International Journal of Geo-Information

MDPI

Article

Synergizing Retrieval and CoT Reasoning via Spatial Consensus for Worldwide Visual Geo-Localization

Yong Tang [1,2], Jianhua Gong [1,2,3], Yi Li [1,*], Jieping Zhou [1] and Jun Sun [1]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; tangyong21@mails.ucas.ac.cn (Y.T.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Zhejiang-CAS Application Center for Geoinformatics, Jiaxing 314100, China
* Correspondence: liyi@aircas.ac.cn

Abstract

Worldwide visual geo-localization aims to predict the geographic coordinates of an image capture location from visual content alone, posing unique challenges due to the vast scale of the Earth's surface and pervasive visual ambiguity across distant regions. Existing approaches face distinct limitations as follows: retrieval-based methods demand massive geo-tagged databases and scale poorly; alignment-based models lack interpretability and are vulnerable to visually similar scenes; and large vision-language models (LVLMs) offer semantic reasoning but suffer from hallucination. A natural solution is retrieval-augmented generation (RAG), yet we observe that directly injecting retrieved candidates as context causes severe context poisoning. To address this, we propose HybridGeo, a dual-stream late-fusion framework that decouples retrieval from reasoning. A retrieval stream applies continuous alignment with spatial–semantic clustering to produce stable regional anchors; a reasoning stream performs context-free Chain-of-Thought inference to yield an independent coordinate estimate. The two streams are fused only at the decision stage via a spatial–consistency module that triggers weighted averaging under agreement or confidence-based arbitration under conflict. Experiments on Im2GPS3k show that HybridGeo achieves 73.89% Country@750km accuracy, outperforming the retrieval baseline by 7.27% and 8.23%, and surpassing both VLM-only and RAG baselines. These results demonstrate that late fusion effectively avoids context poisoning while enabling complementary benefits from both streams.

Keywords: visual geo-localization, large vision-language models, retrieval-augmented generation, dual-process theory, late fusion

Check for updates

Academic Editors: Maria Antonia Brovelli and Wolfgang Kainz

Received: 27 January 2026
Revised: 30 March 2026
Accepted: 6 April 2026
Published: 9 April 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Background

Visual geo-localization aims to predict the precise geographic coordinates of an image's capture location solely from its visual content, and it remains a fundamental problem at the intersection of computer vision and geographic information science (GIS) [1]. It has broad applicability in autonomous driving [2], image forensics [3], augmented reality [4], and content provenance [5]. Yet worldwide localization is intrinsically difficult: the Earth's surface spans roughly 510 million square kilometers, visually similar scenes recur across distant regions, and the same place may appear very different across time, season, weather, and viewpoint. Consequently, inferring a precise location from a single image is a high-ambiguity problem at global scale.

Existing approaches largely follow a discriminative route or a generative route. Discriminative methods, spanning retrieval-based search [6,7], large-scale classification [8], and continuous image-coordinate alignment [9,10], learn strong spatial priors from massive geo-tagged corpora and can provide stable geographic anchors at worldwide scale. However, because they are driven mainly by visual similarity or embedding alignment, they often underexploit high-level semantic evidence such as text, cultural context, and functional scene cues that human geo-experts routinely use.

Generative approaches based on large vision-language models (LVLMs) shift the focus from pure matching to explicit semantic reasoning [11–14]. They can read road signs, interpret architectural style, and combine visual evidence with world knowledge through Chain-of-Thought inference, while recent agentic systems such as GeoVista [15] further extend this capability through multi-turn tool use. Yet these methods remain vulnerable to hallucination, may overtrust plausible narratives unsupported by geographic priors, and often lack the precise distributional grounding that large-scale geo-spatial retrieval systems naturally provide.

This complementarity makes fusion a natural direction. A straightforward strategy is retrieval-augmented generation (RAG), where retrieved candidate locations are injected into the VLM as context. However, geo-localization is especially prone to context poisoning as follows: standardized chain stores, repeated urban forms, and visually similar natural scenery often cause retrieval systems to return plausible but geographically incorrect candidates. Once such candidates enter the reasoning context, they can act as anchors that bias the model toward an erroneous hypothesis and suppress its independent judgment. Moreover, when the VLM is implicitly constrained to reason within the retrieved set, its broader world knowledge becomes less useful precisely when the true location is missing from the candidate list.

Therefore, the central challenge is not simply how to combine retrieval and reasoning, but how to preserve their complementary strengths without allowing the two components to become mutual sources of noise. More concretely, the core problem we address is how to leverage the global coverage of retrieval systems while preserving the VLM's independent reasoning ability and preventing retrieval noise from contaminating semantic inference.

Figure 1 summarizes the three paradigms that motivate our design. Retrieval-only systems are vulnerable to visual confusion because they rely on nearest visual matches in a worldwide geo-database. Early-fusion RAG introduces retrieved candidates into the reasoning chain, but this can transform retrieval noise into a misleading prior and produce context-poisoned errors. HybridGeo instead adopts a late-fusion design as follows: the retrieval stream and reasoning stream operate independently, and their outputs are validated only at the decision stage through spatial consistency. This decoupling preserves independent reasoning while still allowing the system to exploit retrieval-provided spatial anchors.

To address this challenge, we propose HybridGeo, a visual geo-localization framework inspired by dual-process theory. The central idea is "decoupled execution, consensus fusion", where a retrieval stream (System 1) provides high-confidence regional anchors from large-scale geo-spatial data, a reasoning stream (System 2) performs context-free semantic deduction directly from the image, and a consistency fusion module verifies whether the two streams converge in geographic space before deciding how they should be fused or arbitrated. In this way, fusion becomes late verification rather than prompt-level conditioning.

This design addresses context poisoning from both an engineering and a mechanistic perspective as follows: it keeps retrieval evidence available, but it prevents it from dominating the reasoning process before independent inference is complete. Experiments on Im2GPS3k and OSV-5M show that HybridGeo consistently outperforms existing unimodal methods and conventional RAG baselines, indicating that late collaboration is a more

reliable way to combine distributional priors with semantic reasoning under worldwide visual ambiguity.

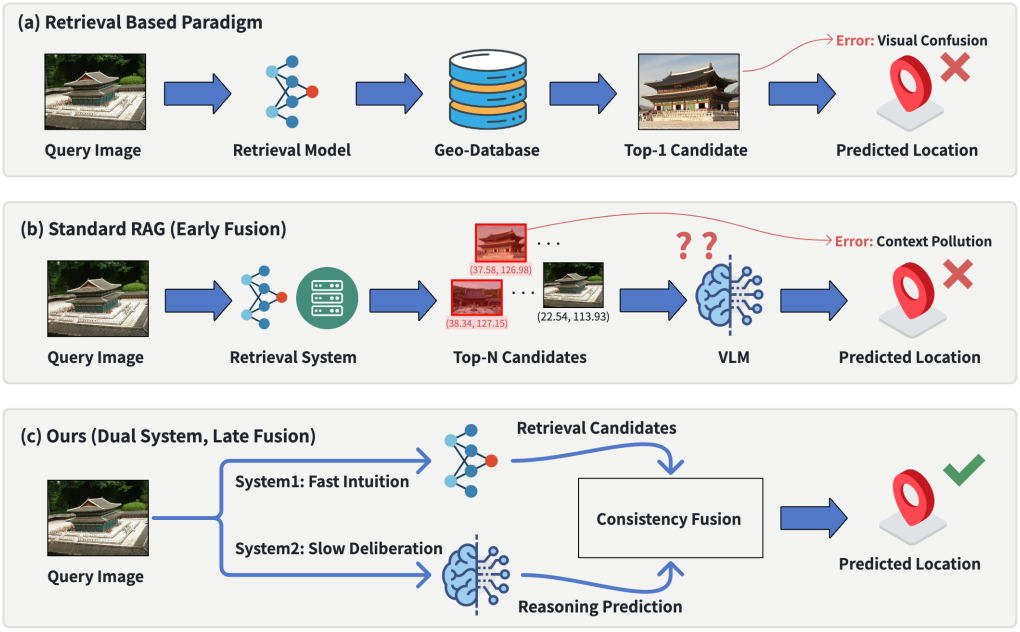


Figure 1. Comparison of Different Geo-localization Paradigms. (**a**) Retrieval paradigm: a query image is fed into a retrieval model to return the Top-1 candidate location from a geo-database, which is vulnerable to visual confusion caused by cross-region appearance similarity; (**b**) Standard RAG (early fusion): Top-N retrieved candidates are provided to the VLM as contextual text/coordinates, where erroneous candidates induce anchoring and lead to context poisoning; (**c**) HybridGeo (late fusion): the retrieval stream and reasoning stream infer independently in parallel and are only validated and fused at the decision stage via a spatial–consistency module, jointly suppressing retrieval noise and semantic hallucination.

Our main contributions are summarized as follows:

(1) We systematically construct and validate a "late-collaboration" geo-localization framework (HybridGeo) inspired by dual-process theory. Unlike traditional early-fusion RAG that directly injects retrieved candidates as VLM context, we reformulate visual geo-localization as an independent parallel inference and consensus fusion process between "fast intuition" and "slow reasoning." This design preserves the VLM's independent reasoning ability while effectively avoiding RAG-style context poisoning via spatial–consistency checks. Our experiments show that RAG Generation can even underperform the VLM-only baseline at the fine-grained City level (37.52% vs. 38.52%), quantitatively demonstrating the severity of context poisoning.

(2) We build a robust retrieval stream via spatial–semantic clustering. To address the high-noise nature of global retrieval, we design a filtering mechanism that combines continuous spatial alignment with HDBSCAN clustering, transforming discrete and unstable Top-k retrieval points into spatially representative regional anchors.

(3) We design a metacognitive arbitration mechanism based on spatial consistency. We propose an adaptive late-fusion strategy that dynamically arbitrates the final prediction by measuring the consensus between the retrieval and reasoning streams in geographic space, effectively filtering visual-ambiguity noise and VLM hallucinations.

(4) We provide comprehensive experimental validation and mechanistic analysis. Experiments on Im2GPS3k and OSV-5M show that HybridGeo significantly outperforms existing unimodal and RAG baselines (7.27%–8.23% absolute gains). Ablations re-

veal that dual-stream decoupling and late fusion are the key pathways for effective collaboration between explicit knowledge and implicit reasoning.

1.2. Related Work

We next situate HybridGeo within three closely related lines of research as follows: discriminative geo-localization, generative geo-localization, and retrieval-augmented fusion strategies.

1.2.1. Discriminative Geo-Localization

Discriminative methods aim to learn a direct mapping between image visual features and geographic location labels. Depending on how geographic space is modeled, this line of work can be broadly categorized into discrete retrieval/classification approaches and continuous space alignment approaches.

