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

Hung-Jen Chen[1], Yu-Heng Ho[⋆1], Ting-Yao Huang[⋆1], Po-Hsiang Hsu[1], Li-Yu Chen[1], Chun-Yi Lee[2], and Min Sun[1]

[1] National Tsing Hua University, Hsinchu, Taiwan
andyqmongo@gapp.nthu.edu.tw, sunmin@ee.nthu.edu.tw
[2] National Taiwan University, Taipei, Taiwan
cylee@csie.ntu.edu.tw

Abstract. Generative vision-language models (VLMs) such as Qwen-VL and LLaVA achieve strong zero-shot performance on tasks overlapping with their pretraining distribution, yet fail on specialized domains where the required discriminative features were never learned, a regime we term *distant* out-of-distribution (OOD). Standard adaptation methods cannot overcome this representational absence because they operate within the encoder's existing feature space. However, VLMs retain a robust *descriptive* capacity even when discrimination collapses: a model that cannot classify a medical scan can still articulate its visual patterns. Exploiting this asymmetry, we introduce **I**nductive **V**isual **L**ogic (IVL), a training-free framework that constructs classification knowledge from the model's surviving descriptive ability. IVL extracts visual traits from few-shot support images through dual-mode prompting, combining semantic descriptions with primitive visual observations, and organizes them into per-class trait dictionaries. At inference, hierarchical filtering identifies spatially grounded trait evidence for classification. Across multiple distant-OOD benchmarks, IVL achieves the highest aggregate accuracy under two VLM backbones while producing interpretable, trait-traceable predictions.

1 Introduction

Generative vision-language models (VLMs) such as Qwen-VL [1] and LLaVA [15] achieve remarkable zero-shot performance on tasks whose visual concepts overlap with web-scale pretraining data [7, 26]. This capability, however, conceals a fundamental limitation. When deployed to specialized domains whose discriminative visual primitives were never encountered during pretraining, these models face not a distributional shift but a *representational absence*, where the features required for classification simply do not exist within the learned representation space. Medical imaging, industrial inspection, and scientific analysis routinely fall into this regime, as their visual taxonomies bear little resemblance to web-crawled natural images. Indeed, the strong few-shot results on standard benchmarks largely reflect prior exposure to those task distributions during pretraining rather than a general capacity for rapid adaptation [13].

⋆ Equal contribution.

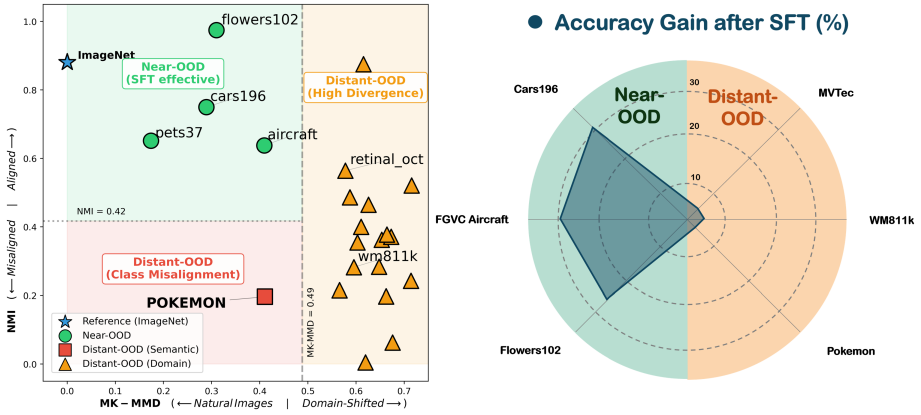

Fig. 1: Distant-OOD vs. near-OOD characterization. (a) Two-dimensional projection based on feature-space divergence (MK-MMD) and latent cluster-label alignment (NMI). Distant-OOD datasets fall outside the low-divergence, high-alignment quadrant where parametric methods succeed. (b) Radar chart of accuracy gains after SFT across datasets, showing that fine-tuning yields substantial improvements only in near-OOD settings but only marginal gains on specialized domains.

This *distant out-of-distribution (OOD)* regime defeats every major adaptation paradigm because all share a common assumption: discriminative visual representations already exist within the model and merely require refinement. Gradient-based methods, including SFT, LoRA [6], and Visual-RFT [16], optimize within the subspace spanned by the encoder's existing features; when the optimal decision boundary requires directions orthogonal to this subspace, gradient descent cannot easily recover the missing structure. This failure extends to RL-based adaptation: Although RL-based post-training generalizes better than SFT in standard settings [3, 16], Visual-RFT offers no advantage over SFT on distant-OOD tasks (Table 1), suggesting that exploration does not reliably discover features the encoder never learned. As Fig. 1 quantifies, SFT yields negligible gains on distant-OOD domains despite full parameter updates. Prompt-based [41, 42] and attribute-guided [20, 25, 33] approaches remain anchored to the pretrained semantic space; when that space encodes no knowledge of the target domain, generated descriptors correlate spuriously with categories [28]. In-context learning [21] reweights existing representations without structural modification, offering no recourse when the features are absent. Distant-OOD adaptation demands the *construction* of new visual knowledge, not refinement of existing representations.

A crucial observation motivates a departure from this paradigm. Although a VLM's discriminative representations collapse in distant-OOD domains, its *generative descriptive capacity* remains largely intact. A model that cannot classify a semiconductor wafer map can still describe "a ring-shaped concentration along the perimeter" and "scattered clusters near the center"; a model that fails to type Pokémon can nonetheless report coloration patterns, body shapes, and

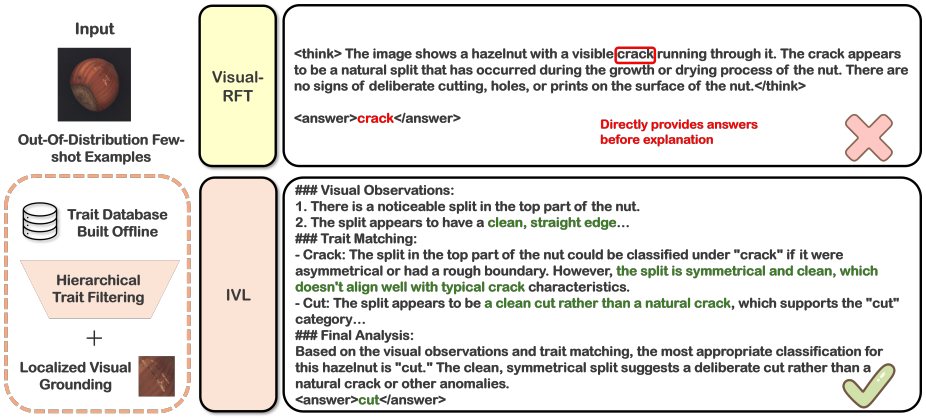

Fig. 2: Reasoning comparison between gradient-based methods and IVL. Visual-RFT produces answers before articulating reasoning, deviating from human cognitive patterns. IVL mirrors human inductive learning by first observing visual traits, comparing them against a structured traits dictionary, refining evidence with prior knowledge, and arriving at an evidence-grounded classification decision.

