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

Shambhavi Shanker*
Indian Institute of Technology Bombay
Mumbai, India
21d070066@iitb.ac.in

Manikandan Padmanaban
IBM Research
Bangalore, India
manipadm@in.ibm.com

Jagabondhu Hazra
IBM Research
Bangalore, India
jahazra1@in.ibm.com

Abstract

Geospatial chain of thought (CoT) reasoning is essential for advancing Visual Question Answering (VQA) on satellite imagery, particularly in climate related applications such as disaster monitoring, infrastructure risk assessment, urban resilience planning, and policy support. Existing VQA models enable scalable interpretation of remote sensing data but often lack the structured reasoning required for complex geospatial queries. We propose a VQA framework that integrates CoT reasoning with Direct Preference Optimization (DPO) to improve interpretability, robustness, and accuracy. By generating intermediate rationales, the model better handles tasks involving detection, classification, spatial relations, and comparative analysis, which are critical for reliable decision support in high stakes climate domains. Experiments show that CoT supervision improves accuracy by 34.9% over direct baselines, while DPO yields additional gains in accuracy and reasoning quality. The resulting system advances VQA for multispectral Earth observation by enabling richer geospatial reasoning and more effective climate use cases.

1 Introduction

The growing impacts of climate change, including floods, wildfires, droughts, and extreme weather, demand accurate interpretation of Earth observation data. Satellite imagery provides rich multispectral and temporal views critical for disaster monitoring, risk assessment, and climate resilience, but manual analysis is resource intensive and traditional machine learning pipelines remain narrow and task specific. Vision language models (VLMs) address this gap by enabling natural language queries on imagery with grounded responses, making climate information more accessible and actionable. This is especially vital for domains such as flood mapping Rahnemoonfar et al. [2021], wildfire monitoring Hong et al. [2023], and climate adaptation planning Rolnick et al. [2022], where timely insights support disaster response and resilience.

Recent advances in multimodal learning have accelerated the integration of VLMs into Earth observation. Models such as EarthDial Soni et al. [2025] highlight this progress by enabling dialogue over remote sensing data and demonstrating strong results in classification, detection, captioning, change detection, and visual question answering. Other efforts, including RemoteCLIP Liu et al. [2024a],

*Work done during an internship at IBM Research India. Correspondence to: Shambhavi Shanker <21d070066@iitb.ac.in>.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

RS5M Zhang et al. [2024a], and GeoChat Kuckreja et al. [2023], contribute multimodal datasets and benchmarks for geospatial tasks. Collectively, these developments mark a shift from task-specific pipelines toward general-purpose frameworks capable of handling diverse geospatial queries. Such capabilities are especially relevant for climate applications, where decision makers must synthesize heterogeneous information under uncertainty. For instance, flood response requires reasoning about water, infrastructure, and settlements, while wildfire monitoring demands temporal comparisons of vegetation indices to assess burn severity and risks.

Despite this progress, existing VLMs often lack explicit reasoning. Trained primarily to provide direct answers, they struggle with multi-step inference, causal reasoning, or comparative analysis. In climate related decision making, where outputs must be accurate and trustworthy, this poses risks. Prior work shows that chain-of-thought reasoning improves interpretability and robustness Wang et al. [2023], Wei et al. [2022], Kojima et al. [2022], while reinforcement learning methods such as DPO Rafailov et al. [2023] and RLHF Christiano et al. [2017] align reasoning with human preferences. In multimodal settings, models like LLaVA-CoT Liu et al. [2023] and Multimodal-CoT Zhang et al. [2024b] demonstrate that reasoning traces enhance performance and interpretability. However, these advances remain underexplored in remote sensing, where reasoning over spatial relationships, multispectral features, and temporal change is essential.

Our work addresses this gap by unifying two complementary directions in remote sensing VLM research: reasoning-augmented supervision and preference-based alignment. Bringing these together in the geospatial context not only enables more interpretable and trustworthy models but also establishes a foundation for next-generation climate intelligence systems capable of supporting decision-making in high-stakes, complex scenarios.