Discrete Retrieval and Classification

Early research on visual geo-localization predominantly formulated the task as image retrieval. NetVLAD [7] is a seminal work in this area; it introduced a differentiable VLAD layer (Vector of Locally Aggregated Descriptors) that aggregates local features extracted by CNNs into global descriptors robust to illumination and viewpoint changes through end-to-end training. To mitigate the high computational cost of triplet mining in large-scale retrieval training, CosPlace [16] reformulated place recognition training as a classification problem by partitioning the dataset into multiple geographic groups and performing within-group classification, learning highly discriminative feature embeddings. MixVPR [17] targeted the efficiency of feature aggregation and proposed an all-MLP architecture that uses a "feature mixing" mechanism to interact across channel and spatial dimensions, achieving strong retrieval performance with a small parameter budget.

Another class of discriminative methods follows a classification-based paradigm, discretizing the continuous Earth surface into a finite set of grid cells. Google's PlaNet [8] used the S2 Geometry library to partition the Earth into tens of thousands of cells with varying sizes, where cell density is proportional to photo density (e.g., denser cells in urban areas). It employed an Inception backbone to predict the cell class and introduced an LSTM module to leverage temporal information from photo albums to refine localization results. TransLocator [18] adopted a Transformer architecture and designed a dual-branch network for joint geo-location prediction and scene recognition; self-attention enables the model to focus on fine-grained discriminative regions (e.g., distinctive vegetation or architectural details), substantially improving localization accuracy.

Although these methods perform well in certain scenarios, discretized modeling has the following inherent drawbacks: it breaks the continuity of geographic space by forcing physically adjacent locations into different class labels, making it difficult for models to preserve and learn meaningful distance metrics between locations.

Continuous Space Alignment

To overcome the accuracy bottleneck introduced by discretization, recent work has shifted toward continuous space alignment. GeoCLIP [9] established this paradigm. Inspired by CLIP-style alignment, GeoCLIP learns to align image features with geographic coordinates within a shared continuous embedding space. Its key contribution is a location encoder based on Random Fourier Features (RFFs) that maps latitude and longitude (lat, lon) into high-dimensional sine/cosine embeddings. This encoding both preserves the geometry of spherical distance and captures fine-grained geographic differences via high-frequency components. Pretrained with contrastive learning on MP-16 (4.7M image–GPS pairs), GeoCLIP demonstrates strong zero-shot localization capability.

Building upon GeoCLIP, PIGEON [10] further improves spatial modeling. It critiques PlaNet's rigid grid partitioning and proposes semantic geocells, dynamically generating cells by combining administrative hierarchies (e.g., country–province–city) with data density. More importantly, PIGEON introduces a multi-task learning objective that requires the model to predict not only location but also climate zone, season, and compass direction. These auxiliary tasks encourage attention to deeper semantic cues in the environment (e.g., inferring season from vegetation phenology and orientation from shadow direction), enabling performance that surpasses top human players in the GeoGuessr game. PIGEON also adopts label smoothing based on the Haversine distance, dynamically weighting the loss according to physical distance between prediction and ground truth to better model spatial continuity. PIGEOTTO [10] subsequently extends this framework to more general image sources (e.g., Flickr and Wikipedia), demonstrating strong generalization beyond street-view imagery.

Compared with traditional retrieval/classification methods, these approaches substantially expand the operational range of geo-localization and remove the need for an explicit large-scale geo-visual database. However, they still rely heavily on visual features and can overlook world-knowledge cues such as text and human geography, which limits localization accuracy and general applicability.

1.2.2. Generative Geo-Localization

With the rapid development of large vision-language models (LVLMs), the focus of geo-localization research has gradually shifted from pure visual feature matching to higher-level semantic reasoning. This paradigm emphasizes mimicking human expert cognition: "observe cues → retrieve knowledge → logical reasoning → verify hypotheses."

Vision-Language Reasoning

Although general-purpose models such as GPT-4V [11] possess broad world knowledge, they are not designed for geo-localization and frequently exhibit hallucinations. To address this limitation, researchers have begun to "distill" human expert geo-reasoning capabilities into models.

GeoReasoner [13] is a representative work in this direction. It first points out that existing street-view datasets contain many low-quality images lacking salient geographic cues, and proposes a "locatability" metric to filter high-quality samples. More importantly, GeoReasoner constructs an instruction-tuning dataset with rich reasoning logic. The dataset is derived from commentary by top GeoGuessr players, converting expert intuition into structured Chain-of-Thought (CoT) trajectories (e.g., "yellow rear plate and left-hand traffic → infer the UK or Hong Kong → confirm the UK based on road-sign typography"). Supervised fine-tuning (SFT) on these data enables the model to learn human-like deductive reasoning patterns.

The Navig [14] framework further refines this process. It builds the NaviClues dataset by automatically analyzing expert commentary videos and extracting triplets of "visual cue → reasoning step → location." Navig shows that explicitly teaching models to recognize specific fine-grained cues (e.g., utility pole shapes, road marking colors, and soil tones across countries) can allow a general VLM to outperform specialized models trained with large-scale data, even under low-data training regimes.

After training or fine-tuning, these methods can surpass discriminative models; however, hallucinations still persist, and fine-tuning/training VLMs typically requires substantial computational resources.

Agentic Geo-location

A recent research trend is to upgrade VLMs into tool-using agents that solve complex localization tasks through multi-turn interaction with the environment.

GeoVista [15] proposes an agentic framework based on an "observe–act–think" loop. Unlike traditional models that passively consume an image, GeoVista can proactively invoke tools to acquire additional information; through multiple rounds of interaction using active observation and information-retrieval tools, it gradually narrows the search space and validates hypotheses. While such agent-based methods can achieve strong performance, they typically depend on complex tooling, network access, and repeated interactions, resulting in higher latency and resource cost.

1.2.3. Retrieval Augmentation and Fusion Strategies

To combine the knowledge breadth of discriminative models with the reasoning depth of generative models, retrieval-augmented generation (RAG) [19] is a natural choice. RAG was originally proposed in NLP to retrieve relevant documents from an external knowledge base as context, compensating for LLMs' limitations in long-tail knowledge recall and timeliness. This paradigm effectively mitigates hallucination and makes generated content more factual and traceable.

In geo-localization, RAG has been introduced to enhance a model's geographic awareness. A straightforward early-fusion strategy is to feed the retrieved Top-k candidate locations as context into a VLM and ask it to choose among them. However, both our experiments and recent studies show that this strategy often performs poorly in geo-localization, mainly due to context poisoning: worldwide, many scenes are visually similar but geographically far apart (e.g., standardized chain stores), and retrieval systems often return high-confidence but incorrect candidates. The anchoring effect induced by such "noise" can severely suppress the VLM's independent reasoning, causing it to abandon correct judgment and blindly follow retrieval outputs.

Different from the above, our proposed HybridGeo adopts a late-fusion strategy: the retrieval stream and reasoning stream run independently and are fused only at the decision stage through spatial–consistency checks. This design preserves the VLM's independent reasoning ability while leveraging retrieval-provided precise spatial anchors, effectively avoiding context poisoning.

2. Materials and Methods

This section details the design principles and implementation details of the HybridGeo framework.

2.1. Method Overview

Given an input image I, the goal of visual geo-localization is to predict the geographic coordinates of its capture location, $(\hat{lat}, \hat{lon})$. During evaluation, localization accuracy is measured by the Haversine distance between the predicted coordinates and the ground truth (lat_{gt}, lon_{gt}).

Different from conventional RAG, we are inspired by dual-process theory and propose a "divide-and-conquer, late collaboration" perspective. As illustrated in Figure 1, HybridGeo consists of three key components: a retrieval stream as the intuitive system, a reasoning stream as the deliberative system, and a consistency fusion module as an arbiter.

Figure 2 illustrates the full architecture of HybridGeo. The input image is fed into two streams in parallel. The upper retrieval stream uses the retrieval encoder to obtain Top-N candidates on a "global map" and produces representative regions via a clustering module. The lower reasoning stream uses an LVLM to analyze visual evidence, infers country and

city through CoT prompting, and parses coordinates P_{vlm}. The final consistency fusion module performs a distance check as follows: if the distance is $< \tau$, the streams are deemed in agreement and a weighted average is applied; otherwise, a conflict is declared and an arbitration strategy is used.

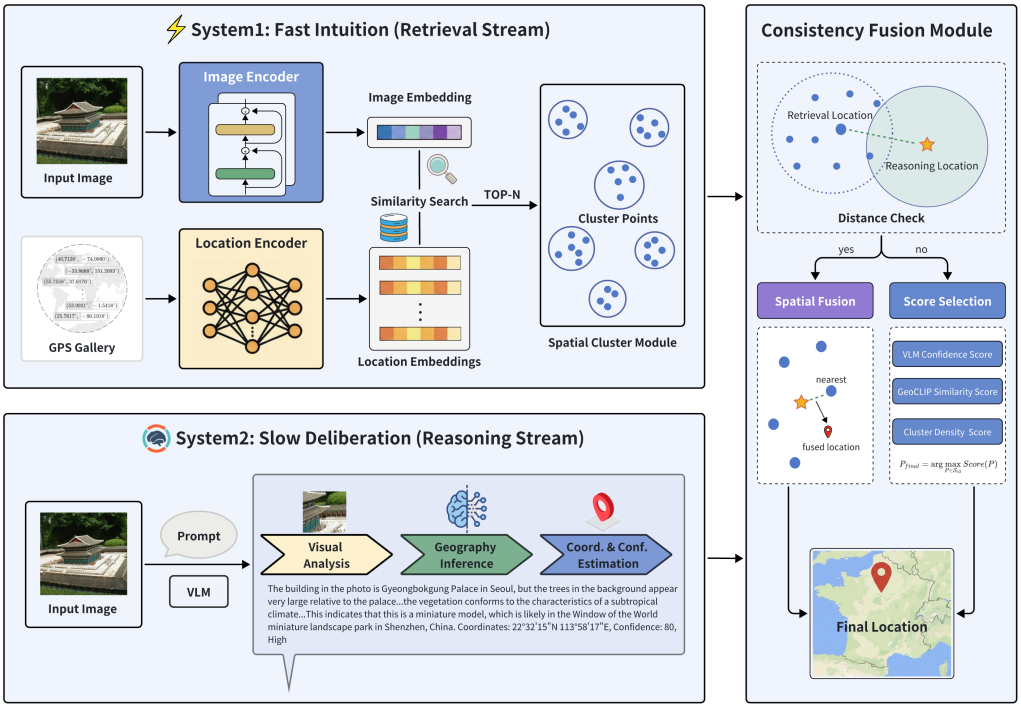


Figure 2. Overview of HybridGeo framework. Given an input image I, the retrieval stream (System 1) performs similarity search using the image encoder and location encoder, returns Top-N candidate locations, and produces K regional anchors $\{P_{ret}^{(k)}\}$ via spatial–semantic clustering. The reasoning stream (System 2) performs CoT reasoning under a no-retrieval-context setting, outputting a coordinate prediction P_{vlm} together with a self-assessed confidence. The consistency fusion module computes $D_{min} = \min_k \text{Haversine}(P_{vlm}, P_{ret}^{(k)})$. If $D_{min} < \tau$, it performs spatial weighted fusion of P_{vlm} with the nearest anchor; otherwise, it arbitrates by scoring candidates based on VLM confidence, retrieval similarity, and cluster density to obtain the final prediction P_{final}.

