DialogueVPR: Towards Conversational Visual Place Recognition

Yukun Song[1,*] Changwei Wang[2,*] Xingtian Pei[1,*] Shibiao Xu[1]†
Wenhao Xu[1] Shunpeng Chen[1] Yu Zhang[3] Ke Zhang[1]
Rongtao Xu[4] Xuxiang Feng[5,6]† Pengyang Wang[5]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology [3] Macquarie University
[4] Spatialtemporal AI [5] University of Macau [6] Aerospace Information Research Institute
{shibiaoxu@bupt.edu.cn, fengxx@aircas.ac.cn}

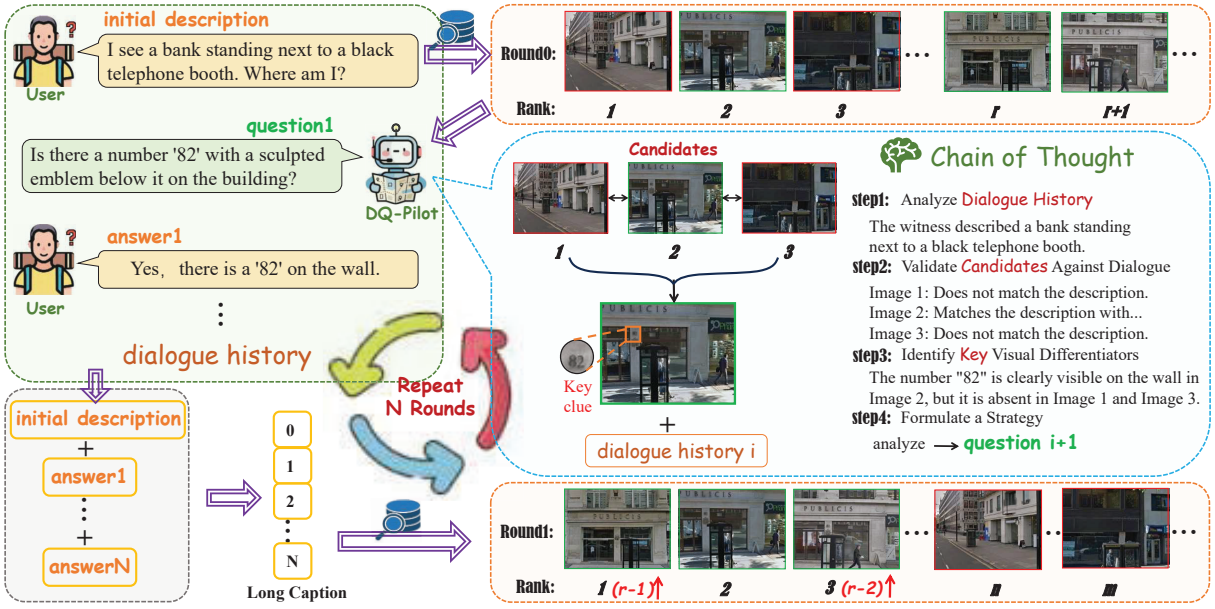

Figure 1. **An Illustration of the Dialogue Place Recognition (DlgPR) Framework.** The key component, DQ-Pilot, functions as a reasoning agent that transforms geolocalization from a simple one-shot "retrieval" into a sophisticated "reasoning-based retrieval" process: the user provides an "initial description", and the retriever CMPL performs a preliminary retrieval (Round 0) to generate multiple visually similar candidate locations. DQ-Pilot formulates high-information-gain questions regarding the "candidate locations" and "dialogue history" through a chain of thought. The user responds to this question with crucial new information, and such feedback is integrated into the dialogue history to form a more detailed context for the subsequent retrieval round. This iterative loop of "analysis-questioning-optimization" enables the system to progressively resolve ambiguities and accurately identify the target location.

Abstract

Inspired by how humans communicate spatial information, language-guided geo-localization has gained significant traction for its intuitive and practical value. Despite this progress, most methods still rely on a static, one-shot retrieval paradigm, which fails to handle the ambiguity and incompleteness inherent in real-world natural language descriptions. We propose a paradigm shift to reasoning retrieval and introduce Dialogue Place Recog-

†Corresponding authors. *Equal contribution.

nition (DlgPR), which casts localization as an interactive, dialogue-driven reasoning process. To support this new task, we present DlgQuest-Cities, the first large-scale dialogue-based benchmark for place recognition, and a unified reasoning framework that couples a cross-modal multi-level retriever with an intelligent questioner, DQ-pilot. DQ-pilot is trained in a curriculum: supervised fine-tuning on a curated DQ-cities-20k subset followed by reinforcement refinement on a harder DQ-cities-10k split via GRPO. Two task-aligned metrics guide learning: a Discriminative Difficulty Index (DDI) for curriculum sampling and a Positional Retrieval Gain (PRG) reward that directly measures retrieval improvement induced by a question. Experiments show this reasoning-based approach significantly outperforms baselines. The code and model are available at https://github.com/Graysonggg/DlgPR.

1. Introduction

Accurately perceiving and determining one's location remains a fundamental challenge for both humans [38] and intelligent agents [32]. Solving this problem underpins a wide range of applications, including precise pedestrian navigation in urban environments, autonomous robot operation in dynamic scenes, and localization correction in GPS-denied areas such as urban canyons [42]. Motivated by these demands, community's recent research has explored a more intuitive paradigm—place recognition driven by natural language descriptions [12, 17, 42, 44, 48]. Reflecting everyday human interactions, these approaches holds strong practical value: a passenger verbally guiding a taxi driver [50], identifying a place through spoken directions, or describing the surroundings in an emergency call [6], or commanding a home service robot through natural language [52].