surface textures. This asymmetry arises because descriptive generation draws on the language decoder's domain-agnostic visual vocabulary, encompassing colors, textures, shapes, and spatial relations, which is reinforced across the full breadth of pretraining and is not tied to any classification task. Human visual learning exploits precisely this distinction: when encountering unfamiliar categories, humans engage in explicit inductive reasoning, systematically observing examples, identifying salient attributes, and constructing classification rules from these traits [11, 30]. As Fig. 2 illustrates, existing methods instead force models to produce answers from opaque internal representations, often committing to a classification before articulating any reasoning. Channeling VLMs' surviving descriptive ability into structured trait-level reasoning offers a principled path to bypass the representational bottleneck.

Building on this insight, we introduce **I**nductive **V**isual **L**ogic (IVL), a training-free framework that transforms VLMs' surviving descriptive capacity into explicit trait-based classification. IVL extracts discriminative traits from few-shot support images through dual-mode prompting, combining semantic descriptions with low-level observations that remain reliable even when semantic understanding collapses. Extracted traits are organized into per-class dictionaries, and a hierarchical filtering mechanism at inference identifies spatially grounded trait evidence for classification. Unlike attribute-based methods [20, 25, 33] that generate descriptors solely from pretrained semantic knowledge, IVL extracts traits from visual observations of support images and grounds every prediction in localized visual evidence. Across distant-OOD benchmarks spanning medical imaging, industrial inspection, and synthetic reasoning, IVL achieves the highest aggregate accuracy under both Qwen-VL and LLaVA in 1-shot and 8-shot settings while remaining entirely training-free. The principal contributions are:

- **Representational absence**, as distinct from distributional shift, is identified as the fundamental barrier to VLM adaptation in specialized domains. A two-dimensional characterization based on feature-space divergence (MK-MMD [17]) and latent cluster-label alignment (NMI) provides a principled criterion for identifying distant-OOD regimes.
- **Explicit trait-based reasoning**, built upon VLMs' surviving descriptive capacity, proves effective precisely where parametric methods fail, achieving broad improvements across diverse benchmarks and two VLM backbones with interpretable, evidence-grounded predictions.
- **Dual-mode extraction**, combining semantic and primitive visual prompting, is essential for robust trait construction. The resulting framework requires no weight updates and enables immediate deployment.

2 Related Work

2.1 Generative Vision-Language Models

Generative VLMs [37] such as LLaVA [15] and Qwen-VL [1] have demonstrated remarkable zero-shot [4] and chain-of-thought reasoning [40] capabilities through LLM backbones that generate textual responses conditioned on visual inputs. Despite their broad knowledge foundation, these models exhibit significant performance degradation on specialized domains where high-level semantic patterns from natural images fail to transfer [14]. Parameter-efficient methods like LoRA [6] show limited effectiveness under large domain gaps, motivating few-shot adaptation strategies tailored for generative VLMs.

2.2 Few-Shot Domain Adaptation with VLMs

Few-shot domain adaptation for VLMs spans several paradigms. Prompt learning methods such as CoOp [42] and CoCoOp [41] learn continuous prompts that adapt to novel domains, while attribute-guided approaches like ArGue [33] align models with primitive visual attributes. Parameter-efficient fine-tuning methods, including LoRA [6], MMA [36], and PACE [23], insert lightweight adapters with minimal overhead; Tip-Adapter [39] eliminates backpropagation entirely by constructing adapter weights from a key-value cache. Visual-RFT [16] applies RL-guided fine-tuning for rapid adaptation. Nevertheless, recent studies [13] and our experiments reveal that these methods universally fail on distant-OOD tasks. Prompt tuning assumes compositional recombination of existing features [42], PEFT methods rely on adaptation within the pretrained subspace [6,39], and even Visual-RFT performs worse than SFT on most distant-OOD benchmarks. This universal failure suggests that distant-OOD adaptation demands fundamentally different approaches beyond parameter refinement.

2.3 Trait-Based and Neuro-Symbolic Approaches

Attribute-based recognition has long explored interpretable alternatives to end-to-end learning. Classical work on visual attributes [5, 12] demonstrated that

modeling explicit visual properties enables zero-shot recognition through attribute composition. Concept bottleneck models [10] enforce interpretability by predicting human-understandable concepts before classification, though they require extensive annotations. In the VLM era, [20] queried GPT-3 for visual descriptors to construct zero-shot classifiers, and CuPL [25] showed that VLM-generated descriptions improve CLIP's zero-shot accuracy, though neither addresses few-shot learning from novel domains. ArGue [33] demonstrated that aligning VLMs with explicit attributes via prompt tuning mitigates distribution shifts, but recent work [32] reveals that VLMs over-rely on spuriously correlated attributes, leading to poor OOD generalization. More fundamentally, VLMs exhibit systematic compositional failures [18, 31, 38] that compound these limitations: models process visual-linguistic inputs through bag-of-words representations rather than genuine compositional understanding, making the construction of new visual concepts from primitives unreliable. Unlike these approaches, IVL grounds both semantic and low-level traits in direct observation of the support images rather than relying on the model to compose discriminative concepts internally, with the low-level mode targeting the domain-agnostic vocabulary that Proposition 1(ii) identifies as surviving in the distant-OOD regime.

3 Methodology

3.1 Problem Formulation

Consider a generative VLM f with vision encoder ϕ_v and language decoder ϕ_l, pretrained on a large-scale distribution $\mathcal{P}_{\text{train}}$. Given a target distribution $\mathcal{P}_{\text{target}}$ whose discriminative visual primitives lie outside the support of $\mathcal{P}_{\text{train}}$, and a support set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n \times C}$ containing n labeled examples for each of C classes, the objective is to classify query images x_q into one of the C classes without updating any parameters of f.

IVL reformulates classification through explicit trait-based reasoning. A *trait* t is a natural-language description of a visual attribute (*e.g.*, "red wings," "spotted pattern," "layered horizontal bands"). From $\mathcal{D}$, IVL constructs a trait database $\mathcal{T} = \bigcup_{c=1}^{C} \mathcal{T}_c$, where each class c is represented by discriminative traits $\mathcal{T}_c = \{t_1^c, t_2^c, \ldots, t_{m_c}^c\}$ extracted from its support images. Classification of x_q then reduces to identifying which traits in $\mathcal{T}$ are visually grounded in x_q and mapping matched traits to a class prediction.