As shown in Figure 2, HybridGeo does not directly feed retrieved candidates into the VLM. Instead, it strictly decouples two evidence chains and runs them in parallel as follows: the retrieval stream outputs the spatial distribution of candidates and compresses discrete Top-N points into stable regional anchors via clustering, while the reasoning stream outputs an independent coordinate estimate based on semantic cues. The fusion module aligns the two sources using great-circle distance as a shared physical metric, enabling cross-validation between "retrieval noise" and "semantic hallucination," and thus producing more robust decisions under global-scale visual ambiguity.

From a probabilistic perspective, visual geo-localization can be formulated as finding the location L that maximizes the posterior probability $P(L|I)$.

- Early Fusion (Traditional RAG) essentially estimates the conditional probability $P(L|I, \mathcal{R})$, where $\mathcal{R}$ denotes the retrieved candidate set. This formulation implicitly assumes that retrieval contains the correct answer or useful evidence. However, for geo-localization, visual ambiguity often makes $P(\mathcal{R}|I)$ highly noisy. When noisy candidates are injected as priors into the VLM's reasoning process, they can cause posterior collapse, i.e., the probability mass is incorrectly concentrated around erroneous retrieved candidates—precisely the "context poisoning" phenomenon we observe.

- Late Fusion (HybridGeo) instead decomposes geo-localization into two independent estimation processes: $P(L|I_{vis})$ based on global visual features (retrieval stream) and $P(L|I_{sem})$ based on semantic logical reasoning (reasoning stream). In the fusion stage, we approximately compute the joint posterior as follows:

$$P(L|I) \propto P(L|I_{vis}) \cdot P(L|I_{sem})$$

This Product of Experts perspective takes into account that the errors of the two streams have a certain degree of conditional independence. Considering that the retrieval stream mainly relies on texture matching, while the reasoning stream mainly relies on semantic understanding, their error patterns are indeed significantly different. Therefore, Late Fusion can effectively suppress the unique noise of a single stream through the intersection of the two streams, thereby obtaining a more robust estimation than any single modality.

Table 1 summarizes the key differences between HybridGeo and conventional RAG strategies as follows:

Table 1. Comparison of HybridGeo with RAG Strategies.

Characteristic	RAG Selection	RAG Generation	HybridGeo (Ours)
VLM Input	Image + Candidate List	Image + Candidate List	Image Only
VLM Task	Select from Candidates	Generate Coordinates Referring to Candidates	Independent Reasoning to Generate Coordinates
Fusion Timing	Early Fusion	Early Fusion	Late Fusion
Noise Impact	High (Candidates May Mislead)	Medium	Low (Independent Reasoning)
VLM Capability	Limited by Candidate Range	Influenced by Candidate Distribution	Fully Exploited

2.2. Retrieval Stream

The retrieval stream serves as a "broad search" component in the framework. Its goal is to quickly filter plausible candidate regions from massive geographic space and provide high-confidence spatial anchors for subsequent fusion decisions. To mitigate the instability of single-point predictions, we design a two-stage pipeline consisting of global retrieval and spatial–semantic clustering.

2.2.1. Global Retrieval

We adopt GeoCLIP [9] as the retrieval backbone, which embeds images and geographic coordinates into a shared 512-dimensional feature space via contrastive learning as follows:

$$f_I = \text{ImageEncoder}(I) \in \mathbb{R}^{512}$$

$$f_G = \text{LocationEncoder}(lat, lon) \in \mathbb{R}^{512}$$

Given an input image I, the retrieval model returns the Top-N candidate locations that are most similar to the image embedding, together with confidence scores, as follows:

$$\{(lat_i, lon_i, s_i)\}_{i=1}^{N}, \quad s_i = \cos(f_I, f_{G_i})$$

In our experiments, we set $N = 200$.

2.2.2. Spatial–Semantic Clustering

Because the Top-N candidates may contain substantial redundancy and noise, using them directly is suboptimal. We apply a spatial–semantic clustering algorithm to compress the candidates into K representative locations.

First, we convert each candidate's geographic coordinates to 3D Cartesian coordinates to avoid longitude wrap-around issues as follows:

$$x = \cos(lat) \cdot \cos(lon)$$

$$y = \cos(lat) \cdot \sin(lon)$$

$$z = \sin(lat)$$

Next, we construct a hybrid feature vector that combines spatial location and visual semantics as follows:

$$\mathbf{h}_i = [\lambda \cdot \mathbf{p}_i; (1-\lambda) \cdot \mathbf{v}_i]$$

Here, $\mathbf{p}_i$ denotes the normalized spatial coordinate, $\mathbf{v}_i$ is the location embedding of candidate (lat_i, lon_i) produced by the LocationEncoder, and λ is the spatial–semantic balance coefficient (set to 0.6). HDBSCAN clusters $\mathbf{h}_i$ in the concatenated feature space using the L_2 norm as the distance metric.

We apply HDBSCAN [20] for density-based clustering. Compared with K-Means and DBSCAN, HDBSCAN offers three advantages well-suited to geo-localization as follows: it automatically determines the number of clusters without a preset K, robustly filters outliers as noise, and adapts to varying density distributions—properties that match the heterogeneous spatial spread of retrieval candidates across queries.

We set `min_cluster_size`=3 and `min_samples`=2. The clustering result is denoted as follows:

$$\mathcal{C} = \text{HDBSCAN}(\{\mathbf{h}_i\}_{i=1}^{N})$$

Here, $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$ denotes the K valid clusters identified by the algorithm, where K is determined automatically and typically ranges from 1 to 5.

Finally, for each cluster, we select as the representative the candidate that is closest to the cluster center and has the highest confidence as follows:

$$P_{ret}^{(k)} = \arg\max_{i \in \mathcal{C}_k} s_i \cdot \exp(-d_i)$$

Here, d_i is the spatial distance from the candidate to the cluster center. We compute the mean vector of the cluster center $\bar{\mathbf{p}}_k$ on unit-sphere Cartesian coordinates and normalize it, and we then measure proximity using the L_2 distance between $\mathbf{p}_i$ and $\bar{\mathbf{p}}_k$. We set the maximum number of clusters to $K = 5$.

2.3. Reasoning Stream

The reasoning stream aims to extract deep semantic information from the image and perform independent localization by mimicking the cognitive process of human experts. Unlike the retrieval stream, which relies on similarity matching of visual features, the reasoning stream focuses on logical deduction over fine-grained cues such as culture, climate, vegetation, and text. To fully elicit the geo-reasoning capability of large vision-language models (LVLMs), we adopt a Chain-of-Thought (CoT) strategy.

We design prompts that guide the LVLM to conduct structured geo-reasoning. The prompts instruct the model to analyze the image following the steps below:

The core idea of this prompt design is to explicitly decompose the complex "image-to-coordinate" inference into controllable sub-steps as follows: first enumerate verifiable visual cues (text, architecture, vegetation, traffic, and cultural elements), then progressively map these cues to geographic hypotheses (country/region/city), and finally output coordinates with a self-assessed confidence. Such stepwise constraints encourage evidence-centered

reasoning and reduce unsupported leaps, thereby lowering the likelihood of semantic hallucination in settings with sparse cues or conflicting evidence.

```
System Prompt:
You are a professional geographic location analyst, skilled at precise
geo-localization through visual cues in images.  Please analyze the image
and infer geographic coordinatesfollowing the procedure below:
1.  Visual Feature Extraction:  - Architectural style (European, Asian,
American, etc.)  - Vegetation type (tropical, temperate, desert, etc.)
- Textual information (road signs, store names, language) - Traffic
characteristics (vehicle types and driving direction) - Cultural elements
(attire, signage style)
2.  Geographic Inference:  - Narrow down the possible geographic range based
on visual features - Identify the likely country/region - Further localize to
city/area
3.  Coordinate Prediction:  - Output the geographic coordinatesyou are most
confident about - Format:  latitude, longitude (decimal degrees)
4.  Confidence Assessment:  - Please provide a confidence score from 0-100 and
a qualitative assessment of "high, medium, or low" - Explain why you assigned
this score
```

```
User Prompt:
Please analyze this image and infer the geographic coordinatesof its capture
location.  Requirements:  1.  Describe in detail the geographic cues you
observe 2.  Explain your reasoning process 3.  Provide the coordinates in
the last line, formatted as:  Final coordinates:  latitude, longitude
```

Moreover, the fixed output structure—especially the "last-line coordinate" requirement and the confidence score—improves system usability. On the one hand, it enables robust downstream coordinate parsing and reduces extraction failures caused by unconstrained free-form text; on the other hand, confidence serves as an explicit signal for the consistency fusion module to arbitrate and weight predictions, yielding more robust decisions when trading off between "correct but uncertain" and "confident but potentially wrong" reasoning.

During inference, we convert the LVLM free-text response into numerical coordinates by parsing the final line that follows the required template "Final coordinates: latitude, longitude." Once a valid decimal latitude–longitude pair is matched, it is directly converted into the coordinate prediction P_{vlm} for downstream fusion. If the expected pattern is missing or the coordinate pair cannot be parsed successfully, we treat the reasoning-stream output as unavailable and fall back to the retrieval prediction $P_{ret}^{(1)}$, consistent with Algorithm 1. This simple rule-based parser is intentionally kept explicit so that the transformation from free-text reasoning output to numeric geo-coordinates is reproducible and malformed responses do not propagate into the fusion stage.

2.4. Consistency Fusion Module

The consistency fusion module is the core innovation of HybridGeo and is responsible for dynamically fusing predictions from the two streams.

Figure 3 depicts a two-stage decision procedure where great-circle distance serves as a "metacognitive verification signal." In Case A, P_{vlm} falls within the spatial–consistency region, indicating that the reasoning and retrieval streams reach consensus at city scale; we then perform a weighted average between P_{vlm} and the nearest cluster center to reduce random error. In Case B, the two streams diverge substantially in geographic space; the

system switches to arbitration, computes a composite score over $\{P_{vlm}\} \cup \{P_{ret}^{(k)}\}$, and selects the highest-scoring candidate as P_{final} (as illustrated by the score list on the right). The composite score jointly considers the VLM self-assessed confidence, retrieval similarity, and cluster density, thereby more robustly suppressing retrieval noise and semantic hallucination in conflict scenarios.