2 Datasets

We conduct experiments on **RSVQA** (Low- and High-Resolution) and **FloodNet**, A.1 with dataset-specific question types and answer formats summarized in Table 2. **RSVQA-LR** is built from *Sentinel-2* imagery over the Netherlands at 10 m resolution Lobry et al. [2020], while **RSVQA-HR** uses 15 cm aerial RGB images from the *USGS HRO* collection covering U.S. urban areas. **FloodNet** contains high-resolution UAV imagery collected with *DJI Mavic Pro* quadcopters during the Hurricane Harvey response (Texas/Louisiana, 2017), offering unique fidelity as real disaster-response imagery Rahnemoonfar et al. [2021].

3 Methodology

Our pipeline consists of three main stages: (A) **CoT Data Distillation**, (B) **Supervised Fine-Tuning (SFT)**, and (C) **Reinforcement Learning with DPO**. Each stage builds on the previous one, progressively improving the model's ability to reason over geospatial imagery while producing reliable and interpretable responses.

A. Chain-of-Thought Data Distillation The limited availability of **CoT-annotated data** for geospatial VQA poses a significant challenge. To address this, we employ a data distillation strategy that leverages existing direct answer data while enriching it with rationales. Specifically, the *InternVL2-LLaMA3-76B* model Chen et al. [2024] is prompted with the image, question, and ground-truth answer to generate step-by-step reasoning A.2 that leads to the correct prediction. This process allows us to construct synthetic rationale-augmented training data from otherwise rationale-free supervision.

To ensure quality, we introduce a verification step using *Qwen2-VL-72B-Instruct* Wang et al. [2024] as an evaluator A.4. This model scores the rationales according to both their quality—defined in terms of coherence, factual grounding, and logical consistency—and their completeness. Samples with low-scoring rationales or with rationales that yield an incorrect final answer are discarded. The result is a curated dataset of high-quality rationale-augmented examples suitable for downstream fine-tuning.

B. SFT For fine-tuning, we use *LLaMA3-LLAVA-NeXT-8B*, initialized with *Open LLaVA-NeXT* weights Liu et al. [2024b]. The training corpus contains two types of supervision: direct ques-

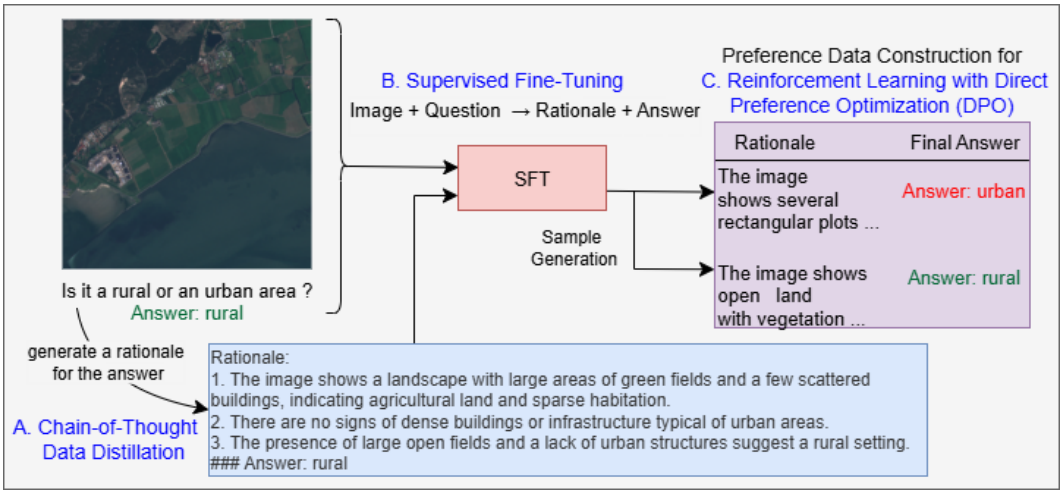

Figure 1: Workflow diagram showing the methodology: (A) Chain-of-Thought Data Distillation, (B) SFT, (C) Preference Data Construction for Reinforcement Learning with DPO.

tion–answer pairs and question–rationale–answer triples. We experiment with two parameter-efficient strategies. In the first, only the projection layer is updated while the backbone and the vision tower remain frozen. In the second, we unfreeze all parameters, including those of the vision tower (*CLIP-ViT-L/14@336px* Radford et al. [2021]), allowing for end-to-end optimization.

