Department of Electrical Engineering,
National Tsing Hua University

Special Topic on Implementation
Research Report

Inductive Visual Logic for Few-Shot Out-Of-Distribution Adaptation in VLMs

以視覺邏輯歸納演算法實現視覺語言模型之少樣本分佈外適應

專題領域：資工領域

組別：B548

指導教授：孫民 副教授

組員姓名：何宇恆

研究期間：2025 年 3 月至 2025 年 11 月止，共 9 個月

Abstract

Modern vision-language models (VLMs) such as Qwen excel in zero-shot tasks and benefit from few-shot visual fine-tuning, where models learn from limited examples while adapting to domains that may be out of their pre-training distributions (OOD). Despite these advantages, when confronting niche, *distant* OOD datasets, traditional supervised few-shot fine-tuning would provide models no improvements, sometimes even deteriorates their performance. To address the limitation, we introduce *Inductive Visual Logic* (IVL), a training-free framework that extracts visual traits through a two-step prompting process, applies inductive-deductive reasoning and concludes with interpretable trait explanation. Through reasoning over traits rather than memorizing features, IVL not only improves few-shot performance but also produces more interpretable predictions, highlighting a complementary path toward OOD adaptation in foundation VLMs.

Introduction and Motivation

The deployment of vision-language models (VLMs) in specialized real-world applications is currently bottlenecked by performance failures in "distant out-of-distribution" (OOD) settings, where target concepts like rare diseases or industrial defects are absent from pre-training data. Standard adaptation methods fall short in these scenarios: gradient-based fine-tuning often leads to catastrophic overfitting on limited data, while In-Context Learning (ICL) fails to generate visual primitives missing from the model's internal representations. To overcome these limitations, we formally characterize the distant-OOD problem and propose **Inductive Visual Logic (IVL)**, a cognitive-inspired framework that splits adaptation into two interpretable phases. By mimicking human learning, IVL employs an *Inductive Stage* to generalize visual "traits" from examples and a *Deductive Stage* to apply these explicit rules for classification, enabling robust adaptation without the need for parameter updates.

Methodology

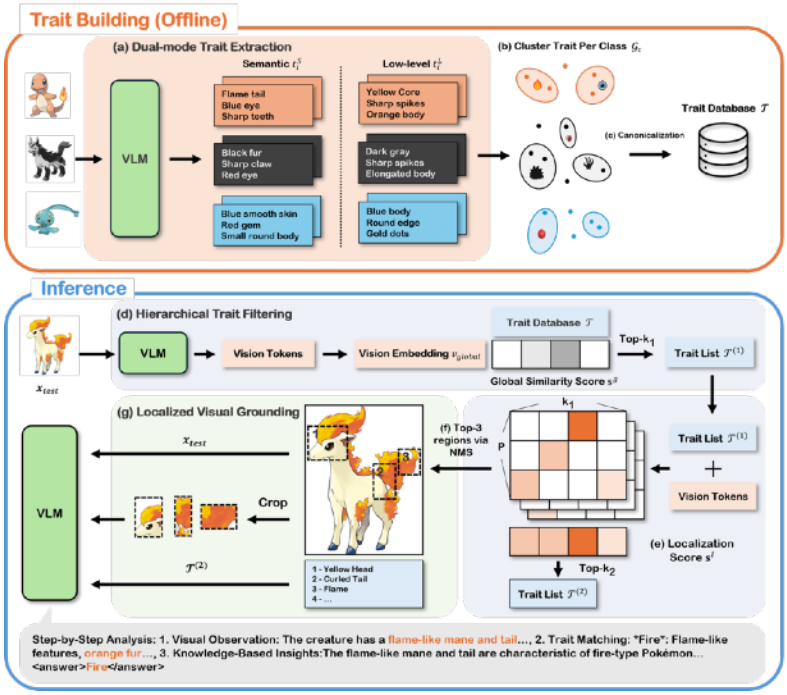

Figure 1. IVL Framework Overview

To complement the existing weakness of VLMs, we proposed **Inductive Visual Logic (IVL)** framework, as shown in *Figure 1*. IVL addresses the distant-OOD challenge by splitting the learning process into two interpretable stages. Instead of updating weights, we explicitly construct a knowledge base. The framework operates first through an **Inductive Stage**, where discriminative visual patterns are extracted and organized into a database. This is followed by a **Deductive Stage** (inference), where specific traits are retrieved and verified against test images to render a decision.

