# Decomposing Visual Classification: Assessing Tree-Based Reasoning in VLMs

**Anonymous submission**


## Abstract

Vision language models (VLMs) excel at zero-shot visual classification, but their performance on fine-grained tasks and large hierarchical label spaces is understudied. This paper investigates whether structured, tree-based reasoning can enhance VLM performance. We introduce a framework that decomposes classification into interpretable decisions using decision trees and evaluates it on fine-grained (GTSRB) and coarse-grained (CIFAR-10) datasets. Although the model achieves 98.2% accuracy in understanding the tree knowledge, tree-based reasoning consistently underperforms standard zero-shot prompting. We also explore enhancing the tree prompts with LLM-generated classes and image descriptions to improve alignment. The added description enhances the performance of the tree based and zero-shot methods. Our findings highlight limitations of structured reasoning in visual classification and offer insights for designing more interpretable VLM systems.


**Code** —
https://anonymous.4open.science/r/VLM_Classification-29CC

## Introduction

Vision language models (VLMs) have revolutionized visual classification by leveraging powerful image-text pretraining to generalize across novel categories without task-specific supervision (Radford et al. 2021; OpenAI 2023; Bai et al. 2023). However, as these models are increasingly deployed in real-world applications, two critical limitations have emerged that challenge their practical utility.

First, VLMs struggle with fine-grained classification tasks involving large, hierarchically organized label spaces (Zhang et al. 2024b). While models like CLIP and GPT-4V excel at distinguishing between semantically distinct categories (e.g., "car" vs. "airplane"), they often fail when faced with subtle visual distinctions within specialized domains, such as differentiating between traffic sign variants or medical imaging subcategories (Kim and Ji 2024; Wei 2024). This limitation is particularly problematic in safety-critical applications where precise classification is essential.

Second, the black-box nature of current VLM decision-making processes limits their interpretability and debuggability. When a model misclassifies an input, practitioners have limited insight into the reasoning pathway that led to the error, making it difficult to identify systematic biases or improve model performance. This opacity is especially concerning in domains requiring accountability and transparency.

To address these challenges, recent work has proposed incorporating structured reasoning approaches, particularly hierarchical decision trees, into VLM inference pipelines (Yellinek, Karlinsky, and Giryes 2025; Ding et al. 2025). The theoretical appeal is compelling as it suggests decomposing complex tasks into a series of interpretable decisions, tree-based reasoning promises to improve accuracy and explainability. Each decision node in the tree can be designed to capture specific visual or semantic distinctions.

However, the practical effectiveness of tree-based reasoning for classification remains an open empirical question. While such methods have shown promise in tasks like image captioning, it is unclear whether the added structural complexity benefits modern VLMs in classification tasks or if it simply introduces new sources of error. Notably, the hierarchical nature of decision trees makes them prone to error propagation, where a single incorrect decision at a higher-level node can compromise the entire classification path. In addition, Hierarchical classification and interpretability, its real-world effectiveness is still under debate. Ongoing work is needed to overcome error propagation, optimize feature alignment, and validate empirical gains in diverse, safety-critical settings.

In this paper, we conduct a systematic empirical evaluation of tree-based reasoning in VLMs, comparing hierarchical decision-making against standard zero-shot image classification across multiple models, datasets, prompts and settings. Our investigation reveals a surprising finding: Contrary to theoretical expectations, tree-based reasoning consistently underperforms direct zero-shot classification, often by substantial margins. Through detailed error analysis, we identify the key factors contributing to this performance degradation, including error propagation through hierarchical structures, sensitivity to question formulation, and depth-related accuracy penalties.

Our contributions are:

1. We provide a comprehensive empirical assessment of tree-based reasoning in multiple VLMs across two classification tasks under diverse settings, which ensures the generalization of the experiments.
2. We identify and analyze the specific failure modes

that limit the effectiveness of hierarchical reasoning approaches.

3. We offer design recommendations for future structured reasoning systems based on our findings.

These results have important implications for the development of interpretable VLM systems and highlight the drawbacks of the current usage of hierarchical reasoning methods.

## Related Work

**Vision-Language Models and Zero-Shot Classification.** VLMs such as CLIP (Radford et al. 2021), LLaVA (Liu et al. 2023a), Qwen-VL (Bai et al. 2023), and GPT-4V (OpenAI 2023) have driven substantial progress in multimodal understanding and zero-shot classification (Li et al. 2025; Cao et al. 2025). These models demonstrate impressive generalization to novel visual categories without task-specific training. However, their performance degrades on fine-grained tasks involving many semantically similar classes (Zhang et al. 2024b; Kim and Ji 2024; Wei 2024). Even with advances in prompt learning (Jha 2024) and model adaptation (Cho, Kim, and Kim 2023; Schwonberg et al. 2024), challenges persist under dense class distributions, domain shifts, and nuanced language constructs like negation or ambiguity (Alhamoud et al. 2025; Anis, Ali, and Sarfraz 2025; Gou et al. 2024). These limitations motivate structured strategies to enhance model robustness and interpretability.