Different prompt templates are used depending on the supervision type 5. When training on direct QA data, the model is instructed to `Answer the question with a short answer`. In contrast, when training on rationale-augmented data, the prompt is modified to `Generate a reason first and then output a short answer`, where the rationale is followed by the answer in the format `### Answer: <final_answer>`. The model is trained for two epochs in both settings, ensuring exposure to both direct and rationale-augmented supervision.

C. Reinforcement Learning with DPO While SFT enables the model to produce rationales, it does not guarantee that these rationales are consistent with user preferences or that they reliably guide the model to correct answers. To further improve alignment, we employ DPO Rafailov et al. [2023], a reinforcement learning method that directly optimizes the policy model by contrasting preferred and non-preferred outputs. DPO formulates training as a binary cross-entropy objective that compares the model's likelihood of generating positive versus negative responses. In doing so, it increases the probability of producing coherent and correct rationales while discouraging poor reasoning.

To construct preference pairs, we use the SFT model itself as the policy generator. For each input, eight candidate responses are sampled: four with a decoding temperature of 0.2 and four with a temperature of 0.6, ensuring a balance between determinism and diversity. Each response is evaluated against the ground-truth answer 3, and data points lacking at least one correct and one incorrect response are removed. From the remaining pool, one correct rationale–answer pair and one incorrect rationale–answer pair are randomly selected, forming a positive–negative pair for DPO training. This strategy yields a dataset that captures the range of plausible model behaviors while providing a clear supervision signal.

Formally, the DPO training dataset is defined as

$$\mathcal{D}_{\text{DPO}} = \{(I, x, y^+, y^-)\},$$

where I denotes the input image, x the question, y^+ the preferred (positive) rationale–answer, and y^- the non-preferred (negative) counterpart. The optimization objective is

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(I,x,y^+,y^-)\sim\mathcal{D}_{\text{DPO}}}\left[\log\sigma\left(\beta\left[\log\frac{\pi_\theta(y^+\mid x, I)}{\pi_{\text{ref}}(y^+\mid x, I)} - \log\frac{\pi_\theta(y^-\mid x, I)}{\pi_{\text{ref}}(y^-\mid x, I)}\right]\right)\right], \tag{1}$$

where π_θ is the policy model to be optimized, π_{ref} is the reference model initialized with SFT weights, σ is the logistic function, and β is a scaling hyperparameter set to 0.1 in our experiments. This formulation encourages the model to assign higher probability to preferred rationales and lower probability to less desirable ones, thereby aligning the reasoning process more closely with correctness and user expectations.

4 Results

Our experiments demonstrate the effectiveness of incorporating *CoT* supervision into vision-language models for remote sensing question answering. Compared to the direct SFT baseline, CoT-based SFT on the projection layer improved overall accuracy by **18.19%**, highlighting the benefits of explicit reasoning signals. Applying Direct Preference Optimization (DPO) on top of CoT data provided a further **5.67%** improvement, confirming the role of preference alignment in refining reasoning quality. The most substantial gains came from fine-tuning the entire model (vision encoder, projection layer, and language model), which resulted in a **34.9%** improvement over the initial zero-shot baseline, achieving an overall accuracy of **82.77%**.

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 1: 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.

Our best-performing model also generates interpretable reasoning traces, which enhances user trust and improves transparency. This is particularly valuable in remote sensing applications, where understanding the rationale behind a prediction is as important as the prediction itself.

Despite these improvements, the model continues to struggle on counting-based questions, with type-wise accuracy lagging behind other categories. We hypothesize that this is due to limitations in text-based CoT explanations, which may not adequately represent precise numerical reasoning. Addressing this limitation could require integrating explicit counting modules or grounding mechanisms that align object-level detections with reasoning steps.