2.1 Inductive Stage (Offline Trait Building)

This stage transforms raw few-shot examples into a structured "dictionary" of visual traits.

- **Dual-Mode Extraction:** We prompt the VLM in two ways. *Semantic prompting* captures high-level concepts (useful for known domains), while *Low-level prompting* forces the model to describe primitive features like colors, shapes, and textures (crucial for novel domains where semantic vocabulary fails).

- **Clustering & Canonicalization:** Raw descriptions are noisy, hence we extract multiple traits per image and use Sentence Transformers to cluster them. Semantically similar traits are merged into canonical descriptors (e.g., merging "crimson legs" and "reddish limbs" into "red limbs").

- **Filtering:** We apply quality controls, filtering out irrelevant semantic hallucinations while preserving low-level visual observations, resulting in a comprehensive Trait Database.

2.2 Deductive Stage (Inferencing)

During inference, the model applies the learned traits to new data through a hierarchical process:

- **Global Retrieval:** We compute the cosine similarity between the test image's global embedding and the trait database, selecting the top-k most relevant traits.

- **Localized Visual Grounding:** To prevent false positives, we verify traits by calculating attention scores on the test image. We check if specific image patches actually align with the textual traits.

- **Reasoning & Classification:** VLM is presented with the original image, the verified trait list, and cropped regions of interest. It uses this multi-modal context to "deduce" the final class, grounding its answer in the explicit visual evidence it has gathered.

Experiments

3.1 Setup

We utilized Qwen2.5-VL-7B as the backbone for all experiments to ensure a fair comparison. We benchmarked IVL against Zero-shot, Supervised Fine-Tuning (SFT), LoRA, Visual-RFT, and In-Context Learning (ICL). Experiments focused on 1-shot and 8-shot scenarios to evaluate efficiency under extreme data constraints.

3.2 Results on Public OOD Dataset

We defined "distant OOD" quantitatively as datasets where standard SFT yields less than a 15% accuracy gain over zero-shot baselines. Based on this, we selected challenging datasets including **Retinal OCT** (medical), **WM811k** (wafer map defects), and **MVTec AD** (anomaly detection). As shown in *Figure 2.*, traditional fine-grained OOD datasets such as FGVC Aircraft, Flowers102 and Cars196 gain more than 20% in accuracy after SFT, whereas the accuracy gains after SFT on distant-OOD datasets of our choice are in general less than 10%, exhibiting limitations of traditional SFT method.

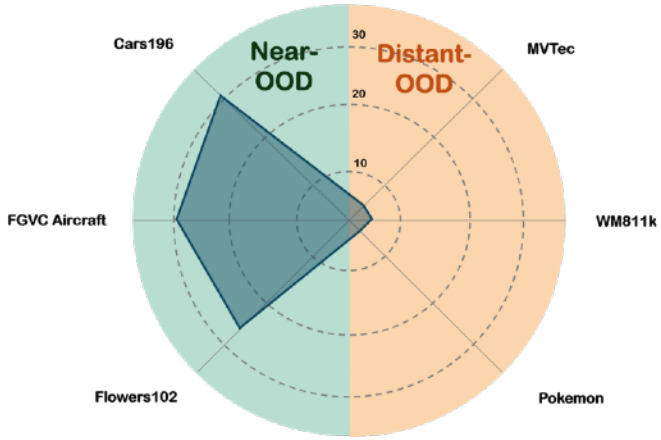

Figure 2. Accuracy Gain after SFT Differences between Near-OOD and Distant-OOD

Results: In these distant domains, standard methods collapsed. For example, on the Retinal OCT dataset (1-shot), standard SFT dropped to nearly 0% accuracy due to severe overfitting to spurious noise. In contrast, IVL demonstrated robust performance, effectively doubling the adaptation gains compared to gradient-based methods. On MVTec AD, IVL outperformed zero-shot baselines by nearly 6% in 1-shot settings, proving that explicit trait reasoning prevents the overfitting common in parametric tuning. Detailed experiment results are recorded in *Table 1*.