**Multi-label and Hierarchical Classification.** A promising direction for improving visual classification involves structuring label spaces hierarchically to better guide visual decision-making and support the understanding of Compositional Language Concepts (CLC) in images. Tree-based and taxonomy-aware approaches break down complex objects into their attributes, states, and relationships with other objects in the scene, enhancing both prediction accuracy and interpretability in complex visual settings (Yellinek, Karlinsky, and Giryes 2025; Ding et al. 2025). Structured label hierarchies allow models to reason step by step across different levels of abstraction (Zhang et al. 2024a). Despite this potential, hierarchical method remain relatively underexplored in the context of image classification. Complementary strategies enrich label semantics by incorporating natural language descriptions generated by large language models (LLMs), which have been shown to improve zero-shot generalization in fine-grained classification tasks (Saha, Horn, and Maji 2024). Nevertheless, VLM continue to face challenges in distinguishing subtly different or compositional classes, particularly in real-world scenarios involving occlusion, ambiguity, or overlapping class boundaries (Xu et al. 2024).

**Structured Reasoning and Prompt Engineering** To address these shortcomings, recent work has explored integrating structured reasoning and prompt engineering into VLM pipelines. Hierarchical or tree-structured reasoning provides a framework for decomposing classification into interpretable decision paths (Besta et al. 2025; Yellinek, Karlinsky, and Giryes 2025; Wang et al. 2024). Prompt-based techniques, such as using category-specific templates or multi-step inference chains, help align model predictions with human-understandable logic (He et al. 2024). In particular, prompt tuning enables VLMs to better navigate complex visual scenes, improving resilience to occlusion, perspective shifts, and context ambiguity (Yellinek, Karlinsky, and Giryes 2025). Recent studies further demonstrate that combining image-text reasoning with structured label spaces (e.g., via intermediate textual descriptions) can improve the accuracy over direct label prediction pipelines. These approaches support the development of debuggable, interpretable systems that maintain hierarchical consistency and allow for targeted analysis of model errors.

## Problem Statement

The deployment of VLMs in real-world applications has led to increased interest in structured reasoning approaches that theoretically promise both improved interpretability and potentially enhanced performance. However, the practical effectiveness of these approaches compared to standard zero-shot classification methods remains an open empirical question.

**Challenge 1: Tree-based vs. Zero-shot Classification.** While VLMs have demonstrated impressive zero-shot capabilities, there is ongoing debate about whether incorporating explicit structural guidance through hierarchical reasoning (such as decision trees) provides additional benefits. Tree-based approaches offer theoretical advantages including interpretability, systematic error analysis, and the ability to decompose complex decisions into simpler choices. However, they also introduce potential complications including error propagation through hierarchical structures, increased inference complexity, and sensitivity to tree design choices.

**Challenge 2: Performance-Interpretability Trade-offs.** The assumption that structured reasoning approaches necessarily improve performance while providing interpretability benefits has not been thoroughly validated empirically. It remains unclear whether the additional complexity introduced by hierarchical structures compensates for any potential accuracy gains, particularly given the strong performance of direct zero-shot methods.

**Challenge 3: Validation of Structured Knowledge.** Even when VLMs demonstrate understanding of the hierarchical knowledge embedded in decision trees, it is unclear whether this understanding translates into improved classification performance. The relationship between knowledge verification accuracy and actual classification effectiveness requires empirical investigation.

**Research Questions.** Our investigation centers on the following empirical questions:

1. **Performance Comparison:** How does tree-based hierarchical reasoning compare to standard zero-shot prompting in terms of classification accuracy across fine-grained and coarse-grained datasets?

2. **Knowledge Translation:** Is VLMs accurate in reasoning and answering visual questions about the datasets' classes? If yes, Does this reasoning or knowledge trans-

late to improved classification performance in hierarchical reasoning systems?

3. **Error Analysis:** What are the primary failure modes of tree-based reasoning systems compared to direct zero-shot approaches?
4. **Robustness Assessment:** How do tree-based and zero-shot approaches compare in terms of consistency across different types of classification tasks?

