RACE: URBAN SPATIAL MULTI-HOP REASONING VIA ROUTE-AWARE CUMULATIVE EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban spatial multi-hop reasoning combines visual observations and road geometry to infer complete routes and spatial answers from sequential instructions. Each location choice changes the spatial context and evidence available at subsequent hops. Similar buildings and road patterns can sustain local self-consistency along incorrect routes, making early errors difficult to detect through stepwise checks. Comparing these routes requires evaluating evidence from different spatial contexts on a common scale. To address these challenges, we propose Route-Aware Cumulative Evidence (RaCE). RaCE maintains feasible candidate routes and dynamically updates spatial contexts according to each branch's decision history. It integrates visual and geometric evidence into support scores comparable across branches and accumulates this evidence along complete routes, allowing later information to inform the reassessment of earlier decisions. The framework selects the best-supported route and final spatial answer through route-level comparison and answer-level evidence aggregation, respectively. Adaptive perception then leverages score bounds to acquire visual evidence selectively, yielding the same route and spatial answer as full acquisition. Experiments across five cities and four general-purpose models show that RaCE achieves mean route and answer accuracies of 77.7% and 92.1%, respectively, while reducing inference latency by 67.7% on average compared with existing methods. The data and code are available at https://anonymous.4open.science/r/RaCE-ECF0.

1 INTRODUCTION

Urban spatial reasoning requires models to identify a sequence of locations and infer a spatial answer from natural-language instructions, visual observations, and road topology (Feng et al., 2025; Zhou et al., 2025). Visual observations support semantic matching, while geometric information verifies spatial constraints. Unlike sequential reasoning with fixed inputs, each location choice changes the spatial context and evidence available at subsequent hops (Zhang et al., 2025). A locally plausible choice may therefore lead to a route that fails to satisfy the complete instruction.

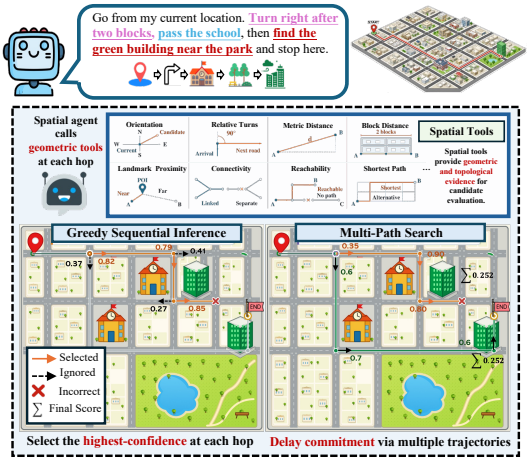

Figure 1: A wrong hop rewrites subsequent context in urban spatial multi-hop reasoning.

Figure 1 illustrates this phenomenon with the instruction: "After two blocks, turn right, pass a school, and proceed to the green building near a park." A school may match the current description, yet continuing from that location may not reach the target building while satisfying the remaining requirements. Once the wrong school is selected, the model continues reasoning around an incorrect spatial anchor. Similar building types and road patterns can make subsequent observations appear consistent even along an incorrect route (Baghaei et al., 2025; Ko et al., 2025). This local self-consistency makes early errors difficult

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to detect from subsequent observations, resulting in routes that are locally plausible but globally incorrect. Resolving this ambiguity requires retaining alternative early choices and evaluating how well their resulting routes satisfy the instruction as a whole (Shi et al., 2025; Song et al., 2025).

To improve multi-hop spatial reasoning in vision-language models (VLMs), existing methods use spatial post-training to strengthen relational composition (Lee et al., 2026; Shen et al., 2025), employ chain-of-thought (CoT) reasoning or spatial agents equipped with geometric tools for stepwise reasoning and state updates (Guo et al., 2026; Bao et al., 2026), or retain multiple candidate routes through search. However, sequential reasoning remains vulnerable to locally self-consistent errors, while search can delay premature commitment to a single route without automatically ensuring cross-branch evidence comparability. Different preceding decisions lead to different observations and candidate sets, so local rankings reflect only a candidate's relative advantage within a particular set and cannot directly measure overall support across complete routes (Lin et al., 2025).

To address this challenge, we formulate urban spatial multi-hop reasoning as a global structured inference problem and propose Route-Aware Cumulative Evidence (RaCE). RaCE explicitly maintains multiple candidate routes, dynamically constructs spatial contexts from their respective decision histories, and integrates visual and geometric evidence into candidate support scores that are comparable across branches. It then accumulates evidence along complete routes and performs global comparison, allowing later evidence to reassess earlier decisions and thereby infer the best-supported route and final spatial answer. To control the perceptual cost of branch expansion, RaCE further adopts a decision-boundary-aware adaptive perception mechanism that reduces visual queries while matching the route and answer predictions obtained with full visual acquisition.

In summary, our main contributions are as follows:

- We formalize inter-hop dependencies in urban spatial multi-hop reasoning by modeling decision-dependent spatial contexts, providing a basis for reassessing early decisions using evidence from subsequent reasoning steps.
- We propose RaCE to accumulate comparable visual and geometric evidence along complete routes and reassess early choices using later evidence. Adaptive perception uses score bounds to reduce visual queries while preserving both route and answer predictions of exhaustive inference.
- Our experiments show that geometric consistency and history-dependent context are central to route accuracy, while later evidence helps resolve ambiguous early matches. Across four general-purpose models, RaCE achieves mean route and answer accuracies of 77.7% and 92.1%, while reducing inference latency by 67.7% on average over existing methods.

2 Related Work

Urban Spatial Reasoning. Existing research covers visual spatial question answering, vision-and-language navigation, and geographic localization and retrieval (Cepeda et al., 2023). CityGPT (Feng et al., 2024) enhances geographic question answering, trajectory generation, and navigation through an urban world model, while UrbanLLaVA (Feng et al., 2025) integrates structured geographic data, trajectories, and multi-source imagery for diverse urban spatial tasks. These efforts improve multi-modal urban understanding and extend spatial reasoning from single-step to sequential reasoning.

Vision-Language Navigation and Planning. Existing approaches primarily focus on improving model capabilities, interactively updating states, and searching over multiple candidates (Lin et al., 2025; Zhang et al., 2025; Shi et al., 2025). The first category strengthens relational composition through CoT, multi-hop training, or preference optimization, as exemplified by spatial post-training on MultiHopSpatial (Lee et al., 2026) and SpatialReasoner-R1 (Shen et al., 2025). The second category uses tools or maps to guide spatial reasoning: MSSR (Guo et al., 2026) acquires spatial evidence on demand, SpatialAgent (Bao et al., 2026) composes executable geospatial workflows from spatial concepts, MapGPT (Chen et al., 2024) uses map-guided prompting for adaptive path planning, and DUET (Chen et al., 2022) combines local observations with topological maps for global planning; however, state updates may still be misled by locally self-consistent evidence. The third category includes beam search, instantiated here as Beam-4 (Freitag & Al-Onaizan, 2017), which retains competing decision histories to delay early commitment. However, jointly controlling visual evaluation costs and maintaining a common scoring scale across branches remains a key challenge when comparing multiple complete routes.

