GeoHiPO: Geolocalization via Hierarchical Preference Optimization

David Rivard
Lake Superior State University
drivard@lssu.edu

Rohit Gupta
University of Central Florida
rohit.gupta@ucf.edu

Parth Parag Kulkarni
University of Central Florida
parthparag.kulkarni@ucf.edu

Mubarak Shah
University of Central Florida
shah@crcv.ucf.edu

Abstract

We introduce GeoHiPO, a hierarchical preference optimization framework to train vision-language models (VLMs) for image geolocalization. GeoHiPO addresses limitations of current methods by leveraging hierarchical filtering, direct preference optimization (DPO), and spatial reasoning. We selected a difficulty-balanced subset of MediaEval Placing 2016 and Mapillary data based on model-specific performance, augmented by reasoning-rich data generated with Kimi-VL. Our hierarchical filtering bins training examples into easy, medium, and hard categories, enforcing model generalization. Evaluations on the IM2GPS3K dataset showed improvements, with GeoHiPO achieving a 25km accuracy increase of 1.74% over the Qwen2.5-VL-7B baseline model.

1. Introduction

Geolocalization is the task of determining the geographical location from image data and is essential for applications ranging from disaster response, navigation, and defense to tourism and social media geolocation. Current image-based geolocalization methods predominantly rely on either flat classification frameworks or direct coordinate regression, trained with large-scale geo-tagged datasets. While these approaches achieve reasonable accuracy within well-documented regions and landmarks, they struggle with generalizing to diverse, ambiguous, or less frequently photographed locations. Additionally, these methods rarely incorporate explicit spatial reasoning, thus limiting their ability to interpret and reason about visual geographical indicators as humans do. Existing vision-language models (VLMs), despite demonstrating substantial success in diverse multimodal tasks, similarly face critical limitations in the geolocation domain:

- **Poor Generalization:** Performance drastically drops in less-represented regions, especially where geo-tagged data is sparse or biased.
- **Lack of Spatial Reasoning:** Most of the current models don't explicitly reason about geographic features and spatial relationships between visual cues and places.
- **Inefficient Training:** Retraining specialized models from scratch for geolocalization tasks often requires extensive, carefully annotated datasets, which is impractical and resource-intensive.

Motivated by these limitations, we introduce GeoHiPO (Geolocalization via Hierarchical Preference Optimization)-a novel hierarchical training framework designed to enhance spatial reasoning in vision-language models through structured preference optimization. Rather than traditional class-based labels or raw regression losses, GeoHiPO leverages direct preference optimization (DPO) to guide models toward spatially accurate reasoning. The approach uniquely combines hierarchical data filtering based on difficulty and explicit reasoning annotations, allowing models to incrementally learn from progressively challenging spatial scenarios. GeoHiPO is model-agnostic, meaning it can be applied across a variety of existing multimodal architectures without extensive modification. This flexibility comes from its use of preference-based training, which does not directly rely on specific architectural components. By aligning model predictions with human-like spatial reasoning and directly incorporating ground-based (geographical distance) errors, GeoHiPO aims to improve predictive accuracy, calibration, and generalization simultaneously. Specifically, our research seeks to address the question: *Can we improve geolocation accuracy by teaching vision-language models to reason spatially and hierarchically, in a manner analogous to human reasoning?* Our contributions are as follows:

- Introduction of a hierarchical preference optimization strategy to improve spatial reasoning in multimodal geolocation.
- Development of hierarchical filtering techniques to systematically select training samples based on model difficulty (easy, medium, hard).
- Integration of detailed reasoning annotations via a structured reasoning dataset, enhancing model interpretability and accuracy.
- Demonstration of the effectiveness of our framework on the IM2GPS3K benchmark, achieving accuracy improvements compared to traditional training methods.

2. Related Work

Classical geolocalization. Early work framed image geolocalization either as (a) *classification* into discrete geocells [9, 10] or (b) *regression* to latitude-longitude pairs [11, 12]. Classification systems achieve high accuracy in densely photographed areas but degrade sharply in the long-tail; regression models avoid discretization yet still require massive, geo-balanced supervision and rarely provide explanatory cues.