All models were trained on the **RSVQA-LR** and **RSVQA-HR** subsets and evaluated on the same benchmarks. 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%**, suggesting improved transferability with rationale-based training but also highlighting challenges in adapting to disaster-specific imagery. Training convergence was efficient, with the model stabilizing within two epochs.

5 Conclusion

This work shows that chain-of-thought supervision enhances geospatial VQA by improving both accuracy and interpretability. On RSVQA, CoT supervision yielded a 34.9% accuracy gain over direct baselines, with DPO providing further refinement. On FloodNet, CoT also improved transfer performance (59.1% $\rightarrow$ 67.4%), highlighting stronger generalization to disaster imagery. Beyond accuracy, reasoning traces promote transparency and trust, which are vital in climate and disaster response. Remaining challenges, especially in counting and cross-dataset adaptation, suggest the need for explicit numerical reasoning and stronger transfer methods. Overall, structured reasoning emerges as a promising pathway toward reliable and trustworthy geospatial AI.

References

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. doi: 10.1109/ACCESS.2021.3090981.

Ziliang Hong, Emadeldeen Hamdan, Yifei Zhao, Tianxiao Ye, Hongyi Pan, and A. Enis Cetin. Wildfire detection via transfer learning: A survey. *arXiv preprint arXiv:2306.12276*, 2023. URL https://arxiv.org/abs/2306.12276.

David Rolnick, Priya Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Krishna Sankaran, Alexandra S Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys*, 55(2):1–96, 2022.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, and Salman Khan. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *arXiv preprint arXiv:2412.15190*, 2025. URL https://arxiv.org/abs/2412.15190.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2024a. URL https://arxiv.org/abs/2306.11029.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision- language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024a. doi: 10.1109/TGRS.2024.3449154.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. *arXiv preprint arXiv:2311.15826*, 2023. URL https://arxiv.org/abs/2311.15826.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. URL https://arxiv.org/abs/2304.08485.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2024b. URL https://arxiv.org/abs/2302.00923.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, December 2020. ISSN 1558-0644. doi: 10.1109/tgrs.2020.2988782. URL http://dx.doi.org/10.1109/TGRS.2020.2988782.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. URL https://arxiv.org/abs/2312.14238.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

A Appendix

A.1 Datasets: Distribution and Visual Examples

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

Table 2: Overview of datasets used in this work, including total number of samples, supported question types, and corresponding answer formats.

Figure 2: RSVQA-LR

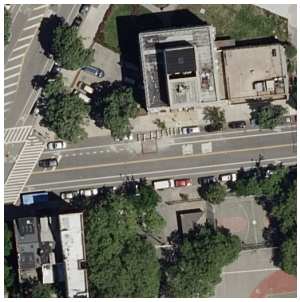

Figure 3: RSVQA-HR

Figure 4: FloodNet

A.2 Prompt Template for CoT Generation

Prompt for Rationale Generation

```
<|image|>
You are given a satellite image and a question, along with a
reference answer. Your task is to generate a detailed rationale
that uses the visual content of the image to derive the answer.

## Question
{question}

## Reference Answer
{answer}

## Instructions
- Do not state an answer at the beginning
- Use detailed reasoning grounded directly in what you see in the
  image.
- Justify your conclusion clearly with observations.
- The reasoning should be clear, grounded in the visual content,
  and should not speculate or estimate.
- Do not comment on the reference answer. Do not state whether the
  reference answer is correct or consistent to your finding.
  You are writing your own solution.
- Finish with a final derived one word answer on a new line as:


### Answer: <your_answer>
```