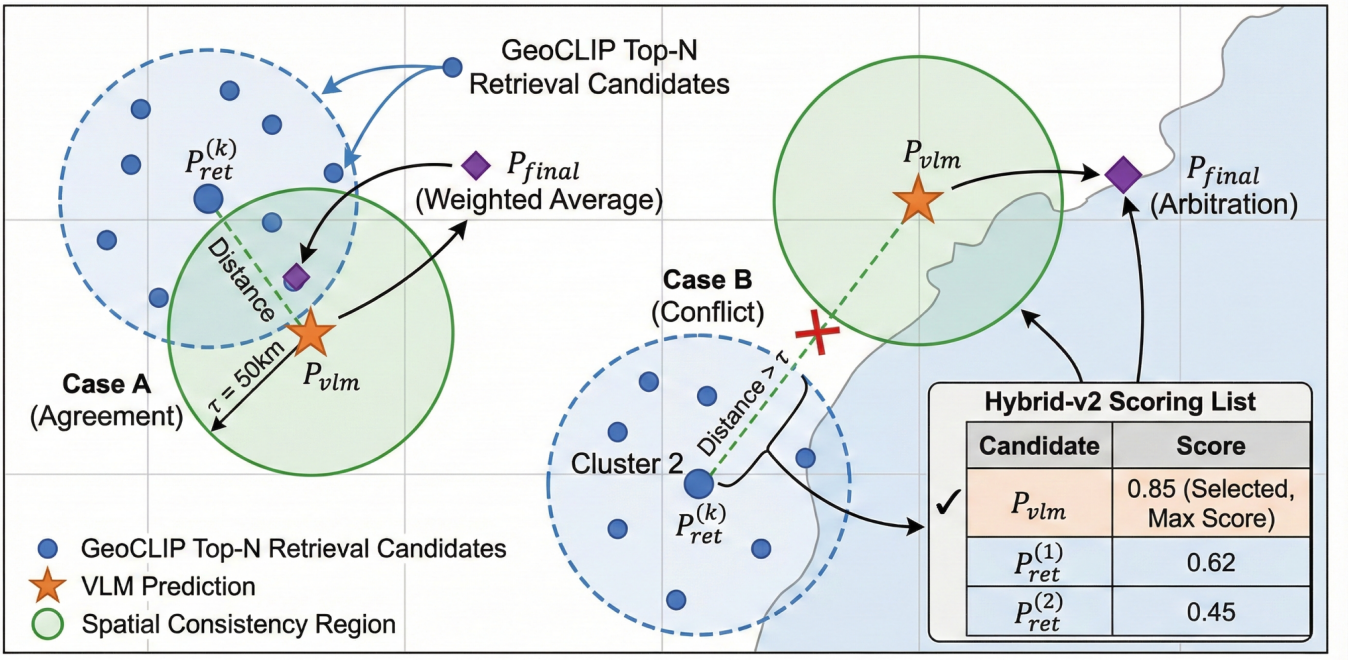


Figure 3. Consistency fusion strategy. Blue dots denote Top-N retrieved candidates; dashed circles indicate candidate clusters and their centers $P_{ret}^{(k)}$; the orange star denotes the VLM prediction P_{vlm}; the green circle is the spatial–consistency region with radius τ; and the purple diamond denotes the final prediction P_{final}. The left panel shows the Agreement case ($D_{min} < \tau$) where P_{vlm} is fused with the nearest anchor; the middle panel shows the Conflict case ($D_{min} \geq \tau$) where arbitration is triggered; and the right panel illustrates an example of HybridGeo-v2 composite scoring.

2.4.1. Consistency Check

First, we compute the minimum distance between the reasoning-stream prediction P_{vlm} and the cluster centers from the retrieval stream as follows:

$$D_{min} = \min_{k=1}^{K} \text{Haversine}(P_{vlm}, P_{ret}^{(k)})$$

Based on the relationship between the distance and the threshold τ, we determine whether the two streams are consistent as follows:

$$\text{Consistency} = \begin{cases} \text{Agreement} & \text{if } D_{min} < \tau \\ \text{Conflict} & \text{otherwise} \end{cases}$$

In our experiments, we set $\tau = 50$ km, corresponding to a city-scale spatial range.

The role of τ is to judge whether the two streams reach city-level spatial consensus, rather than to require street-level coincidence. A very small threshold such as 1 km is too strict for this purpose; even when retrieval and reasoning refer to the same city or landmark area, their predictions may still differ by several kilometers because one stream outputs a cluster representative while the other outputs a single coordinate estimate. In contrast, a much larger threshold such as 200 km is too loose, because geographically distinct cities or neighboring regional centers may then be incorrectly treated as agreement, allowing false

consensus to pass into the fusion stage. We therefore use 50 km as a city-scale compromise as it is broad enough to recognize same-city agreement under realistic prediction noise, yet still restrictive enough to separate genuinely different urban hypotheses.

2.4.2. Fusion Strategy

Agreement Case (Dual-Stream Consensus):

When the two stream predictions are spatially close, we apply weighted-average fusion as follows:

$$P_{final} = \omega \cdot P_{vlm} + (1 - \omega) \cdot P_{ret}^{nearest}$$

Here, $P_{ret}^{nearest}$ is the retrieval cluster center nearest to P_{vlm}, and ω is the VLM weight (set to 0.6).

In practice, we first convert coordinates to unit-sphere Cartesian coordinates, perform the weighted average there, and we then convert back to geographic coordinates to handle longitude wrap-around as follows:

$$\mathbf{u}(\text{lat}, \text{lon}) = (\cos \text{lat} \cos \text{lon},\ \cos \text{lat} \sin \text{lon},\ \sin \text{lat})$$

Here, $\mathbf{u}(\text{lat}, \text{lon})$ denotes the mapping from geographic coordinates to unit-sphere Cartesian coordinates. We define $\mathbf{u}_{vlm} = \mathbf{u}(P_{vlm})$ as the transformed VLM prediction and $\mathbf{u}_{ret} = \mathbf{u}(P_{ret}^{nearest})$ as the transformed nearest retrieval anchor.

$$\mathbf{u}_{final} = \frac{\omega \cdot \mathbf{u}_{vlm} + (1 - \omega) \cdot \mathbf{u}_{ret}}{\|\omega \cdot \mathbf{u}_{vlm} + (1 - \omega) \cdot \mathbf{u}_{ret}\|_2}$$

$$lat_{final} = \arcsin(u_z), \quad lon_{final} = \text{atan2}(u_y, u_x)$$

In the inverse transform, (u_x, u_y, u_z) are the three components of the normalized vector $\mathbf{u}_{final}$.

Conflict Case (Dual-Stream Divergence):

When the two streams diverge substantially, an arbitration decision is required. We design two conflict-resolution strategies (HybridGeo-v1 and HybridGeo-v2) as follows:

(a) HybridGeo-v1: Binary Selection

HybridGeo-v1 adopts a simplified strategy that chooses only between the VLM-predicted coordinate and the Top-1 retrieved result.

$$\mathcal{S}_{v1} = \{P_{vlm}, P_{ret}^{(1)}\}$$

The arbitration rule is defined as follows:

$$P_{final} = \begin{cases} P_{vlm} & \text{if } c_{vlm} = \text{“high”} \\ P_{ret}^{(1)} & \text{otherwise} \end{cases}$$

where c_{vlm} denotes the VLM self-assessed confidence level.

(b) HybridGeo-v2: Cluster-aware Selection

HybridGeo-v2 adopts a more robust strategy, using the VLM prediction together with all K cluster centers as the candidate set.

$$\mathcal{S}_{v2} = \{P_{vlm}\} \cup \{P_{ret}^{(k)}\}_{k=1}^{K}$$

Arbitration is based on a composite confidence score $Score(P)$. We define a normalized scoring function that jointly considers VLM reasoning confidence, retrieval visual similarity, and spatial clustering consistency as follows:

$$Score(P) = w_{vlm} \cdot \hat{S}_{vlm}(P) + w_{ret} \cdot \hat{S}_{ret}(P) + w_{clus} \cdot \hat{S}_{clus}(P)$$

$$P_{final} = \arg\max_{P \in \mathcal{S}_{v2}} Score(P)$$

where the following hold:

- $\hat{S}_{vlm}(P)$ is the VLM self-assessed confidence (normalized to $[0, 1]$). For candidates that only appear in the retrieval stream, this term is 0.
- $\hat{S}_{ret}(P)$ is the retrieval cosine similarity score (min–max normalized).
- $\hat{S}_{clus}(P)$ is the (normalized) cluster density score of the candidate, reflecting the strength of local spatial consensus.
- $w_{vlm}, w_{ret}, w_{clus}$ are weighting coefficients; via validation search, we set $w_{vlm} = 0.4, w_{ret} = 0.4, w_{clus} = 0.2$.

Normalization details: $\hat{S}_{vlm}$ linearly rescales the model-reported 0–100 confidence to $[0, 1]$; $\hat{S}_{ret}$ and $\hat{S}_{clus}$ are rescaled using global min–max ranges computed on the validation set to ensure comparability across samples.

Confidence calibration: to alleviate over-confidence in VLM self-assessment, we apply temperature scaling to $\hat{S}_{vlm}$ on the validation set. Specifically, let z be the pre-normalization confidence logit (obtained via a linear mapping from the reported confidence); the calibrated confidence is $\sigma(z/T_{vlm})$, where T_{vlm} is obtained by minimizing the negative log-likelihood on the validation set. The calibrated $\hat{S}_{vlm}$ is used for conflict arbitration scoring to improve confidence comparability across samples.

The advantage of HybridGeo-v2 is that it provides more "recovery" opportunities for weaker VLMs—even if the VLM prediction is wrong, the weighted composite score may still select the correct answer from the K clustered candidates.

2.4.3. Inference Algorithm

Algorithm 1 summarizes the complete HybridGeo inference procedure. Given an input image, the algorithm first retrieves Top-N candidates and clusters them into K regional anchors (lines 1–7), then independently obtains a VLM prediction (lines 8–11). Based on the spatial distance between the VLM output and the nearest anchor, the algorithm either performs weighted fusion under agreement (lines 14–17) or triggers confidence-based arbitration under conflict (lines 18–22).

The evaluation thresholds used throughout this paper are summarized in Table 2.

Table 2. Evaluation threshold definitions.