# Methodology

To evaluate the tree-based inference approach, we construct decision trees for image classification across two real-world tasks. We experiment incorporating the history of previously traversed questions and answers into the tree model's prompt. We further augment the model input with class descriptions and image visual description generated by an LLM to provide richer semantic grounding. To assess the quality and interpretability of the generated decision trees, we conduct a verification study examining the VLM's understanding of tree structure and branching logic. Figure 1 presents an overview of our experimental framework, including zero-shot classification, hierarchical reasoning, and the construction and use of decision trees for step-by-step visual classification.

## Datasets

We evaluate the approach on two benchmark datasets, each representing a real-world image classification task. These datasets were chosen to balance task complexity with practical relevance:

- **German Traffic Sign Recognition Benchmark (GTSRB)** (Stallkamp et al. 2012):
  - 43 traffic sign classes
  - Over 50,000 images of varying sizes
  - Sequential captures (multiple frames per sign)

  This dataset introduces a fine-grained image classification task, requiring models to distinguish between visually similar German traffic signs under real-world, noisy conditions. Its relevance to safety-critical domains, such as autonomous driving, makes it a valuable testbed for evaluating hierarchical reasoning. To reduce redundancy, we sample **901 representative images** by selecting one random frame per sequence from the training set, including all the sequences inside the dataset.
- **CIFAR-10** (Krizhevsky and Hinton 2009):
  - 10 object classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck)
  - **1,000 balanced samples** (100 per class)

  This dataset presents a coarse-grained classification task involving general object categories commonly encountered in daily life. It is designed to assess visual generalization capabilities across diverse semantic classes and tests the model's ability to differentiate between broad, well-known object types.

## Models

We conduct the evaluation experiments on three state-of-the-art VLMs, chosen to represent a range of architectural paradigms, size and capability levels. All models were used via API calls:

- **GPT-4o**
  *Rationale*: As OpenAI's most advanced multimodal model, GPT-4o provides a strong baseline for cutting-edge zero-shot visual reasoning. Its architecture excels at contextual understanding, making it ideal for testing hierarchical (tree-based) versus flat prompting strategies.
  *Role in Study*: Serves as our primary model for cross-dataset comparison due to its robust performance.
- **LLaMA-3.2 11B Vision Instruct**
  *Rationale*: This open-weight model balances scale and efficiency. Its decoder-only design contrasts with GPT-4o's hybrid architecture, testing generalization across model families.
  *Role in Study*: Isolates the impact of model family (decoder-only vs. hybrid) and the model size variation.
- **Qwen-VL MAX**
  *Rationale*: A leading multilingual VLM with strong visual grounding capabilities, particularly in Asian contexts. Its performance on traffic signs (often designed with Unicode-like symbols) tests cultural and domain biases.
  *Role in Study*: Clarify dataset-specific biases (e.g., GTSRB's Western Unicode-like symbols) from model-specific limitations by cross-referencing performance against GPT-4o.

## Implementation Details

Our methodology started with the construction of a hierarchical multi-decision tree tailored to each dataset to guide the classification process via a sequence of interpretable questions. The tree is traversed using depth-first search, and each leaf node maps to one final class label (e.g., ``20 kph speed limit'' or ``airplane''). The models' performance was rigorously evaluated by comparison with zero-shot prompting baselines corresponding to each model, as illustrated in Figure 1.

## Tree Structure

The initial version of the tree is automatically generated using an LLM (We used both Deepseek and GPT-4.1), guided by class metadata and linguistic descriptions. The model is prompted to produce the tree in text format, with constraints ensuring no repeated questions along a single path and exactly one class at each leaf node. A human annotator then reviews, traverses, and refines the generated tree to ensure semantic accuracy and consistency.

A partial subtree [1] for both datasets used in the experiments is illustrated below:

```
GTSRB Tree:
(Max Depth: 16, Number of nodes: 65)
```

[1] The complete tree structure is available in the provided project code.