3.2 Theoretical Foundations

The design of IVL rests on two empirically motivated observations about VLM behavior in distant-OOD regimes. Gradient-based adaptation is fundamentally limited when the required discriminative features are absent from the learned representation space, and the model's capacity to generate visual descriptions survives even when its discriminative capacity collapses. These observations are formalized below as propositions; their formal justifications appear in the

Supp. §A, and their connection to the measurable diagnostic introduced in Section 3.3 is made explicit throughout.

Proposition 1 (Descriptive-Discriminative Asymmetry). *Let $f = (\phi_v, \phi_l)$ be a VLM with vision encoder ϕ_v and language decoder ϕ_l, pretrained on distribution $\mathcal{P}_{\text{train}}$. Define the* discriminative capacity $D(f, \mathcal{P}_{\text{target}})$ *as the mutual information $I\big(\phi_v(x); y\big)$ between the encoder's representation and the task label y under target distribution $\mathcal{P}_{\text{target}}$, and define the* descriptive capacity $G(f, \mathcal{P}_{\text{target}})$ *as the expected recall of visually grounded primitive attributes (*e.g.*, colors, textures, shapes, spatial relations) in descriptions generated by ϕ_l for images drawn from $\mathcal{P}_{\text{target}}$. If ϕ_l was pretrained on a corpus whose visual vocabulary spans general primitive attributes, then:*

(i) $D(f, \mathcal{P}_{\text{target}})$ degrades as the feature-space divergence $\delta\big(\mathcal{P}_{\text{train}}, \mathcal{P}_{\text{target}}\big)$ increases, approaching chance-level performance (assumptions in Supp. §A);

(ii) $G(f, \mathcal{P}_{\text{target}}) \geq G_0 > 0$ for all $\mathcal{P}_{\text{target}}$, where G_0 depends only on the breadth of the vision-language backbone's visual vocabulary and is independent of δ.

Proposition 2 (Gradient Futility under Representational Absence). *Let $\phi_v : \mathcal{X} \to \mathbb{R}^d$ be a frozen vision encoder and let $\Phi = \text{span}\big\{\phi_v(x) : x \in \mathcal{X}_{\text{train}}\big\}$ denote the subspace spanned by training-distribution representations. For a target distribution $\mathcal{P}_{\text{target}}$, let $w^\dagger \in \mathbb{R}^d$ be the unconstrained population minimizer of a squared-loss linear surrogate, and decompose it as $w^\dagger = w^\dagger_\parallel + w^\dagger_\perp$, where $w^\dagger_\parallel \in \Phi$ and $w^\dagger_\perp \in \Phi^\perp$. Under the assumptions in Supp. §A, any linear classifier $\hat{w}$ obtained by optimization over Φ satisfies*

$$\mathcal{R}_{\text{sq}}(\hat{w}) \;\geq\; \mathcal{R}_{\text{sq}}(w^\dagger) \;+\; \Omega\left(\frac{\|w^\dagger_\perp\|^2}{\|w^\dagger\|^2}\right), \tag{1}$$

where $\mathcal{R}_{\text{sq}}(\cdot)$ denotes squared-loss risk. The excess risk is irreducible with respect to the number of gradient steps, as $w^\dagger_\perp$ lies permanently outside the optimization domain Φ.

Remark 1 (Implications for IVL and SFT). Propositions 1 and 2 together provide a theoretical grounding for two complementary observations in practice.

SFT addresses the symptom, not the cause. Supervised fine-tuning repairs the degraded discriminative capacity $D(f, \mathcal{P}_{\text{target}})$ by re-learning task-specific decision boundaries on labeled target-domain data. However, this approach is costly and ignores a crucial asymmetry: the discriminative capacity D is bottlenecked by the *vision encoder* ϕ_v, whose frozen representations become increasingly uninformative as feature-space divergence $\delta(\mathcal{P}_{\text{train}}, \mathcal{P}_{\text{target}})$ grows, not by the language decoder ϕ_l, which retains descriptive capacity $G \geq G_0$ regardless of domain shift. Proposition 2 establishes that any classifier confined to the pretrained encoder's representation space incurs irreducible excess risk proportional to the missing subspace; Table 1 confirms that encoder-updating methods (SFT, LoRA, Visual-RFT) exhibit the same failure empirically, expending their supervision budget on structure the encoder never encoded.

IVL exploits the preserved descriptive capacity. Rather than repairing the degraded vision encoder ϕ_v, IVL routes reasoning through the guaranteed lower bound on G. Concretely, IVL: (i) uses the vision-language decoder ϕ_l to ground images in primitive attribute descriptions, leveraging $G \geq G_0$; (ii) organizes these descriptions into structured per-class trait dictionaries; and (iii) classifies novel examples by retrieving and spatially grounding matched traits, bypassing the encoder's degraded discriminative associations. This yields a few-shot pathway that is robust to feature-space divergence $\delta(\mathcal{P}_{\text{train}}, \mathcal{P}_{\text{target}})$ by design, directly addressing the structural limitation identified in Proposition 2. Where SFT requires labels to reconstruct D, IVL substitutes structured reasoning over the preserved descriptive capacity G, making the two approaches complementary. The formal derivation, including the information-theoretic argument and the conditions under which G_0 is maximized, appears in Supp. §A.1.

Taken together, Propositions 1 and 2 delineate both the failure mode of existing methods and the opportunity that IVL exploits. Parametric adaptation is structurally limited by the representational gap (Proposition 2), yet the VLM retains sufficient descriptive capacity to articulate visual observations about novel domains (Proposition 1). The following section introduces a two-dimensional diagnostic that determines, before any adaptation attempt, which structural properties of a target dataset govern whether gradient-based adaptation succeeds or fails, using only frozen encoder representations and ground-truth class labels.

3.3 Characterizing Representational Absence

We define the *distant-OOD regime* for a frozen encoder ϕ_v as the setting in which the encoder's representation space fails to support the target classification task along one or both of two axes: (i) the target images occupy regions of the representation space poorly covered by the pretraining distribution, and (ii) the encoder's internal grouping of target images bears little correspondence to class boundaries. This characterization depends only on the frozen encoder and the target data, not on any downstream adaptation procedure. To make it measurable, we introduce two complementary metrics computed entirely from frozen vision encoder embeddings and ground-truth class labels.