State-of-the-art decoder and retrieval models. Recent architectures have moved beyond coarse cells toward richer sequence generation and retrieval: *GeoDecoder* [18] introduces a query-based transformer decoder that autoregressively predicts fine-grained geographic tokens, achieving SOTA results on multiple benchmarks. *GeoCLIP* [19] pioneers image-to-GPS retrieval with a CLIP-style contrastive objective, bridging vision embeddings and geographic coordinates in a single shared space. More recently, *PIGEON* [20] combines geocell classification, cluster refinement, and contrastive pretraining to surpass human expert performance in GeoGuessr-style evaluation.

Vision-language geolocators. Parallel to these purely visual systems, multimodal approaches such as *GeoReasoner* [2] and *GAEA* [21] leverage large VLMs to generate free-form place descriptions with explicit spatial reasoning. These models highlight the promise of natural-language explanations but have, to date, relied on conventional supervised objectives rather than preference alignment.

Preference learning and curricula. Outside geolocation, Direct Preference Optimization (DPO) [1] and SimPO [4] demonstrate that aligning language models with relative human preferences yields better factuality and controllability. Curriculum strategies such as self-paced learning [5] show that gradually increasing sample difficulty promotes generalization.

Position of GeoHiPO. GeoHiPO unifies these strands by (i) adopting a model-agnostic VLM backbone, (ii) introducing a hierarchical *easy* $\rightarrow$ *hard* curriculum inspired by self-paced learning, and (iii) optimizing with DPO over structured reasoning pairs-thereby complementing decoder-based (GeoDecoder), retrieval-based (GeoCLIP), and hybrid (PIGEON) advances with explicit spatial reasoning and preference alignment. To our knowledge, GeoHiPO is the first framework to combine hierarchical curricula, multimodal reasoning, and distance-grounded preference optimization for global image geolocalization.

3. Dataset Preparation

A well-curated dataset is crucial for training effective vision-language models (VLMs), especially for tasks such as geolocalization that demand nuanced spatial reasoning. In the GeoHiPO framework, we primarily utilized the MediaEval Placing 2016 (MP16) dataset [13], supplemented by additional imagery from the Mapillary dataset [14], ensuring diverse geographical coverage and scene variety.

3.1. Hierarchical Difficulty Filtering

We implemented a hierarchical filtering strategy to systematically select images based on model performance, categorizing them into distinct difficulty levels-Easy, Medium, and Hard-to guide effective incremental training. **Easy Images** were defined as those correctly predicted by both the Qwen2.5-VL-3B and Qwen2.5-VL-7B models. These images were excluded from the final training dataset due to their simplicity, as both shallow and deeper models successfully identified their locations. **Medium Images** included images partially identified by either model, typically where predictions captured either city-level or country-level correctness but not both simultaneously. These images were central to maintaining a balanced dataset, challenging models to generalize across ambiguous visual contexts. **Hard Images** consisted of scenarios where only the deeper Qwen2.5-VL-7B model correctly identified the location, while the shallower 3B model failed. This subset represents challenging visual contexts and helps the model develop robust spatial reasoning capabilities. This hierarchical filtering allowed GeoHiPO to incrementally expose models to progressively challenging spatial reasoning scenarios, closely mimicking a curriculum learning approach.

3.2. Structured Reasoning Data Generation

In addition to hierarchical filtering, we generated structured reasoning annotations to enhance model interpretability and spatial reasoning accuracy. For this, we employed the Kimi-VL model [8], a powerful vision-language model specialized in producing detailed reasoning-based predictions. We processed selected subsets of images, generating explicit structured reasoning statements about geographic context, architectural cues, and environmental indicators. These structured reasoning annotations were crucial for two primary training stages:

1. **Thinking-SFT Stage:** Fine-tuning the base Qwen2.5-VL model using single correct predictions accompanied by explicit reasoning to develop the model's reasoning capability.