Level	Threshold (km)	Description
Street	1	Street-level
City	25	City-level
Region	200	Region-level
Country	750	Country-level
Continent	2500	Continent-level

Algorithm 1 HybridGeo Inference

Require: Image I; retrieval model; VLM model; hyperparameters $N, K, \lambda, \tau, \omega, w_{vlm}, w_{ret}, w_{clus}$
Ensure: Final predicted coordinate P_{final}

Compute image embedding $f_I \leftarrow \text{ImageEncoder}(I)$
Retrieve Top-N candidates $\{(lat_i, lon_i, s_i)\}_{i=1}^{N}$ via retrieval model
Construct hybrid features $\mathbf{h}_i \leftarrow [\lambda \cdot \mathbf{p}_i; (1-\lambda) \cdot \mathbf{v}_i]$ for $i = 1, \ldots, N$
Apply HDBSCAN clustering: $\{\mathcal{C}_k\}_{k=1}^{K} \leftarrow \text{HDBSCAN}(\{\mathbf{h}_i\}_{i=1}^{N})$
for each cluster $\mathcal{C}_k$, $k = 1, \ldots, K$ **do**
    Select representative: $P_{ret}^{(k)} \leftarrow \arg\max_{i \in \mathcal{C}_k} s_i \cdot \exp(-d_i)$
end for
Call VLM to obtain P_{vlm} and confidence score c_{vlm}
if VLM call fails **then**
    return $P_{ret}^{(1)}$
end if
Compute minimum distance: $D_{min} \leftarrow \min_{k=1}^{K} \text{Haversine}(P_{vlm}, P_{ret}^{(k)})$
if $D_{min} < \tau$ **then** ▷ Agreement case: dual-stream consensus
    Find nearest anchor: $P_{ret}^{nearest} \leftarrow \arg\min_k \text{Haversine}(P_{vlm}, P_{ret}^{(k)})$
    Perform weighted fusion on unit sphere:
        $\mathbf{u}_{final} \leftarrow \frac{\omega \cdot \mathbf{u}(P_{vlm}) + (1-\omega) \cdot \mathbf{u}(P_{ret}^{nearest})}{\|\omega \cdot \mathbf{u}(P_{vlm}) + (1-\omega) \cdot \mathbf{u}(P_{ret}^{nearest})\|_2}$
    Convert back to GPS: $P_{final} \leftarrow \mathbf{u}^{-1}(\mathbf{u}_{final})$
else ▷ Conflict case: arbitration via composite scoring
    Construct candidate set: $\mathcal{S} \leftarrow \{P_{vlm}\} \cup \{P_{ret}^{(k)}\}_{k=1}^{K}$
    for each candidate $P \in \mathcal{S}$ **do**
        Compute composite score: $Score(P) \leftarrow w_{vlm} \cdot \hat{S}_{vlm}(P) + w_{ret} \cdot \hat{S}_{ret}(P) + w_{clus} \cdot \hat{S}_{clus}(P)$
    end for
    Select best candidate: $P_{final} \leftarrow \arg\max_{P \in \mathcal{S}} Score(P)$
end if
return P_{final}

2.5. Experimental Setup

2.5.1. Datasets

To comprehensively validate the effectiveness of the HybridGeo framework, we evaluate on two widely used visual geo-localization benchmarks. The first is Im2GPS3k [6], which contains 3000 Flickr test images covering worldwide locations and diverse scenes (natural landscapes, urban street views, architectural landmarks, etc.), each annotated with precise GPS coordinates. The second is OSV-5M [21], a large-scale street-view dataset consisting of OpenStreetView imagery. We evaluate on its test subset; because it mainly contains outdoor images captured by vehicles and often lacks salient landmark cues, it is relatively more challenging.

2.5.2. Evaluation Metrics

We use distance error and multi-level accuracy as evaluation metrics. Distance error is measured by the Haversine distance between predicted and ground-truth coordinates. Accuracy is defined as the proportion of samples whose prediction error is below a given distance threshold. Specifically, we use five thresholds at different granularities as follows: street-level (Street@1km), city-level (City@25km), region-level (Region@200km), country-level (Country@750km), and continent-level (Continent@2500km). This multi-granularity protocol provides a comprehensive assessment across spatial scales.

2.5.3. Implementation Details

For the retrieval stream, we use GeoCLIP (ViT-L/14) [9] as the retrieval backbone and retrieve the Top-200 candidate locations for each query image. We then apply HDBSCAN density clustering for spatial–semantic clustering, set the spatial–semantic balance coefficient to $\lambda = 0.6$, and limit the maximum number of representative centers to five. For the reasoning stream, we use Qwen3-VL-235B-A22B as the primary LVLM. In the consistency fusion module, we set the best spatial–consistency threshold to $\tau = 50$ km; when the two streams agree, we set the VLM weight to $\omega = 0.6$; under conflict, we use $(w_{vlm}, w_{ret}, w_{clus}) = (0.4, 0.4, 0.2)$ for candidate arbitration. To improve confidence comparability, we calibrate the VLM self-assessed confidence on the validation set via temperature scaling (T_{vlm}), and we use the calibrated confidence for fusion scoring. Across all experiments, retrieval inference is run on an NVIDIA RTX 5090 GPU, while VLM inference is run on a multi-GPU NVIDIA A100 server.

From a practical perspective, the extra overhead introduced by the dual-stream design is small once a large VLM is already in the loop. In our implementation, the retrieval stream requires only about 0.2 seconds per sample on an NVIDIA RTX 5090 GPU, whereas the Qwen3-VL-235B reasoning stream typically requires about 10–30 seconds per sample on a multi-GPU A100 server, depending on scene difficulty and reasoning length. Therefore, compared with VLM-Only and RAG-style baselines that use the same reasoning backbone, HybridGeo remains in essentially the same latency and resource regime; the end-to-end cost is dominated by the VLM rather than by retrieval. The retrieval-only baseline is substantially cheaper, but it also gives up the semantic reasoning ability that drives the main accuracy gains. Importantly, HybridGeo still operates as a single-round pipeline without external tool calls or multi-turn interaction, which keeps its practical cost much closer to standard VLM inference than to agentic geo-location systems.

2.5.4. Baseline Methods

We use GeoCLIP [9] as the retrieval backbone throughout this work. Table 3 summarizes the baseline methods we compare against.

Table 3. Baseline Methods.

Method	Description
Top-1	Highest-confidence retrieved candidate
Top-K Oracle	Best of Top-K candidates (theoretical upper bound)
VLM-Only	VLM prediction without retrieval context
RAG Selection	VLM selects from retrieved candidates
RAG Generation	VLM generates coordinates referencing candidates
HybridGeo	Our dual-stream late-fusion method

To ensure fair comparisons, all methods share the same Top-200 retrieval results and HDBSCAN clustering parameters (min_cluster_size=3, min_samples=2). All VLM-based methods use identical prompts and output parsers, differing only in how retrieval context is injected. Key hyperparameters (RAG candidate count $K \in \{3, 5, 7\}$, VLM temperature $T \in \{0.0, 0.3, 0.7\}$, etc.) are grid-searched on the validation set; final choices are $K = 5$ and $T = 0.3$. HybridGeo hyperparameters (λ, τ, ω, fusion weights) are tuned on the same validation set.

3. Results

3.1. Main Experimental Results

This section is organized as follows: we first report overall comparisons with SOTA methods on Im2GPS3k (Section 3.1.1), then verify cross-domain generalization on the more challenging OSV-5M dataset (Section 3.1.2), and finally analyze the decision distribution (Section 3.1.3). We then present key ablations and strategy comparisons in Section 3.2, followed by qualitative examples and failure analyses in Section 3.3.

3.1.1. Results on Im2GPS3k Dataset

To comprehensively evaluate HybridGeo, we first compare it with existing SOTA methods, including both discriminative and generative approaches. Table 4 reports results on the Im2GPS3k dataset.

Table 4. Comparison with SOTA methods on Im2GPS3k. *Note*: "—" indicates that the corresponding metric was not reported in the original paper or the method was not evaluated on the Im2GPS3k dataset. Boldface indicates the best result in each metric column.

Method	Type	Street@1	City@25	Region@200	Country@750	Continent@2500
Discriminative Methods						
PlaNet [8]	Classification	8.5%	24.8%	34.3%	48.4%	64.6%
ISNs [22]	Classification	10.5%	28.0%	36.6%	49.7%	66.0%
CPlaNet [23]	Classification	10.2%	26.5%	34.6%	48.6%	64.6%
TransLocator [18]	Classification	11.8%	31.1%	46.7%	58.9%	80.1%
GeoCLIP (Original Paper) [9]	Alignment	14.11%	34.47%	50.65%	69.67%	83.82%
GeoCLIP (Our Reproduction)	Alignment	13.12%	32.24%	48.16%	66.62%	82.34%
PIGEON [10]	Alignment	11.3%	36.7%	53.8%	72.4%	85.3%
Generative Methods						
GeoReasoner [13]	VLM+CoT	8.77%	33.99%	—	67.43%	82.02%
NAVIG [14]	VLM+CoT	5.5%	28.9%	49.1%	68.3%	84.0%
HybridGeo (Ours)	**Hybrid**	**16.72%**	**40.47%**	**56.54%**	**73.89%**	**86.51%**

Table 4 provides an overall comparison between HybridGeo and representative SOTA methods across multiple spatial scales. HybridGeo achieves the best performance on all threshold metrics, with particularly pronounced advantages on City@25km and Region@200km, which are more practically meaningful. This suggests that "independent dual-stream inference + late-stage consistency fusion" not only improves coarse-grained recall, but also significantly reduces fine-grained bias, reflecting strong complementarity between the spatial anchors provided by the retrieval stream and the semantic constraints provided by the reasoning stream.

Comparing discriminative and generative baselines further reveals the following: classification-based paradigms (e.g., PlaNet variants) are largely limited by discretization errors at global scale; continuous alignment methods (GeoCLIP, PIGEON) generalize better worldwide but remain vulnerable to visual ambiguity; VLM-based reasoning methods (GeoReasoner, NAVIG) can be powerful when semantic cues are rich, yet their stability and coverage are still constrained. HybridGeo's consistent lead in Table 4 indicates that, for worldwide visual geo-localization with substantial ambiguity, combining "interpretable semantic inference" with "measurable spatial consistency" via late fusion provides a more robust system-level solution.

Table 5 further reports ablation comparisons of HybridGeo components, including performance under different fusion strategies. The results shown here are based on the Qwen3-VL-235B model.

Table 5. Ablation results on Im2GPS3k. Boldface highlights the main proposed variants and their corresponding key values.

Method	Mean (km)	Median (km)	Street	City	Region	Country	Continent
Top-1	1763.96	240.70	13.12%	32.24%	48.16%	66.62%	82.34%
VLM-Only (Qwen-235B)	1835.55	143.09	14.02%	38.52%	53.08%	69.77%	81.23%
RAG (Selection)	1526.50	170.20	11.65%	36.18%	51.74%	70.99%	84.81%
RAG Generation	1685.32	158.47	12.85%	37.52%	52.15%	70.12%	83.45%
HybridGeo-v1 (Ours)	1547.73	128.61	14.43%	39.19%	54.43%	71.44%	84.73%
HybridGeo-v2 (Ours)	**1352.29**	**107.48**	**16.72%**	**40.47%**	**56.54%**	**73.89%**	**86.51%**
Top-*K* Discrete Oracle	928.08	52.10	18.83%	45.07%	62.89%	79.66%	90.30%

Note: Top-K Discrete Oracle selects, from the Top-5 candidates, the discrete candidate with the smallest distance to the ground truth. Since HybridGeo performs continuous coordinate fusion in the Agreement case, it may obtain a coordinate closer to the ground truth via interpolation and can therefore surpass this oracle in principle.

In addition, to more finely examine the distribution of localization errors, we plot cumulative distribution function (CDF) curves.