Feature-Space Divergence. For a target dataset and a reference natural-image distribution (ImageNet-1K), embeddings are extracted from the frozen vision encoder ϕ_v and the Multi-Kernel Maximum Mean Discrepancy (MK-MMD) [17] is computed between the two sets of vision embeddings. MK-MMD directly measures the feature-space divergence $\delta(\mathcal{P}_{\text{train}}, \mathcal{P}_{\text{target}})$ from Proposition 1(i). Both distributions comprise vision encoder outputs only; no text or class-label information enters the computation. A high MK-MMD indicates that the encoder maps target images to regions of $\mathbb{R}^d$ that are poorly populated by pretraining data, reducing the discriminative utility of the resulting representations.

Latent Cluster–Label Alignment. The frozen vision encoder embeddings of the target dataset are clustered via k-means (k=C, number of target classes,

n_{init}=10, Euclidean distance), and the Normalized Mutual Information (NMI) between the resulting cluster assignments and ground-truth class labels is computed. NMI $\in [0, 1]$ quantifies whether the encoder's internal grouping of target images already reflects task-relevant class boundaries, approximating the discriminative capacity $D(f, \mathcal{P}_{\text{target}})$ from Proposition 1. A dataset falls in the distant-OOD regime when it exhibits either high MK-MMD or low NMI; each axis captures a structurally distinct failure mode, as illustrated by Pokémon, which exhibits low MK-MMD (visually familiar cartoon images) yet low NMI (type labels have no grounded correspondence in the encoder's latent structure). Cross-model robustness and clustering-method sensitivity analyses appear in Supp. §B.

MK-MMD and NMI identify when discriminative capacity $D(f)$ has collapsed. Descriptive capacity $G(f) \geq G_0 > 0$ is guaranteed by Proposition 1(ii) independently of domain shift, and the trait extraction results in Section 4 empirically corroborate this. IVL helps only when the regime is also *trait-separable*, that is, when the primitives the decoder can articulate differ across classes (Section 4.3).

Projecting datasets onto these two axes yields the landscape in Fig. 1. Standard FGVC benchmarks occupy the low-divergence, high-alignment quadrant where parametric methods succeed; specialized domains fall into the high-divergence or low-alignment quadrants. Regime boundaries are set by a two-stage variant of Otsu's method [24] that thresholds MK-MMD first and then NMI on the low-divergence subset (Supp. §B), yielding $\tau_{\text{MMD}} = 0.49$ and $\tau_{\text{NMI}} = 0.42$, derived purely from frozen encoder geometry. Dataset groupings remain stable as τ_{NMI} varies across $[0.30, 0.60]$ (Supp. §B). Near-OOD datasets show consistently positive 1-shot SFT gains (mean $\Delta = +18.6\%$), while distant-OOD datasets show near-zero mean with high variance (Δ from -27.5% to $+36.9\%$), consistent with Proposition 2.

3.4 Framework Overview

Guided by the theoretical analysis above, IVL instantiates the descriptive-discriminative asymmetry (Proposition 1) in a two-stage pipeline. Fig. 3 provides an architectural overview. The first stage, **Offline Trait Building** (Section 3.5), prompts the language decoder in two complementary modes to extract fine-grained visual traits from support images and consolidates them into a compact per-class trait database $\mathcal{T}$. The second stage, **Inference** (Section 3.6), classifies a query image by identifying which traits in $\mathcal{T}$ are visually grounded in the query through hierarchical filtering, bypassing the discriminative bottleneck identified in Proposition 2. The entire pipeline requires no parameter updates and produces interpretable predictions traceable to specific traits and their spatial grounding.

3.5 Trait Building Stage

To operationalize the surviving descriptive capacity established in Proposition 1, IVL constructs an explicit trait vocabulary from the support set through three phases illustrated in Fig. 3 (a–c): dual-mode trait extraction, semantic clustering with canonical naming, and quality filtering.

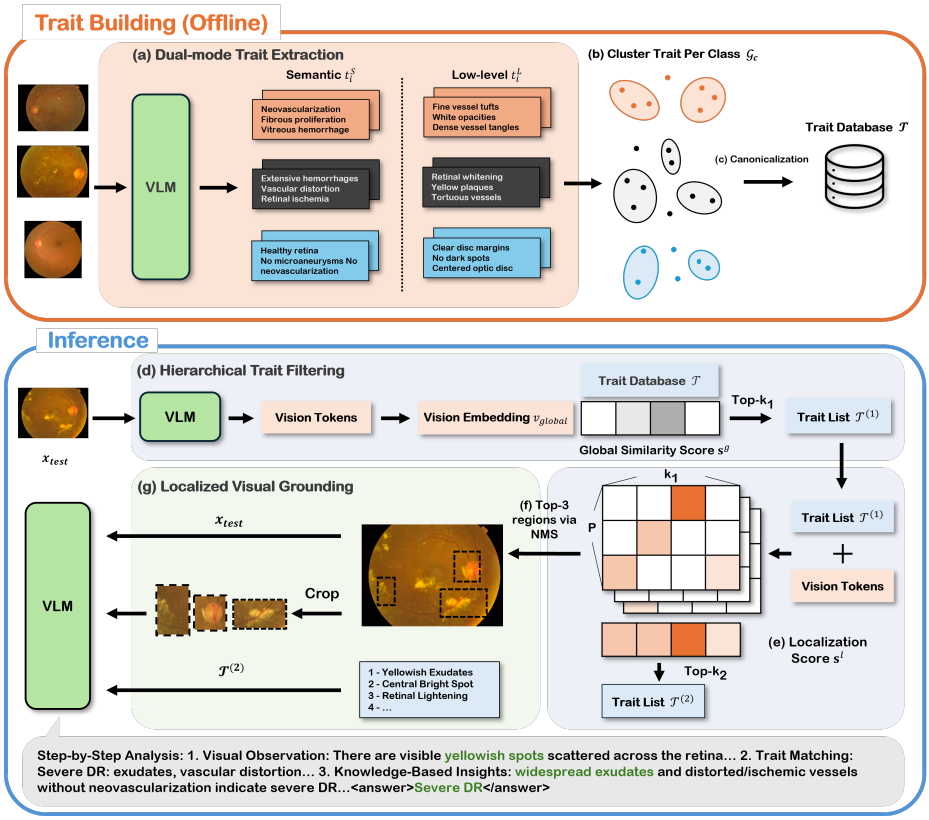

Fig. 3: Offline Trait Building: (a) The system extracts semantic (t_i^S) and low-level (t_i^L) traits from support images (b) clusters them per class and (c) assigns canonical names, and stores them in database $\mathcal{T}$. **Inference**: Given test image x_{test}, hierarchical filtering is performed: (d) compute $\mathbf{s}^g = \hat{\mathbf{E}}_{\mathcal{T}}^{\top} \hat{\mathbf{v}}_{\text{global}}$ to select top-k_1 traits $\mathcal{T}^{(1)}$, (e) compute image patch-trait attention scores for localized grounding while selecting top-k_2 traits $\mathcal{T}^{(2)}$ and (f) extracting top-3 regions via NMS, and (g) prompt the VLM with the original image, cropped regions, and the refined trait set $\mathcal{T}^{(2)}$ for final classification.