*Due to the limited number of defect samples in MVTec AD dataset, only 8-shot training wasn't conducted.

3.3 Pokemon Dataset

To rigorously investigate adaptation dynamics under controlled conditions, we constructed a synthetic "distant-OOD" benchmark centered on Pokémon type classification. The objective of this task is to predict a Pokémon creature's elemental classification (e.g., Fire, Water, Grass) relying exclusively on observable visual attributes. This benchmark was designed to provide three distinct methodological advantages for systematic evaluation:

1. **Graduated Reasoning Complexity:** The dataset establishes a spectrum of visual-semantic mappings, ranging from elementary associations—such as red and orange coloration indicating "Fire" types—to complex morphological patterns required to distinguish between similar categories like "Psychic" and "Fairy". This variation enables us to decompose performance across different levels of reasoning difficulty.

2. **Human Baseline Comparison:** The task effectively highlights the gap between human cognitive efficiency and VLM capabilities. While human children can rapidly internalize these visual-type associations through brief exposure, current VLMs struggle to adapt, serving as a critical test for few-shot learning potential.

3. **Diagnostic Reliability:** Unlike real-world datasets where salient features can be ambiguous, this controlled environment provides known ground-truth discriminative features. This allows for precise verification of whether adaptation methods are genuinely learning the intended visual logic or merely overfitting to spurious correlations.

To benchmark model performance against human cognitive capabilities, we established human performance baselines through a stratified questionnaire study. Participants were categorized into three expertise levels, each serving as an analog to a specific stage of model training:

- **Beginners**: Individuals familiar only with iconic entities (e.g., Pikachu), representing the baseline capabilities of the pre-trained Qwen2.5-VL model.

- **Intermediates**: Participants possessing implicit knowledge of visual-type associations (such as color schemes and morphological patterns), serving as the human equivalent to few-shot adaptation.

- **Experts**: Individuals with extensive domain knowledge derived from completing Pokémon games, analogous to fully trained models.

To ensure balanced evaluation, the dataset was partitioned into three questionnaire versions, each containing approximately 341–342 samples with a controlled distribution of well-known Pokémon. Data collection was conducted over a one-week period, yielding 47 total responses (16 beginners, 16 intermediates, and 15 experts).

We also developed specific accuracy criteria to address the prevalence of dual-type creatures, which constitute approximately half of the dataset. For single-type Pokémon, predictions required an exact match to be considered correct; for dual-type Pokémon, the identification of either valid type was accepted as a correct prediction

Results: As shown in *Figure 3*., human "beginners" quickly achieved high accuracy while standard VLMs struggled significantly. Fine-tuning methods like LoRA and SFT showed minimal improvement, often failing to capture the underlying visual logic. IVL, however, bridged this gap. By explicitly extracting traits (e.g., "flame-like tail" → Fire type), IVL

achieved 56.3% accuracy in 8-shot settings, significantly outperforming the 49% baseline. This confirms that the performance gap is representational, not just optimization-based. Detailed experiment results are recorded in *Table 1*.