Recent language-driven localization methods, such as Text2Pose [17] and Text2Loc [44], primarily focus on identifying individual locations within 3D point clouds. However, constructing and storing large-scale 3D maps remains costly, hindering practical deployment. Instead, recent work [28, 50] frames the problem as a large-scale retrieval task by correlating natural language with expansive, readily available visual data like satellite or street-view images. Despite progress, most language-guided localization methods still follow a static retrieval paradigm, where a fixed textual query is processed once to return the best-matching location. The fundamental limitation of this design lies in its *passivity*: it fails to handle the ambiguity inherent in real-world descriptions. When the initial input is vague or incomplete—such as an imprecise verbal account (the "user description dilemma") or an erroneous recollection—these systems cannot actively seek clarification or gather additional information. Consequently, single-turn, non-interactive retrieval remains fragile in dynamic, real-world scenarios.

To transcend these constraints, we argue that geo-localization should evolve from passive retrieval to an advanced paradigm of reasoning Retrieval. An intelligent agent must move beyond passive matching toward active understanding, reasoning, and interaction with uncertain environments and ambiguous human instructions.

To drive this paradigm shift, we introduce Dialogue Place Recognition (DlgPR)—a new task that reformulates localization as an iterative, collaborative dialogue. In DlgPR, the system transforms from a passive retriever into an active reasoner: it analyzes candidate locations, proactively engages the user with targeted questions to obtain discriminative evidence, and incrementally refines its belief about the correct place as the dialogue history becomes richer and the information more complete. Specifically, we develop a unified reasoning framework composed of a Cross-Modal Progressive Learning (CMPL) Retriever and an intelligent Multimodal Large Language Model, Dialogue-Quest-Pilot (DQ-pilot). The CMPL retriever is responsible for iteratively integrating information from the evolving dialogue to refine its search and retrieve relevant candidate locations. These candidates are then passed to DQ-pilot, which acts as the reasoning core—diagnosing ambiguity and generating questions to maximize information gain. This synergy transforms the system from a passive retriever into an active reasoner, enabling efficient and precise localization by incrementally refining its belief as the dialogue unfolds.

Our main contributions are summarized as follows:

- We propose a novel task, Dialogue Place Recognition (DlgPR), which shifts the paradigm from static retrieval to active, dialogue-driven reasoning. To facilitate research on this new task, we construct DlgQuest-Cities (DQ-cities), the first large-scale benchmark dataset for dialogue-based place recognition.
- We develop DlgQuest, a unified and effective reasoning framework featuring a cross-modal retriever (CMPL) and an MLLM agent (DQ-Pilot). Crucially, to train this framework, we introduce a novel curriculum learning strategy guided by two task-aligned metrics—a Discriminative Difficulty Index (DDI) and a Positional Retrieval Gain (PRG)—enabling the agent to learn progressively from basic perception to advanced reasoning. Extensive experiments demonstrate the superiority of our approach.

2. Related Work

2.1. Natural Language-Driven Visual Perception and Localization

Geo-localization [2–4, 8, 9, 13, 14, 22, 25, 26, 30, 36, 39, 45, 49] predicts a query's location by retrieving similar images from a geo-tagged database. Recently, multi-modal

retrieval incorporating natural language has emerged in this field [11, 12, 28, 33, 37, 40]. For example, [50] introduces scene text, breaks through the limitation of text length, and, for the first time, introduces an interpretability framework, ensuring the localization process is no longer a black box. Meanwhile, [7] enhances the model's spatial perception capabilities by learning phrases that describe fine-grained spatial relationships in natural language through Blending Spatial Matching. In 3D localization, [17] uses natural language instructions for position matching in point clouds, while [44] advances this by directly fusing textual semantics with geometric features for end-to-end position regression. For indoor recognition, [37] refines ranking using discriminative text filtered from images. Despite their success, these text-driven geolocation tasks remain largely static and lack dynamic interaction capabilities.

2.2. Interactive Retrieval

Cross-modal interactive retrieval has been actively explored in text-to-image [20, 21, 27, 53] and text-to-video domains [23, 29], encompassing various interaction formats [5, 18, 19]. For example, [23] diversifies question generation, while PlugIR [20] decouples dialogue understanding from retrieval via LLMs, enabling compatibility with black-box models. Furthermore, LLaVA-ReID [27] generates questions maximizing information gain through forward-looking supervision.

Ultimately, interactive retrieval aims to replicate human-like logical reasoning. However, current multi-turn dialogue methods primarily perform reactive information aggregation based on explicit feedback, lacking deeper proactive reasoning capabilities. In contrast, our work pioneers the first multi-modal interactive reasoning task in the field of geolocation.

2.3. Visual Reinforcement Learning

The advent of the OpenAI's o1 [15] and DeepSeek-R1 reasoning model [10] introduced the paradigm of incorporating visual reasoning into visual tasks. Reinforcement learning (RL) is pivotal for endowing models with reasoning capabilities, and Group Relative Policy Optimization (GRPO) [34], characterized by its verifiable rewards, has emerged as a prominent RL methodology. Building on this, VLM-R1 [35] developed multiple verifiable reward functions to fine-tune Vision-Language Models (VLMs). Subsequently, Visual-RFT [24] formulated simple yet effective reward functions for diverse visual tasks, further enabling efficient learning under data-scarce conditions. Existing research demonstrates that, compared to Supervised Fine-Tuning (SFT), GRPO facilitates deeper reasoning, offers greater interpretability through its reasoning process, and exhibits superior generalization under limited supervision. Therefore, our proposed framework, DlgQuest, employs both SFT and GRPO to achieve active, reasoning-based geolocation.