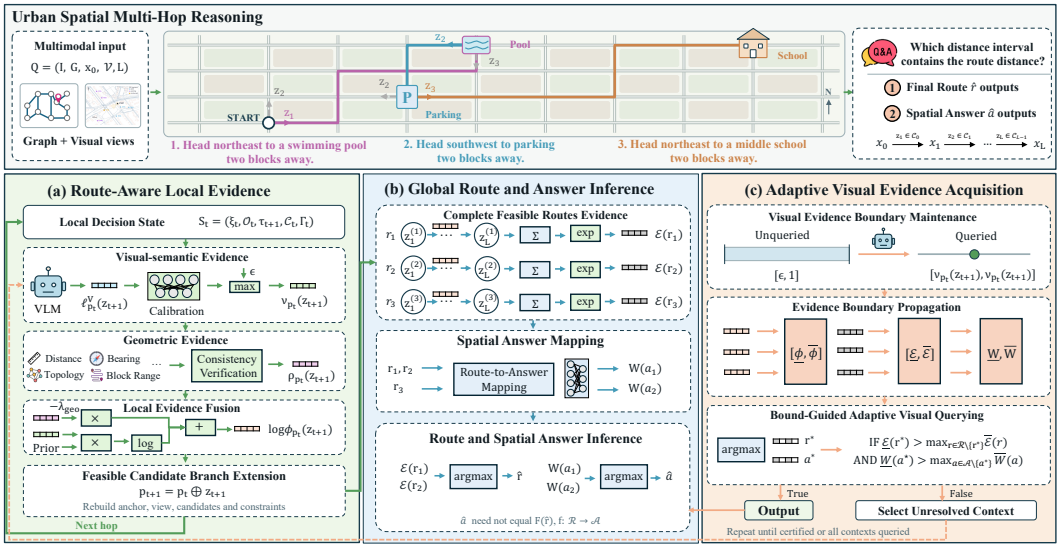

Figure 2: Overview of RaCE, which fuses visual and geometric evidence over candidate histories and adaptively updates route and answer scores.

3 PRELIMINARIES

An urban spatial multi-hop reasoning instance $\mathcal{Q} = (I, G, x_0, \mathcal{V}, L)$ consists of a natural-language navigation instruction I, an urban road graph G, a starting location x_0, location-indexed visual observations $\mathcal{V}$, and a reasoning depth L.

Feasible Reasoning Route. A reasoning route $r = (z_1, \ldots, z_L)$ records the locations selected successively at each hop. The decision history $p_t = (z_1, \ldots, z_t)$ determines the current spatial anchor $x(p_t)$, with $p_0 = \emptyset$ and $x(p_0) = x_0$. Let $\mathcal{C}(p_t)$ denote the set of next-hop candidates reachable from $x(p_t)$ along the road network G. The feasible route set is:

$$\mathcal{R} = \{r = (z_1, \ldots, z_L) \mid z_{t+1} \in \mathcal{C}(p_t),\ t = 0, \ldots, L-1\}, \tag{1}$$

here, feasibility requires only road-network reachability; distance and direction constraints are evaluated by geometric scoring rather than imposed as conditions for inclusion in $\mathcal{R}$.

Route-Aware Decision State. For a decision history p_t, in addition to the current spatial anchor $x(p_t)$, we record the incoming heading h_t^{in} upon arrival at the anchor and the direction in which the road at the current anchor extends, h_t^{road}. These three components form the spatial reference context ξ_t, which, together with the visual observation, next-hop semantic target, candidate set, and spatial constraints, defines the decision state S_t:

$$\xi_t = \big(x(p_t), h_t^{\text{in}}, h_t^{\text{road}}\big), \qquad S_t = \big(\xi_t, \mathcal{O}_t, \tau_{t+1}, \mathcal{C}_t, \Gamma_t\big), \tag{2}$$

where $\mathcal{O}_t$ is the visual observation from $\mathcal{V}$, τ_{t+1} specifies the next-hop target category or appearance, $\mathcal{C}_t = \mathcal{C}(p_t)$ is the reachable candidate set, and Γ_t denotes spatial constraints. Absolute directions use world coordinates, while incoming- and road-relative instructions use h_t^{in} and h_t^{road} as references; their construction and conversion to Γ_t are detailed in Appendix B.

Problem Statement. Given an instance $\mathcal{Q}$, the reasoning method F combines navigation instructions, visual observations, and road geometry to predict a complete route $\widehat{r}$ and a spatial answer $\widehat{a}$:

$$(\widehat{r}, \widehat{a}) = F(\mathcal{Q}) \in \mathcal{R} \times \mathcal{A}, \tag{3}$$

here, $\mathcal{A}$ is the set of distance intervals, and $f : \mathcal{R} \to \mathcal{A}$ maps each route to the interval containing its total road-network distance. Since different routes may map to the same interval, we evaluate complete routes and final answers separately.

4 PROPOSED METHOD

RaCE consists of (i) Route-Aware Local Evidence, (ii) Global Route and Answer Reasoning, and (iii) Adaptive Visual Evidence Acquisition. Figure 2 shows how spatial relations and on-demand visual scores are accumulated over feasible decision histories for route and answer prediction.

4.1 ROUTE-AWARE LOCAL EVIDENCE

RaCE maintains a decision state S_t for each feasible branch and updates it after every location choice. Given p_t, it uses the current anchor $x(p_t)$ to retrieve local observations $\mathcal{O}_t$ and reachable candidates $\mathcal{C}_t$, while the next-hop instruction specifies the semantic target τ_{t+1} and spatial constraints Γ_t for visual and geometric scoring.

Visual and Geometric Decomposition. In state S_t, RaCE separately assesses each candidate's visual-semantic match and compliance with spatial constraints. The VLM uses local observations $\mathcal{O}_t$ to independently judge whether candidate z_{t+1} matches the semantic target τ_{t+1}. Let $y \in \{0, 1\}$ indicate whether a match holds. The visual log odds are defined as:

$$\ell^{\mathrm{V}}_{p_t}(z_{t+1}) = \log \frac{P_{\mathrm{VLM}}(y = 1 \mid \mathcal{O}_t, \tau_{t+1}, z_{t+1})}{P_{\mathrm{VLM}}(y = 0 \mid \mathcal{O}_t, \tau_{t+1}, z_{t+1})}. \tag{4}$$

Their extraction is detailed in Appendix B. The log-odds $\ell^{\mathrm{V}}_{p_t}$ provide uncalibrated visual evidence: positive values favor a match and negative values a mismatch. Geometry and later evidence help resolve visual ambiguity.

Geometric assessment uses the route-aware state ξ_t and road network G. For each constraint $\gamma \in \Gamma_t$, let $d_\gamma(z_{t+1}; \xi_t, G) \geq 0$ denote its normalized violation, yielding the geometric deviation:

$$\rho_{p_t}(z_{t+1}) = \sum_{\gamma \in \Gamma_t} d_\gamma(z_{t+1}; \xi_t, G)\,, \tag{5}$$

here, $\rho_{p_t}(z_{t+1}) = 0$ indicates full constraint satisfaction, while larger values indicate greater violations. Bearing and distance violations are defined in Appendix B. Road reachability restricts candidates, while geometric deviation measures spatial compliance.

Local Scoring and State Update. To fuse the two heterogeneous evidence sources, RaCE calibrates the visual log odds $\ell^{\mathrm{V}}_{p_t}$ to better reflect matching confidence and maps them to a visual matching score:

$$\nu_{p_t}(z_{t+1}) = \max\left\{\epsilon,\, \sigma\left(\frac{\ell^{\mathrm{V}}_{p_t}(z_{t+1})}{T} + b\right)\right\} \in [\epsilon, 1], \tag{6}$$

here, $T > 0$ and b are the temperature and bias, fitted on the calibration split and fixed during evaluation; σ is the sigmoid function. Since sigmoid outputs can approach zero, $\max$ clips visual scores from below at $\epsilon \in (0, 1)$, keeping their logarithms at least $\log \epsilon$. RaCE then combines the calibrated visual score, candidate prior, and geometric deviation into a local evidence score:

$$\log \phi_{p_t}(z_{t+1}) = \log \pi(z_{t+1} \mid p_t) + \log \nu_{p_t}(z_{t+1}) - \lambda_{\mathrm{geo}} \rho_{p_t}(z_{t+1}), \tag{7}$$

here, $\pi(z_{t+1} \mid p_t) = 1/|\mathcal{C}(p_t)| > 0$ is uniform within each candidate set, and $\lambda_{\mathrm{geo}} \geq 0$ is the geometric penalty coefficient. To preserve differences in support strength across branches, the combined scores are accumulated across hops without further normalization within candidate sets.