Figure 4 shows the CDFs of prediction errors for different methods. HybridGeo-v2 (red curve) stays above unimodal baselines across the entire error range, indicating that it not only increases the fraction of low-error samples but also substantially suppresses long-tail large-error cases.

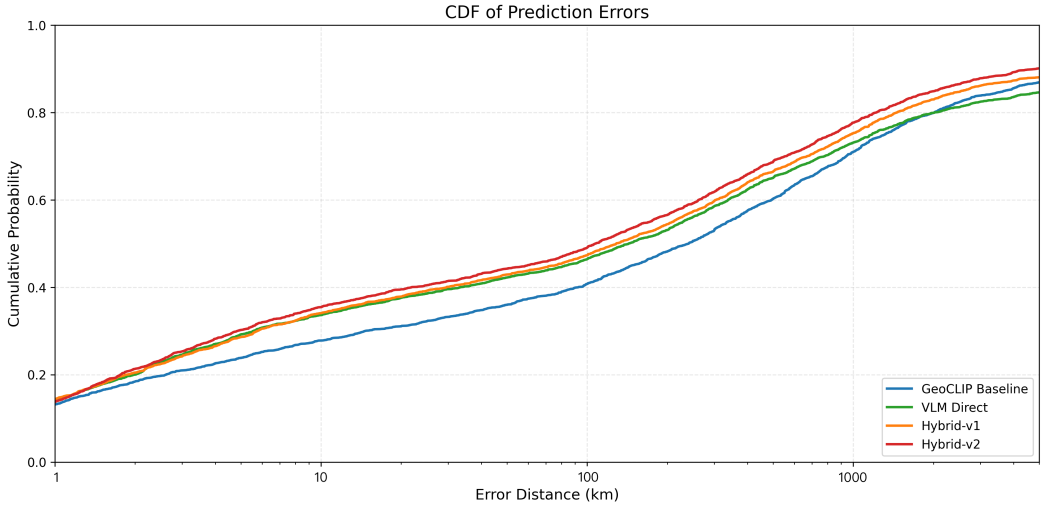


Figure 4. Cumulative Distribution Function (CDF) of prediction errors. HybridGeo demonstrates optimal performance across all error ranges.

From the results in Table 5 and Figure 4, we draw the following:

First, HybridGeo substantially outperforms unimodal baselines. Compared to the retrieval Top-1, which relies only on visual matching, HybridGeo-v2 improves City by 8.23% (32.24% $\rightarrow$ 40.47%); compared to VLM-Only, which relies only on semantic reasoning, it gains 1.95%. This strongly validates the effectiveness of dual-stream fusion: it successfully combines the spatial alignment capability of the retrieval stream with the semantic understanding of the reasoning stream.

Second, the framework approaches the theoretical performance ceiling. We treat the Top-5 Oracle (assuming the optimal choice can always be selected from Top-5 candidates) as the ceiling under the current retrieval capability. The gap between HybridGeo-v2's City accuracy (40.47%) and the oracle (45.07%) is reduced to 4.6 percentage points. This suggests

that HybridGeo's fusion mechanism can efficiently identify the correct target from the candidate set, effectively unlocking the potential of existing models.

Finally, the cluster-aware strategy (v2) is more robust. Comparing the two fusion variants, HybridGeo-v2 outperforms HybridGeo-v1 across metrics, with a particularly notable 2.45% gain at Country@750km. This indicates that, compared with a simple binary arbitration, introducing spatial clustering information as a reference frame provides richer decision cues and enables more accurate arbitration under stream conflict.

3.1.2. Results on OSV-5M Dataset

Table 6 reports results on OSV-5M (a 3000-sample subset).

Table 6. Results on OSV-5M. Boldface highlights the main proposed variant (HybridGeo-v2) and its corresponding values.

Method	Mean (km)	Median (km)	City	Region	Country	Continent
Retrieval Baseline	2361.52	1116.48	1.00%	14.00%	41.00%	76.00%
VLM-Only	3881.16	1310.89	1.45%	13.04%	34.78%	68.12%
HybridGeo-v1	3194.53	1265.97	1.08%	12.90%	37.63%	73.12%
HybridGeo-v2	**2551.10**	**925.56**	**3.23%**	**15.05%**	**43.01%**	**75.27%**
Top-*K* Oracle	1700.61	662.76	2.15%	20.43%	52.69%	84.95%

Because OSV-5M consists of a large number of randomly collected street-view images, it is substantially more challenging than Im2GPS3k, and all methods exhibit a significant drop in accuracy. From Table 6, VLM-Only has a median error as high as 1310.89 km and performs poorly at the country level (34.78%), largely because street views often lack clear cultural landmarks or language cues, causing purely semantic reasoning to degenerate into "blind guessing" under information scarcity. Nevertheless, even under such stringent conditions, the proposed HybridGeo-v2 still exhibits strong synergy as follows: its median error (925.56 km) is reduced by about 191 km compared to the retrieval baseline, and its city-level accuracy more than doubles (1.00% $\rightarrow$ 3.23%). This again supports the generality of the late-fusion design; even for "hard samples" where both streams are severely limited, spatial–consistency checks can act as a reliable metacognitive gate that preserves a baseline level of accuracy by filtering mutually incompatible unimodal errors.

It is noteworthy that HybridGeo-v2 achieves city-level performance (3.23%) that even surpasses the Top-K Discrete Oracle (2.15%). This is not contradictory but rather an inherent advantage of continuous space fusion as follows: the Oracle can only select one from five discrete candidates, whereas HybridGeo performs a weighted average between the VLM prediction and the retrieval cluster center in Agreement scenarios, thereby obtaining continuous coordinates closer to the ground truth than any single discrete candidate through coordinate interpolation. This phenomenon is particularly pronounced when the ground truth lies between two candidates.

3.1.3. Decision Distribution Analysis

In this subsection, we report statistics on the Im2GPS3k test set to understand HybridGeo's decision tendencies under agreement and conflict cases.

Table 7 shows the decision distribution of two HybridGeo variants using Qwen3-VL-235B:

By analyzing the decision distribution in Table 7, it is found that the spatial consistency agreement rate of the HybridGeo-v2 version is significantly higher than that of the v1 version (48.3% vs 32.6%). This is attributed to the high-dimensional clustering mechanism introduced in v2, which expands the single Top-1 retrieval point into five more representative spatial centers, thereby greatly increasing the probability that the two streams reach a

consensus in geospace. When dealing with conflict samples, the v1 version shows a strong preference for VLM inference results (55.4%), which reflects that, in a simple two-choice configuration, the semantic reasoning of VLM often exhibits higher individual confidence than a single retrieval point. In contrast, by introducing more abundant geographical candidates, the v2 version increases the winning proportion of the retrieval stream in conflict arbitration from 12.0% to 30.6%, which greatly balances the decision weights of the two systems. This more balanced arbitration mechanism provides more possibilities for correcting the hallucinations or logical errors of VLM, thus achieving more robust positioning performance in complex scenarios.

Table 7. Decision distribution of HybridGeo.

Decision Type	v1 Ratio	v2 Ratio	Description
Agreement	32.6%	48.3%	VLM and Retrieval Agree
Conflict→VLM	55.4%	21.1%	Select VLM after Conflict
Conflict→Retrieval	12.0%	30.6%	Select Retrieval after Conflict

In addition, it can be observed from the decision distribution that the agreement rate between VLM and retrieval results in both v1 and v2 versions is lower than 50%, which also confirms the discussion in Section 2 is about the independence of the retrieval stream and the inference stream, that is, the error patterns of the two methods are significantly different. Modeling them as two conditionally independent processes can suppress the impact of noise more effectively than the RAG mode.

As can be seen from Table 8, the consistency fusion module can lock in the agreement part with a high success rate by determining the spatial consensus between the following two different modes: the success rates of v1 and v2 in Country are 96.6% and 92.6%, respectively. This once again confirms that the retrieval stream and the reasoning stream do have a high degree of independence. Therefore, the success rate of their spatial consensus is much higher than their average success rate. Otherwise, if the error patterns of the two were very similar, the success rate after their spatial consensus would not differ much from the average success rate.

Table 8. Decision success rates of HybridGeo.

Decision Type	HybridGeo-v1			HybridGeo-v2		
	Street	City	Country	Street	City	Country
Agreement	33.50%	82.80%	96.60%	24.00%	68.40%	92.60%
Conflict→VLM	5.30%	19.80%	62.60%	6.80%	21.60%	60.90%
Conflict→Retrieval	4.20%	8.90%	43.00%	2.70%	10.70%	54.10%

3.2. Ablation Studies

3.2.1. Impact of Consistency Threshold τ

Figure 5 compares the impact of different consistency thresholds on Country@750 accuracy across Im2GPS3k and OSV-5M.

Figure 5 shows a consistent inverted-U pattern on both datasets. On Im2GPS3k, Country@750 peaks at 73.89% when $\tau = 50$ km and then declines as the threshold becomes either too strict or too permissive. On OSV-5M, the curve exhibits the same trend, as follows: performance improves from the small-threshold regime, reaches its best value around 50 km, remains close at 100 km, and then degrades as τ becomes larger. These results indicate that $\tau = 50$ km is not a fragile single-point choice, but a stable city-scale operating point across both datasets. Intuitively, a very small threshold makes agreement too strict and prevents the two streams from benefiting from consensus fusion, whereas

an overly large threshold introduces false agreement between geographically distinct hypotheses and passes additional noise into the fusion stage.

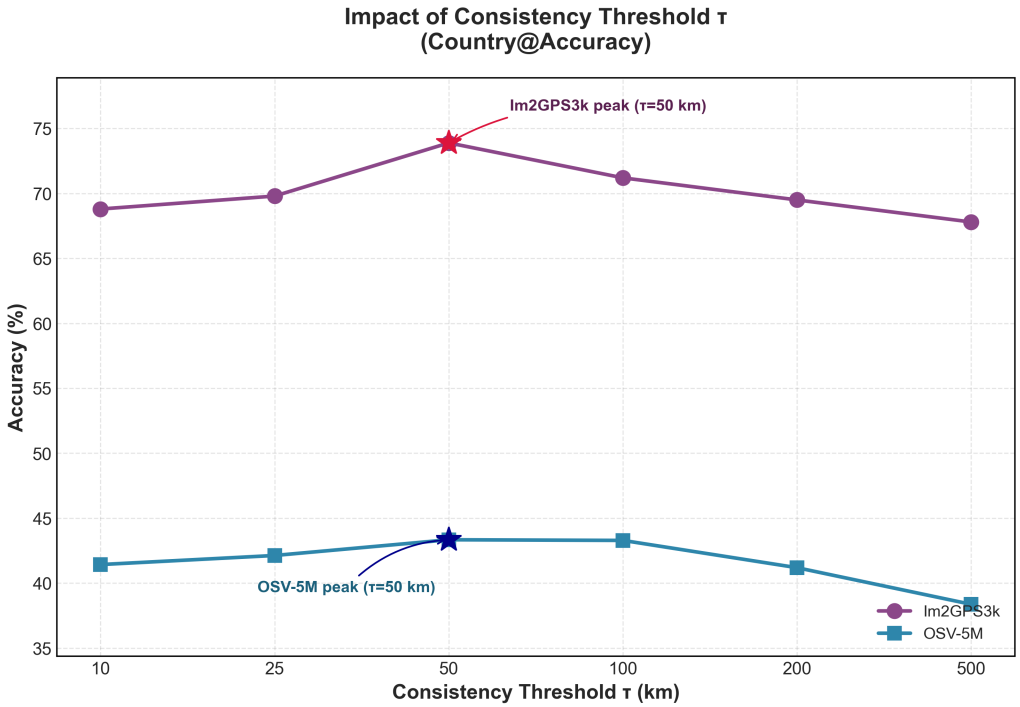


Figure 5. Ablation study on consistency threshold τ. Country@750 accuracy under different consistency thresholds on Im2GPS3k and OSV-5M. The purple curve denotes Im2GPS3k and the blue curve denotes OSV-5M. Both datasets achieve their best operating point at $\tau = 50$ km, while the surrounding range remains comparatively stable.