A.3 Input-Output Example for Model Evaluation

Input Image + Question	Output " Generate a reason first and then output a short answer"	Output " Answer the question with a short answer"
 **Is there a rectangular building?**	1. The image shows an aerial view of a city block with various structures. 2. One of the buildings in the image has a distinct rectangular shape with straight edges and right angles. 3. The building is located towards the bottom right of the image and is surrounded by other buildings and streets. ### Answer: yes	**yes**
 **What is the overall condition of the given image?**	1. The image shows a large area of land with a significant portion covered in water, indicating that the area is submerged. 2. The water appears to have spread across a wide region, suggesting that the land is not just wet but heavily inundated. 3. The presence of water extending over what seems to be a field or grassy area indicates that the normal land use is disrupted by the water coverage. 4. The water's extent and the lack of dry ground in the majority of the visible area imply a severe water accumulation. ### Answer: flooded"	**flooded**

Figure 5: Prompt-based Model Input and Output

A.4 Prompt Template for Rationale Evaluation

Prompt for Rationale Scoring

```
You are a visual reasoning expert. Your task is to verify if a given
rationale (chain-of-thought) explanation is logically coherent based on
the provided image, question, and answer.

Please follow these steps:
1. Read the question and analyze the image carefully.
2. Evaluate the rationale for:
   - Logical consistency
   - Visual grounding in the image
   - Relevance and coherence
   - Completeness (did it miss anything important?)
3. Determine whether the answer derived from the rationale follows
   naturally and reasonably and is valid based on the image and question.
4. Compare the provided answer with the ground truth answer.
5. If the answers differ, decide which answer is more likely
   correct.
6. Provide a quality score (0 to 10) indicating how confident you
   are in the rationale.
7. Poor structure, vague reasoning, or lack of detail should lower
   the score.

Format as JSON:
  "rationale_logically_valid": true/false,
  "rationale_consistent_with_image": true/false,
  "rationale_correctness_score": float,
  "rationale_quality_score": float,
  "answer_follows_from_rationale": true/false,
  "answer_matches_ground_truth": true/false,
  "correct_answer": "model"/"ground_truth",
  "explanation": "Short explanation"
```

A.5 Example: DPO Training Data Construction

Image + Question:

Are there more grass areas than roads? Answer in lower case. Generate a reason first and then output a short answer.

Ground Truth Rationale + Answer:

The image shows a large highway with multiple lanes of traffic, indicating a significant amount of road space. The grass areas are limited to small patches on the sides and median of the highway, which are much smaller in comparison to the road area. The majority of the image is dominated by the road, with vehicles and lanes clearly visible, while the grassy areas are minimal and scattered.

`### Answer: no`

The image shows a road with multiple lanes and a grassy area on the side. The road occupies a significant portion of the image, while the grassy area is relatively smaller. The road is clearly the dominant feature, with multiple lanes and vehicles, indicating it is the primary focus. The grassy area is limited to the side of the road and does not cover a large portion of the image. **### answer: no**
The image shows a road with multiple lanes and a grassy area on the side. The road occupies a larger portion of the image compared to the grassy area. The grassy area is limited to the sides of the road, while the road stretches across the image, indicating a greater area of road compared to grass. **### answer: no**
The image shows a large road with multiple lanes and a few grass areas. The road occupies a significant portion of the image, while the grass areas are relatively small. The road is clearly more extensive than the grass areas. **### answer: no**
The image shows a highway with multiple lanes and a grassy area alongside it. There are more visible lanes than grassy areas. **### answer: no**
The image shows a large road with multiple lanes and several vehicles. The surrounding area has a significant amount of grass, including a large green space and smaller patches of grass. The road occupies a significant portion of the image, while the grassy areas are more prominent and cover a larger area. **### answer: no**
The image shows a highway with multiple lanes, and there are several grass areas visible on both sides of the road. The grass areas are more extensive than the road itself. **### answer: yes**
The image shows a highway with multiple lanes and several vehicles. To the left of the highway, there is a grassy area with trees. The grassy area is significantly larger than the road lanes. The grass area occupies more space than the roads. **### answer: yes**
The image shows a large area of grass on the left side and a long road on the right side. The grass area is significantly larger than the road area. **### answer: yes**

Table 3: Example of eight candidate responses for DPO preference data construction. Final answers are highlighted in green (no) and red (yes).