For each feasible candidate, RaCE updates the decision history of the corresponding branch. If the instruction is not yet complete, it constructs the next-hop state using the new spatial anchor:

$$p_{t+1} = p_t \oplus z_{t+1}, \qquad S_{t+1} = \big(\xi_{t+1}, \mathcal{O}_{t+1}, \tau_{t+2}, \mathcal{C}(p_{t+1}), \Gamma_{t+1}\big), \qquad 0 \leq t < L - 1, \tag{8}$$

here, $\oplus$ appends a location choice. The updated history determines $x(p_{t+1})$ and its reference headings. At the final hop, it simply becomes $r = p_L$, with no next-hop state.

4.2 GLOBAL ROUTE AND ANSWER INFERENCE

Because spatial references vary across branches, local support alone cannot assess whether a complete route follows the instruction. Route-Aware Cumulative Evidence (RaCE) therefore sums route-

aware local log scores $\log \phi_{p_t}(z_{t+1})$ along each complete route $r = (z_1, \ldots, z_L)$ and pools the resulting evidence across routes mapped to the same answer, yielding answer support $W(a)$:

$$\log \mathcal{E}(r) = \sum_{t=1}^{L} \log \phi_{p_{t-1}}(z_t), \qquad W(a) = \sum_{r \in \mathcal{R}_a} \mathcal{E}(r), \tag{9}$$

here, $\mathcal{E}(r) > 0$ denotes route evidence, and $\mathcal{R}_a = \{r \in \mathcal{R} \mid f(r) = a\}$ is the set of feasible routes corresponding to answer a. This accumulation allows weak matches at later steps to reduce the overall support for a route, thereby enabling early choices to be reassessed.

RaCE separately maximizes route evidence and answer support to obtain $\widehat{r}$ and $\widehat{a}$:

$$\widehat{r} = \arg\max_{r \in \mathcal{R}} \mathcal{E}(r), \qquad \widehat{a} = \arg\max_{a \in \mathcal{A}} W(a). \tag{10}$$

Ties in both $\arg\max$ operations follow fixed total orders. Since multiple routes may support the same answer, $\widehat{a}$ need not equal $f(\widehat{r})$. For later use, let $s(r) = \log \mathcal{E}(r)$ denote the route log-evidence.

4.3 Adaptive Visual Evidence Acquisition

As reasoning deepens, RaCE maintains exponentially more feasible branches, making exhaustive visual acquisition require many VLM calls. Even at admissible score extremes, some routes cannot alter the globally optimal outputs. Starting at S_0, RaCE maintains visual-score bounds per branch and queries the VLM only when they cannot determine the best route and answer.

For a local candidate $z \in \mathcal{C}_t$, let $[\underline{\nu}_{p_t}(z), \overline{\nu}_{p_t}(z)]$ denote its visual-score interval, initialized as $[\epsilon, 1]$ before querying and collapsed to $[\nu_{p_t}(z), \nu_{p_t}(z)]$ afterward. Combining this interval with the candidate prior and geometric deviation yields the local evidence interval:

$$\left[\underline{\phi}_{p_t}(z), \overline{\phi}_{p_t}(z)\right] = \pi(z \mid p_t) e^{-\lambda_{\text{geo}} \rho_{p_t}(z)} \left[\underline{\nu}_{p_t}(z), \overline{\nu}_{p_t}(z)\right]. \tag{11}$$

Multiplying hop-wise lower and upper bounds yields the route-evidence interval $[\underline{\mathcal{E}}(r), \overline{\mathcal{E}}(r)]$. Summing these route bounds by answer yields the answer-support interval $[\underline{W}(a), \overline{W}(a)]$. In each round, RaCE selects the route and answer with the largest lower bounds as the current leaders:

$$r^\star = \arg\max_{r \in \mathcal{R}} \underline{\mathcal{E}}(r), \qquad a^\star = \arg\max_{a \in \mathcal{A}} \underline{W}(a). \tag{12}$$

RaCE stops early only if each leader's lower bound exceeds its competitors' upper bounds:

$$\underline{\mathcal{E}}(r^\star) > \max_{r \in \mathcal{R} \setminus \{r^\star\}} \overline{\mathcal{E}}(r), \qquad \underline{W}(a^\star) > \max_{a \in \mathcal{A} \setminus \{a^\star\}} \overline{W}(a). \tag{13}$$

When both certificates hold, RaCE returns $\widehat{r} = r^\star$ and $\widehat{a} = a^\star$; otherwise, it queries an unresolved history by the priority rule, with fixed-order fallback. Let $\mathcal{P}$ denote the finite set of nonterminal decision histories along feasible routes. Each query resolves one such history, so RaCE terminates within $|\mathcal{P}|$ queries and, with identical scores and tie-breaking, matches exhaustive querying. Algorithm 1 gives the procedure; Appendix B proves correctness and termination.

5 Experiments

Our experiments address the following six research questions:

- **RQ1 (Overall Performance):** How does RaCE compare with general-purpose models, spatially post-trained models, and tool-augmented agents in reasoning accuracy and inference cost?
- **RQ2 (Component Analysis):** How do local evidence construction, global inference, and adaptive perception affect RaCE's reasoning performance?
- **RQ3 (Local Ambiguity):** Can RaCE use subsequent evidence to improve early decisions when local candidates are difficult to distinguish?
- **RQ4 (Cross-City Generalization):** Can RaCE maintain stable performance in new cities by directly applying calibration parameters and the geometric weight set in other cities?
- **RQ5 (Parameter Sensitivity):** Does a higher geometric weight always improve RaCE's reasoning? How do reasoning depth and candidates per hop affect its performance?
- **RQ6 (Case Study):** How does RaCE select routes and query visual evidence in a specific case?

5.1 EXPERIMENTAL SETTINGS

We evaluate RaCE's performance in predicting complete routes and final spatial answers on a shared urban navigation task. Details of baseline adaptation, metric definitions, and hyperparameter settings are provided in the appendix.

Datasets. Our dataset contains about 10,000 urban navigation instances across New York, Sydney, Singapore, Berlin, and Tokyo. Each instance covers about $4\,\mathrm{km}^2$, with roads and POIs from OpenStreetMap. Instructions involve three successive decisions, each with four reachable candidates and route visuals, and specify targets using POI descriptions, distance, direction, turns, and connectivity.

Baseline Methods. *General Vision-Language Models*: Using the same vision-language backbone, we compare RaCE with CoT and Beam-4. CoT selects one candidate per hop, while Beam-4 retains the four routes with the highest cumulative scores. *Spatially Post-Trained Models*: MultiHopSpatial (Lee et al., 2026) learns multi-hop spatial relation composition, while SpatialReasoner-R1 (Shen et al., 2025) combines search trajectories with preference learning. We integrate RaCE with both to evaluate its complementarity with spatial post-training. *Tool-Augmented Agents*: MSSR (Guo et al., 2026) adaptively acquires additional observations, while SpatialAgent (Bao et al., 2026) organizes geospatial tool operations into concept-guided workflows. We compare both with RaCE on accuracy and query cost.