Dual-Mode Trait Extraction The dual-mode decomposition follows directly from the capacity structure in Proposition 1. A VLM's descriptive capacity $G(f)$ encompasses two qualitatively different resources. The first is residual domain knowledge, which allows the model to produce semantically rich descriptions when it retains partial familiarity with the target domain. The second is the domain-agnostic visual vocabulary guaranteed by Proposition 1(ii), encompassing primitive descriptions of colors, textures, and shapes that remain available regardless of domain shift. IVL targets each resource with a dedicated extraction mode, ensuring that trait quality degrades gracefully rather than collapsing entirely as the target domain moves further from the pretraining distribution.

Semantic traits t_i^S are elicited through a prompt p^S that requests knowledge-based descriptions, leveraging whatever domain understanding the model retains.

Low-level traits t_i^L are elicited through a prompt p^L that constrains responses to primitive visual features and explicitly prohibits category labels, targeting the $G_0 > 0$ floor from Proposition 1(ii). Rather than producing the generic label "retinal tissue," a low-level prompt yields descriptions such as "white wrinkled regions in center" or "dark void between horizontal bands." This decomposition distinguishes IVL from prior attribute-based methods [20, 25, 28, 33], which generate descriptors entirely from pretrained semantic knowledge without examining the actual support images. When that semantic knowledge is absent, such methods produce spurious or generic descriptors. IVL instead extracts traits from visual observations of the support set, grounding every descriptor in image content rather than in the model's prior beliefs.

Both prompts follow a unified, domain-agnostic template; the only dataset-specific tokens are the domain and class names. The extraction performs K independent runs per image (typically K=5), each generating 6–8 traits per mode following the self-consistency principle [34]. Semantic trait coverage degrades with feature-space divergence, whereas low-level coverage remains stable because primitive visual vocabulary is domain-agnostic (*cf*. Proposition 1). The combined set $\mathcal{T}^S \cup \mathcal{T}^L$ therefore achieves coverage strictly higher than either mode alone whenever the two sets are not fully redundant, which is the typical case for distant-OOD domains. Full prompt templates appear in Supp. §J.

Clustering, canonicalization, and refinement. Raw traits are lexically diverse and redundant. For each class c, extracted traits are embedded using Sentence Transformers [27] and clustered via HDBSCAN [19], with isolated traits retained as rare discriminative cues. Each cluster is assigned a canonical name through VLM compositional reasoning (*e.g.*, "red legs" and "crimson arms" become "red limbs"). Sentence Transformers are preferred over VLM text encoders because the latter over-compress textual semantic structure (Supp. §E). Canonicalized semantic traits undergo relevance filtering through VLM common-sense reasoning; low-level traits bypass this step, as the model cannot reliably assess primitive features in unfamiliar domains (*cf*. Proposition 1). Final trait embeddings are computed using the VLM's text encoder, yielding $\mathbf{E}_{\mathcal{T}} = [\mathbf{e}_1, \ldots, \mathbf{e}_{|\mathcal{T}|}] \in \mathbb{R}^{d \times |\mathcal{T}|}$.

3.6 Inference Stage

Given $\mathcal{T}$ and its embedding matrix $\mathbf{E}_{\mathcal{T}}$, inference classifies a query image by identifying visually grounded traits through hierarchical filtering.

Global Trait Retrieval As illustrated in Fig. 3 (d), given test image x_{test}, the VLM vision encoder produces P patch tokens $\{v_1, \ldots, v_P\}$, each $v_i \in \mathbb{R}^d$. A global embedding $\mathbf{v}_{\text{global}} = \frac{1}{P}\sum_{i=1}^{P} v_i$ is obtained via mean pooling, and global similarity scores $\mathbf{s}^g = \hat{\mathbf{E}}_{\mathcal{T}}^{\top}\hat{\mathbf{v}}_{\text{global}} \in \mathbb{R}^{|\mathcal{T}|}$ are computed via L2-normalized cosine similarity. The top-k_1 traits form $\mathcal{T}^{(1)} \subset \mathcal{T}$.

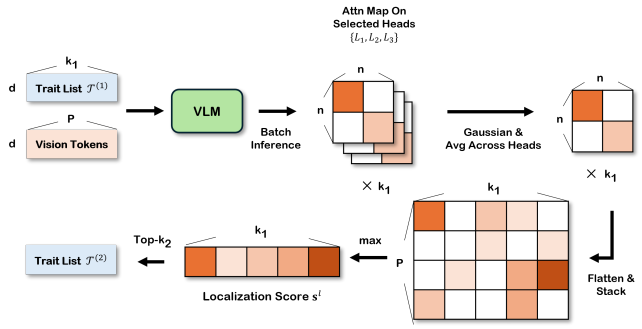

Fig. 4: Localized Trait Refinement. Given the test image and the globally filtered trait list $\mathcal{T}^{(1)}$, the VLM produces cross-modal attention maps from the grounding heads $\{L_1, L_2, L_3\}$. After averaging and spatial smoothing, each trait's attention map is flattened into a P-dimensional patch-level vector. The localization score s_i^l is computed as the maximum attention over patches, and the top-k_2 traits are selected to form $\mathcal{T}^{(2)}$.

Localized Trait Refinement Global retrieval cannot verify spatial grounding. Following Kang *et al.* [8], IVL extracts cross-modal attention maps from three grounding heads $\{L_1, L_2, L_3\}$ (Fig. 4). For each trait $t_i \in \mathcal{T}^{(1)}$, maps are averaged and flattened into the patch-trait attention matrix $A \in \mathbb{R}^{P \times k_1}$. A localization score $s_i^l = \max_p A_{p,i}$ captures whether the trait has concentrated evidence in at least one region. The top-k_2 traits form $\mathcal{T}^{(2)} \subset \mathcal{T}^{(1)}$.