3. DlgQuest-Cities

3.1. Overview

To support the dialogical reasoning required by our proposed DlgPR task, we construct the DlgQuest-Cities (DQ-cities) dataset. This new benchmark is built upon the widely-used GSV-Cities collection [1], augmenting its rich geo-tagged imagery with multi-layered annotations tailored for dialogue-based localization. Each location in DlgQuest-Cities is annotated with information specifically designed for interactive spatial reasoning. Specifically, the dataset includes: (1) Initial ambiguous place captions, simulating users' vague or uncertain verbal queries based on incomplete memories; (2) Fine-grained place descriptions, offering comprehensive visual semantic details that serve as the factual foundation for multi-turn reasoning; (3) Region-level annotations, where bounding boxes are paired with corresponding textual descriptions to provide localized evidence for spatial grounding; and (4) Multi-turn, goal-oriented dialogues, in which each question is purposefully designed to differentiate visually similar locations and progressively resolve ambiguity. DQ-cities in total consists of 106,880 location images and 30k carefully selected conversation samples. Each fine-grained description has an average of 154.6 words, with the maximum reaching up to 262 words.

3.2. Dataset Construction

The rich annotations in DlgQuest-Cities are generated via an automated, multi-stage pipeline designed to produce the textual and dialogical data needed to train DQ-pilot. This pipeline, illustrated in Fig.2, is specifically engineered to synthesize strategy-aware dialogues for each place. It consists of four principal stages: (1) text-modality expansion, (2) text-driven region-level annotation, (3) Chain-of-Thought (CoT) based dialogue generation guided by GPT-4o and (4) Curriculum sampling based on discrimination difficulty.

Step 1: Text Modality Expansion. This initial stage is responsible for creating the foundational textual layers for each place. To emulate a user's initial query, the pipeline first generates an initial ambiguous caption (e.g., "I see a bank with a telephone booth beside it."). Following this, the system expands the caption into a long-form, fine-grained place description. This detailed narrative serves as a fact-rich foundation for subsequent dialogue generation. The generation process is constrained by a structured prompt (see Appendix) with task-specific rules: it focuses strictly on static elements (e.g., buildings, signage, spatial relationships), disregards transient objects (e.g., cars), and empha-

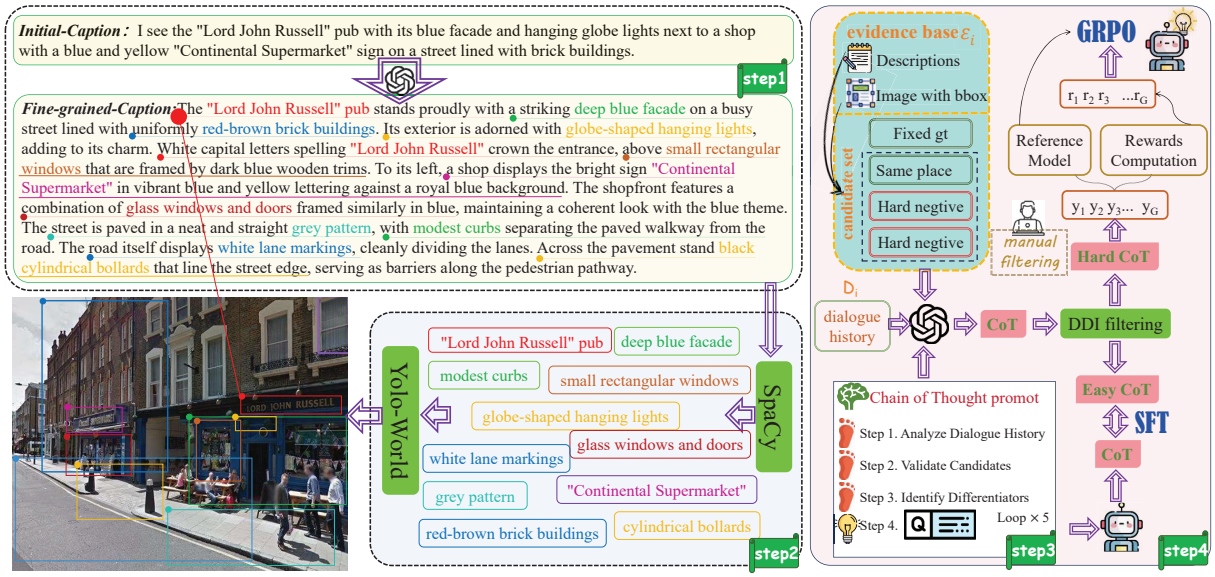

Figure 2. Dataset Construction Flowchart. It is mainly divided into 4 parts: Text modality expansion; Region-level visual evidence construction; Chain-of-Thought dialogue generation; Discriminative difficulty-aware sampling.

sizes features informative for place recognition. This procedure ensures the resulting descriptions provide reliable, factually grounded information for downstream reasoning.

Step 2: Region-Level Visual Evidence Construction To ground the dialogue in specific visual details, DlgQuest-Cities incorporates region-level annotations. We begin by extracting salient noun phrases from the fine-grained descriptions. Unlike approaches such as FG-CLIP [47] that often rely on simple nouns, we employ a greedy expansion strategy with *spaCy* to capture maximally descriptive phrases, including rich adjectival modifiers and prepositional clauses (e.g., "the red brick bank with green awnings"). These descriptive phrases serve as more effective text prompts for an open-vocabulary detector (YOLO-World), enabling it to localize the corresponding objects with greater precision. The final output is a set of structured annotations, where each annotation links a specific image region (the bounding box) to its corresponding textual phrase. This step enriches the visual evidence base of the place, effectively avoiding the omission of key details by the teacher model.

Step 3: Chain-of-Thought Dialogue Generation This stage constructs the interactive reasoning samples that power the DlgPR training process. For each place, we synthesize a five-round dialogue sequence, where each round simulates one reasoning–questioning cycle of the teacher model.