2. **HiPO-DPO Stage:** Creating reasoning-based preference pairs, consisting of one chosen (correct) reasoning prediction and one rejected (incorrect) reasoning prediction. These pairs directly guided preference optimization during the final fine-tuning step.

3.3. Evaluation Dataset

For evaluation, we primarily used the IM2GPS3K benchmark [15]. This dataset contains images from diverse global locations, making it ideal for assessing the generalization capability and accuracy of our trained models. Although we initially intended to evaluate on additional benchmarks such as GWS15K, practical constraints limited our evaluations to IM2GPS3K, providing a rigorous yet manageable benchmark scenario. Through this comprehensive and structured dataset preparation, GeoHiPO successfully leveraged hierarchical filtering, explicit spatial reasoning annotations, and structured preference optimization to significantly enhance multimodal geolocation performance.

4. Methodology

GeoHiPO introduces a hierarchical preference optimization framework specifically designed to enhance spatial reasoning in vision-language models (VLMs) for geolocalization tasks. Our approach uniquely integrates direct preference optimization (DPO), hierarchical training data filtering, and structured reasoning annotations to systematically improve geolocation predictions. Below, we detail each component and training stage in our methodological pipeline.

4.1. Model Selection and Initial Setup

We began by selecting suitable base models capable of multimodal reasoning. After evaluating various options, we primarily chose: **Qwen2.5-VL (3B and 7B)** [6]: A powerful and versatile vision-language model providing strong baseline performance in multimodal tasks. **Kimi-VL (A3B)** [8]: Chosen specifically for generating detailed structured reasoning annotations due to its sophisticated textual reasoning capabilities.

4.2. Hierarchical Spatial Reasoning Dataset Creation

Structured reasoning annotations were generated using Kimi-VL. For each image, Kimi-VL produced detailed reasoning descriptions highlighting visual cues, geographic contexts, environmental conditions, architectural features, and spatial indicators. These reasoning annotations served two critical roles: **Thinking-SFT Stage:** Training the Qwen2.5-VL base models explicitly on correct reasoning predictions to build foundational spatial reasoning skills. **Preference Pair Construction:** Creating hierarchical reasoning-based preference pairs used directly for DPO training.

4.3. Preference Pair Generation and Hierarchical Filtering

Preference pairs consisted of two structured reasoning predictions for the same image: a chosen prediction (accurate spatial reasoning and correct location) and a rejected prediction (plausible reasoning yet geographically inaccurate). Preference pairs were systematically filtered based on explicit geographic distance thresholds to ensure meaningful learning signals: **Easy Pairs:** Predictions differed significantly in geographic accuracy (e.g., correct continent vs. incorrect). **Medium Pairs:** Moderate geographic distinctions (e.g., correct country but incorrect city-level predictions). **Hard Pairs:** Closely contested predictions requiring nuanced spatial understanding (e.g., correct city vs. neighboring city). Pairs failing these criteria were excluded to ensure training examples consistently emphasized meaningful differences in spatial reasoning accuracy.

4.4. Direct Preference Optimization (DPO)

Preference optimization training was conducted using the TRL framework [7]. Unlike traditional supervised fine-tuning, DPO directly optimized models toward choosing predictions deemed more spatially accurate based on explicit human-like reasoning preferences. This approach bypassed the need for carefully annotated data during training, emphasizing relative correctness, and significantly improving generalization and calibration.

4.5. Incremental Training Strategy

GeoHiPO employed a progressive, incremental training strategy across several clearly defined stages to gradually enhance the model's spatial reasoning capabilities:

1. **Geo-SFT:** Initially fine-tuning the Qwen2.5-VL models on standard geolocalization tasks without explicit reasoning annotations, serving as a robust baseline.

2. **Thinking-SFT:** Training models explicitly with structured reasoning annotations, enhancing their capability to reason spatially about geographic contexts.

3. **Self-DPO:** Directly applying preference optimization on the original Qwen2.5-VL model using basic preference pairs without explicit reasoning, introducing the concept of preference-based optimization.