Evaluation Protocol. We split instances into training, calibration, validation, and test sets in a 6:2:1:1 ratio. The calibration set is used to calibrate visual scores, and the validation set is used to select prompts and hyperparameters. All methods are evaluated on the same test set, reporting accuracy for the first decision (Hop-1 ACC), the first two decisions (First-2 ACC), complete routes (Route ACC), and spatial answers (Answer ACC). Calls denotes the mean number of logical visual-query events per instance; for RaCE, one event evaluates all candidates at a decision history. Latency is mean inference time (s); VRAM is peak GPU memory usage (GiB).

Implementation Details. RaCE requires no additional fine-tuning of the vision-language backbone. We use Qwen3.5-9B (Qwen3.5) as the primary backbone and evaluate generalization with Qwen2.5-VL-7B (Qwen2.5), InternVL3.5-8B (InternVL), and Kimi-VL-A3B-Instruct (Kimi-VL). All models use 512×410 visual observations, with $\lambda_{\mathrm{geo}} = 16$. Experiments run on four NVIDIA RTX 4090 GPUs with 24 GiB.

5.2 OVERALL PERFORMANCE (RQ1)

Table 1 compares different reasoning paradigms on urban spatial multi-hop reasoning. RaCE achieves the best complete-route and answer accuracy. Across the four general-purpose VLMs, CoT reaches an average Hop-1 ACC of 50.9% but only 5.1% Route ACC, showing that correct early decisions do not guarantee reliable complete routes. Beam-4 retains multiple high-scoring partial routes, but its reliance on local relative ranking and irreversible pruning still leads to low route accuracy. Tool-augmented methods such as SpatialAgent achieve at most 4.1% Route ACC, indicating that additional observations and tool calls do not directly translate into reliable route selection. In contrast, RaCE explicitly maintains decision branches, integrates visual and geometric evidence, and uses later evidence to reassess earlier choices, achieving average Route ACC and Answer ACC of 77.7% and 92.1%, respectively. For spatially post-trained models, incorporating RaCE further improves Route ACC by 24.1 percentage points on average, showing that our framework can further leverage their existing spatial capabilities to improve multi-hop route reasoning. In terms of efficiency, RaCE reduces inference latency by 67.7% on average compared with existing methods. Relative to Beam-4, its latency varies across backbones, while consistently achieving higher route and answer accuracy; its peak GPU memory usage remains comparable to baselines using the same backbone. Overall, RaCE provides a favorable balance between reasoning accuracy and efficiency.

5.3 COMPONENT ANALYSIS (RQ2)

Table 2 presents the component ablation results of RaCE. In Table 2(a), removing geometric evidence or fixing the spatial reference causes Route ACC to decrease by 86.4% on average relative to full RaCE, while joint visual-geometric scoring decreases it by 54.7%. These results indicate that route-aware geometric evaluation, spatial reference updating, and evidence separation are par-

Table 1: Performance of different reasoning paradigms on urban spatial multi-hop reasoning.

Model	Method	Hop-1 ACC ↑	First-2 ACC ↑	Route ACC ↑	Answer ACC ↑	Calls ↓	Time ↓	VRAM ↓
Qwen2.5	CoT	48.0	13.7	3.4	41.5	3.0	38.2	19.2
	Beam-4	51.2	14.0	3.6	34.9	9.0	**1.3**	10.2
	MSSR	24.9	6.4	2.0	41.8	14.7	75.1	10.0
	SpatialAgent	27.5	6.8	2.0	40.4	14.7	62.9	10.0
	RaCE	**93.0**	**85.9**	**75.9**	**92.0**	4.0	1.8	10.0
Qwen3.5	CoT	66.4	24.2	9.7	49.6	3.0	59.0	10.9
	Beam-4	53.1	24.0	11.1	50.6	9.0	2.9	11.0
	MSSR	26.6	6.7	2.2	40.2	9.6	137.8	11.1
	SpatialAgent	26.3	6.6	4.1	48.7	15.0	106.5	11.2
	RaCE	**96.5**	**95.1**	**91.1**	**95.7**	3.3	**1.3**	11.5
InternVL3.5	CoT	42.4	12.8	4.1	41.7	3.0	46.6	10.0
	Beam-4	42.1	13.3	5.1	45.0	9.0	**2.1**	9.8
	MSSR	25.2	6.5	2.8	34.6	18.4	216.2	11.1
	SpatialAgent	26.3	6.8	1.7	39.2	18.0	176.9	13.2
	RaCE	**89.4**	**79.9**	**70.5**	**90.0**	4.3	3.4	9.8
Kimi-VL	CoT	46.9	10.4	3.0	38.1	3.0	77.1	17.8
	Beam-4	31.3	11.0	4.6	47.8	9.0	**6.9**	19.5
	MSSR	27.4	7.0	1.9	38.5	9.2	165.0	18.7
	SpatialAgent	26.7	6.6	2.4	41.0	21.9	361.9	21.7
	RaCE	**90.6**	**83.6**	**73.4**	**90.8**	4.5	10.8	20.3
MultiHopSpatial	Base	87.3	76.4	63.2	85.1	3.0	15.5	10.1
	+ RaCE	**97.4**	**96.5**	**92.5**	**97.2**	3.2	**1.4**	10.3
SpatialReasoner-R1	Base	89.8	80.8	68.5	86.9	3.0	10.8	10.4
	+ RaCE	**95.2**	**92.4**	**87.4**	**96.4**	3.5	**2.3**	9.9

ACC in %, time in s, VRAM in GiB.

Table 2: Component analysis of RaCE with Qwen3.5-9B .

(a) Local Evidence

Variant	Route ACC ↑	Answer ACC ↑
RaCE	**91.1**	95.7
w/o Geo.	10.1	59.4
w/o Update	14.7	55.6
w/o Split	41.3	81.9
w/o Calib.	90.6	**96.6**

(b) Global Reasoning

Variant	Route ACC ↑	Answer ACC ↑
RaCE	**91.1**	95.7
w/o Agg.	20.6	68.8
w/ Local	84.8	94.3
w/o Branch	89.6	96.1
w/o Answer Pool	**91.1**	**96.8**

(c) Adaptive Acquisition

Variant	Route ACC ↑	Answer ACC ↑	Calls ↓	Time ↓	VRAM ↓
RaCE (Dual)	**91.1**	95.7	3.3	**1.3**	11.52
w/o Adaptive Acq.	**91.1**	95.7	21.0	11.5	11.52
w/o Query Priority	**91.1**	95.7	14.9	7.8	11.52
w/o Route Cert.	87.2	95.7	**3.0**	1.7	11.52
w/o Answer Cert.	**91.1**	**96.5**	3.1	1.7	11.52

* *Local evidence:* w/o Geo. removes geometry; w/o Update fixes the spatial reference at the start; w/o Split uses joint VLM scoring; w/o Calib. removes visual-score calibration. *Global reasoning:* w/o Agg. uses only final-hop evidence; w/ Local normalizes within candidate sets; w/o Branch keeps only the locally best branch; w/o Answer Pool directly maps the selected route to its answer. *Adaptive acquisition:* w/o Adaptive Acq. queries all contexts; w/o Query Priority queries randomly; w/o Route Cert. and w/o Answer Cert. remove the respective certificate.

ticularly important for local reasoning; in contrast, visual-score calibration has a relatively small effect. In Table 2(b), using only final-hop evidence causes a relative Route ACC decrease of 77.4%, while local normalization or retaining only the locally optimal branch leads to an average decrease of 4.3%. This shows that evidence accumulation, cross-branch scoring, and candidate preservation jointly support global reasoning. Notably, removing answer pooling leaves Route ACC unchanged while slightly increasing Answer ACC. This is related to the fact that multiple routes may correspond to the same answer, and also shows that route and answer prediction are not fully synchronized. In Table 2(c), random querying increases VLM Calls to 4.5 times the original value, validating the role of query priority in reducing visual overhead. Using only a single certificate can slightly reduce queries but cannot simultaneously guarantee both route and answer outputs; as discussed earlier, since an answer is jointly determined by evidence from multiple routes, identifying the optimal route does not imply that the answer has also been determined. Therefore, separate stopping decisions are still required for the two outputs.