At each dialogue round i, the pipeline assembles a decision-making context composed of: (1) a compact and distinctive candidate set—comprising the target image I_{t}, positive samples I_{p} from the same location, and two challenging negatives $I_{\mathrm{n1}}, I_{\mathrm{n2}}$ retrieved by trained CMPL; (2) the evidence base $\mathcal{E}_i = \{(I_{\mathrm{t}}, t_{\mathrm{t}}, B_{\mathrm{t}}), (I_{\mathrm{p}}, t_{\mathrm{p}}, B_{\mathrm{p}}), (I_{\mathrm{n}}, t_{\mathrm{n}}, B_{\mathrm{n}})\}$, where t and B denote the textual descriptions and bounding boxes obtained in Steps 1 and 2; and (3) the accumulated dialogue history D_i.

Next, the teacher model (GPT-4o) is prompted to execute a four-step chain-of-thought before composing the next question:

- **Analyze Dialogue History:** summarize confirmed and ruled-out evidence contained in D_i;
- **Validate Candidates Against Dialogue:** compare each candidate's evidence in $\mathcal{E}_i$ with D_i and eliminate inconsistent ones;
- **Identify Key Visual Differentiators:** examine the remaining candidates within $\mathcal{E}_i$ to pinpoint region-grounded, text-anchored cues that most clearly distinguish them;
- **Formulate a Strategy:** Design a question that targets the most decisive visual uncertainty to maximize information gain.

Finally, the teacher's internal deliberation and proposed question are wrapped in $\langle think \rangle \langle /think \rangle$ and $\langle question \rangle \langle /question \rangle$.

Step 4: Discriminative Difficulty-Aware Curriculum Sampling To ensure DQ-pilot learns progressively from simple to complex scenarios, we introduce a curriculum-aware sampling strategy. This strategy is guided by a unified Discriminative Difficulty Index (DDI), a weighted

score combining two complementary metrics: Semantic Ambiguity (SA) and Retriever-Informed Difficulty (RID).

Semantic Ambiguity (SA). SA quantifies the intrinsic ambiguity of a candidate set. Given positive and negative textual embeddings t_{p} and t_{n}, and their corresponding visual embeddings, we compute:

$$\begin{aligned}\mathrm{SA} = \alpha \cdot \mathrm{sim}\big(\phi_T(t_{\mathrm{t}}), \phi_T(t_{\mathrm{n}})\big) \\ + (1-\alpha) \cdot \big(1 - \mathrm{sim}(\phi_T(t_{\mathrm{t}}), \phi_T(t_{\mathrm{p}}))\big).\end{aligned} \tag{1}$$

where $\phi_T(\cdot)$ is the text encoder of CMPL. A higher SA indicates stronger semantic overlap and thus greater ambiguity among candidates.

Retriever-Informed Difficulty (RID). RID measures the empirical difficulty of a dialogue turn by quantifying the rank improvement of positive samples after answering the generated question. Let $r_j^{(i-1)}$ and $r_j^{(i)}$ be the rank of a positive item $j \in \mathcal{P}$ before and after dialogue round i. The *Positional Retrieval Gain* (PRG) normalizes the observed rank improvement against the maximum possible improvement:

$$\mathrm{PRG}_i = \frac{G^{(i)} - G^{(i-1)}}{G^* - G^{(i-1)}}, \tag{2}$$

where gain G is the sum of nDCG-style [43] contributions $c(r) = 1/\log_2(r+1)$ over all items in $\mathcal{P}$, and G^* represents the ideal total gain if all positive items occupied the top ranks ($G^* = \sum_{k=1}^{|\mathcal{P}|} c(k)$). We then set $\mathrm{RID}_i = 1 - \mathrm{PRG}_i$, so that minimal rank improvement (low PRG) corresponds to high empirical difficulty.

DDI-based Curriculum Sampling. We first filter out low-quality dialogues (e.g., with minimal rank changes, $\mathrm{PRG}_i < \tau_1$) and overtly noisy ones using automated metrics. This automated screening is complemented by a brief manual inspection, primarily focused on borderline cases, to ensure overall data integrity. For the resulting filtered pool, we compute the final difficulty score:

$$\mathrm{DDI} = w_{\mathrm{sa}} \cdot \mathrm{SA} + w_{\mathrm{rid}} \cdot \mathrm{RID}. \tag{3}$$

Using a threshold on the DDI score, we construct a two-stage curriculum:

- **Stage 1 (For Supervised Fine-Tuning):** We sample 20k instances as DQ-cities-20k, prioritizing low-DDI samples ($\sim$70%). This stage focuses on learning fundamental visual grounding and core reasoning patterns.
- **Stage 2 (For Reinforcement Learning):** We sample 10k instances as DQ-cities-10k, prioritizing high-DDI samples ($\sim$70%). This stage challenges the model with highly ambiguous and hard-to-distinguish cases.

This entire pipeline, from description generation to curriculum sampling, produces the final 30k dialogue rounds in DQ-Cities. The resulting dataset is not only rich in content but also structured to facilitate progressive learning, advancing the model from basic visual grounding to robust, evidence-backed reasoning. More dataset statistics and construction details are provided in Appendix.

4. Method

4.1. The Dialogue Place Recognition Framework

The DlgPR framework reframes place recognition as a dynamic, interactive reasoning process, departing from traditional static retrieval. It orchestrates two core components: a multi-modal retriever, CMPL, that iteratively refines the search and a dialogue agent, DQ-pilot, that generates discriminative questions to resolve ambiguity.