4. **HiPO-DPO (Final Stage):** Culminating in hierarchical preference optimization using structured reasoning-based preference pairs. This stage fine-tunes the Thinking-SFT model, training it to explicitly reason spatially in alignment with human judgment.

The incremental approach facilitated gradual skill acquisition in spatial reasoning, hierarchical generalization, and preference alignment, closely mirroring curriculum learning principles [5].

4.6. Computational Details

Training was conducted using NVIDIA H100 and A100 GPUs available at the Newton and CRCV computing clusters at UCF, respectively, utilizing FlashAttention2 [16] and Deepspeed [17] libraries for memory-efficient training and optimal performance. We systematically tuned hyperparameters, particularly learning rate and preference weighting parameters, to achieve optimal model convergence and performance.

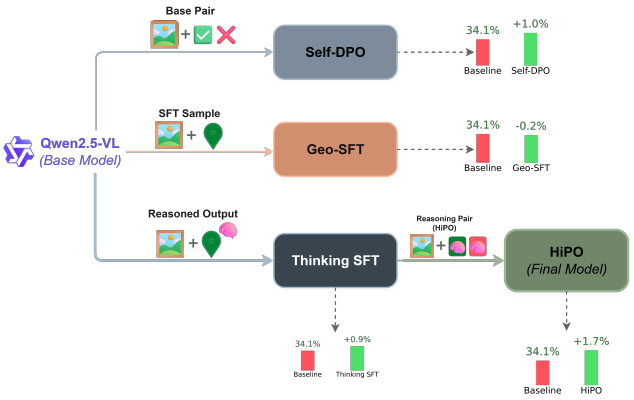


Figure 1. GeoHiPO Model Evolution.

5. Experimental Setup

In this section, we provide a detailed description of our experimental environment, model training configuration, evaluation protocols, and the practical considerations taken into account during the GeoHiPO project's execution.

5.1. Model Training Pipeline

Our training pipeline followed a multi-stage incremental approach, carefully structured to progressively introduce complexity and reasoning capability to the Qwen2.5-VL-3B and Qwen2.5-VL-7B models:

Baseline Geo-SFT Training:

- Purpose: Establish initial geolocation capabilities using standard supervised fine-tuning without explicit reasoning annotations.

Thinking-SFT (Structured Reasoning):

- Purpose: Train models explicitly on structured reasoning annotations generated by Kimi-VL, enabling the model to develop explicit spatial reasoning skills.

Self-DPO Training:

- Purpose: Introduce basic direct preference optimization using preference pairs without explicit structured reasoning annotations.

HiPO-DPO:

- Models: Final refinement of model that finetuned the Thinking-SFT model with DPO.

- Purpose: Utilize structured hierarchical reasoning-based preference pairs to explicitly optimize spatial reasoning capabilities, integrating hierarchical filtering and explicit spatial preference distinctions.

5.2. Evaluation Protocol

Our evaluation primarily utilized the IM2GPS3K benchmark [15], chosen for its challenging geographic diversity and established benchmark status within the geolocalization community. Although we intended to perform additional evaluations on the GWS15K benchmark, practical constraints led us to focus solely on IM2GPS3K for this project. Evaluation metrics included:

- Accuracy at Geographic Thresholds: Specifically reported at 25km (city-level) and 250km (country/regional-level) accuracies.

- Median and Mean Haversine Error: For comprehensive understanding of geographic precision.

- Qualitative Analysis: Analysis of prediction patterns, spatial reasoning quality, and common error types (e.g., frequently confused locations or ambiguous scenarios).

5.3. Hyperparameter Tuning Strategy

We employed systematic hyperparameter tuning informed by model performance trends:

- Learning rate, batch size, and preference optimization β-parameter were iteratively adjusted.

- Performance on validation subsets of MP16 and IM2GPS3K guided hyperparameter adjustments.

- Final hyperparameters were selected based on optimal validation performance and model stability.