Visual region extraction and classification. Attention maps from $\mathcal{T}^{(2)}$ are reshaped into 2D heatmaps, smoothed with a Gaussian filter, and bounding boxes are extracted around the top-3 local maxima (Fig. 3 (f)). The VLM receives three complementary inputs for classification: the original image x_{test}, the refined trait list $\mathcal{T}^{(2)}$ as textual context, and the three cropped regions for focused inspection (Fig. 3 (g)). By design, classification routes through the decoder's preserved descriptive capacity $G(f)$ while bypassing the encoder's degraded discriminative features (Propositions 1 and 2). The entire pipeline is training-free and produces interpretable predictions traceable to traits and spatial grounding.

4 Experimental Results

4.1 Experimental Setup

IVL is designed for the distant-OOD regime, where pretrained representations lack discriminative structure. Benchmarks are selected through the two-metric protocol of Section 3.3: datasets exhibiting high MK-MMD or low NMI relative to Otsu-derived thresholds are retained, yielding 18 classification tasks: Retinal OCT [22], WM811k [35], MVTec AD [2] (15 categories evaluated independently), and Pokémon. The Pokémon benchmark exhibits low visual divergence yet low semantic alignment (elemental type labels have no grounded correspondence in

the VLM's label space), isolating semantic collapse from visual unfamiliarity. Full per-dataset metrics appear in Supp. §B. MVTec AD serves as a distant-OOD stress test for general-purpose generative VLMs, not as a comparison with specialized anomaly detection pipelines that leverage task-specific architectures unavailable to general-purpose models.

Baselines span the major adaptation paradigms: MajorVote (k=5) for test-time scaling, ICL for few-shot demonstration, CuPL [25] and Menon *et al.* [20] for attribute-based zero-shot classification, Tip-Adapter [39] and CoOp [42] for adapter-based and prompt-tuning approaches, Visual-RFT [16] for RL-based few-shot learning, and SFT/SFT+LoRA for gradient-based fine-tuning. DIFO [29] is evaluated in Supp. §I. All methods share the same VLM backbone (Qwen2.5-VL-7B or LLaVA-1.6-Mistral-7B) with a fixed few-shot support set across all methods. Additional details appear in Supp. §J.

4.2 Quantitative Results

The most striking pattern in Table 1 is the failure of gradient-based adaptation on distant domains. On Retinal OCT under 1-shot Qwen2.5-VL, SFT achieves only 14.77% (+0.41 pp over zero-shot) despite full parameter updates, and Visual-RFT degrades to 14.00%; even at 8-shot, neither exceeds 16.00%. On WM811k, both SFT+LoRA and Visual-RFT fall below the zero-shot baseline, confirming the gradient futility phenomenon formalized in Proposition 2. Test-time scaling (MajorVote) and in-context learning exhibit analogous instability, with MajorVote degrading on WM811k under LLaVA (6.39% vs. 10.30% zero-shot) despite marginal gains elsewhere. IVL avoids these failure modes by building visual knowledge from the VLM's surviving descriptive capacity, achieving the best result on Retinal OCT (20.83%), WM811k (16.51%), and Pokémon (50.30%).

Attribute-based methods that generate descriptors from pretrained semantic knowledge without observing target-domain images inherit the model's blind spots, as reflected by CuPL's 20.1% and Menon *et al.*'s 32.8% on MVTec AD compared to IVL's 46.1%. Across all 18 tasks, IVL places in the top two for 12 under Qwen2.5-VL and 10 under LLaVA, the widest coverage of any single method under both backbones. These improvements hold under both backbones: under LLaVA-1.6, IVL achieves the highest accuracy on Pokémon, WM811k, and MVTec AD mean, and matches the best-performing method on Retinal OCT (33.70% vs. 34.30%) without any parameter updates.

Computational analysis (Supp. §H) shows that IVL's trait building requires only a single 32 GB VRAM GPU and 67% fewer GPU-seconds than SFT, which demands a multi-GPU cluster. Per-image inference incurs a moderate overhead, dominated by the VLM processing enriched trait-grounded prompts rather than by the retrieval pipeline itself.

4.3 Failure Modes

IVL's errors are interpretable from its trait rationale and fall into three modes, each arising at a different level of the pipeline.

Table 1: Experimental results on distant-OOD benchmarks. Zero-shot and description-based methods (CuPL, Menon) use 0-shot. **Bold**: best per dataset within each backbone. Underline: second best. All few-shot results use n=1 unless marked 8-shot.