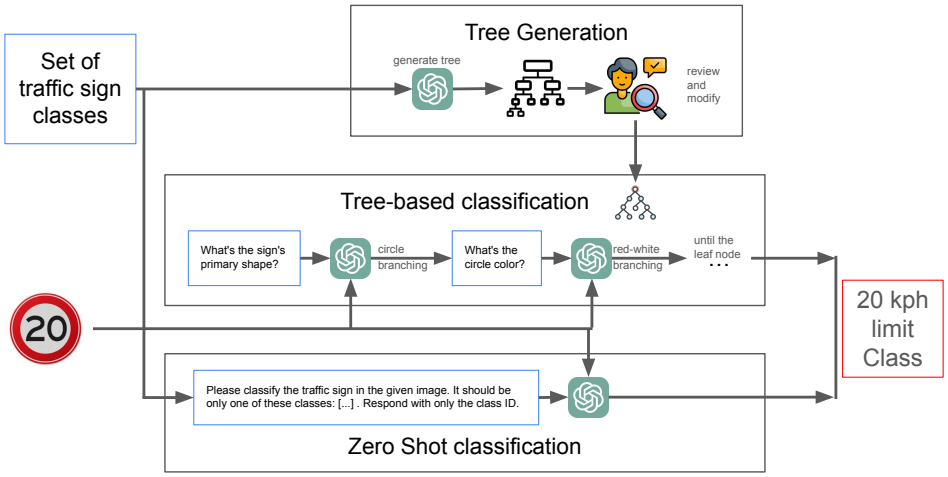


Figure 1: Experimental framework setup illustrating both zero-shot classification and hierarchical tree-based reasoning. It also shows the construction and use of decision trees for step-by-step visual classification.

```
[L0] Q: What's the sign's primary shape?
  -> triangle:
    [L1] Q: Does the triangle have an
        exclamation mark?
      -> yes: Exclamation mark warning
          ([18])
      -> no:
        [L2] Q: Does it depict a left curve?
          -> yes: Left curve warning ([19])
          -> no: (Continue traversal...)

CIFAR-10 Tree
(Max Depth = 5, Number of nodes = 19)
[L2] Q: Is the animal typically shown with
    prominent hooves?
  -> Yes (hooves):
    [L3] Q: Does the animal have antlers
        visible in the image?
      -> Yes (antlers): [L4] Leaf Node: deer
          (ID: 4)
      -> No (no antlers): [L4] Leaf Node:
          horse (ID: 7)
  -> No (paws): (Continue traversal...)
```

## Decision Tree Knowledge Verification

Before experimenting with tree-based classification, we first assessed two key components: the visual understanding capabilities of the strongest baseline VLM (GPT-4o) and the semantic coherence of our decision tree structure. To do this, we experimented using the decision tree built for the GTSRB dataset . In this setup, the model is given the ground truth class label and prompted to sequentially answer each question along the corresponding path in the tree that leads to this ground truth class. This allows us to evaluate both the model's grasp of class-specific visual and semantic attributes, and the overall quality and structure of the decision tree, specifically, whether its questions are meaningful and well-organized for hierarchical classification.

For example, for the class "20 kph speed limit", the model interaction proceeds as follows:

```
"20 kph speed limit": [
    {"[L0] Q: What's the sign's primary
        shape?": "circle"},
    {"[L1] Q: What's the circle color?": "
        red-white"},
    {"[L2] Q: Does it contain numbers?": "
        yes"},
    {"[L3] Q: What is the number?": "20"}
]
```

The model is given the class name up front and retains access to the full question-answer history to simulate a coherent reasoning process. We measure per-class accuracy as the number of correctly answered questions divided by the total number of questions along the path.

## Prompting Strategies

We tested two main prompting strategies for visual classification across both datasets and all selected three VLMs, as illustrated below:

- **Baseline (Zero-shot Classification)**:
  This is a standard prompting strategy that directly asks the model to classify the image by selecting a single class from the full classes set. The model receives the image as input along with the following prompt:

  ```
  "Please classify the [GTSRB: traffic sign
      | CIFAR-10: object] in the given
      image. It should be only one of these
      classes: {class_ids_and_names}.
      Respond with only the class ID."
  ```

  We also extend this method by providing Deepseek-generated textual descriptions for each class and a generated image caption by the tested model within the prompt context. We refer to this variation as **Zero-shot with Description Classification**.
- **Tree-Based Classification**:
  This is the primary strategy evaluated in our study. It in-

volves traversing a decision tree from the root to a leaf node, where the classification decision is made. At each node, the prompt includes the node-specific question and a set of possible branching answers. The model is expected to choose one answer based on the image, which determines the next node to visit, continuing this process until reaching the final classification.

The prompt template used at each node is as follows:

```
"{tree.question} Choose one of these
    answers: {tree.answers}."
```

An example from the GTSRB dataset tree:

- **Question:** What's the sign's primary shape?
- **Answers:** ["triangle", "circle", "diamond", "inverted-triangle", "octagon"]

Furthermore, we extended the tree-based strategy by introducing two additional variations:

- **Tree-Based Classification with History**:
  In this variation, we append the sequence of prior decisions made by the model to each prompt. This provides additional context about the path taken so far.
- **Tree-Based Classification with Descriptions**:
  Similar to the Zero-shot with Description strategy, we include LLM-generated textual descriptions of the classes and the target image caption as part of the prompt context. This information is provided before the model selects a branching answer, offering semantic guidance during classification.

