stop prematurely before adequate exploration has been completed. Conversely, when the search limit increases to 4 or 5 steps, all three metrics deteriorate, with GP dropping markedly to 33.7 under the 5 step setting. This trend suggests that excessive search depth accumulates geometric drift and decision errors, causing the predicted trajectory to deviate from the true path. Overall, the 3 step configuration achieves the best trade-off between sufficient exploration and limiting error accumulation, thereby yielding the most stable navigation performance.

4. Results of Remote Sensing Large Models

As shown in Table 3, we investigate the performance of MLLM (GeoGround [7]) and vision foundation model (RemoteSAM [6]) that are specifically trained for remote sensing scenarios when applied to the training-free AVDN task. The experiments are conducted on the ANDH dataset. We find that even models specifically trained on remote sensing data still struggle to effectively perform the training-free AVDN task when directly given natural language navigation instructions. This validates the effectiveness of our proposed three-stage reasoning pipeline and structured spatial memory module.

5. Results of different MLLMs

As shown in Table 4, applying our PSC-AVDN framework to different MLLMs consistently improves their performances. This demonstrates the strong generalizability of our proposed CoT reasoning and spatial memory, with the improvement being particularly pronounced on specialized VL models (*e.g.*, Qwen-VL in our paper and Seed-VL).

6. Results of different Models in the Parse Stage

The parsing stage relies solely on textual instructions without visual input. Therefore, a text-only LLM is more appropriate, as it avoids cross-modal interference and focuses on linguistic reasoning. As shown in Table 5, replacing DeepSeek-V3 with Qwen-VL results in consistent performance degradation. On the Unseen Val. split of ANDH, SR drops by 4.4 % , with similar declines in SPL and GP.

Method	ANDH					
	Seen Val.			Unseen Val.		
	SPL	**SR**	**GP**	**SPL**	**SR**	**GP**
GPT-4o [5]	2.6	2.7	-9.5	3.4	3.9	-11.8
Ours (GPT-4o)	**6.0**	**9.5**	**27.7**	**7.6**	**10.9**	**35.2**
Seed-VL [3]	3.5	4.3	-4.1	4.0	4.6	-5.5
Ours (Seed-VL)	**10.2**	**13.5**	**25.7**	**12.9**	**16.1**	**27.9**

Table 4. Experiments with different MLLMs on the ANDH dataset.

Figure 1. Failure case.

Model	SPL	SR	GP
Qwen-VL [1]	**17.8**	**22.6**	**39.2**
Deepseek-V3 [4]	15.7	19.2	22.9

Table 5. Comparison of Different Models in the Parse Stage

7. Detailed Prompts

Tables 6 and 7 provide detailed prompt specifications for S-CoT and C-CoT.

8. Failure case

Many failure cases typically arise from interference by visually similar cues. As shown in Figure 1, the dark green target building is confused with another building of a similar color, leading to trajectory drift.

9. More Visualization Results

We present additional visualization examples in Figure 2 on the ANDH and ANDH-Full datasets. Our model can accurately interpret the given instructions, navigate to the target regions, and generate reliable navigation predictions even under challenging visual conditions.

10. Limitations

(1) Compared with lightweight models, our MLLM-based models incur higher inference latency and token costs. (2) Training-based approaches can leverage supervised data to achieve more concentrated navigation behavior, whereas our training-free framework lacks task-specific adaptation and is therefore more prone to error accumulation and trajectory drift, which ultimately leads to lower GP scores.

References

[1] S. Bai, K. Chen, X. Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2

[2] Y. Fan, W. Chen, T. Jiang, and et al. Aerial vision-and-dialog navigation. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2023. 1

[3] Dong Guo, Feng Wu, Fei Zhu, et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 2

[4] A. Liu, B. Feng, B. Xue, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2

[5] S. Shahriar, B. Lund, N. R. Mannuru, M. Arshad, K. Hayawi, R. Varma Kumar Bevara, A. Mannuru, and L. Batool. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *arXiv preprint arXiv:2407.09519*, 2024. 2

[6] L. Yao, F. Liu, D. Chen, et al. Remotesam: Towards segment anything for earth observation. In *Proceedings of the 33rd ACM International Conference on Multimedia(ACM MM)*, 2025. 2

[7] Y. Zhou, M. Lan, X. Li, et al. Geoground: A unified large vision-language model for remote sensing visual grounding. *arXiv preprint arXiv:2411.11904*, 2024. 2

