Geospatial Chain of Thought Reasoning for Enhanced Visual Question Answering on Satellite Imagery

SHAMBHAVI SHANKER
MANIKANDAN PADMANABAN
JAGABONDHU HAZRA

Motivation

Climate-Critical Need

- Climate change impacts (floods, wildfires, droughts) demand fast, accurate interpretation of satellite imagery.
- Manual analysis is slow, expert-dependent, and difficult to scale.

Limitations of Existing VQA Models

- Current remote sensing VQA models give direct answers without structured reasoning.
- Lack robustness for multi-step spatial reasoning, comparative analysis, or causal inference.
- Insufficient interpretability for high-stakes climate tasks.

Why Chain-of-Thought Reasoning?

- Provides transparent, step-wise rationales for complex queries.
- Enhances trust, interpretability, and reliability in decision-critical applications.
- Improves generalization to unseen geospatial scenarios.

Real-World Impact

- Rapid decision support for emergency responders.
- Expert-free insights; non-experts can query satellite imagery naturally.
- Supports urban planning, policy, and climate resilience.

Proposal

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

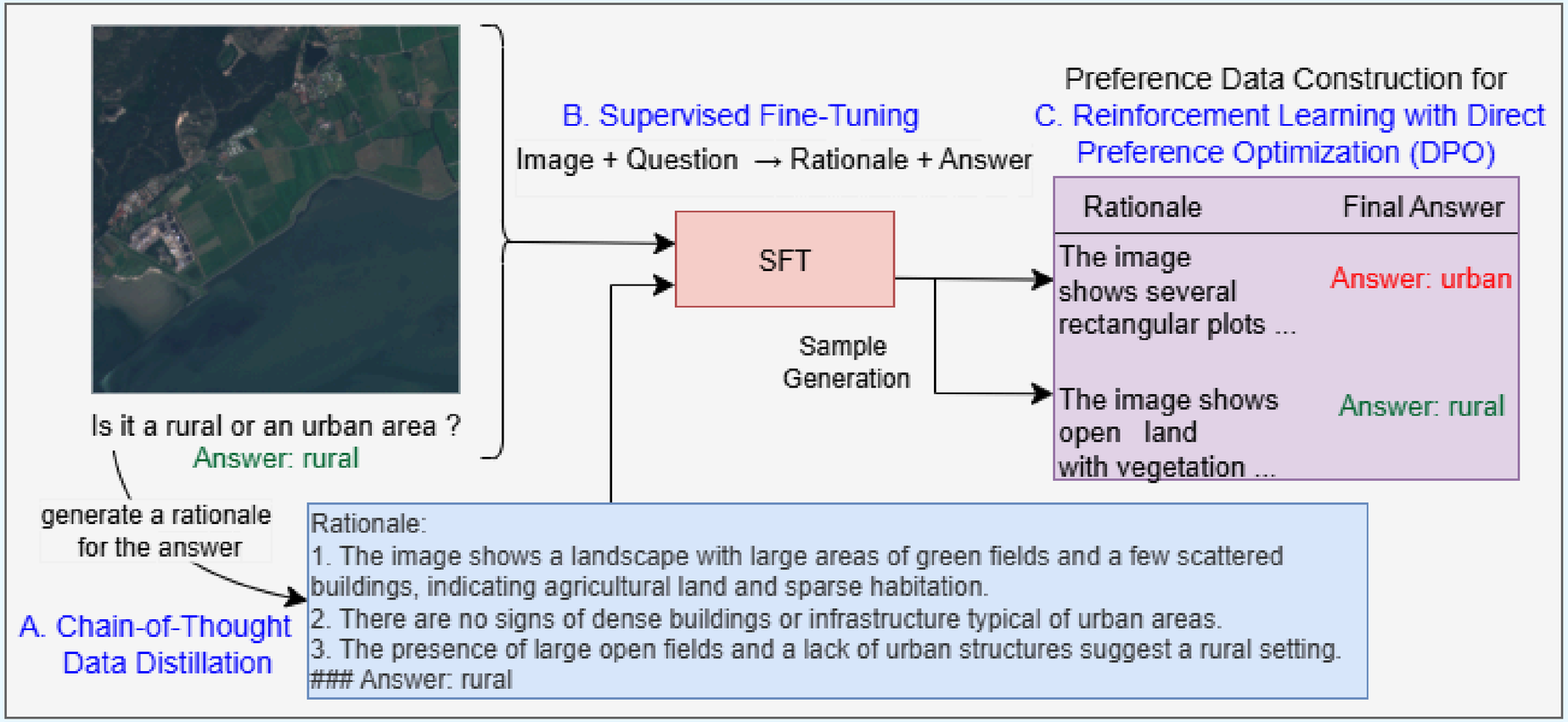
Key Components-

- 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.
- Together, they yield significant accuracy gains.