5.4. Debugging and Practical Considerations

Throughout the training pipeline, extensive debugging was required, particularly during:

- **Insight-V to TRL Transition:** Initial experimentation with Insight-V revealed stability issues, prompting a transition to TRL's more robust multimodal DPO implementation.
- **Structured Reasoning Integration:** Ensuring correct integration and alignment between visual and textual reasoning annotations required careful handling.
- **Memory Management:** Regularly monitoring GPU memory usage and optimizing batch sizes with FlashAttention2 and Deepspeed was crucial for stable training.

By carefully addressing these practical challenges and systematically optimizing our experimental setup, we successfully trained robust vision-language models with enhanced hierarchical spatial reasoning and generalization capability.

6. Results

We compared our final HiPO-DPO model against several baselines, including the original Qwen2.5-VL models, Geo-SFT, Self-DPO, and Thinking-SFT variants. Evaluation metrics primarily focused on accuracy at specific geographic thresholds (25km for city-level accuracy and 250km for country-level accuracy). Table 1 summarizes these quantitative results: The GeoHiPO-trained model

Model Variant	**25km Acc. (%)**	**250km Acc. (%)**
Base Model (Qwen2.5-VL-7B)	34.13	52.22
Geo-SFT	33.97	52.28
Self-DPO	35.16	53.19
Thinking-SFT	35.07	53.49
HiPO-DPO (ours)	**35.87 (+1.74%)**	**54.59 (+2.37%)**

Table 1. Quantitative results of GeoHiPO on IM2GPS3K benchmark dataset.

clearly outperformed all baselines, achieving a significant increase in geographic accuracy at both fine (25km) and broader (250km) geographic scales. Notably, Geo-SFT slightly underperformed compared to the base model, indicating limited effectiveness when standard supervised fine-tuning alone was used. The incremental introduction of reasoning (Thinking-SFT) and preference optimization (Self-DPO) progressively improved performance, culminating in the best results with hierarchical reasoning and direct preference optimization (HiPO-DPO).

6.1. Qualitative Analysis

A qualitative examination of GeoHiPO's predictions revealed improvements in spatial reasoning capabilities. Specifically, the HiPO-DPO model enhanced interpretability of predictions, with model outputs providing detailed reasoning about geographic indicators (architecture, landscape, climate, etc.). Despite these improvements, some common error patterns remained, particularly in highly ambiguous cases (e.g., visually similar cities or landscapes). Future work will address these remaining challenges.

6.2. Summary of Results

In summary, GeoHiPO advanced the Qwen2.5-VL-7B baseline geolocalization performance by explicitly integrating hierarchical spatial reasoning and direct preference optimization. Our method improved geographic accuracy at both fine and broad geographic scales.

7. Architecture Details

GeoHiPO leverages a carefully structured multimodal vision-language architecture designed to explicitly facilitate hierarchical spatial reasoning through direct preference optimization. The architecture builds upon foundational components common in modern VLMs and integrates unique features tailored specifically for our geolocation task. Figure 2 illustrates the high-level overview of the GeoHiPO architecture, highlighting its modular components and interaction.

7.1. Vision Encoder

The vision encoder extracts visual features from input images, producing rich, descriptive embeddings that capture visual context essential for spatial reasoning. We used the pretrained vision encoder component of the Qwen2.5-VL models [6], which relies on robust vision transformer architectures (ViT), known for their strong representation learning capabilities in multimodal contexts.

7.2. Text Encoder and Reasoning Module

A key innovation in GeoHiPO is the explicit integration of structured reasoning annotations generated using Kimi-VL [8]. The text encoder processes these structured reasoning annotations, converting textual descriptions into detailed semantic embeddings. These embeddings explicitly encode spatial and geographic reasoning cues, such as architectural styles, landscape features, environmental context, and human-related indicators, providing critical reasoning context.

7.3. Multimodal Fusion Module

At the core of GeoHiPO lies the multimodal fusion module, responsible for combining visual embeddings (from

the vision encoder) and semantic reasoning embeddings (from the text encoder). The fusion mechanism uses cross-attention layers to integrate multimodal signals effectively, enabling the model to explicitly reason about geographic relationships and spatial indicators described in the textual annotations relative to visual information in the images.