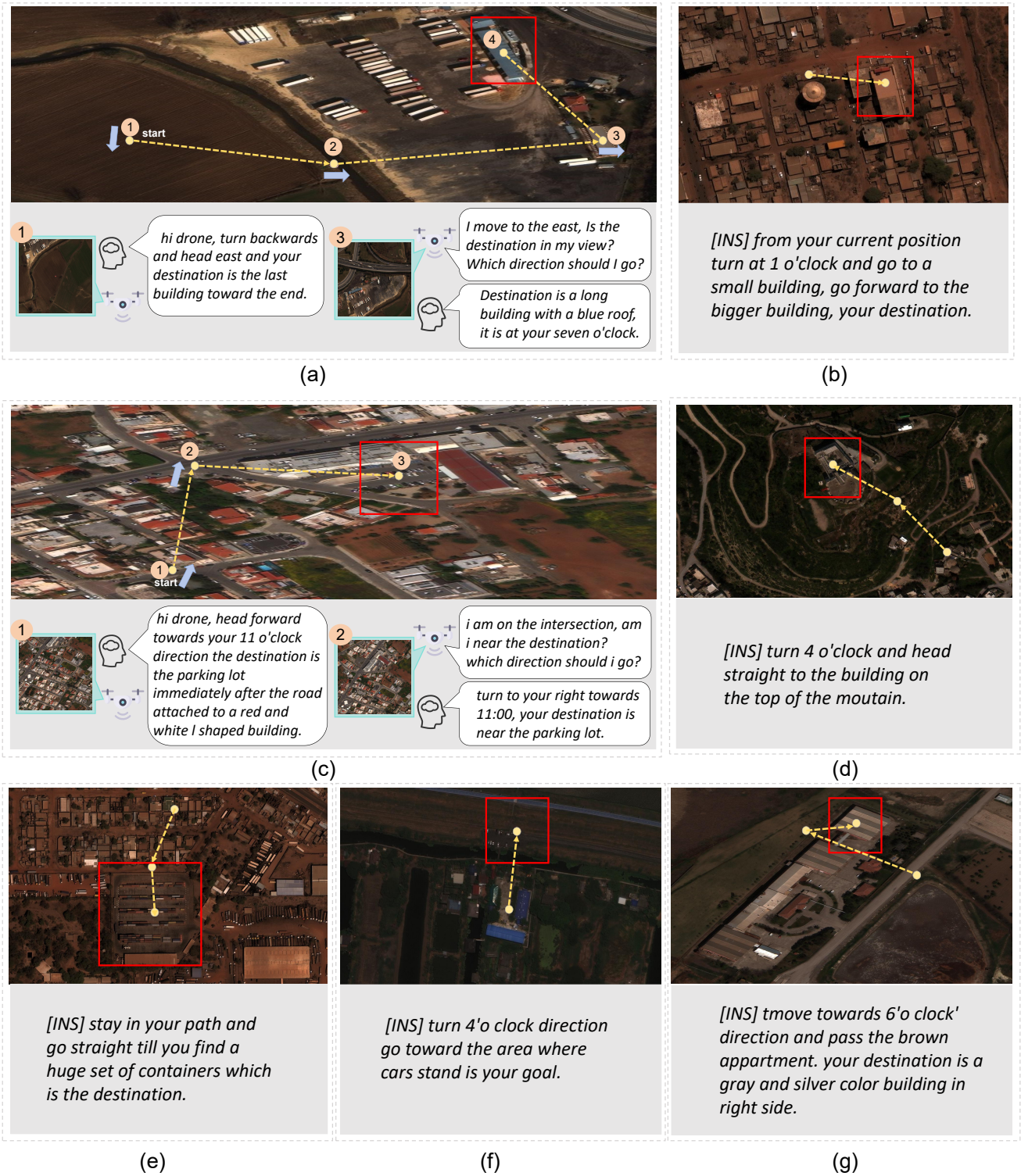

Figure 2. Additional visualization results of our method on the ANDH and ANDH-Full datasets. (a) and (c) are examples from ANDH-Full, while (b), (d), (e), (f), and (g) are examples from ANDH.

S-CoT

Views:
Image 1: main view (`FIXED_CROP_SIDE` × `FIXED_CROP_SIDE`), top indicates the current heading.
This is the only image you may output a bounding box for.
Image 2: same center and heading as Image 1 but with a narrower, more zoomed-in view of the same moment (local detail context).
Image 3: same center and heading as Image 1 but with a wider, more zoomed-out view of the same moment (broader surrounding context).
Image 4: north-up global map; red dot represents the current position, red arrow denotes the current heading, yellow trail shows the recent trajectory.