5.4 Local Ambiguity (RQ3)

When candidate regions exhibit similar building distributions and road-connectivity patterns, local evidence often struggles to distinguish among them.

To analyze this ambiguity, we use multiple VLMs to independently evaluate the first-hop candidates, retain samples jointly identified as ambiguous, and group them by the score gap between the highest- and second-highest candidates to examine how accumulated route evidence affects candidate preferences.

As shown in Figure 3, score differences in the high-ambiguity group are concentrated near zero, while some incorrect early preferences are gradually corrected as later evidence accumulates. This occurs because RaCE does not commit to a single candidate when local evidence is insufficient; instead, it preserves candidate branches and uses subsequent visual and geometric evidence to reassess earlier choices. Table 3 further shows that RaCE reaches a final Hop-1 ACC of 91.9% in the high-ambiguity group, an improvement of 3.2 percentage points over the early decision. In contrast, gains are smaller in the medium- and low-ambiguity groups, indicating that RaCE is particularly beneficial in urban settings with high visual and road-network homogeneity and greater local decision uncertainty.

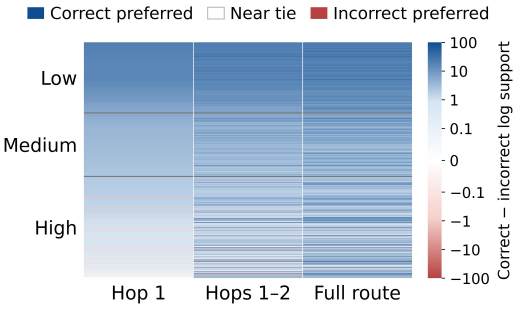

Figure 3: First-hop preference as route evidence accumulates across ambiguity levels.

Table 3: RaCE accuracy as route evidence accumulates across local-ambiguity groups.

	Low		Medium		High	
Evidence stage	Hop-1 ACC	Route ACC	Hop-1 ACC	Route ACC	Hop-1 ACC	Route ACC
Hop 1 only	100.0	—	100.0	—	88.7	—
Hops 1–2	100.0	—	100.0	—	89.1	—
Full route	**100.0**	**94.0**	**100.0**	**92.9**	**91.9**	**88.0**

—: Route ACC (%) is not reported before the full decision sequence is completed.

5.5 Cross-City Generalization (RQ4)

Figure 4 presents RaCE's transfer results on unseen cities. As shown in Figure 4(a), the mean Route ACC across the five cities reaches 89.9%, with overall performance remaining stable and showing no evident degradation. Notably, accuracy is relatively lower in Sydney, where spatially proximate locations are more often topologically separated than in cities such as New York and Berlin. Such locations typically require detours, increasing the discrepancy between geometric proximity and actual road distance and making route assessment more difficult. Figure 4(b) further illustrates this phenomenon: the straight-line distance between a parking lot and a fire station is only 65.9 m, whereas the shortest road distance reaches 401.6 m, 6.1 times larger. Such detour structures make scores calibrated on the other four cities less able to reflect the true road cost, leading to some transfer-performance degradation.

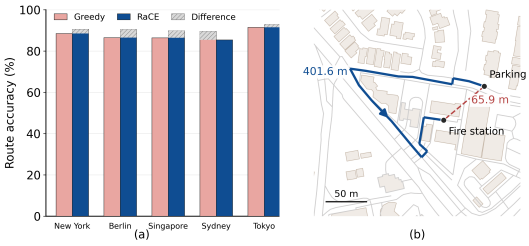

Figure 4: Cross-city transfer to Sydney and a road-detour case. (a) Route accuracy. (b) Road-detour case.

5.6 Parameter Sensitivity (RQ5)

As discussed earlier, the geometric weight determines the contribution of map constraints to hop-wise scoring. It should be noted that although OSM is generally reliable, local regions may still be affected by update delays or annotation errors, and an excessively large weight may amplify their impact on complete-route decisions. To evaluate RaCE's robustness to GIS noise, we therefore apply mild perturbations to the geometric inputs. Table 4(a) shows that increasing the geometric weight λ_{geo} from 1 to 16 improves Route ACC by 12.6% and 10.6% under the original and perturbed settings, respectively, while reducing Calls by 56.7%. Increasing it further to 32 yields only a marginal improvement under the original setting, while providing no further benefit and even causing a decline under perturbations, indicating that the geometric weight should remain moderate to avoid amplifying local GIS deviations in road direction and distance over multi-hop decisions. Tables 4(b)

Table 4: Parameter sensitivity and controlled scaling results.

(a) Geometric Weight

λ_{geo}	Origin Route ACC	Shifted Route ACC	Calls
1	80.9	78.3	7.623
4	90.7	85.8	3.579
16	91.1	**86.6**	3.298
32	**91.7**	86.5	3.252

(b) Reasoning Depth

L	Routes	SR-R1	RaCE
1	4	79.0	**95.0**
3	64	29.0	**89.0**
4	256	19.0	**88.0**
5	1,024	17.0	**90.0**

(c) Candidate Count

K	Routes	SR-R1	RaCE
2	8	73.0	**98.0**
4	64	29.0	**89.0**
6	216	15.0	**75.0**
8	512	13.0	**62.0**

* Shifted averages Route ACC under 5-degree heading, 5-m location, and 5% distance perturbations. SR-R1 denotes SpatialReasoner-R1; L is the reasoning depth, K is the number of candidates per hop, and Routes is the number of complete candidate routes.

and 4(c) further examine RaCE's scaling behavior under different reasoning depths and candidate sizes. When the reasoning depth increases from 1 to 5 and the number of candidate routes expands from 4 to 1,024, SpatialReasoner-R1's Route ACC decreases by 78.5%, whereas RaCE decreases by only 5.3% and remains at or above 88.0%, showing that RaCE effectively mitigates error accumulation in multi-hop decision-making. In addition, as the number of candidates per hop increases from 2 to 8 and the number of candidate routes grows from 8 to 512, the Route ACC of RaCE and the baseline decreases by 36.7% and 82.2%, respectively. Despite the continued expansion of the candidate space, RaCE maintains a clear advantage, demonstrating stronger robustness to scaling.

5.7 Case Study (RQ6)

This section uses a three-hop Berlin case to show how RaCE corrects an early route choice using later information. In Figure 5(a), CoT selects G8A, V8A, and Y2A according to the highest local scores, whereas RaCE retains the initially weaker N8A branch and continues through C3A and Z4A to obtain the correct route. Since both branches satisfy the geometric constraints specified by the instruction, the early evidence alone is insufficient to rule out the incorrect route. Figure 5(b) shows the resulting reversal: the cumulative score difference between the correct and CoT routes is -0.205 and -0.194 after the first two hops, but rises to $+1.031$ after the third-hop visual evidence is incorporated, making the correct route preferred. Figure 5(c) shows the on-demand querying process. RaCE prioritizes candidate contexts that may change the current route ranking and reevaluates the route scores after each query. After five queries, the lower bound on the current route's advantage over the strongest competitor reaches $+0.657$. Thus, even if all unqueried evidence favors the competing route, the current route remains ahead and querying can stop. RaCE therefore queries only 5 of 21 candidate contexts, reducing visual queries by 76.2%.