### Answer Extraction

Due to the inherently non-deterministic nature of VLM responses, we cannot fully control their output format. To reliably extract answers, we employ regular expression pattern matching, which captures the first occurrence of any predefined answer candidate from the model's response.

## Evaluation and Results

A comprehensive evaluation of the decision tree-based methodology is conducted and benchmarked against a zero-shot image classification baseline. Further details are provided in the following sections.

### Decision Tree Knowledge Verification

We tested the GPT-4o model using the decision tree developed for the GTSRB dataset. The model achieved **100% accuracy** on all decision tree questions for **39 out of 43** traffic sign classes, with an overall average accuracy of **98.20%**. Crucially, the VLMs demonstrated comprehensive knowledge, correctly answering every tree branch question for each specific class. This validates that the curated decision tree's structure and question order are optimally suited for the task.

The four classes with slightly lower performance—*"End of restriction"*, *"Right-of-way at intersection"*, *"Double curve warning"*, and *"Ice/snow warning"*—exhibit high visual or semantic similarity to other signs, leading to occasional confusion. These results confirm the model's strong reasoning capabilities and the decision tree's alignment with the classification challenge. The minimal errors likely reflect ambiguities in visual design or contextual interpretation of certain signs rather than deficiencies in the tree itself.

### Zero-shot Classification

We began our evaluation with a baseline zero-shot classification experiment. Given the non-deterministic behavior of VLMs and their known sensitivity to prompt phrasing (Fatemi, Halcrow, and Perozzi 2023), we performed the classification using **10 diverse prompts** to assess robustness. The model was run with a **temperature of 0.7** to allow variability in responses. By incorporating prompt variation, we aimed to better understand the model's generalization and reduce performance artifacts caused by prompt-specific biases. Figure 2 illustrates the distribution of zero-shot classification accuracy across different prompts.

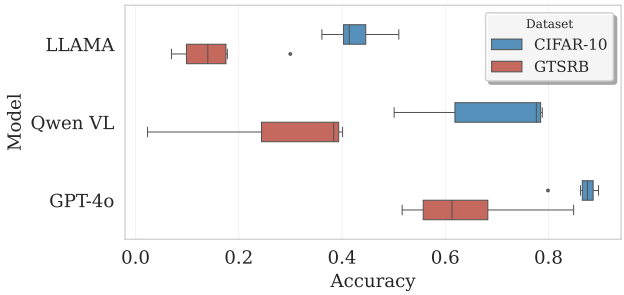


Figure 2: Distribution of zero-shot classification accuracy over 10 prompt variations, evaluated across all datasets.

We then selected a **representative prompt** with average performance (based on mean accuracy across experiments) and conducted three runs for each of two temperature settings: **0.7 and 0**. The results were consistent across experiments, showing little to no variance. Figure 3 presents the mean zero-shot accuracy across all classes.

Additionally, we ran a variant of the experiment that included **class descriptions** as part of the prompt. While this slightly improved performance for the LLaMA model on CIFAR-10, it did not benefit other experiments. For the GTSRB dataset, incorporating textual descriptions for the large number of classes increased context length **without meaningful accuracy gains**, suggesting that excessive prompt context may hinder performance in fine-grained classification tasks.

### Tree-Based Classification

We evaluated the selected models on the generated decision trees using two temperature settings: **0.7 and 0**. For each setting, we recorded the **mean accuracy** across all classes. Figure 3 presents the results for the tree-based classification approach.

- **GPT-4o** consistently achieved the highest performance across both datasets and temperature settings, reaching **52.05% accuracy** on GTSRB and **75.40%** on CIFAR-10 (temperature 0). These results highlight the increased