The process begins when an initial user query, d_0, yields a coarse set of candidate locations C_0 via the CMPL retriever. To disambiguate these candidates, the system enters an iterative loop. At each round t, DQ-pilot analyzes the current candidates C_t to formulate an optimal question q_t. Upon receiving the user's answer a_t, the framework aggregates the dialogue history into an enriched textual query $d_{t+1} = \mathrm{concat}(d_0, a_1, \ldots, a_t)$ for the CMPL retriever. This updated query d_{t+1} enables CMPL to perform a more informed retrieval, producing a refined candidate set C_{t+1}. This cycle of question-answering and retrieval progressively narrows the search space, achieving robust localization by resolving ambiguities through natural conversation.

4.2. Cross-Modal Progressive Learning Retriever

To support dialogue-driven reasoning, our Cross-Modal Progressive Learning (CMPL) retriever incrementally refines visual-textual alignment from local to global granularity.

Progressive Feature Alignment. We extract hierarchical visual patches $V^{(l)}$ and text tokens $T^{(l)}$ from intermediate layers $P = \{p_3, p_6, p_9, p_{12}\}$. To highlight geographically relevant cues, $V^{(l)}$ is refined into $V_s^{(l)}$ via a saliency filtering module (SFM) that dynamically selects discriminative tokens based on attention weights, supervised by an auxiliary loss L_{vpr} [41].

To bridge modality structures, we introduce a shared fine-grained extractor E_f and learnable *instance-concept queries* $Q^{(l)}$. Acting as semantic anchors, they distill $V_s^{(l)}$ and $T^{(l)}$ into unified representations:

$$F_v^{(l)} = E_f(Q^{(l)}, V_s^{(l)}), \quad F_t^{(l)} = E_f(Q^{(l)}, T^{(l)}). \tag{4}$$

Hierarchical Similarity Distribution Matching. We apply an SDM loss [16] at multiple granularities to minimize the bidirectional KL-divergence between the predicted similarity distribution p and the ground-truth q. For an image anchor $F_{v,i}$, its predicted distribution across B batch texts is:

$$p_{v\to t,i,j} = \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{B} \exp(s_{i,k}/\tau)}, \tag{5}$$

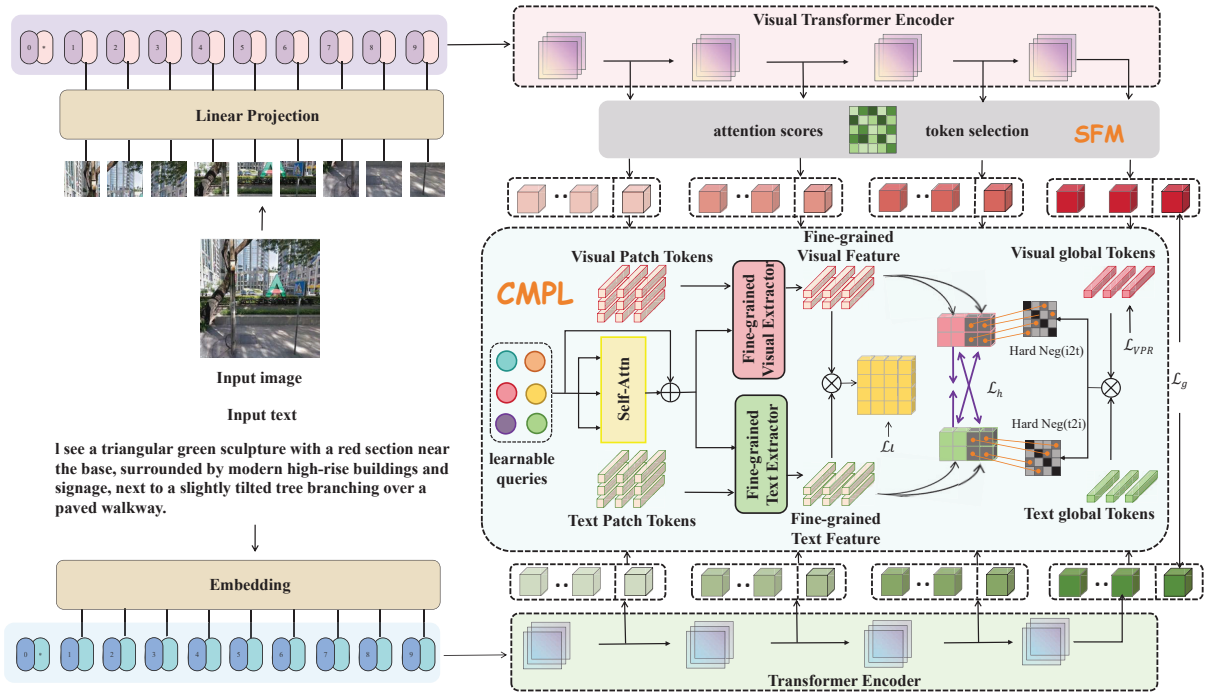

Figure 3. Flowchart of the proposed cross-modal progressive learning retriever. The core is the cross-modal progressive learning (CMPL) module, which aligns the global and local information of multi-level visual and textual features respectively, and mines hard negative samples for triplet loss learning.

where $s_{i,j}$ is the similarity score and τ is the temperature. The target q is normalized from binary batch labels.

Hard-Negative Isolation (HI). To further improve geometric separability, we propose a Hard-Negative Isolation (HI) loss that applies localized repulsion to the most confusing negatives within each batch. For an image–text pair $(F_{v,i}, F_{t,i})$, the hardest negatives j^* and k^* are selected by similarity, and the margin-based triplet objective $L_{\text{hi}} = [d(F_{v,i}, F_{t,i})^2 - d(F_{v,i}, F_{t,j^*})^2 + \alpha]_+ + [d(F_{t,i}, F_{v,i})^2 - d(F_{t,i}, F_{v,k^*})^2 + \alpha]_+$ enforces discriminative separation across modalities.

