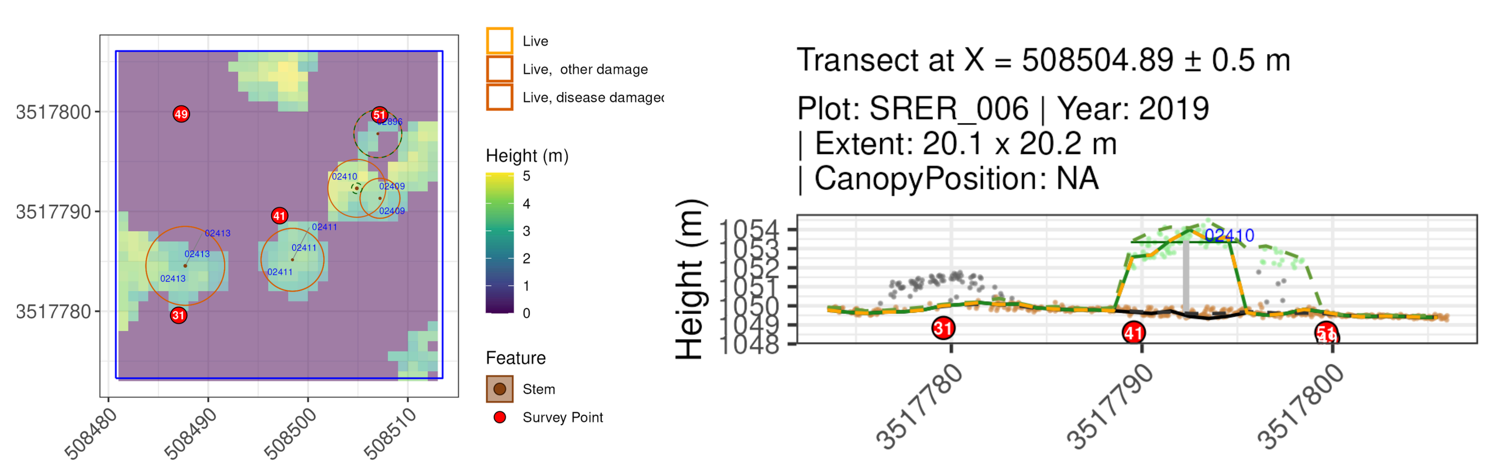
TreeAgent: A Generalizable Multi-Agent Framework for Automated Bias Labeling in Forestry via Compiled Expert Rules and Vision-Language Models

Shiyi Chen , Nicholas Saban , Collin Hargreaves , Huiqi Wang

1. University of California, Berkeley

1. Why Scientific Labeling Is Hard?



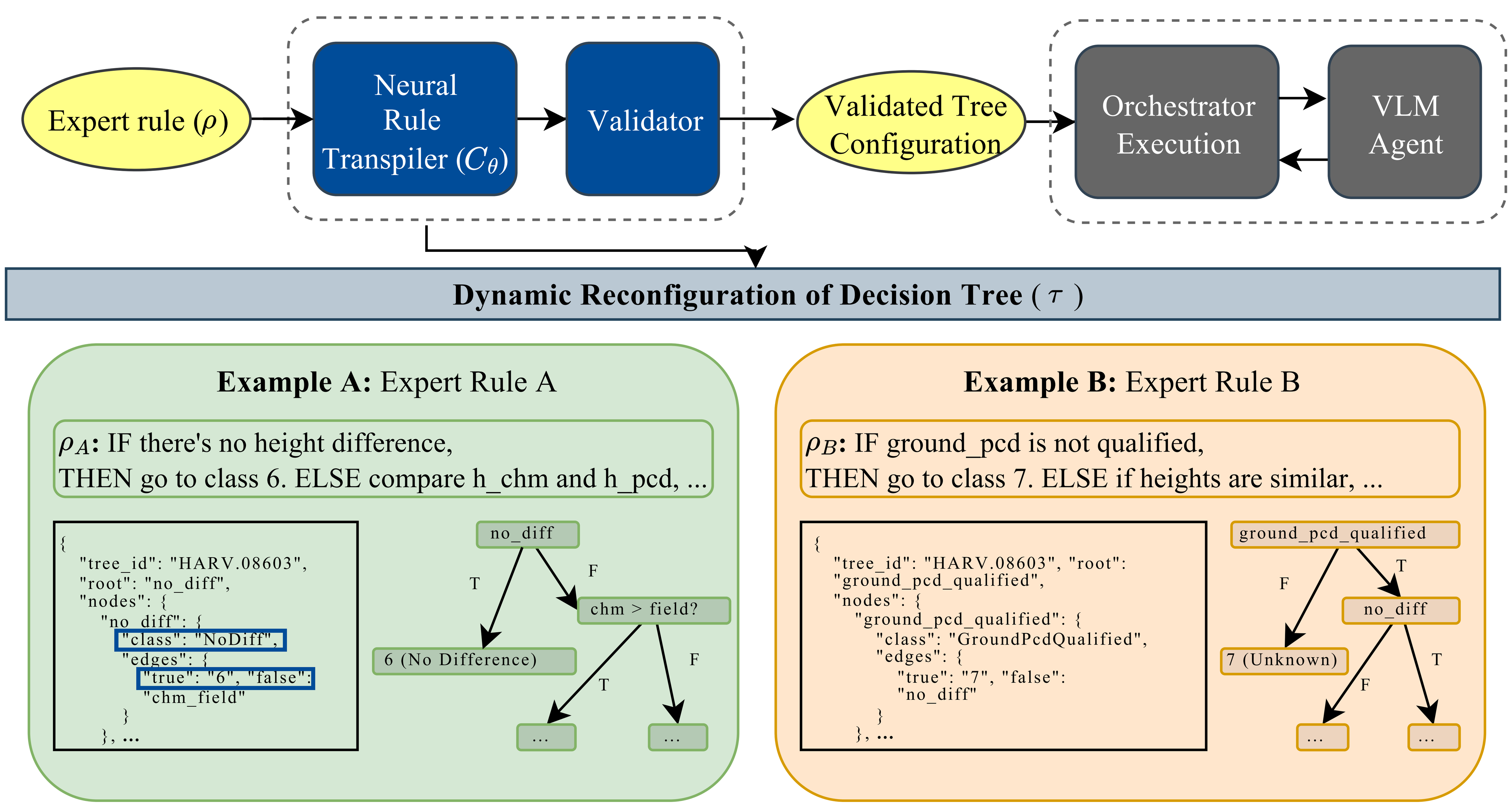
- There is systematic bias in tree height estimation. Labeling the source of bias requires visual interpretation, which is slow and not scalable.
- **RQ1.** Can VLMs reliably detect perceptual cues from scientific figures?
- **RQ2. Can a multi-agent system, given expert rules, classify the bias types end-to-end?**
- **RQ3.** Does the resulting system outperform baseline ML classifiers?

2. Methods for Automated Labeling

Approach	Limitations
Supervised ML	Needs labeled data, not auditable
Single LLM Prompt	Unreliable, Costly
Hard-coded Rule	Difficult to maintain

- Treat the expert rule as a decision tree while VLMs perform **localized semantic perception**.
- The Decoupled Declarative Decision (D3) framework:
  - **Verifiability**: rigidly encode individual nodes and decision tree structure as structural prior.
  - **Generalizability**: use LLMs to compile expert rules described in natural language to decision trees that are executable by the orchestrator agent.

2.1 Decoupled Declarative Decision (D3) Framework



- **Logic Primitive Inventory, LPI**: decouples domain-specific logic into individual nodes (deterministic, *det*; needs VLM, *vlm*; and exit, *exit*).
  - each node is defined by:
    - execution type (e.g., det)
    - field (data or figure)
    - condition (true or false)
- **Neural Rule Transpiler**: transpiles expert rule, ρ, to executable decision tree, using elements from the LPI.
- The boolean output determines where the orchestrator goes next.

3. Results

Table 1. *Prompt variants for VLMs*

Variant	Adds	Acc	Recall	Specificity
Minimal	bare Ques	48%	100%	7%
V1	figure Desc	58%	100%	25%
V2	+ 5-step CoT	52%	95.50%	18%
V3	+ expert framing	62%	90.90%	39%

- VLMs resolve perceptual nodes in decisions requiring visual interpretation.

Table 2. *TreeAgent outperforms ML baselines and reduces labeling time*

Method	Time per Tree (min)	Macro-F1
TreeAgent	0.040 ± 0.007	67.6%
Supervised ML	-	36.2%
Human Labeling	5.000 ± 2.000	-

- Our ML baseline, a class-weighted LightGBM on 25 features, reaches 36.2% Macro-F1. We confirm this with an ablation over engineered features, pseudo-labeling, other model families. ML models are limited by data availability and feature representation of forest conditions.

4. Limitations & Next Steps

- **Sensitivity to the fidelity of expert rule.** The system's accuracy depends vastly on the fidelity between the encoded rule and the rule used by human annotators.
- **Not flexible over edge cases.** Hard-coded thresholds are brittle at boundaries, but softening them with visual adjudication carries a non-trivial computational cost.
- Current VLM limitations on scientific figures suggest that **perception**, rather than orchestration, is now the primary bottleneck.
- TreeAgent suggests that **combining expert rules with foundation models** may be more effective than purely data-driven learning in low-label scientific domains.

Selected Reference. [1] Aidan Z.H. Yang, Y. T. and Brandon Paulsen, Josiah Dodds, D. K. Vert: Verified equivalent rust transpilation with large language models as few-shot learners. URL https://arxiv.org/html/2404.18852v2. **Dataset** drawn from the NEON API (https://data.neonscience.org/data-api/)

Acknowledgement: We thank the support from Trevor Keenan and Berkeley Data Discovery program, Emily Remirez and George Obaido.