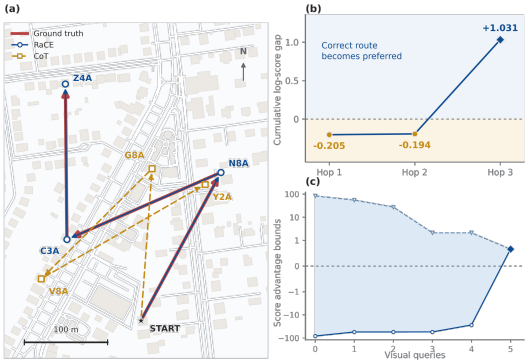

Figure 5: A Berlin three-hop case illustrating route reassessment and adaptive visual querying.

6 Conclusion and Future Work

For urban spatial multi-hop reasoning, we propose RaCE, which accumulates comparable cross-hop evidence over complete routes, allowing later observations to reassess earlier decisions. Experiments across five cities show that RaCE consistently improves route and answer prediction while substantially reducing visual query overhead. These results demonstrate the effectiveness of RaCE under the current urban map setting, while also leaving room to evaluate more complex real-world conditions. Our current benchmark is mainly based on OSM-rendered map scenes and does not fully capture road occlusion, viewpoint variation, or map bias in real environments. Future work will extend RaCE to real satellite and street-view imagery and larger-scale urban search spaces.

USE OF LARGE LANGUAGE MODELS

This study did not use generative AI tools for core research tasks, including developing theoretical frameworks, mathematical derivations, hypothesis formulation, research methodology or experimental design, method implementation, data processing, or result interpretation. Generative AI was used only for language editing to improve readability and fluency, and for limited assistance with debugging experimental scripts, such as identifying potential coding issues, suggesting fixes, and improving code readability; it was not used to design or implement the proposed method.

All AI-assisted content was manually reviewed and verified by the authors. Language edits were checked to ensure that they preserved the intended technical meaning and introduced no unsupported content, while code modifications were manually inspected, executed, and tested. The authors take full responsibility for the final text, claims, code, and other research content.

REFERENCES

Kourosh T. Baghaei, Dieter Pfoser, and Antonios Anastasopoulos. Follow the beaten path: The role of route patterns on vision-language navigation agents generalization abilities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7986–8005. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.406. URL https://aclanthology.org/2025.naacl-long.406/.

Riyang Bao, Cheng Yang, Dazhou Yu, Zhexiang Tang, Gengchen Mai, and Liang Zhao. SpatialAgent: Agentic geo-spatial reasoning with scientific core concepts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14896–14911. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.679. URL https://aclanthology.org/2026.acl-long.679/.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8690–8701, 2023.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. 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)*, pp. 9796–9810. Association for Computational Linguistics, 2024.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16537–16547, 2022.

Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. CityGPT: Empowering urban spatial cognition of large language models. *arXiv preprint arXiv:2406.13948*, 2024.

Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. UrbanLLaVA: A multi-modal large language model for urban intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6209–6219, 2025.

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 56–60. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-3207. URL https://aclanthology.org/W17-3207/.

Yejie Guo, Yunzhong Hou, Wufei Ma, Meng Tang, and Ming-Hsuan Yang. Pursuing minimal sufficiency in spatial reasoning. In *International Conference on Learning Representations*, 2026.

Heeju Ko, Sung June Kim, Gyeongrok Oh, Jeongyoon Yoon, Honglak Lee, Sujin Jang, Seungryong Kim, and Sangpil Kim. Active test-time vision-language navigation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50024–50043, 2025. doi: 10.52202/085713-1490.

Youngwan Lee, Soojin Jang, Yoorhim Cho, Seunghwan Lee, Yong-Ju Lee, and Sung Ju Hwang. MultihopSpatial: Multi-hop compositional spatial reasoning benchmark for vision-language model. *arXiv preprint arXiv:2603.18892*, 2026.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. NavCoT: Boosting LLM-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:5945–5957, 2025. doi: 10.1109/TPAMI.2025.3554559.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James M. Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xiangyu Shi, Zerui Li, Wenqi Lyu, Jiatong Xia, Feras Dayoub, Yanyuan Qiao, and Qi Wu. SmartWay: Enhanced waypoint prediction and backtracking for zero-shot vision-and-language navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 16923–16930, 2025. doi: 10.1109/IROS60139.2025.11246147.

Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12078–12088, 2025. doi: 10.1109/CVPR52734.2025.01128.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. MapNav: A novel memory representation via annotated semantic maps for VLM-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13032–13056. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.638. URL https://aclanthology.org/2025.acl-long.638/.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. UrBench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10707–10715, 2025.

APPENDIX

This appendix supplements RaCE with experimental results, method implementation, and the evaluation protocol. Appendix A reports supplementary experiments and related analyses; Appendix B describes the method implementation and inference procedure, together with correctness and complexity analyses; Appendix C specifies baseline adaptation, the experimental protocol, and the evaluation metrics.

A SUPPLEMENTARY EXPERIMENTS

Benchmark Ranking Consistency. Mainstream leaderboards evaluate foundation models on standardized multimodal tasks, but whether such performance transfers to reliable urban spatial multi-hop reasoning remains unclear. We therefore compare four general-purpose models using their official MMMU and MathVista scores alongside urban spatial reasoning metrics. Table 5 shows that

Table 5: Benchmark ranking consistency across general multimodal tasks and urban spatial reasoning (%).

Backbone	MMMU	MathVista	General Avg.	Hop-1 ACC	Route ACC	Answer ACC	Retention
Qwen2.5	58.6	68.2	63.4	48.0	3.4	41.5	7.1
Qwen3.5	78.4	85.7	82.1	66.4	9.7	49.6	14.6
InternVL3.5	73.4	78.4	75.9	42.4	4.1	41.7	9.7
Kimi-VL	57.0	68.7	62.9	46.9	3.0	38.1	6.4

Route Retention = Route ACC / Hop-1 ACC.

general benchmark performance only partially reflects urban spatial reasoning ability. The Spearman correlation between the general score and Hop-1 ACC is only $\rho = 0.40$; for example, InternVL3.5 ranks second by general score but fourth by Hop-1 ACC. Although the model rankings by general score, Route ACC, and Answer ACC are consistent, the average Route ACC across the four models is only 5.1%. Moreover, Route Retention, defined as the proportion of initially correct decisions that remain correct through the complete route, averages only 9.5%. These results indicate that general capabilities can partly reflect relative differences among models, but are insufficient to ensure stable multi-hop spatial reasoning, highlighting the need to consider changes in spatial reference, cross-hop dependence, and error propagation.

Early-Decision Error Propagation. To analyze early-error propagation, we continue inference from the same decision point under identical prior conditions, after either an incorrect or correct choice, and compare the resulting candidates, downstream routes, and spatial relations.

Table 6: Downstream spatial-state changes induced by an early decision error.

Difference type	Metric	Result
Search space	Candidate-set Jaccard distance	93.7%
	Downstream-route Jaccard distance	94.9%
Geometric relations	Any-candidate feasibility flip rate	40.1%
	Shared-candidate angular-deviation change	48.8°

Table 7: Downstream reasoning performance on matched incorrect and correct branches.