Overall Objective. We apply this hierarchically. The global loss (L_{gs}) uses cosine similarity between [CLS] tokens. For local losses ($L_{ls}^{(l)}$) from a set of intermediate layers $P = \{p_3, p_6, p_9, p_{12}\}$, the score $s_{i,j}^{(l)}$ is the mean similarity across all local tokens. The final training objective integrates hierarchical alignment and hard-negative isolation:

$$L_{\text{total}} = \lambda_{gs} L_{gs} + \lambda_h \sum_{l \in P} \left(L_{ls}^{(l)} + L_{hi}^{(l)} \right) + L_{vpr}. \quad (6)$$

4.3. Intelligent DQ-pilot

The DQ-pilot acts as a strategic visual reasoner, trained to formulate discriminative questions that enhance retrieval performance. Its training proceeds in two progressive stages: (1) Supervised Fine-Tuning (SFT) to establish foundational reasoning abilities, and (2) Reinforcement Learning (GRPO) to refine its question-generation strategy with task-aligned rewards.

Supervised Fine-Tuning (SFT). In the first training stage, DQ-pilot is fine-tuned on a carefully selected DQ-cities-20k subset of the DQ-Cities dataset using the standard next-token prediction objective. Each training instance corresponds to a single dialogue turn, where the input consists of the current dialogue history Q_i, the associated candidate set represented by `<image>` tokens, and an instruction that specifies the Questioner's reasoning goal and response format. The output is a structured reasoning trace followed by a well-formed discriminative question that effectively differentiates visually similar locations. Through this next-token prediction process, DQ-pilot learns to connect accumulated dialogue context with spatial ambiguity and to formulate questions that progressively guide the retriever toward the correct place. This stage establishes the model's foundational reasoning and dialogue abilities, providing a solid initialization for subsequent reinforcement refinement.

Reinforcement Learning via GRPO. To further enhance strategic behavior beyond imitation, we refine the SFT-

Table 1. Interactive multi-round retrieval performance across five representative regions. We report the recall at the 3rd and 5th rounds (initiated from a short initial query), along with the BRI evaluation metric. The best metrics are shown in **red bold**.

Method	Round	LosAngeles		BuenosAires		MexicoCity		Osaka		PRG		BRI↓
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	
Initial	round0	35.9	56.3	39.0	60.8	42.0	64.0	35.6	55.4	52.8	74.2	/
Qwen2.5-VL-7B	round3	42.4	59.3	45.5	64.1	47.2	65.9	41.2	57.9	58.0	74.6	1.58
	round5	43.2	60.1	46.2	64.6	48.2	66.8	42.1	58.5	59.1	74.9	
Qwen2.5-VL-72B	round3	46.1	65.2	49.3	69.2	52.4	71.7	46.2	64.7	62.9	80.0	1.44
	round5	49.5	68.6	51.9	71.4	54.6	74.0	49.1	67.5	65.1	82.1	
PlugIR	round3	48.1	67.0	50.3	70.9	52.5	72.2	47.3	65.6	63.9	80.2	1.41
	round5	51.2	70.3	53.2	72.5	55.7	75.1	50.5	68.8	66.2	83.2	
DlgQuest (SFT)	round3	49.2	68.4	51.8	71.5	53.6	73.7	48.1	66.9	64.1	81.5	1.29
	round5	54.6	73.6	55.9	74.7	57.8	77.4	53.3	71.6	68.4	85.3	
DlgQuest (SFT+GRPO)	round3	52.1	71.6	54.0	75.5	58.0	76.9	52.9	71.9	67.8	84.4	**1.18**
	round5	**58.4**	**76.5**	**59.3**	**79.6**	**61.8**	**80.1**	**58.6**	**76.7**	**71.4**	**86.6**	

initialized model on the more challenging DQ-cities-10k subset using GRPO reinforcement learning.

- **Format Reward (R_{fmt}).** To ensure consistent reasoning structure and interpretability, we define a binary reward verifying adherence to the required `<think></think><question></question>` template:

$$R_{\text{fmt}}(y) = \begin{cases} 1, & \text{if } y \text{ matches the required format,} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

- **Retrieval Reward (R_{prg}).** We reuse the Positional Retrieval Gain (PRG) from Sec 3.2 as a task-aligned measure of how effectively a generated question improves localization. Given the retriever's updated ranks at round t, the retrieval reward is defined as

$$R_{\text{prg}} = \text{PRG}_t, \tag{8}$$

which directly quantifies retrieval improvement induced by the model's question.

Final Objective. The scalar reward used for GRPO optimization is a weighted combination of these two components:

$$R = \alpha\, R_{\text{prg}} + \beta\, R_{\text{fmt}}, \tag{9}$$

where $\alpha, \beta > 0$ balance task performance and structural consistency. This reinforcement phase encourages DQ-Pilot to move beyond supervised imitation—learning to generate concise, discriminative, and retrieval-effective questions that actively steer the reasoning process within DlgPR.

5. Experiments

5.1. Experimental Setup

Dataset. All the experiments are conducted on our proposed DQ-cities dataset, and the evaluation is carried out for five representative cities from various continents. Each sample begins with a vague initial description (in the 0th round), and the questioner completes the retrieval through iterative dialogues. Table 1 reports results up to Round 5.
Evaluation Metrics. We evaluate the performance using cumulative Recall@K up to round r, where $k \in \{1, 5\}$, as the primary evaluation metric. In addition, the BRI index [20] is introduced as an indicator to measure the efficiency of each round of questioning.
Implementation Details. Our CMPL adopts a CLIP ViT-B/16 backbone and is trained using the proposed CMPL framework with fine-grained long descriptions as input, the number of learnable queries for each layer is set to 16. To handle long texts, we apply linear interpolation to the positional embeddings of tokens that exceed the original context length in the text encoder. The DQ-pilot is based on Qwen2.5-VL-7B-Instruct, fine-tuned with LoRA for parameter-efficient adaptation training follows our curriculum learning strategy : (1) Supervised Fine-Tuning (SFT) on low-DDI samples, and (2) GRPO-based reinforcement optimization on high-DDI samples. All the experiments are conducted on two A100s.