7.4. Model Training and Optimization Details

We utilized TRL's multimodal DPO implementation [7], adapting it specifically to handle hierarchical reasoning-based multimodal inputs. Models were fine-tuned incrementally, ensuring stable integration of reasoning annotations and hierarchical preference signals.

7.5. Computational Environment and Efficiency

Our implementation heavily leveraged modern deep learning optimizations, including:

- **FlashAttention2** [16]: significantly improved attention computation efficiency, crucial for training large multimodal transformer models.
- **Deepspeed** [17]: facilitated distributed training across multiple GPUs, enabling stable training of large batch sizes without memory bottlenecks.

The computational environment provided by CRCV allowed efficient experimentation, training, and evaluation of various model configurations, facilitating systematic hyperparameter tuning and model optimization.

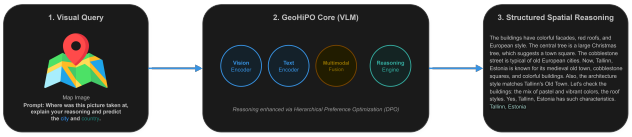


Figure 2. GeoHiPO Core Architecture.

8. Discussion

The GeoHiPO framework demonstrates a clear and consistent improvement in vision-language geolocation accuracy through the integration of hierarchical spatial reasoning and preference-based optimization. Our results suggest several important takeaways about multimodal training for geographic tasks and how structured reasoning can be effectively used to augment model performance in ambiguous real-world scenarios.

8.1. Effectiveness of Hierarchical Preference Optimization

One of the central findings of this work is that preference-based training grounded in geographic distance, rather than solely in classification accuracy or regression loss, leads to more reliable and spatially aligned predictions. The HiPO-DPO model, trained with structured preference pairs and hierarchical filtering, achieved the highest accuracy among all tested variants. Notably, it outperformed both standard supervised fine-tuning (Geo-SFT) and reasoning-only models (Thinking-SFT), highlighting the complementary strengths of both reasoning and optimization via distance-based preferences. This result reinforces our hypothesis: vision-language models can benefit from learning not just to answer correctly, but to answer in a relatively better way-especially when dealing with uncertain or ambiguous geographic inputs.

8.2. Curriculum-Like Training Behavior

Our hierarchical filtering scheme (easy, medium, hard) acted as a form of curriculum learning. By exposing the model to increasingly difficult spatial distinctions, GeoHiPO promoted better generalization and robustness. Interestingly, our filtering strategy also helped avoid overfitting to "easy" visual landmarks commonly overrepresented in geolocation datasets (e.g., Paris, London), a frequent issue in models trained solely on large geo-tagged datasets. Moreover, the structured binning of training samples (e.g., cases where only the 7B model succeeded but 3B failed) allowed us to prioritize "borderline" examples-scenarios that often challenge both humans and machines-ultimately improving the model's decision boundaries.

8.3. Interpretability through Reasoning

By incorporating structured reasoning annotations into the training pipeline (generated using Kimi-VL), GeoHiPO models gained improved interpretability. Predictions were no longer just coordinate estimates-they were reasoned, text-supported explanations grounded in observable cues such as architecture, vegetation, lighting, or infrastructure style. This made debugging, qualitative evaluation, and model accountability far more practical. Sample predictions revealed models using cues like "palm trees along a wide boulevard" or "snow-covered homes with Scandinavian roofing" to support regional predictions. This level of interpretability is particularly valuable for real-world deployment, where trust in AI systems matters as much as raw accuracy.

8.4. Tradeoffs and Limitations

Despite its strengths, GeoHiPO has its limitations:

- **Inference Time and Complexity:** Models that incorporate detailed reasoning text and preference alignment require more complex inputs and incur additional inference time. While this was acceptable in our research setting, scaling this to real-time applications may require optimization or distillation.