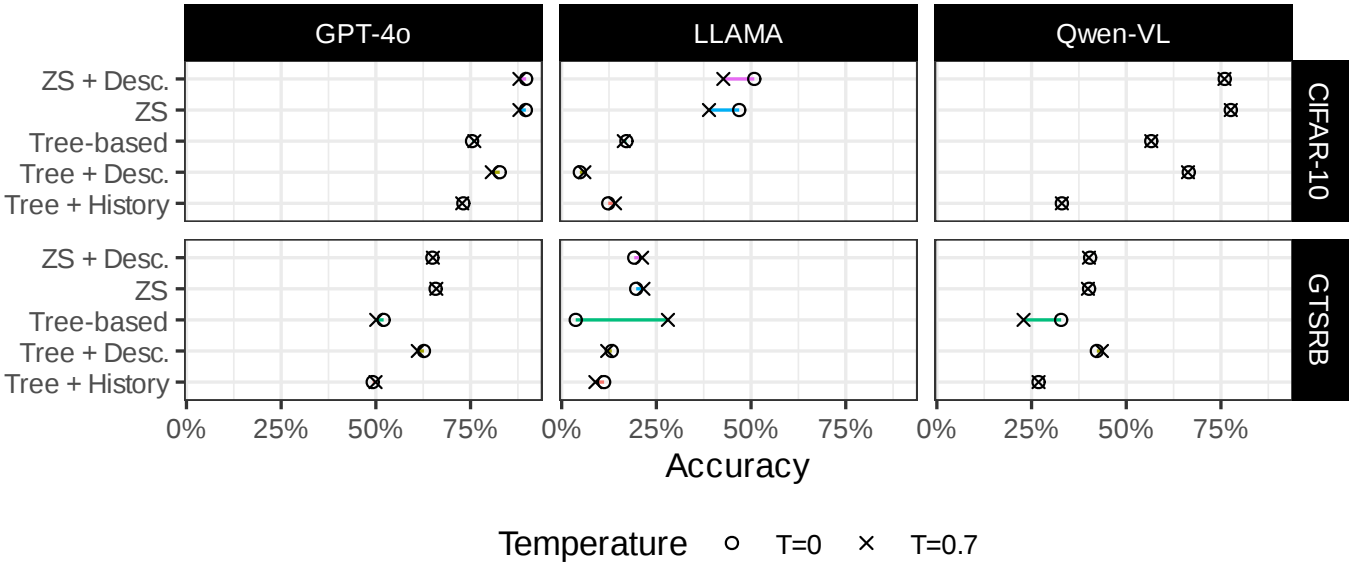


Figure 3: Mean accuracy (%) comparison across prompting strategies, models, datasets, and temperature settings.

complexity of the decision tree in fine-grained classification tasks like GTSRB.

- In contrast, the **LLaMA model** exhibited relatively poor and unstable performance, particularly with varying temperature settings, suggesting lower reliability in following structured reasoning prompts.

Furthermore, Figure 3 illustrates the impact of **two prompting variations** in the tree-based classification setting:

1. **Including full decision path history** negatively affected performance, likely due to context overload with non-essential information.
2. **Adding class descriptions** generally improved performance across datasets and models—except for LLaMA, where it caused a performance drop.

These results suggest that **descriptive context enhances understanding in larger VLMs** capable of effectively leveraging long-context inputs.

### Accuracy Comparison

Contrary to our expectations, the experimental results indicate that incorporating tree-based reasoning structures does not improve performance for any of the evaluated VLMs on either the fine-grained (GTSRB) or coarse-grained (CIFAR-10) classification tasks when compared to the zero-shot baseline, as shown in Figure 3.

For the GTSRB dataset, we observe a substantial performance gap between tree-based and zero-shot classification. The best-performing model, GPT-4o, achieved 65.78% accuracy in the zero-shot setting (temperature 0) compared to just 52.05% with tree-based reasoning—a difference of over 13 percentage points. While including class descriptions and image captions helped narrow this gap, it was insufficient to bridge it entirely. Similarly, the Qwen-VL model showed noticeable zero-shot superiority, with 40.18% accuracy (temperature 0) versus 32.18% in the tree-based setup. These consistent gaps suggest that the models may not effectively leverage structured reasoning and instead rely more on dataset-specific biases than on decomposing and interpreting traffic signs' visual attributes.

Further per-class analysis revealed that GPT-4o (temperature 0 in tree-based configuration) outperformed zero-shot classification in only 11 of 43 classes—a ratio that remained consistent across all experimental settings. This reinforces the limited effectiveness of tree-based classification for most GTSRB categories.

The same trend persists in the coarse-grained CIFAR-10 classification (Figure 3). While overall accuracy increased—with GPT-4o reaching 75.40% in the tree-based setup compared to its GTSRB performance—a performance gap of up to 14 percentage points remained relative to the zero-shot baseline. Notably, per-class analysis shows the zero-shot approach outperformed tree-based classification across all CIFAR-10 classes.

These findings suggest that the additional complexity of hierarchical reasoning structures does not enhance VLM performance in vision-related classification tasks. Instead, simpler zero-shot baseline proves more effective for both fine-grained (GTSRB) and coarse-grained (CIFAR-10) settings. We explore potential explanations for these results in the Discussion section.