5.2. Main Results

Interactive Reasoning Retrieval. As summarized in Table 1, our fine-tuned DQ-pilot markedly surpasses both the original Qwen2.5-VL series and prior interactive retrieval methods. Compared to its 7B backbone, our model improves R@1 by 9.2% and 13.4% after 3 and 5 dialogue

Table 2. Static retrieval performance using fine-grained long descriptions across five representative regions. The best metrics are shown in **red bold**.

Method	LosAngeles			BuenosAires			MexicoCity			Osaka			PRG		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [31]	41.0	60.2	68.2	49.3	70.0	75.9	52.2	69.3	74.2	54.0	71.8	76.7	64.0	77.5	83.7
Long-CLIP [51]	46.8	65.9	73.1	54.9	75.7	80.1	57.0	74.9	79.0	59.2	76.5	82.0	69.1	82.6	88.1
FG-CLIP [47]	56.1	76.6	83.2	65.2	84.0	89.1	66.9	83.5	88.0	68.8	85.3	90.5	78.2	91.0	96.6
Flair [46]	57.9	76.2	79.9	66.1	85.2	90.3	66.5	83.2	88.2	68.3	85.1	89.2	78.1	89.1	95.9
CMPL(Ours)	**71.9**	**88.3**	**92.9**	**69.3**	**89.3**	**93.4**	**69.5**	**87.3**	**92.4**	**72.5**	**90.2**	**94.8**	**82.5**	**95.1**	**97.4**

Table 3. Ablation study on CMPL retriever components. The average value of per-city tests across five representative regions. The best metrics are shown in **red bold**.

Configurations	**R@1**	**R@5**	**R@10**
Baseline	71.6	88.0	95.3
+ Token Selection	71.9	88.5	95.9
+ Progressive hsdm	72.3	89.0	96.4
+ HI (Hard-negative Isolation)	72.7	89.5	97.0
Full (All components)	**73.2**	**90.0**	**97.5**

Table 4. Ablation studies on key components of our DQ-pilot's learning strategy. Final 5-round results are reported. The best metrics are shown in **red bold**.

Setting	R@1	R@5
DQ-pilot	**60.5**	**77.8**
w/o DDI Curriculum (Random)	59.6	77.2
w/o GRPO (SFT-30k)	59.1	76.6
w/o GRPO (SFT)	58.1	75.9

rounds, respectively, and even outperforms the much larger Qwen2.5-VL-72B by 7.3%. This indicates that our progressive alignment strategy and reward-optimized training effectively boost interactive reasoning beyond mere model scaling. We further include specialized interactive retrievers, PlugIR [20] as reference baselines to compare against established dialogue-driven retrieval pipelines. Our method achieves substantial gains in both R@1 and R@5 while maintaining the lowest BRI score, demonstrating superior interaction efficiency. It's also worth noting that the results highlight the advantage of our **SFT+GRPO** fine-tuning strategy: SFT provides structured reasoning alignment from supervised dialogues, while GRPO further promotes the model to achieve deeper reasoning. For more examples, please refer to the appendix.

Retriever Performance. We also evaluate the retriever under ideal conditions using complete long descriptions, representing the upper bound of static retrieval. As shown in Table 2, our retriever significantly surpasses the Clip-based models including the recent state-of-the-art fine-grained image-text alignment models, validating its strong fine-grained cross-modal alignment.

5.3. Ablation Studies

To quantify the contribution of each core component, we conducted detailed ablation experiments. For DQ-pilot, regarding the Discriminative Difficulty-Aware Sampling strategy, we conducted two additional sets of control experiments: The first group trained the SFT model using the combined dataset of the two parts, while the second group performed SFT + GRPO training with the same sample quantity using a random sampling strategy. The results in Table 4 demonstrated that compared to imitation learning (SFT), the GRPO strategy could guide the model to perform deeper reasoning, and the curriculum setting guided by DDI was reasonable. For CMPL, Table 3 reports the results of the ablation experiments. Here, Baseline indicates the use of only the L_{gs} loss. We sequentially add the vpr loss and salient feature selection to the baseline to illustrate the importance of the salient location patches in the scene localization task. Then, we apply the multi-layer progressive CMPL, local-sdm loss, and HI loss, achieving the performance, which indicates that fully exploiting the alignment between fine-grained features is necessary.

6. Conclusion

We present DlgPR, a new paradigm that transforms traditional static geo-localization into an interactive, reasoning-driven process. Built upon our large-scale benchmark DQ-Cities and a curriculum guided by the DDI, our reasoning framework featuring the intelligent questioner DQ-Pilot—learns to iteratively refine spatial understanding through dialogue. Extensive experiments demonstrate that this interactive reasoning approach significantly enhances localization robustness and efficiency, highlighting the importance of active questioning for real-world geo-localization. In future work, we plan to explore more adaptive dialogue policies, tighter retriever–questioner co-training, and real-time deployment strategies for embodied agents in open environments.

7. Acknowledgments

This work was supported by the Beijing Natural Science Foundation (No.JQ23014), National Natural Science Foundation of China (No.62271074), Taishan Scholars Program (No.TSQN202507241), Key R&D Program of Shandong Province, China (No.2025KJHZ013), Shandong Provincial University Youth Innovation and Technology Support Program (No.2022KJ291), Shandong Provincial Natural Science Foundation for Young Scholars Program (No.ZR2025QC1627), and Qilu University of Technology (Shandong Academy of Sciences) Major Project under Grant (No. 2025ZDZX02).