3.2.2. Sensitivity of λ, ω, and Conflict-Arbitration Weights

Beyond the consistency threshold τ, we further examine the sensitivity of several key weighting-related hyperparameters in HybridGeo-v2. Specifically, we vary the spatial–semantic balance coefficient λ in the clustering stage, the agreement-fusion weight ω, and several representative local perturbations around the default conflict-arbitration weights, while keeping the remaining settings fixed. Since these parameters mainly affect the fusion and arbitration behavior rather than the task definition itself, we report Country@750 on Im2GPS3k as a compact indicator in Table 9.

Table 9 shows that all three parameter groups exhibit stable local behavior around the default configuration. For λ, the best result is achieved at 0.6, while the nearby range 0.5–0.7 remains competitive, suggesting that the clustering stage benefits from a balanced spatial–semantic trade-off rather than either extreme. For ω, Country@750 reaches 73.89% at 0.60 and remains unchanged at 0.65, indicating that the agreement-fusion weight is not tuned to a brittle single point. For the conflict-arbitration weights, the default setting $(0.40, 0.40, 0.20)$ performs best among the tested local variants; modest shifts toward either the VLM or retrieval term remain competitive, whereas increasing the cluster-density weight causes a clearer drop. Taken together, these results support the robustness of the default parameter choices used in HybridGeo-v2.

Table 9. Consolidated sensitivity evidence for λ, ω, and the conflict-arbitration weights. Boldface marks the selected default setting(s) used in HybridGeo-v2.

Parameter	Setting	Country@750
λ	0.40	69.32%
λ	0.50	71.74%
λ	**0.60**	**73.89%**
λ	0.70	72.79%
λ	0.80	71.11%
ω	0.50	71.76%
ω	0.55	73.01%
ω	**0.60**	**73.89%**
ω	0.65	73.89%
ω	0.70	72.11%
$(w_{vlm}, w_{ret}, w_{clus})$	**(0.40, 0.40, 0.20)**	**73.89%**
$(w_{vlm}, w_{ret}, w_{clus})$	(0.50, 0.30, 0.20)	72.62%
$(w_{vlm}, w_{ret}, w_{clus})$	(0.30, 0.50, 0.20)	72.48%
$(w_{vlm}, w_{ret}, w_{clus})$	(0.35, 0.35, 0.30)	71.71%

3.2.3. Impact of VLM Model Scale

We conduct comparative experiments with VLMs of different scales and capabilities (Qwen3-VL-8B, Qwen3-VL-32B, Qwen3-VL-235B) to study how VLM capacity affects the performance gains brought by HybridGeo. Table 10 reports results under the HybridGeo-v2 strategy.

Table 10. Impact of VLM model scale on HybridGeo.

VLM Model	VLM-Only@Country	HybridGeo-v2@Country	Gain (Absolute)
Qwen3-VL-8B	60.46%	68.76%	+8.30%
Qwen3-VL-32B	63.56%	70.19%	+6.63%
Qwen3-VL-235B	69.77%	73.89%	+4.12%

As shown in Table 10, as the capabilities of VLMs improve, the benefits of HybridGeo at the Country level show a decreasing trend. For the relatively weaker 8B model, HybridGeo brings a significant performance improvement (+8.30%); for the 32B model, the improvement is +6.63%; and for the stronger 235B model, the improvement further decreases to +4.12%. For weak models, due to their more severe hallucination issues, the 6-choice-1 mechanism of HybridGeo-v2 effectively uses retrieved candidates to correct the errors of VLMs. This finding indicates that the HybridGeo framework is of extremely high value for edge application scenarios with limited computing resources, where only smaller models can be deployed. For more powerful models that may emerge in the future, more powerful retrieval models may be needed to match them.

As can be seen from the results in Table 11, the RAG Selection method does not show a significant improvement compared to VLM-Only. It is speculated that the main reason is that the VLM is restricted to a range of K candidates and must select one even if all candidates are wrong. When there is no correct answer among the candidates, the VLM cannot "step out of the candidates" to make an independent judgment, which severely limits the exertion of its reasoning ability. At the same time, the performance of RAG Generation has not surpassed the VLM-Only baseline (City@25: 37.52% vs 38.52%, Country@750: 70.12% vs 69.77%). Although RAG Generation shows a slight improvement at the Country level, it has a significant decline at the more refined City level. To further quantify the severity of the anchoring effect, we analyzed the spatial relationship between the predicted coordinates output by RAG Generation and the candidate set. Statistics

based on the Im2GPS3k test set show that in the predictions of RAG Generation, 73.5% of the outputs are within 50km of any candidate, while only 42.1% of the predictions of VLM-Only fall within 50km of the same candidates (Table 12). This indicates that when wrong candidates are input as context, they strongly "attract" the output distribution of the VLM, making it abandon independent reasoning and explain around the candidates, even if these candidates themselves are wrong.

3.2.4. Comparison with RAG Strategies

Table 11 provides a detailed comparison between HybridGeo and different RAG strategies.

Table 11. Comparison of different fusion strategies. Boldface indicates the best value in each metric column.

Fusion Strategy	City	Region	Country	Description
VLM-Only	38.52%	53.08%	69.77%	Pure VLM Reasoning
RAG Selection	36.18%	51.74%	70.99%	VLM Selects from Candidates
RAG Generation	37.52%	52.15%	70.12%	VLM Generates Referring to Candidates
HybridGeo	**40.47%**	**56.54%**	**73.89%**	Dual-Stream Independent + Late Fusion

Table 12. Quantitative analysis of context poisoning. Boldface highlights stronger anchoring to retrieved candidates (higher within-threshold ratios and lower nearest-candidate distance).

Method	Within 50km	Within 200km	Avg. Nearest Dist.
VLM-Only	42.1%	68.3%	187.3 km
RAG Generation	**73.5%**	**85.7%**	**68.5 km**

This phenomenon directly verifies the "anchoring effect" hypothesis we put forward in the introduction, outlined as follows: when retrieval candidates are injected into the VLM reasoning chain as prior conditions, it will cause the posterior distribution to collapse towards wrong anchor points, thereby producing systematic biases. In contrast, the VLM in HybridGeo performs independent reasoning without being disturbed by retrieval noise; the system only verifies and fuses through spatial consistency in the later stage. When there is a conflict between the two streams, the fusion module has the opportunity to choose the more accurate one, thus realizing the real complementarity of the advantages of the two streams.

3.3. Qualitative Case Analysis

Figure 6 shows typical behaviors of HybridGeo under different scenarios.
Figure 6 illustrates three typical collaborative scenarios:

- **Case 1 (Agreement)**: This example corresponds to Fenway Park, the home stadium of the Boston Red Sox in Boston, Massachusetts, USA. Because the venue has abundant web photos, the visual-text alignment-based retrieval model can localize it reliably, and the VLM also identifies the stadium via distinctive visual cues and textual hints. As the two streams reach spatial consensus, the final error is only 0.02 km. Below is the VLM's core reasoning:

> *"Fenway Park" is the home stadium of the Boston Red Sox in Boston, Massachusetts, USA. The "Green Monster" is Fenway's most iconic landmark; its height and the arrangement of advertisements match historical photos. The Coca-Cola bottle installation is a signature decoration on top of the left-field wall, established in the early 2000s. The John Hancock Tower (now 200 Clarendon Street) is a prominent element of Boston's skyline, located about 1.5 km northeast of Fenway Park. The scoreboard shows "AL EAST" divisional information, consistent with an MLB American League East home venue. Taken together, these visual cues uniquely identify the location as Fenway Park.*

- **Case 2&3 (VLM Corrects Retrieval)**: In Case 2, the location is Shakespeare's Globe on the south bank of the River Thames in London, UK. The VLM can directly pinpoint the correct site via text recognition and world knowledge, whereas the retrieval model, although it identifies the correct country based on visual style, is less precise at the city/POI level. The consistency fusion module therefore favors the higher-confidence VLM prediction. In Case 3, the image comes from Shenzhen Window of the World, featuring a miniature replica of Gyeongbokgung Palace. The retrieval stream is misled to Seoul, Korea, because it primarily focuses on appearance similarity, while the VLM reasons about deeper visual logic (e.g., scale and vegetation inconsistent with the real palace) and arrives at a more accurate answer. Below is the VLM's core reasoning:

> *The architecture resembles a traditional Korean royal palace, with multi-tiered upturned eaves, blue-tiled roofs, vermilion columns, and green lattice windows. The complex is arranged symmetrically, with the main hall elevated on a platform and surrounded by corridors and courtyards, consistent with layouts such as Gyeongbokgung or Changdeokgung. Dense vegetation with tall trees and neatly trimmed shrubs suggests a temperate climate zone. However, the forecourt contains many miniature figurines, indicating this is likely a miniature landscape model rather than the real palace. Considering well-known miniature theme parks worldwide, the most plausible match is a replica site in Shenzhen, China, such as "Splendid China Folk Village" or Window of the World.*

- **Case 4 (Retrieval Corrects VLM)**: This image is of a bridge over the Mississippi River in Minnesota, USA. The VLM hallucinates and misidentifies it as Boston's Longfellow Bridge, while the retrieval model locks onto the correct location via appearance matching. Since the VLM confidence is low in this case, the fusion module selects the retrieval prediction.