---

1) Destination analysis:

- *Analyze the destination description and extract key visual / physical attributes (shape, texture, material, color).*
- *Extract geometric structure and any directional / positional cues.*

2) Scene understanding:

- *Summarize the scene and spatial layout in Image 1 (main view).*
- *Summarize local spatial / contextual structure seen in Image 2 and Image 3 (they provide tighter or wider context of the exact same place and heading).*
- *Separately summarize global position / heading / trajectory cues from Image 4.*

3) Generate 5×5 reference grid on Image 1:

- *Partition Image 1 into 5 rows × 5 columns (from top to bottom, from left to right).*
- *For each cell, output a label chosen only from this closed set (lowercase; 1–2 words):* `road`, `building`, `container yard`, `parking`, `runway`, `water`, `field`, `forest`, `roof`, `shadow`, `rail`, `vehicle` *(use* `unknown` *if unclear).*
- *Return this grid in JSON under key* `"grid_5x5"` *as a 5×5 array of strings.*

4) Target localization on Image 1 only:

- *First restrict your search to the top half of Image 1 (forward / current heading direction).*
- *Propose candidate regions that satisfy the destination description, using appearance, geometry, texture, shadows, arrangement, and nearby structures.*
- *If nothing in the top half matches, extend to the full Image 1 — but the final chosen region must still come from Image 1.*
- *Choose exactly one final region. If you cannot isolate one unambiguous region, output* `dest_present = false`*.*

5) Bounding box output (Image 1 only):

- *Produce the minimal square bounding box fully enclosing the selected region.*
- *All bounding box coordinates must be given in Image 1's* `FIXED_CROP_SIDE` *pixel grid.*
- *Return* `dest_present` *(true/false),* `bbox_2d`*, confidence in [0.0, 1.0], and a brief reason (<500 chars).*

Rules:

- *Never output or reference a bounding box on Image 2, Image 3, or Image 4; they are context only.*
- *If not sufficiently confident in one clear match, output* `dest_present = false`*.*
- *These are overhead / satellite views; rely on shape, texture, spatial layout, shadows, and stable geometry rather than raw color.*

Table 6. Detailed S-CoT prompts.

C-CoT

Views:

Image 1: main view (`FIXED_CROP_SIDE` × `FIXED_CROP_SIDE`), top indicates the current heading. This is the only image you may output a box for.
Image 2: same center and heading as Image 1 but with a narrower, more zoomed-in view of the same moment (local detail context).
Image 3: same center and heading as Image 1 but with a wider, more zoomed-out view of the same moment (broader surrounding context).
Image 4: north-up global map; red dot represents the current position, red arrow denotes the current heading, yellow trail shows the recent trajectory.

the current heading, yellow trail shows the recent trajectory.

Reasoning Chain:

Step 1 *— Generate an explicit reasoning chain. Break the destination description into ordered, checkable sub-steps. Example: "the vehicle on the far left of the central parking line" → "identify the central parking line" → "scan along that line for the leftmost vehicle."*

Step 2 *— Follow those sub-steps on Image 1. For each sub-step, progressively narrow the candidate area in Image 1, starting with the top half first whenever it applies. Use Image 2 / Image 3 only to clarify local geometry and layout (e.g., which elongated strip is a runway, which dense blob is a parking lot), and use Image 4 only for high-level position / heading / trajectory context. Record at least one concrete visual evidence item for each critical sub-step.*

Generate reference grid map on Image 1:

- *Partition Image 1 into 5 rows × 5 columns (from top to bottom, from left to right).*
- *For each cell, output a label chosen only from this closed set (lowercase; 1–2 words):* `road`, `building`, `container yard`, `parking`, `runway`, `water`, `field`, `forest`, `roof`, `shadow`, `rail`, `vehicle` *(use* `unknown` *if unclear).*
- *Return this grid in json under key* `"grid_5x5"` *as a 5×5 array of strings.*

Bounding box output:

- *Draw one tight square bounding box that encloses the final chosen region on image 1 only.*
- *All bbox coordinates must be reported in Image 1's* `FIXED_CROP_SIDE` × `FIXED_CROP_SIDE` *pixel grid.*
- *Return* `dest_present` *(true/false),* `bbox_2d`*, confidence in [0.0, 1.0], and 1–3 short justification sentences.*
  - *In Simple Direct Localization mode: name the decisive visual attributes.*
  - *In Reasoning Chain mode: cite the key reasoning steps and supporting evidence.*

Critical rules:

- *Do not output bounding boxes for Image 2, Image 3, or Image 4; they are context only.*
- *If you cannot isolate exactly one high-confidence match within Image 1, you must output* `dest_present = false`*.*
- *These are overhead / satellite views; color may be low-contrast. Use structure, texture, shape, arrangement, shadows, and spatial relations, not just raw color.*

Table 7. Detailed C-CoT prompts.