## Discussion

Our experimental results demonstrate consistent performance degradation when employing tree-based reasoning compared to the zero-shot baseline, revealing several critical contributing factors. The tree-based approach exhibits significant sensitivity to errors in upper-level decision nodes, where top-level fragility proves particularly problematic.

In the GTSRB dataset, preliminary questions such as

"Does it contain a number?" disproportionately affect downstream classification when answered incorrectly, while alternative question orderings yield accuracy variations that indicate strong path dependence in the reasoning process. Furthermore, structural deficiencies in tree design introduce substantial limitations, including question hierarchy flaws where parent nodes improperly subsume children (e.g., Layer 10: "Pedestrian?" versus Layer 11: "Pedestrian with child?"), resulting in systematic misclassification of all "pedestrian with child" images as generic pedestrians.

The reasoning process also struggles with contextual integrity, as including conversation history degrades performance by introducing irrelevant information that dilutes classification-relevant features. This finding aligns with (Liu et al. 2023b)'s demonstration that performance drops when input exceeds optimal context length or when critical information is suboptimally positioned.

The least accurate layers predominantly involve higher-level reasoning, where questions pertain to the semantic interpretation of traffic signs, such as their class or function, rather than low-level visual features like shapes or colors. This pattern suggests that the models perform reasonably well on basic perceptual tasks but demonstrate significant weaknesses when reasoning about abstract or contextual aspects, highlighting a fundamental limitation in current vision-language models regarding hierarchical reasoning capabilities.

Additionally, we observe that complex semantic queries (e.g., "What action is prohibited?") consistently underperform simple visual questions (e.g., "Triangular shape?"), further supporting the notion that current models excel at atomic visual recognition but struggle with semantic interpretation tasks requiring deep traversal of hierarchical structures. These findings suggest several architectural improvements for future implementations: breadth-first prioritization where shallow, wide trees outperform deep ones, and visual primitives first approaches where atomic visual queries concerning color and shape should precede semantic interpretation to maximize classification accuracy and minimize error propagation through the reasoning hierarchy.

## Conclusion

Our study demonstrates that tree-based reasoning structures do not improve the performance of VLMs on vision-related classification tasks, as evidenced by experiments on the GTSRB and CIFAR-10 datasets. Instead, we observe a consistent degradation in accuracy compared to zero-shot prompting across all evaluated models. Zero-shot prompting significantly outperforms tree-based reasoning in both fine-grained (GTSRB) and coarse-grained (CIFAR-10) settings, achieving higher mean accuracy and correctly classifying samples that the hierarchical approach misclassifies. These findings suggest that the added complexity of tree-based reasoning does not necessarily enhance VLM performance in such tasks and may even introduce inefficiencies.

Furthermore, our experiments reveal that the tree-based methodology is highly sensitive to structural and question-formatting adjustments, indicating that minor changes in the reasoning path or question phrasing can significantly impact performance. Critically, the knowledge verification confirms VLMs possess a strong understanding of elementary visual attributes (achieving 90.7% perfect accuracy across GTSRB classes), but their inability to reliably use this decomposed knowledge within the reasoning tree to arrive at the correct final classification provides a key explanation for the observed performance degradation. This failure to integrate elementary visual understanding through structured decomposition suggests limitations in the compositional reasoning capabilities of current VLMs for such classification tasks, rather than a simple inefficiency in the method.

### Future Work

To enhance the effectiveness of tree-based reasoning in vision-language tasks, several promising directions warrant further exploration. One approach involves the investigation of hybrid reasoning strategies that integrate zero-shot predictions into the tree-based reasoning pipeline. For example, incorporating the zero-shot label as an additional feature in the decision tree could help reduce misclassifications and capitalize on the complementary strengths of both methods.

In addition, future work could focus on developing automated or semi-automated techniques for generating optimal decision trees. Methods such as reinforcement learning or iterative refinement may prove effective in constructing trees that are both accurate and robust, potentially reducing sensitivity to structural variations and enhancing generalizability. It is also important to explore whether these improvements hold across different multimodal tasks and datasets. Furthermore, alternative reasoning frameworks that strike a balance between interpretability and performance merit investigation. Addressing these challenges will contribute to a deeper understanding of when and how structured reasoning can amplify the capabilities of large language models in vision-language applications.

## References


Alhamoud, K.; Alshammari, S.; Tian, Y.; Li, G.; Torr, P.; Kim, Y.; and Ghassemi, M. 2025. Vision-Language Models Do Not Understand Negation. arXiv:2501.09425.