Metric	Incorrect branch	Correct branch
Next-hop accuracy	5.7%	36.9%
Remaining-route match rate	3.8%	26.7%
Final-answer accuracy	23.9%	25.5%

Table 6 shows that a single incorrect choice can substantially alter the downstream search space: the candidate sets and reachable routes of the incorrect and correct branches are nearly disjoint, with Jaccard distances of 93.7% and 94.9%, respectively. More importantly, in 40.1% of the comparisons, an incorrect spatial anchor directly changes which next-hop candidates satisfy the direction or distance constraints. In other words, an early error not only shifts the current location but also redefines subsequent candidates and their spatial relations, allowing the incorrect branch to remain locally plausible under the new reference frame and continue reasoning. We further compare the downstream reasoning induced by incorrect and correct choices at the same decision point. As

shown in Table 7, continuing along the incorrect branch yields only 5.7% Next-hop Accuracy and 3.8% Remaining-route Match Rate; replacing that decision with the correct choice improves the two metrics by 31.2 and 22.9 percentage points, respectively. In contrast, Final Answer Accuracy improves by only 1.6 percentage points because multiple routes may map to the same answer. This shows that evaluating only the final answer can conceal severe route-level error propagation.

B METHOD DETAILS

Spatial Reference and Geometric Scoring. Bearings are measured clockwise from north. For $p_t \neq \emptyset$, h_t^{in} points from the preceding anchor to $x(p_t)$, while h_t^{road} is the tangent of its associated (otherwise nearest) road edge, oriented toward whichever direction is closer to h_t^{in} (retaining the stored orientation on a tie); at p_0, the stored edge orientation is used when available. Absolute instructions use fixed world bearings. Incoming- and road-relative instructions add $0°$, $+90°$, $180°$, or $-90°$ for ahead, right, behind, or left to the corresponding reference modulo $360°$ to obtain the target direction in Γ_t. A relative constraint is evaluated only when its reference is defined; wrap maps angular differences to $[-180°, 180°)$. These headings depend on selected anchors and local road geometry, not on the shortest path used for the final distance answer.

For candidate z, let $\theta(z)$ denote its bearing from $x(p_t)$ and $\theta^\star$ the target direction. The bearing violation and block count are defined as

$$d_{\text{bearing}} = \max\left\{0, \frac{|\operatorname{wrap}(\theta(z) - \theta^\star)|}{50°} - 1\right\}, \qquad B(z) = \max\left\{1, \left\lceil \frac{\|x(p_t \oplus z) - x(p_t)\|_2}{180\,\text{m}} \right\rceil\right\}. \tag{14}$$

For an instructed distance interval $[b_-, b_+]$, we define the dimensionless distance violation d_{dist} and combine it with the dimensionless bearing violation d_{bearing} to obtain the total geometric deviation:

$$d_{\text{dist}} = \max\{b_- - B(z), 0, B(z) - b_+\}, \qquad \rho_{p_t}(z) = d_{\text{bearing}} + d_{\text{dist}}. \tag{15}$$

Visual Score Extraction. Each candidate is independently classified as support, contradiction, or unknown. Let q_{s}, q_{c}, and q_{u} be the softmax-normalized next-token scores. The binary log odds in Eq. 4 are $\log(q_{\text{s}}/q_{\text{c}})$, excluding the unknown score. Temperature T and bias b are fitted by binary negative log-likelihood on labeled matches and fixed at test time.

RaCE Route–Answer Inference. Algorithm 1 gives the complete inference procedure, where $\mathcal{P}$ contains all nonterminal histories on feasible routes and $\mathcal{U} \subseteq \mathcal{P}$ the unresolved ones. Route and answer maximizations use the same fixed lexicographic order as exhaustive inference only when scores are exactly equal.

To prioritize unresolved histories, RaCE constructs a temporary route score $\widetilde{s}(r)$ using $\widetilde{\nu}_p(z) = \nu_p(z)$ for queried histories and $\widetilde{\nu}_p(z) = \nu_0$ otherwise, where $\nu_0 \in [\epsilon, 1]$ is the empirical match frequency on the calibration set:

$$\widetilde{s}(r) = \sum_{t=1}^{L} \left[\log \pi(z_t \mid p_{t-1}) + \log \widetilde{\nu}_{p_{t-1}}(z_t) - \lambda_{\text{geo}} \rho_{p_{t-1}}(z_t)\right]. \tag{16}$$

$\mathrm{EarliestUnresolved}(p, q)$ selects the earliest unresolved history shared by the two routes, or otherwise the earliest on p and then q, with histories ordered by depth and candidate ID; it returns $\perp$ if none exists. The temporary score affects query order only, while unresolved evidence bounds remain $[\epsilon, 1]$ for stopping decisions.

Correctness and Complexity. For an unresolved history, the visual score lies in $\epsilon, 1$ and collapses to its true value after querying. Since the candidate prior and geometric penalty are fixed, local evidence is monotonic in the visual score, so local, route, and answer evidence remain within their bounds. When both certificates in Eq. 13 hold, no remaining observation can change the best route or answer, with $\max \varnothing = -\infty$ when no competitor exists. Otherwise, each round resolves one history in finite $\mathcal{U}$, so the algorithm terminates within $|\mathcal{P}|$ queries and matches exhaustive acquisition once all histories are queried under identical scoring and tie-breaking.

With at most K candidates per history and depth L, exhaustive acquisition requires at most $1 + K + \cdots + K^{L-1}$ visual queries, with up to K^L complete routes and $O(LK^L)$ explicit scoring cost. Thus, RaCE reduces visual-query cost rather than the worst-case route space.

Algorithm 1 RaCE route–answer inference with adaptive visual evidence acquisition

Require: Instance $\mathcal{Q} = (I, G, x_0, \mathcal{V}, L)$; calibrated VLM; $T, b, \epsilon, \lambda_{\mathrm{geo}}, \nu_0$
Ensure: Predicted route $\widehat{r}$ and answer $\widehat{a}$

Construct the finite, nonempty route set $\mathcal{R}$ from I, G. ▷ equation 1
$\mathcal{P} \leftarrow \{r_{1:t} \mid r \in \mathcal{R},\ 0 \le t < L\}$, with $r_{1:0} = \emptyset$.
for each $p_t \in \mathcal{P}$ **do**
    Construct S_t from p_t. ▷ equation 2, equation 8
    Compute $\pi(z \mid p_t)$ and $\rho_{p_t}(z)$ for all $z \in \mathcal{C}(p_t)$. ▷ equation 5
    Set $[\underline{\nu}_{p_t}(z), \overline{\nu}_{p_t}(z)] \leftarrow [\epsilon, 1]$ for every $z \in \mathcal{C}(p_t)$.
end for
Initialize the unresolved histories $\mathcal{U} \leftarrow \mathcal{P}$.
while true **do**
    Compute local evidence intervals. ▷ equation 7, equation 11
    Accumulate route and answer bounds. ▷ equation 9
    Select leaders $r^\star$ and $a^\star$. ▷ equation 12
    if both certificates hold **then** ▷ equation 13
        return $(\widehat{r}, \widehat{a}) \leftarrow (r^\star, a^\star)$.
    end if
    if $\mathcal{U} = \varnothing$ **then**
        return the exact maximizers $(\widehat{r}, \widehat{a})$. ▷ equation 10
    end if
    Compute temporary route scores $\widetilde{s}(r)$. ▷ equation 16
    if the route certificate fails **then**
        Let $\widetilde{r}$ maximize $\widetilde{s}(r)$ and r^{c} maximize $\overline{\mathcal{E}}(r)$ over $r \neq \widetilde{r}$.
        $p_t^\star \leftarrow \textsc{EarliestUnresolved}(\widetilde{r}, r^{\mathrm{c}})$.
    else
        Let a^{c} maximize $\overline{W}(a)$ over $a \neq a^\star$.
        Let r^+ maximize $\widetilde{s}(r)$ in $\mathcal{R}_{a^\star}$ and r^- maximize $\overline{\mathcal{E}}(r)$ in $\mathcal{R}_{a^{\mathrm{c}}}$.
        $p_t^\star \leftarrow \textsc{EarliestUnresolved}(r^+, r^-)$.
    end if
    if $p_t^\star = \perp$ **then**
        $p_t^\star \leftarrow$ the first history in $\mathcal{U}$.
    end if
    Query $p_t^\star$ and score all candidates. ▷ equation 4, equation 6
    Set $[\underline{\nu}_{p_t^\star}(z), \overline{\nu}_{p_t^\star}(z)] \leftarrow [\nu_{p_t^\star}(z), \nu_{p_t^\star}(z)]$, and $\mathcal{U} \leftarrow \mathcal{U} \setminus \{p_t^\star\}$.
end while

Table 8: Baseline adaptation and shared experimental settings.