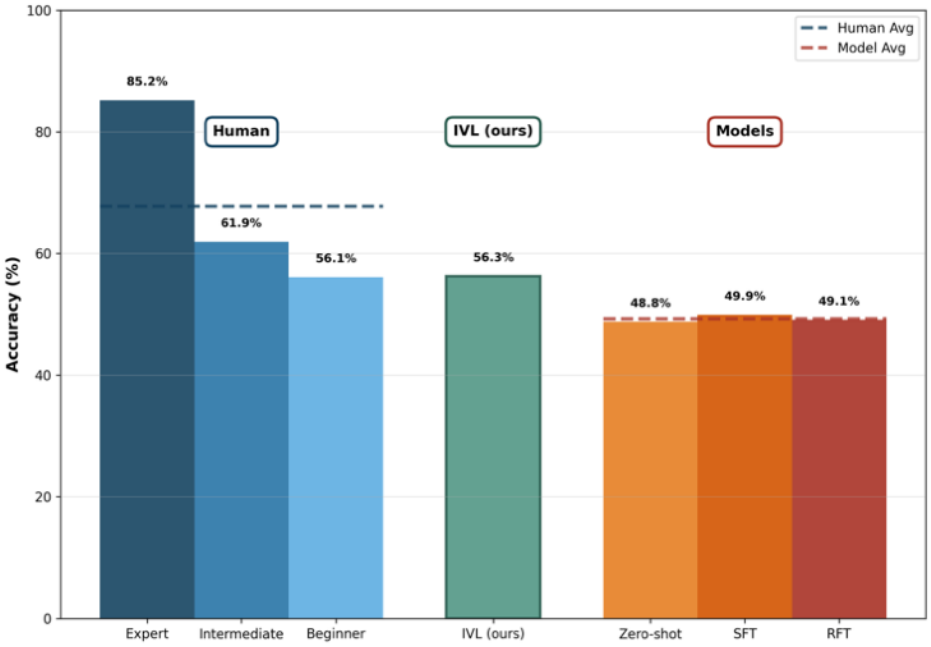

Figure 3. Human vs. Model Performance on Pokémon Dataset

Table 1. Experimental Results: Comparing n-shot performance by methods across 4 distant-OOD datasets

Model	Method	n-shot	Pokemon	Retinal OCT	WM811k	MVTec AD*
Qwen2.5-VL	None	0	48.81	18.75	11.45	34.88
	SFT	1	45.52	0.00	12.41	39.46
		8	49.94	**33.39**	15.90	-
	SFT + LoRA	1	45.59	13.69	9.16	37.16
		8	50.09	14.00	9.16	-
	RFT	1	48.77	14.00	9.52	36.71
		8	49.09	13.14	8.92	-
	ICL	1	22.00	13.75	12.41	24.45
		8	34.24	19.54	14.22	-
	IVL (Ours)	1	**51.00**	**20.83**	**13.01**	**40.69**
		8	**56.30**	23.21	**18.43**	-
ViT-B/32	CLIP	0	42.25	12.29	13.49	39.12

Conclusion

We introduced Inductive Visual Logic (IVL), a training-free framework designed to complement the weakness of modern VLMs on distant out-of-distribution tasks. Our analysis revealed that standard adaptation methods fail because they attempt to tune parameters for visual primitives that simply do not exist in the pre-trained model. IVL circumvents this by explicitly building a visual vocabulary through inductive reasoning and applying it via deductive verification. Our results across medical, industrial, and synthetic domains demonstrate that operationalizing human-like trait reasoning is a scalable and effective path toward robust adaptation in foundation models.

Contribution and Thoughts

I am honored to have conducted research under the guidance of the VS-ELSA Lab and to have contributed to the ongoing work of our doctoral candidate. My primary contributions to this project were twofold:

First, I executed the empirical evaluation of baseline frameworks to identify suitable distant Out-of-Distribution (OOD) datasets. Through this rigorous screening process, I gained significant expertise in dataset management and model monitoring. Furthermore, I developed custom automation pipelines to optimize experimental throughput, allowing us to maximize efficiency despite limited computational resources.

Second, I spearheaded the curation of the synthetic Pokémon benchmark and the design of the associated human baseline study. This required a full-stack approach to data acquisition: I engineered web crawlers to gather high-quality visual and attribute data from online databases and developed a custom web interface to facilitate the collection and analysis of human evaluation metrics.

Beyond these technical competencies, I have developed a deeper understanding of the research mindset. I learned the importance of strategic decision-making: recognizing when to decisively discard a flawed dataset versus when to persist with a novel framework. I realized that even when a proposed method does not immediately yield the intended results, it should not be abandoned prematurely. Instead, the true essence of research often lies in analyzing those setbacks to find the "spark"—the underlying narrative or theoretical insight that drives the scientific story forward. This experience has not only equipped me with the tools to conduct experiments but has also instilled in me the resilience and curiosity required for academic discovery.