Dataset	Zero-Shot	MajorVote	ICL	CuPL [25]	Menon *et al.* [20]	V-RFT [16]	SFT+LoRA	SFT	Tip-Adapter [39]	CoOp [42]	IVL (Ours)
Qwen2.5-VL-7B											
Pokémon	47.47	48.23	22.00	15.28	25.77	48.77	45.59	45.52	9.72	20.01	**50.30**
Retinal OCT	14.36	16.39	13.75	12.50	11.46	14.00	13.69	14.77	12.50	13.39	**20.83**
WM811k	13.13	13.73	12.41	12.77	12.29	9.52	9.16	12.41	11.08	9.30	**16.51**
8-shot											
Pokémon	–	–	34.24	–	–	49.09	50.09	49.94	6.94	29.17	**52.20**
Retinal OCT	–	–	19.54	–	–	13.14	14.00	16.00	20.00	**39.92**	23.21
WM811k	–	–	14.22	–	–	8.92	9.16	15.90	11.08	17.59	**18.43**
MVTec AD – per-category (Qwen2.5-VL, 1-shot)											
Bottle	50.6	54.43	36.7	26.6	**57.0**	51.9	53.2	48.1	31.65	39.2	46.8
Cable	12.8	11.35	14.9	9.2	12.8	9.2	12.8	41.1	11.35	8.5	**45.4**
Capsule	46.0	**47.62**	25.4	14.3	35.7	42.1	40.5	28.6	23.02	17.5	44.4
Carpet	53.2	**61.26**	20.7	16.2	40.5	55.0	58.6	55.0	21.62	14.4	60.4
Grid	36.1	34.72	23.6	20.8	34.7	**52.8**	**52.8**	40.3	27.78	15.3	43.1
Hazelnut	61.9	58.10	41.9	16.2	37.1	59.1	60.0	**77.1**	30.48	21.9	72.4
Leather	44.9	42.37	18.6	13.6	24.6	39.0	40.7	**60.2**	28.81	12.7	48.3
Metal Nut	**47.3**	37.27	26.4	21.8	28.2	37.3	36.4	20.9	17.27	37.3	40.9
Pill	15.1	16.35	13.8	11.3	5.0	13.8	12.6	20.8	11.95	11.9	**22.0**
Screw	26.0	29.22	20.1	14.3	26.0	23.4	23.4	18.8	27.27	20.8	**33.1**
Tile	58.6	62.16	27.0	13.5	50.5	46.9	47.8	49.6	25.23	55.0	**65.8**
Toothbrush	55.0	60.00	52.5	**72.5**	**72.5**	57.5	55.0	27.5	**72.5**	45.0	65.0
Transistor	18.9	11.58	10.5	9.5	9.5	13.7	13.7	55.8	**62.11**	42.1	11.6
Wood	60.3	57.53	46.6	30.1	45.2	42.5	41.1	58.9	24.66	24.7	**65.8**
Zipper	19.6	25.17	11.2	11.2	13.3	**28.0**	**28.0**	17.5	13.99	16.8	26.6
MVTec Mean	40.4	40.61	26.0	20.1	32.8	38.1	38.4	41.3	28.65	25.54	**46.1**
LLaVA-1.6-Mistral-7B											
Pokémon	28.00	33.60	5.37	15.97	8.29	37.18	34.76	38.15	31.94	6.94	**44.60**
Retinal OCT	10.57	10.96	11.46	11.68	12.61	11.18	31.38	**34.30**	24.36	12.39	33.70
WM811k	10.30	6.39	10.72	14.58	11.20	12.41	11.92	12.10	15.90	11.45	**17.21**
8-shot											
Pokémon	–	–	4.39	–	–	41.23	41.47	33.42	34.03	39.92	**48.41**
Retinal OCT	–	–	17.50	–	–	7.23	26.29	**35.20**	12.50	12.89	31.33
WM811k	–	–	12.17	–	–	4.00	8.92	12.81	10.96	**24.11**	19.13
MVTec AD – per-category (LLaVA-1.6, 1-shot)											
Bottle	47.0	39.24	29.1	26.6	45.6	32.4	33.4	42.6	**60.76**	17.72	34.2
Cable	5.4	**39.72**	12.1	6.4	7.8	8.6	15.2	14.2	25.53	8.51	16.4
Capsule	21.4	18.25	23.8	17.5	17.5	44.2	38.1	**46.5**	19.84	15.08	32.2
Carpet	14.4	**41.44**	29.7	16.2	16.2	19.0	13.4	16.4	14.41	14.41	34.2
Grid	10.9	**34.72**	33.3	23.6	29.2	9.5	28.9	27.4	15.28	18.06	25.0
Hazelnut	40.0	34.29	35.2	15.2	41.9	38.1	33.7	27.1	26.67	16.19	**47.8**
Leather	36.1	40.68	26.3	15.3	0.9	62.3	68.3	**73.1**	25.42	11.86	58.1
Metal Nut	31.8	30.91	25.5	20.0	19.1	20.0	21.4	38.4	**54.55**	19.09	31.7
Pill	13.8	13.84	13.8	14.5	5.0	42.1	**58.3**	54.2	21.38	16.35	48.7
Screw	21.7	26.62	21.4	15.6	14.3	24.9	34.1	**38.8**	18.83	15.58	28.7
Tile	33.7	43.24	26.1	13.5	42.3	48.4	22.7	71.3	29.73	15.32	**76.7**
Toothbrush	35.0	62.50	60.0	37.5	30.0	45.0	20.0	40.0	**72.5**	67.5	**72.5**
Transistor	39.3	14.74	17.9	9.5	31.6	26.1	31.7	30.4	14.74	9.47	**43.2**
Wood	8.3	46.58	32.9	27.4	28.8	**64.2**	12.2	16.8	26.03	17.81	54.8
Zipper	38.4	21.68	11.2	11.2	11.2	41.2	32.7	31.1	10.49	13.29	**48.9**
MVTec Mean	26.48	33.90	26.6	18.0	22.8	35.1	31.0	37.9	29.08	18.42	**43.54**

Reference-free vocabulary collapse (mechanism level). Because IVL inspects each query in isolation rather than against a normal reference, benign within-tolerance variation is read as a defect and the trait dictionary collapses toward one dominant concept. The signature is a single trait cited as decisive across many queries, as on MVTec AD Transistor, where most categories are all predicted *bent_lead* and accuracy falls to 11.6%. Scoring candidate traits against support references rather than in isolation would address this (Supp. §G).

Dictionary coverage failure (retrieval level). When hierarchical filtering drops the ground-truth class from the candidate dictionary, no downstream reasoning can recover it, and the signature is a trace whose cited traits never include the true class. This accounts for 48.4% of Pokémon errors and motivates open-vocabulary dictionary construction (Supp. §D).

Table 2: Prompt design ablation on MVTec AD (Qwen2.5-VL, 1-shot, 3-run mean per mode). Semantic: knowledge-driven prompting. Low-level: observation-driven prompting. Δ: dual-mode gain over best single mode.

	Bottle	Cable	Capsule	Carpet	Grid	Hazelnut	Leather	Metal Nut	Pill	Screw	Tile	Toothbrush	Transistor	Wood	Zipper	Mean
Semantic	39.2	14.7	35.7	36.9	34.3	52.7	28.0	26.7	5.9	32.3	57.7	51.7	**15.4**	40.2	17.0	32.6
Low-level	40.5	30.3	35.7	50.8	36.6	46.7	39.5	31.2	10.5	30.3	52.6	**72.5**	14.4	47.9	**29.3**	37.9
Dual	**46.8**	**45.4**	**44.4**	**60.4**	**43.1**	**72.4**	**48.3**	**40.9**	**22.0**	**33.1**	**65.8**	65.0	11.6	**65.8**	26.6	**46.1**
Δ	+6.3	+15.1	+8.7	+9.6	+6.5	+19.7	+8.8	+9.7	+11.5	+0.8	+8.1	−7.5	−3.8	+17.9	−2.7	+7.2

Table 3: Component ablation on the Pokémon benchmark (144-sample balanced subset, Qwen2.5-VL, 1-shot). Δ: change relative to full pipeline.

Component	Configuration	Acc. (%)	Δ
Extraction (Section 3.5)	Semantic only	32.6	−6.2
	Low-level only	27.8	−11.1
Filtering (Section 3.6)	Global only	39.6	+0.7
	Local only	37.5	−1.4
	No filtering	36.8	−2.1
Quality (Section 3.5)	w/o salience	36.3	−2.6
Grounding (Section 3.6)	w/o localized	34.7	−4.2
Full pipeline		**38.9**	—

Trait-separability boundary (applicability level). IVL can only succeed when the classes differ in traits the model can actually name. When they do not, there is nothing for descriptive reasoning to match on, and accuracy stays near chance. FractalDB [9] is the clearest case, since its labels come from the generative parameters that produced each image rather than from any visible attribute. This is not a defect IVL can fix but a genuine limit of trait-based reasoning, marking where parametric or hybrid methods are still needed (Supp. §G).