- **Evaluation Constraints:** While IM2GPS3K served as a rigorous benchmark, we were not able to evaluate on GWS15K or other global-scale datasets due to time constraints. Further benchmarking is needed to confirm generalizability across wider datasets.

- **Quality of Reasoning Annotations:** The effectiveness of GeoHiPO is partly dependent on the quality of structured reasoning annotations. While Kimi-VL produced strong outputs, some noise and verbosity were present. Future work may explore human-verified reasoning annotations or iterative refinement (e.g., self-critique modules).

- **Ambiguous Regions and Lookalike Cities:** As expected, even our best model struggled in cases where visual features between cities or countries are nearly indistinguishable (e.g., rural France vs. rural Quebec). Addressing these challenges may require external geographic priors or metadata in future systems.

8.5. Broader Impact and Applicability

GeoHiPO presents a new direction for geolocalization systems: one that combines human-like spatial reasoning, preference alignment, and hierarchical learning. Because our approach is model-agnostic, it can be adapted to a variety of foundation models or vision-language systems. Furthermore, the structured reasoning outputs provide an additional layer of interpretability that makes the method suitable for high-stakes applications (e.g., disaster response or humanitarian aid) where prediction justification is essential.

9. Future Work

Future research directions include:

- Scaling and geographically diversifying datasets.

- Evaluating on additional benchmarks (e.g., GWS15K).

- Implementing contrastive learning approaches.

- Introducing stricter hierarchical filtering.

- Generalizing GeoHiPO across various VLM architectures.

10. Conclusion

GeoHiPO demonstrates a clear pathway toward improved image geolocalization by integrating hierarchical spatial reasoning and preference optimization. Our model-agnostic approach presents opportunities for broader adoption across multimodal architectures.

References

[1] R. Rafailov *et al.*, "Direct Preference Optimization," *NeurIPS*, 2023.

[2] J. Li *et al.*, "GeoReasoner," *ICML*, 2024.

[3] A. Coker *et al.*, "GaGA," *arXiv preprint arXiv:2412.08907*, 2024.

[4] D. Simianer, F. Févote, "SimPO," 2024.

[5] J. Turrin *et al.*, "Self-Paced Curriculum Learning," *ICML*, 2019.

[6] Alibaba Cloud, "Qwen-VL: Vision-Language Models," 2023.

[7] HuggingFace, "Transformers Reinforcement Learning (TRL)," 2023.

[8] MoonshotAI, "Kimi-VL Vision-Language Model," 2023.

[9] T. Weyand *et al.*, "PlaNet: A deep planner for photo geolocation," *ECCV*, 2016.

[10] N. Vo *et al.*, "Revisiting the revisit problem: A case for a unified evaluation framework," *CVPR Workshops*, 2017.

[11] R. Muller *et al.*, "Geolocation estimation with deep learning," *arXiv preprint arXiv:1804.01757*, 2018.

[12] M. Izbicki *et al.*, "Exploiting feature hierarchies for image-based localization," *WACV*, 2019.

[13] B. L. D. S. *et al.*, "The MediaEval 2016 Placing Task," *MediaEval*, 2016.

[14] G. Neuhold *et al.*, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," *ICCV*, 2017.

[15] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," *CVPR*, 2008.

[16] T. Dao, "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning," *arXiv preprint arXiv:2307.08691*, 2023.

[17] J. Rasley *et al.*, "DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters," *KDD*, 2020.

[18] N. Clark, S. Vogt, and J. Philbin. Where We Are and What We're Looking At: Query-based GeoDecoder for fine-grained image geolocalization. In *CVPR*, 2023.

[19] Z. Zhong, J. Geng, and M. Li. GeoCLIP: Vision–language contrastive pre-training for image-to-GPS retrieval. *arXiv:2309.16020*, 2023.

[20] L. Haas, L. Skreta, S. Alberti, and C. Finn. PIGEON: Predicting image geolocations. In *CVPR*, 2024.

[21] A. Coker, A. Geli, X. Li, *et al.* GAEA: Geo-aware vision-language model with OpenStreetMap grounding. *arXiv:2503.16423*, 2025.