Approach


Is it a rural or an urban area ?
Answer: rural
generate a rationale for the answer
A. Chain-of-Thought Data Distillation
Rationale:
1. The image shows a landscape with large areas of green fields and a few scattered buildings, indicating agricultural land and sparse habitation.
2. There are no signs of dense buildings or infrastructure typical of urban areas.
3. The presence of large open fields and a lack of urban structures suggest a rural setting.
Answer: rural
B. Supervised Fine-Tuning
Image + Question → Rationale + Answer
SFT
Sample Generation
Preference Data Construction for
C. Reinforcement Learning with Direct Preference Optimization (DPO)
Rationale
Final Answer
The image shows several rectangular plots ...
Answer: urban
The image shows open land with vegetation ...
Answer: rural

Dataset

- RSVQA-LR (a): Sentinel-2 imagery over the Netherlands at 10 m resolution
- RSVQA-HR (b): 15 cm aerial RGB images from the USGS HRO collection covering U.S. urban areas.
- FloodNet (c): high-resolution UAV imagery collected with DJI Mavic Pro quadcopters during the Hurricane Harvey response (Texas/Louisiana, 2017)

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

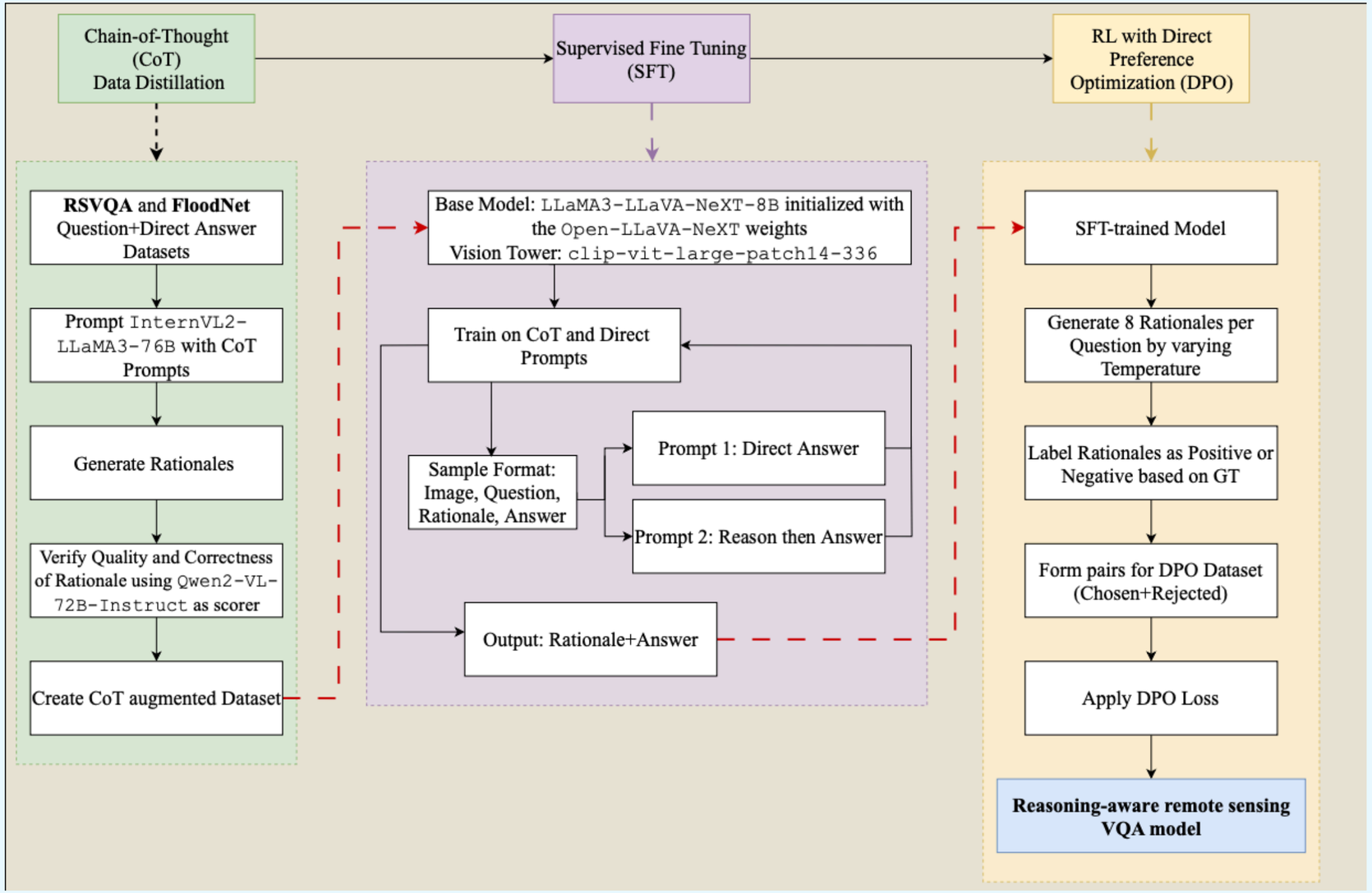
Table: Overview of datasets used in this work, including total number of samples, supported question types, and corresponding answer formats.

(a)

(b)

(c)

Methodology



Results

- 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 (vision encoder, projection layer, and language model), resulted in a **34.9%** improvement over the initial zero-shot baseline achieving an overall accuracy of **82.77%**
- We also evaluated our fine-tuned models on the *FloodNet* dataset. The 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%**

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**

Table: 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.

Looking Forward

- We built an end-to-end geospatial VQA pipeline for remote sensing data integrating rationale generation and verification to enable chain-of-thought reasoning with SFT and DPO to align the model's reasoning with high-quality answers.
- Beyond significant improvement in results over direct baselines, our model is also capable of generating interpretable reasoning, enhancing user trust and understanding.
- Moving forward, we aim to investigate the lower accuracies in numerical questions and address them; Incorporation of multispectral satellite data for richer reasoning capabilities, is a future goal.

THANK YOU!

Shambhavi Shanker- 21d070066@iitb.ac.in

Manikandan Padmanaban- manipadm@in.ibm.com

Jagabondhu Hazra- jahazra1@in.ibm.com