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 3

[2] Amar Ali-Bey, Brahim Chaib-draa, and Philippe Giguère. Global proxy-based hard mining for visual place recognition. *arXiv preprint arXiv:2302.14217*, 2023. 2

[3] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17794–17803, 2024.

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[5] Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. Ask&confirm: active detail enriching for cross-modal retrieval with partial query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1835–1844, 2021. 3

[6] Jiaqi Chen, Daniel Barath, Iro Armeni, Marc Pollefeys, and Hermann Blum. "where am i?" scene retrieval with language. In *European Conference on Computer Vision*, pages 201–220. Springer, 2024. 2

[7] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pages 213–231. Springer, 2024. 3

[8] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023. 2

[9] Sourav Garg, Madhu Vankadari, and Michael Milford. Seqmatchnet: Contrastive learning with sequence matching for place recognition & relocalization. In *Conference on Robot Learning*, pages 429–443. PMLR, 2022. 2

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[11] Ziyang Hong, Yvan Petillot, David Lane, Yishu Miao, and Sen Wang. Textplace: Visual place recognition and topological localization through reading scene texts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2861–2870, 2019. 3

[12] Jingqi Hu, Chen Mao, Chong Tan, Hui Li, Hong Liu, and Min Zheng. Progeo: Generating prompts through image-text contrastive learning for visual geo-localization. In *International Conference on Artificial Neural Networks*, pages 448–462. Springer, 2024. 2, 3

[13] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[14] Sergio Izquierdo and Javier Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In *Computer Vision – ECCV 2024*, pages 240–257, Cham, 2025. Springer Nature Switzerland. 2

[15] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 3

[16] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2787–2797, 2023. 5

[17] Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6687–6696, 2022. 2, 3

[18] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2):185–210, 2015. 3

[19] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021. 3

[20] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. *arXiv preprint arXiv:2406.03411*, 2024. 3, 7, 8

[21] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36: 61437–61449, 2023. 3

[22] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17397–17407, 2023. 2

[23] Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11091–11101, 2023. 3

[24] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 3

[25] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[26] Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. *Advances in Neural Information Processing Systems*, 37:5789–5816, 2024. 2

[27] Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025. 3

[28] Zonglin Lyu, Juexiao Zhang, Mingxuan Lu, Yiming Li, and Chen Feng. Tell me where you are: Multimodal llms meet place recognition. *arXiv preprint arXiv:2406.17520*, 2024. 2, 3

[29] Avinash Madasu, Junier Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 356–365, 2022. 3

[30] Qibo Qiu, Shun Zhang, Haiming Gao, Honghui Yang, Haochao Ying, Wenxiao Wang, and Xiaofei He. Emvp: Embracing visual foundation model for visual place recognition with centroid-free probing. *Advances in Neural Information Processing Systems*, 37:120928–120950, 2024. 2

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 8

[32] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019. 2

[33] Tianyi Shang, Zhenyu Li, Pengjie Xu, Jinwei Qiao, Gang Chen, Zihan Ruan, and Weijun Hu. Bridging text and vision: A multi-view text-vision registration approach for cross-modal place recognition. *arXiv preprint arXiv:2502.14195*, 2025. 3

[34] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[35] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 3

[36] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 2

[37] Huaqi Tao, Bingxi Liu, Calvin Chen, Tingjun Huang, He Li, Jinqiang Cui, and Hong Zhang. Textinplace: Indoor visual place recognition in repetitive structures with scene text spotting and verification. *arXiv preprint arXiv:2503.06501*, 2025. 3

[38] Huilin Tian, Jingke Meng, Wei-Shi Zheng, Yuan-Ming Li, Junkai Yan, and Yunong Zhang. Loc4plan: Locating before planning for outdoor vision and language navigation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4073–4081, 2024. 2

[39] Changwei Wang, Shunpeng Chen, Yukun Song, Rongtao Xu, Zherui Zhang, Jiguang Zhang, Haoran Yang, Yu Zhang, Kexue Fu, Shide Du, et al. Focus on local: Finding reliable discriminative regions for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7536–7544, 2025. 2

[40] Teng Wang, Lingquan Meng, Lei Cheng, and Changyin Sun. Lvlm-empowered multi-modal representation learning for visual place recognition. *arXiv preprint arXiv:2407.06730*, 2024. 3

[41] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5022–5030, 2019. 5

[42] Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems*, 36:5301–5319, 2023. 2

[43] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR, 2013. 5

[44] Yan Xia, Letian Shi, Zifeng Ding, Joao F Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from natural language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14958–14967, 2024. 2, 3

[45] Zimin Xia, Yujiao Shi, Hongdong Li, and Julian FP Kooij. Adapting fine-grained cross-view localization to areas without fine ground truth. In *European Conference on Computer Vision*, pages 397–415. Springer, 2024. 2

[46] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24884–24894, 2025. 8

[47] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-

clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025. 4, 8

[48] Junyan Ye, Jun He, Weijia Li, Zhutao Lv, Jinhua Yu, Haote Yang, and Conghui He. Skydiffusion: Street-to-satellite image synthesis with diffusion models and bev paradigm. *arXiv e-prints*, pages arXiv–2408, 2024. 2

[49] Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network. In *European Conference on Computer Vision*, pages 74–90. Springer, 2024. 2

[50] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Qi Zhu, Conghui He, and Weijia Li. Where am i? cross-view geo-localization with natural language descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5890–5900, 2025. 2, 3

[51] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024. 8

[52] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024. 2

[53] Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 978–987, 2024. 3