4.4 Ablation Studies

Dual-mode prompt design. Since the dual-mode prompt strategy is the central design choice distinguishing IVL from prior attribute-based methods, Table 2 provides a systematic per-category evaluation on MVTec AD. Three patterns emerge. First, low-level prompting outperforms semantic prompting in 10 of 15 categories, confirming that on distant-OOD domains the VLM's primitive visual vocabulary is more reliable than its domain-specific knowledge, as predicted by Proposition 1. Second, dual-mode extraction surpasses the better single mode on 12 of 15 categories, demonstrating that the two prompt strategies contribute non-redundant discriminative information. Third, the three exceptions are interpretable: Transistor suffers from vocabulary collapse (Supp. §G), while Toothbrush and Zipper are visually simple categories where low-level descriptors alone suffice. This dual-mode design is the key distinction from CuPL [25] and Menon *et al.* [20], which rely exclusively on semantic prompting without visual grounding and underperform IVL's dual-mode results on all 15 categories.

Component ablation. Table 3 isolates the remaining components on the Pokémon benchmark. The relative ranking of extraction modes reverses com-

pared to MVTec AD: semantic-only (32.6%) outperforms low-level-only (27.8%), because the VLM retains partial semantic knowledge of cartoon imagery but lacks grounded type-label associations. This reversal confirms that dual-mode extraction is necessary because the dominant failure mode varies across domains. Removing visual grounding degrades accuracy by -4.2 pp, confirming that the VLM requires direct observation of localized regions where discriminative traits manifest. Disabling salience filtering costs -2.6 pp, and removing all hierarchical filtering degrades accuracy by -2.1 pp; the marginal global-only gain ($+0.7$ pp) reflects Pokémon's whole-body structure, where local refinement is less critical than on spatially complex datasets. Detailed experiments appear in Supp. §E.

5 Conclusion

This paper introduces IVL, a training-free framework that addresses VLM failures on distant-OOD tasks through trait-based reasoning rather than parameter adaptation. The central insight is that these failures stem from structurally absent visual primitives, not from optimization limitations; gradient-based adaptation cannot construct discriminative features that the encoder never learned. IVL constructs such features explicitly through dual-mode trait extraction and hierarchical filtering, grounding each classification decision in spatially localized visual evidence. Experiments across distant-OOD benchmarks confirm that this approach succeeds where parametric methods fail, establishing trait-based reasoning as a principled and immediately deployable path for VLM adaptation in specialized domains. Extending trait-based reasoning toward open-set recognition, where the target class may lie outside the support set, and compositional generalization is a promising direction for future work.

Acknowledgement

The authors gratefully acknowledge the support from the National Science and Technology Council (NSTC) in Taiwan under grant numbers NSTC 113-2221-E-007-105-MY3, NSTC 114-2634-F-002-004, NSTC 114-2221-E-002-069-MY3, NSTC 113-2221-E-002-212-MY3, NSTC 115-2634-F-002-012, and NSTC 115-2218-E-A49-010, as well as the support from the Academia Sinica Scholar Award (ASSA) under grant number AS-ASSA-115-02, NTU Artificial Intelligence Center of Research Excellence, and Taiwan Centers of Excellence in Artificial Intelligence. This research was also supported by the NVIDIA Academic Grant Program. The authors would also like to express their appreciation for the hardware grant donation of the GPUs from NVIDIA Corporation and NVIDIA AI Technology Center (NVAITC) used in this work. Furthermore, the authors extend their gratitude to the National Center for High-Performance Computing (NCHC) for providing computational and storage resources. The authors also thank the NVIDIA Taipei-1 supercomputer for providing essential computing resources.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. International Journal of Computer Vision **129**(4), 1038–1059 (2021)
3. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)
4. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
5. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1778–1785 (2009). https://doi.org/10.1109/CVPR.2009.5206772
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
7. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
8. Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9339–9350 (2025)
9. Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., Satoh, Y.: Pre-training without natural images. In: Proceedings of the Asian Conference on Computer Vision (2020)
10. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
11. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
12. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 951–958 (2009). https://doi.org/10.1109/CVPR.2009.5206594
13. Li, C., Flanigan, J.: Task contamination: Language models may not be few-shot anymore. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18471–18480 (2024)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
16. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)

17. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015)
18. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023)
19. McInnes, L., Healy, J., Astels, S., et al.: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
20. Menon, S., Vondrick, C.: Visual classification via description from large language models. arXiv preprint arXiv:2210.07183 (2022)
21. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: Rethinking the role of demonstrations: What makes in-context learning work? arXiv preprint arXiv:2202.12837 (2022)
22. Naren, O.S.: Retinal oct image classification - c8 (2021). https://doi.org/10.34740/KAGGLE/DSV/2736749, https://www.kaggle.com/dsv/2736749
23. Ni, Y., Zhang, S., Koniusz, P.: Pace: Marrying generalization in parameter-efficient fine-tuning with consistency regularization. Advances in Neural Information Processing Systems **37**, 61238–61266 (2024)
24. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
25. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15691–15701 (2023)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
27. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
28. Roth, K., Kim, J.M., Koepke, A.S., Vinyals, O., Schmid, C., Akata, Z.: Waffling around for performance: Visual classification with random words and broad concepts. arXiv preprint arXiv:2306.07282 (2023)
29. Tang, S., Su, W., Ye, M., Zhu, X.: Source-free domain adaptation with frozen multimodal foundation model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23711–23720 (2024)
30. Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D.: How to grow a mind: Statistics, structure, and abstraction. Science **331**, 1279 – 1285 (2011), https://api.semanticscholar.org/CorpusID:469646
31. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
32. Tian, X., Zou, S., Yang, Z., He, M., Zhang, J.: Black sheep in the herd: Playing with spuriously correlated attributes for vision-language recognition. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 33527–33560 (2025), https://proceedings.iclr.

cc/paper_files/paper/2025/file/530eb593743559fa3c926a1fc1f520e9-Paper-Conference.pdf
33. Tian, X., Zou, S., Yang, Z., Zhang, J.: Argue: Attribute-guided prompt tuning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28578–28587 (2024)
34. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
35. WM811K: Wm811k dataset. https://universe.roboflow.com/wm811k-paasr/wm811k (mar 2023), https://universe.roboflow.com/wm811k-paasr/wm811k, visited on 2025-09-25
36. Yang, L., Zhang, R.Y., Wang, Y., Xie, X.: Mma: Multi-modal adapter for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23826–23837 (June 2024)
37. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
38. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=KRLUvxh8uaX
39. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
40. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
41. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
42. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) (2022)