Item	Protocol
Output	The model outputs a candidate identifier that determines the next spatial reference; unparseable, incomplete, or out-of-set outputs are counted as incorrect.
Beam-4	Ranks routes by accumulated stepwise candidate probabilities from the backbone, without RaCE route or answer scores.
Spatially post-trained models	Retain their original spatial reasoning procedures and use a shared adapter for candidate parsing.
Tool-augmented agents	Retain their original image, bearing, and distance tool interactions.
Common settings	Use deterministic decoding with identical image preprocessing and candidate representations.

C Experimental Protocol

Implementation and Measurement Protocol. All methods share identical instructions, decision histories, local observations, reachable candidates, and road geometry. Table 8 summarizes baseline adaptations and shared settings; methods differ only in candidate selection, tool interaction, and visual evidence acquisition.

The answer map f sums road-network shortest-path distances along the selected anchors from x_0, including snap-to-graph offsets, and maps the total to A–D using $[0, 600)$, $[600, 900)$, $[900, 1200)$,

and $[1200, \infty)$ m. RaCE uses $\epsilon = 10^{-12}$, with $\nu_0 = 0.5433747412$ for Qwen3.5. Histories are ordered by depth and candidate ID; exact route and answer ties are broken lexicographically, and both certificates use the strict comparisons in Eq. 13. For RQ3, the first-hop top-two local-score gap is split by validation-set tertiles at 2.1066 and 5.4608.

Latency is measured from the first visual query to final route and answer generation, excluding model loading and warm-up, while Peak VRAM is the maximum GPU memory used during inference. VLM Calls counts logical visual-query events. For RaCE, each queried history counts once and evaluates all its candidates, so Calls equals the number of distinct queried histories but not necessarily model forward passes; repeated visits by other methods are counted separately. Text processing, geometric computation, and cache access are excluded.

Evaluation Metrics. Let the evaluation set contain N instances. For instance i, let $\widehat{r}_i = (\widehat{z}_{i,1}, \ldots, \widehat{z}_{i,L})$ denote the predicted route and $\mathcal{R}_i^\star$ the set of valid reference routes. Let $H_i^{(k)}$ indicate whether the length-k decision history of the predicted route matches that of any valid reference route:

$$H_i^{(k)} = \mathbf{1}[\exists r \in \mathcal{R}_i^\star : \widehat{r}_{i,1:k} = r_{1:k}], \qquad \mathrm{ACC}^{(k)} = 100\frac{1}{N}\sum_{i=1}^{N} H_i^{(k)}, \tag{17}$$

where $\mathbf{1}[\cdot]$ is the indicator function, and $k = 1$, 2, and L correspond to Hop-1 ACC, First-2 ACC, and Route ACC, respectively. Let $\widehat{a}_i$ and $a_i^\star$ denote the predicted and reference answers. Route Retention measures the proportion of first-hop-correct predictions that remain correct through the complete route. Answer ACC and Route Retention are

$$\text{Answer ACC} = 100\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\widehat{a}_i = a_i^\star], \qquad \text{Route Retention} = 100\frac{\text{Route ACC}}{\text{Hop-1 ACC}}. \tag{18}$$

Answer ACC is evaluated independently of Route ACC because multiple valid routes may map to the same answer.

Let Q_i denote the number of logical visual-query events, t_i the measured inference time, and $m_{i,s}(u)$ the GPU memory usage of instance i in run s at time u. We report

$$\text{VLM Calls} = \frac{1}{N}\sum_{i=1}^{N} Q_i, \qquad \text{Latency} = \frac{1}{N}\sum_{i=1}^{N} t_i, \qquad \text{Peak VRAM} = \max_{i,s,u} m_{i,s}(u). \tag{19}$$

For adaptive acquisition, let $(\widehat{r}_i^{\mathrm{ada}}, \widehat{a}_i^{\mathrm{ada}})$ and $(\widehat{r}_i^{\mathrm{full}}, \widehat{a}_i^{\mathrm{full}})$ denote the route and answer outputs of adaptive and exhaustive acquisition, respectively. Their agreement rates are

$$\text{Route Agreement} = 100\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\widehat{r}_i^{\mathrm{ada}} = \widehat{r}_i^{\mathrm{full}}], \qquad \text{Answer Agreement} = 100\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\widehat{a}_i^{\mathrm{ada}} = \widehat{a}_i^{\mathrm{full}}]. \tag{20}$$

Let Q_i^{ada} and Q_i^{full} denote their respective query counts. Query reduction is

$$\text{Query Reduction} = 100\left(1 - \frac{\sum_{i=1}^{N} Q_i^{\mathrm{ada}}}{\sum_{i=1}^{N} Q_i^{\mathrm{full}}}\right). \tag{21}$$

For early-error propagation, let N_{err} denote the number of matched correct–incorrect branch pairs, with $+$ and $-$ denoting the correct and incorrect branches, respectively. For pair j, let $\mathcal{C}_j^+$ and $\mathcal{C}_j^-$ be the next-hop candidate sets, and $\mathcal{T}_j^+$ and $\mathcal{T}_j^-$ the reachable downstream route-suffix sets. For nonempty sets A and B, define their Jaccard distance as

$$d_{\mathrm{Jac}}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}. \tag{22}$$

The mean candidate- and route-space divergences are

$$D_{\mathrm{cand}} = \frac{1}{N_{\mathrm{err}}}\sum_{j=1}^{N_{\mathrm{err}}} d_{\mathrm{Jac}}(\mathcal{C}_j^+, \mathcal{C}_j^-), \qquad D_{\mathrm{route}} = \frac{1}{N_{\mathrm{err}}}\sum_{j=1}^{N_{\mathrm{err}}} d_{\mathrm{Jac}}(\mathcal{T}_j^+, \mathcal{T}_j^-). \tag{23}$$

Let $F_j^{\pm}$ indicate whether the corresponding branch contains at least one candidate satisfying both direction and distance constraints. The feasibility-flip rate is

$$R_{\text{flip}} = \frac{1}{N_{\text{err}}} \sum_{j=1}^{N_{\text{err}}} \mathbf{1}[F_j^+ \neq F_j^-]. \tag{24}$$

For the shared candidate set $\mathcal{H}_j = \mathcal{C}_j^+ \cap \mathcal{C}_j^-$, let $\delta_j^{\pm}(z)$ denote candidate z's circular deviation from the target direction, in degrees, under the two branches. Let $N_{\text{sh}} = |\{j : \mathcal{H}_j \neq \varnothing\}|$. The mean directional change is

$$\Delta_{\text{ang}} = \frac{1}{N_{\text{sh}}} \sum_{j:\mathcal{H}_j \neq \varnothing} \frac{1}{|\mathcal{H}_j|} \sum_{z \in \mathcal{H}_j} \left| \delta_j^+(z) - \delta_j^-(z) \right|. \tag{25}$$

Jaccard distances and feasibility-flip rates are reported as percentages.