Geospatial Chain-of-thought Reasoning for Enhanced VQA on Satellite Imagery

Shambhavi Shanker, Manikandan Padmanaban, Jagabondhu Hazra

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation & Impact

Climate change impacts such as floods, wildfires, and droughts require fast, accurate interpretation of satellite imagery. Manual analysis is expert-dependent, slow, and hard to scale, motivating the need for Geospatial VQA systems. However, current models often provide direct answers without structured reasoning, limiting robustness, multi-step spatial analysis, and interpretability for high-stakes climate decisions.

Our Approach: Chain-of-Thought–Enhanced Geospatial VQA

- Adds transparent, step-wise rationales for complex spatial queries
- Improves interpretability and reliability in decision-critical scenarios
- Enhances generalization to unseen real-world geospatial conditions

This helps provide rapid decision support for emergency responders, expert-free insights allowing non-experts to query satellite imagery naturally, and supports urban planning, policy, and climate resilience.

Proposed Framework & Approach

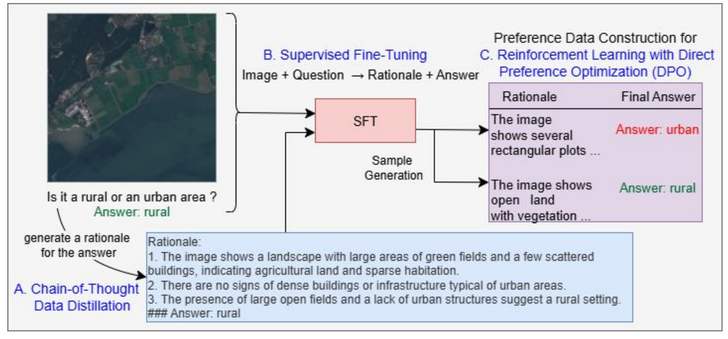
A geospatial VQA framework that integrates Chain-of-Thought (CoT) reasoning with Direct Preference Optimization (DPO).

Key Components-

- *CoT-augmented supervision*: Model generates intermediate reasoning steps to handle detection, classification, spatial relations, & comparison.
- *DPO-based preference alignment*: Reinforces high-quality reasoning and penalizes incorrect or weak rationales, improving robustness and interpretability.

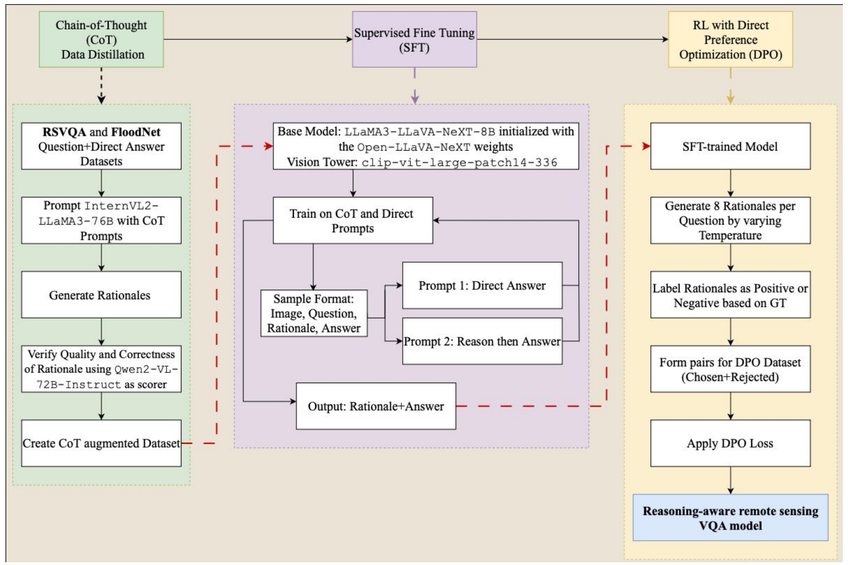
Motivation-

- CoT breaks down complex queries into intermediate steps, improving accuracy & interpretability.
- DPO aligns model outputs with preferred reasoning patterns.



Workflow diagram showing (A) Chain-of-Thought Data Distillation, (B) Supervised Fine Tuning (SFT), (C) Preference Data Construction for Reinforcement Learning with Direct Preference Optimization (DPO)..

Methodology



Results

Approach	Total Samples	Correct	Overall Accuracy	Type-wise Accuracy			
				Comp	Count	Presence	Rural/Urban
Initial Zero-Shot	14339	6858	0.4783	0.3987	0.1757	0.6091	0.6667
SFT with Direct Data	14339	6940	0.4840	0.4397	0.2619	0.5611	0.8333
SFT with CoT Data (Projection Layer Only)	14339	9548	0.6659	0.6881	0.1178	0.6890	0.7222
DPO on SFT CoT Data (Projection Layer Only)	14339	10362	0.7226	0.7099	0.1652	0.7915	0.8333
SFT with CoT Data (All Weights Unfrozen)	14339	11868	**0.8277**	**0.8532**	**0.2337**	**0.8511**	**0.7778**

Performance comparison across fine-tuning strategies on RSVQA-LR and RSVQA-HR datasets. Results are reported as overall and type-wise accuracy for different question categories.

- Compared to the direct SFT baseline, CoT-based SFT on the projection layer improved overall accuracy by **18.19%**.
- Applying DPO on top of CoT data provided a further **5.67%** improvement.
- Fine-tuning the entire model resulted in a **34.9%** improvement over the initial zero-shot baseline achieving an overall accuracy of **82.77%**.
- Zero-shot Evaluation on the *FloodNet* dataset: model fine-tuned on direct data achieved an accuracy of **59.1%**, while the model fine-tuned on CoT data achieved a higher accuracy of **67.4%**.

The Role of Reasoning in Satellite Image Understanding

Q. Is the number of flooded houses greater than the number of non-flooded houses?

Direct Ans. **Yes**

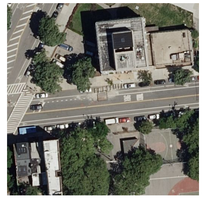
To arrive at the answer, we need to **detect** the houses, **classify** them as flooded or non-flooded, **count** them individually and **compare** them.

- Training VLMs on short answers doesn't generalise well to reasoning tasks that require more detailed responses.
- Reasoning helps the model break down complex geospatial questions into interpretable steps like object identification, spatial analysis, and comparison.
- Direct-answer training encourages shortcut learning, while CoT training provides richer contextual signals that teach the model the actual geospatial reasoning steps.

Dataset

Dataset	Total Samples	Question Types	Answer Format
FloodNet	4510	Simple/Complex Counting Condition Recognition Yes-No	Numerical Flooded / Non-Flooded Yes / No
RSVQA-HR (15 cm)	62554	Comparison (Comp) Presence	Yes / No Yes / No
RSVQA-LR (10 m)	10004	Comparison (Comp) Presence Counting Rural/Urban	Yes / No Yes / No Numerical Rural / Urban

FloodNet



RSVQA-HR

RSVQA-LR

References

Rahnemoonfar, M. et al. (2021). FloodNet: High-resolution aerial imagery for post-flood understanding. *IEEE Access.*
Lobry, S. et al. (2020). RSVQA: Visual question answering for remote sensing. *IEEE TGRS.*
Zhang, R. et al. (2024). Improving VLM Chain-of-Thought reasoning. *arXiv:2410.16198.*
Soni, S. et al. (2025). EarthDial: Multi-sensor Earth observation to interactive dialogue. *arXiv:2412.15190.*