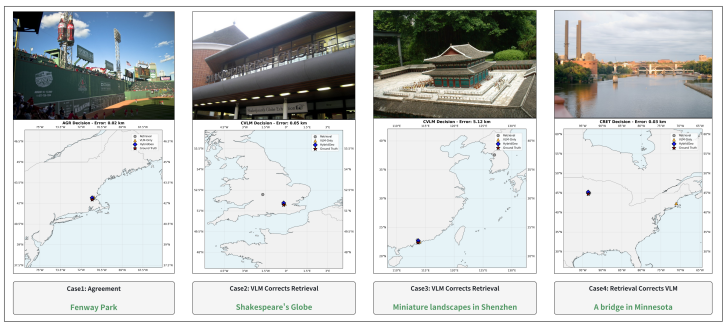


Figure 6. Qualitative examples of HybridGeo performance. Each row shows the input image, a map visualization, and a textual explanation. In the maps, the green star denotes the ground truth, blue dots denote retrieval candidates/clustered anchors, the orange triangle denotes the VLM prediction, and the red pin denotes the final HybridGeo output; "Err" indicates the great-circle distance error between a prediction and the ground truth. The figure covers the Agreement case and multiple Conflict correction scenarios, illustrating stream complementarity and how the consistency fusion module makes decisions across different situations.

These cases indicate that HybridGeo's gains do not come from naive averaging, but from a dynamic decision process triggered by distance consistency as follows: when the two streams converge within scale τ, fusion reduces random error; when they diverge significantly, arbitration selects the more reliable side based on confidence and retrieval distribution shape, effectively correcting typical unimodal failures (visual confusion in retrieval and semantic hallucination in VLMs).

To complement the successful cases above, Figure 7 presents a representative false-consensus failure under the agreement branch.



Figure 7. A false-consensus failure in the agreement branch. The query depicts a Chinese temple scene whose Guanyin statue, traditional Chinese couplets, and Southeast Asian decorative cues jointly bias both streams toward a Penang-like hypothesis. The ground truth is Bangkok, Thailand, yet the retrieval anchor, the VLM-only prediction, and the final HybridGeo-v2 output all collapse near the same wrong location, each with an error of about 925 km. Because the retrieval and reasoning predictions are only 0.90 km apart, the agreement test is satisfied and the consistency signal becomes a false confirmation rather than a correction mechanism. In the map, the red star denotes the ground truth, the gray circle the retrieval anchor, the yellow triangle the VLM-only prediction, and the blue diamond the HybridGeo-v2 output.

In this sample, the saved result is assigned to the agreement branch with a recorded consistency distance of 0.898 km. The ground truth is Bangkok, but both streams independently converge near Penang/George Town, producing errors of 925.59 km for retrieval, 925.38 km for VLM-only reasoning, and 925.46 km for HybridGeo-v2. The VLM reasoning focuses on a Guanyin statue, traditional Chinese couplets, and Southeast Asian Chinese temple decor, all of which are genuinely informative but not sufficiently distinctive at the city level. Meanwhile, the retrieval stream is attracted by the same semantic prototype. As a result, the late-fusion module receives two mutually consistent yet jointly wrong hypotheses, showing that when the two evidence chains share the same erroneous prior, post hoc spatial consistency alone cannot guarantee correction.

4. Discussion

Building on the empirical findings above, this section focuses on what the results imply mechanistically, where the framework still fails, and which directions are most promising for improvement.

4.1. Mechanistic Interpretation

Mechanistically, our emphasis is not that "retrieval + VLM is always better," but that the fusion timing determines whether retrieved evidence becomes a gain or a source of noise. When candidates are injected into the reasoning chain via early fusion/RAG, erroneous candidates can trigger anchoring and context poisoning. In contrast, late fusion separates the two evidence chains and allows spatial consistency to act as a "consensus gate" as follows: it performs continuous fusion under agreement to reduce random bias, and it enables arbitration under conflict to correct failures, yielding a more robust trade-off between retrieval confusion due to visual similarity and model hallucination due to semantic reasoning.

4.2. False Consensus and Mitigation

A more subtle failure mode is false consensus, where retrieval and reasoning are not complementary errors but instead collapse to the same wrong hypothesis. In this situation, the agreement signal no longer acts as evidence verification; it becomes an error-confirmation signal because both branches inherit the same semantic prior. Figure 7 illustrates this pattern; the temple image evokes a shared Southeast Asian Chinese temple prototype, so both streams converge near Penang although the ground truth is Bangkok. Importantly, this is a relatively rare but important failure mode rather than the dominant behavior of the agreement branch, which still accounts for many of our strongest predictions.

Mitigating false consensus likely requires evidence that is not redundant with the current two streams. One direction is to add a third evidence stream, such as remote-sensing imagery, map structure, POI or knowledge-graph evidence, and other geospatial priors, to cross-validate candidate regions from a different modality. A second direction is to introduce agreement-side uncertainty checks that become more conservative when agreement occurs in highly confusable semantic categories such as temples, chain stores, or visually repeated landmarks. A third direction is to use retrieval diversity diagnostics as follows: if multiple clusters remain plausible or the retrieved evidence is overly concentrated around a single semantic prototype, the system can treat the apparent agreement as fragile rather than fully trustworthy.

4.3. Limitations

We also acknowledge several limitations. The consistency threshold essentially defines the spatial scale of "consensus" and may need re-calibration under different data distributions and evaluation granularities. When cues are scarce or scenes are highly template-like, the two streams may share the same wrong hypothesis, making spatial consistency unable to filter it. Arbitration signals such as confidence and density can be biased under long-tail uncertainty and landmark popularity. Moreover, real-world deployment must balance cost, latency, and compliance, and it must prevent misuse of geo-location capabilities for privacy invasion or sensitive inference.

4.4. Future Directions

Looking forward, a practical direction is to improve adaptivity and reliability while keeping the "late verification" principle. This includes upgrading rule-based arbitration to learnable fusion so that the system can adaptively decide whom to trust and when to

abstain, improving candidate diversity/recall while explicitly addressing popularity bias and uncertainty, and introducing remote-sensing imagery, maps, and geospatial knowledge graphs as a "third-party evidence chain." Such external evidence can validate or refute hypotheses from the two streams via cross-view alignment (ground-level vs. overhead), map-structure constraints (road geometry, coastline shape, and POI distribution), or region-level priors (climate zones and terrain), further reducing false consensus and long-tail errors without relying on multi-turn interaction.

5. Conclusions

We propose HybridGeo, a dual-stream collaborative framework for worldwide visual geo-localization. Inspired by dual-process theory, we treat the retrieval stream as "fast intuition" (System 1) that rapidly recalls stable spatial anchors at global scale and the reasoning stream as "slow logic" (System 2) that performs context-free semantic deduction to produce an independent coordinate estimate. We decouple the two evidence chains and run them in parallel, compress discrete retrieval candidates into regional anchors via spatial–semantic clustering, and apply spatial–consistency verification and arbitration at the decision stage, enabling interpretable cross-filtering of retrieval noise and semantic hallucination under a shared great-circle distance metric.

Experiments on Im2GPS3k and OSV-5M show that HybridGeo consistently outperforms unimodal and RAG baselines across multiple scales, with notable gains at the city level. These results validate a system-level recipe of "independent dual-stream inference + late-stage consistency fusion" as follows: it preserves the VLM's context-free reasoning ability while leveraging retrieval-provided spatial anchors to reduce long-tail large-error cases under worldwide visual ambiguity.

Author Contributions: Yong Tang: Conceptualization; Methodology; Software; Validation; Data curation; Writing; Visualization. Jianhua Gong: Writing—review and editing; Supervision; and Project administration. Yi Li: Supervision; Project administration; and Funding acquisition. Jieping Zhou: Writing— review and editing. Jun Sun: Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Yunnan Sci-Tech Talent and Platform Plan Project (Grant No. 202605AF350036).

Data Availability Statement: The data presented in this study are openly available in the public domain. Specifically, the Im2GPS3k dataset is available at https://github.com/lugiavn/revisiting-im2gps (accessed on 30 March 2026), and the OSV-5M dataset is available at https://github.com/gastruc/osv5m (accessed on 30 March 2026).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Gong, J.H.; Lin, H.; Xu, B.L.; Li, W.H.; Zhang, G.Y. Primary exploration of geographic metaverse from the perspective of virtual geographic environment. *Natl. Remote Sens. Bull.* **2024**, *28*, 5–20. [CrossRef]
2. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2938–2946.
3. Huh, M.; Liu, A.; Owens, A.; Efros, A.A. Fighting fake news: Image splice detection via learned self-consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–117.
4. Middelberg, S.; Sattler, T.; Untzelmann, O.; Kobbelt, L. Scalable 6-DOF localization on mobile devices. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 268–283.
5. Cao, J.; Li, Y.; Zhang, Z. Automatic geolocation estimation of images. *IEEE Access* **2018**, *6*, 53878–53888.
6. Hays, J.; Efros, A.A. IM2GPS: estimating geographic information from an image collection. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.

7. Arandjelović, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
8. Weyand, T.; Kostrikov, I.; Philbin, J. Planet-photo geolocation with convolutional neural networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 37–55.
9. Luo, C.; Li, L.; Wen, K.; Yan, X.; Zhang, A.; Nie, L. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36.
10. Haas, L.; Schurholt, K.; Klemmer, K.; Borth, D. PIGEON: Predicting Image Geolocations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 18861–18871.
11. OpenAI. GPT-4V(ision) System Card. 2023. Available online: https://openai.com/research/gpt-4v-system-card (accessed on 30 March 2026).
12. DeepMind, G. Gemini: A Family of Highly Capable Multimodal Models. 2023. Available online: https://deepmind.google/technologies/gemini/ (accessed on 30 March 2026).
13. Li, L.; Ye, Y.; Jiang, B.; Zeng, W. GeoReasoner: Geo-localization with Reasoning in Street Views using a Large Vision-Language Model. *arXiv* **2024**. [CrossRef]
14. Zhang, Z.; Wang, R.E.; Kabir, T.; Boyd-Graber, J. NAVIG: Natural Language-guided Analysis with Vision Language Models for Image Geo-localization. *arXiv* **2025**. [CrossRef]
15. Wang, Y.; Liu, Z.; Wang, Z.; Liu, P.; Hu, H.; Artzi, Y. GeoVista: Web-Augmented Agentic Visual Reasoning for Geolocalization, *arXiv* **2025**. [CrossRef]
16. Berton, G.; Masone, C.; Caputo, B. Rethinking visual geo-localization for large-scale applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 4878–4888.
17. Ali-bey, A.; Chaib-draa, B.; Giguère, P. MixVPR: Feature mixing for visual place recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 2998–3007.
18. Zhu, S.; Yang, T.; Chen, C. TransLocator: Global-Scale Visual Geo-Localization with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1928–1938.
19. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems, Virtual Conference, 6–12 December 2020; Volume 33, pp. 9459–9474.
20. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based spatial clustering of applications with noise. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]
21. StreetView 5M Dataset. 2024. Available online: https://github.com/gastruc/osv5m (accessed on 30 March 2026).
22. Muller-Budack, E.; Pustu-Iren, K.; Ewerth, R. Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification. In *Computer Vision—ECCV 2018, Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland; pp. 575–592.
23. Seo, P.; Weyand, T.; Sim, J.; Han, B. CPlaNet: Enhancing Image Geo-localization by Combinatorial Partitioning of Maps. In *Computer Vision—ECCV 2018, Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland; pp. 544–560.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.