Anis, A. M.; Ali, H.; and Sarfraz, S. 2025. On the Limitations of Vision-Language Models in Understanding Image Transforms. arXiv:2503.09837.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Besta, M.; Memedi, F.; Zhang, Z.; Gerstenberger, R.; Piao, G.; Blach, N.; Nyczyk, P.; Copik, M.; Kwaśniewski, G.; Müller, J.; Gianinazzi, L.; Kubicek, A.; Niewiadomski, H.; O'Mahony, A.; Mutlu, O.; and Hoefler, T. 2025. Demystifying Chains, Trees, and Graphs of Thoughts. arXiv:2401.14295.

Cao, Q.; Chen, Y.; Lu, L.; Sun, H.; Zeng, Z.; Yang, X.; and Zhang, D. 2025. Generalized domain prompt learning for accessible scientific vision-language models. *Nexus*, 2(2): 100069.

Cho, E.; Kim, J.; and Kim, H. J. 2023. Distribution-Aware Prompt Tuning for Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22004–22013.

Ding, T.; Li, W.; Miao, Z.; and Pfister, H. 2025. Tree of Attributes Prompt Learning for Vision-Language Models. arXiv:2410.11201.

Fatemi, B.; Halcrow, J.; and Perozzi, B. 2023. Talk like a Graph: Encoding Graphs for Large Language Models. arXiv:2310.04560.

Gou, C.; Felemban, A.; Khan, F. F.; Zhu, D.; Cai, J.; Rezatofighi, H.; and Elhoseiny, M. 2024. How Well Can Vision Language Models See Image Details? arXiv:2408.03940.

He, M.; Shen, Y.; Zhang, W.; Tan, Z.; and Lu, W. 2024. Advancing Process Verification for Large Language Models via Tree-Based Preference Learning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2086–2099. Miami, Florida, USA: Association for Computational Linguistics.

Jha, A. 2024. In the Era of Prompt Learning with Vision-Language Models. arXiv:2411.04892.

Kim, J.; and Ji, H. 2024. Finer: Investigating and Enhancing Fine-Grained Visual Concept Recognition in Large Vision Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6187–6207. Miami, Florida, USA: Association for Computational Linguistics.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Li, Z.; Wu, X.; Du, H.; Liu, F.; Nghiem, H.; and Shi, G. 2025. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. arXiv:2501.02189.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. arXiv:2304.08485.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023b. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Saha, O.; Horn, G. V.; and Maji, S. 2024. Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions. arXiv:2401.02460.

Schwonberg, M.; Werner, C.; Gottschalk, H.; and Meyer, C. 2024. A Study on Unsupervised Domain Adaptation for Semantic Segmentation in the Era of Vision-Language Models. *CoRR*, abs/2411.16407.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32: 323–332.

Wang, Z.; Yu, S.; Stengel-Eskin, E.; Yoon, J.; Cheng, F.; Bertasius, G.; and Bansal, M. 2024. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos.

Wei, C. 2024. Enhancing Fine-Grained Image Classifications via Cascaded Vision Language Models. arXiv:2405.11301.

Xu, Z.; Zhu, Y.; Deng, T.; Mittal, A.; Chen, Y.; Wang, M.; Favaro, P.; Tighe, J.; and Modolo, D. 2024. Benchmarking Zero-Shot Recognition with Vision-Language Models: Challenges on Granularity and Specificity. arXiv:2306.16048.

Yellinek, N.; Karlinsky, L.; and Giryes, R. 2025. 3VL: Using Trees to Improve Vision-Language Models' Interpretability. arXiv:2312.17345.

Zhang, R.; Zhang, B.; Li, Y.; Zhang, H.; Sun, Z.; Gan, Z.; Yang, Y.; Pang, R.; and Yang, Y. 2024a. Improve Vision Language Model Chain-of-thought Reasoning. arXiv:2410.16198.

Zhang, Y.; Unell, A.; Wang, X.; Ghosh, D.; Su, Y.; Schmidt, L.; and Yeung-Levy, S. 2024b. Why are Visually-Grounded Language Models Bad at Image Classification? arXiv:2405.18415.

## Reproducibility Checklist

---

**1. General Paper Structure**

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

**2. Theoretical Contributions**

2.1. Does this paper make theoretical contributions? (yes/no) yes

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) yes

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) yes

2.4. Proofs of all novel claims are included (yes/partial/no) yes

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) yes

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) yes

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) yes

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) yes

**3. Dataset Usage**

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

**4. Computational Experiments**

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) partial

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes