AMCI: Unlock the Potential of Large Multimodal Models for Fine-grained Open-world Classification via Adaptive Memory Context Injection

Xiao Liu[1], Haiyang Zheng[2], Nan Pu[1*], Wenjing Li[1], Nicu Sebe[2], and Zhun Zhong[1*]

[1] Hefei University of Technology, Hefei, China
[2] University of Trento, Trento, Italy

Abstract. Open World Classification (OWC) with Large Multimodal Models (LMMs) is an emerging and promising task that moves beyond the closed-world setting by enabling direct answer generation in response to open-ended queries. However, we observe that the OWC performance of LMMs significantly degrades on fine-grained tasks. Empirically, we find that their performance is sensitive to contextual information in the query. To address this limitation, we propose Adaptive Memory Context Injection (AMCI), a training-free and model-agnostic framework that dynamically restores contextual information during OWC inference. AMCI maintains a non-parametric memory that aligns visual indices with their corresponding attribute-augmented descriptions. By employing an adaptive similarity threshold to preserve memory diversity and utilizing a refinement LLM to transform accumulated context into a structured prompt, AMCI transforms the initial zero-shot description into a calibrated fine-grained prediction without parameter updates. Furthermore, we construct a specialized low-resource fine-grained benchmark to simulate extreme data scarcity and introduce a new metric that penalizes overly generic predictions, tailored for fine-grained OWC evaluation. Extensive experiments demonstrate that AMCI consistently improves performance by a substantial margin while remaining computationally efficient and compatible with various LMMs.

Keywords: Open-World Classification · Large Multimodal Models · Test-Time Adaptation

1 Introduction

Image classification, which assigns semantic labels to images, is a fundamental task in computer vision [20, 25]. Traditional approaches typically rely on a closed-world assumption, where the category set is fixed and known during training [4]. However, this assumption does not reflect real-world environments,

[2] *Corresponding authors.

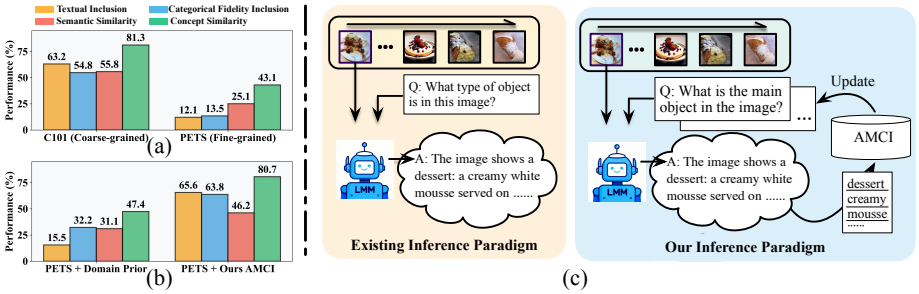

Fig. 1: (a) LMMs show the obvious performance gap between coarse-grained and fine-grained OWC. (b) Incorporating contextual information (e.g., domain priors) in query benefits fine-grained OWC. Furthermore, by applying automatic AMCI in a training-free manner, the performance is significantly improved. (c) Existing instance-by-instance inference paradigm *vs.* our streaming inference paradigm.

where new categories continuously emerge [42]. Recent advances in Large Multimodal Models (LMMs), such as LLaVA [33], BLIP-2 [30], and Qwen-VL [3], challenge this paradigm by enabling open-ended visual recognition through joint vision–language modeling [39]. By answering natural language queries (e.g., "What is the object in the image?"), LMMs can identify semantic concepts beyond predefined categories. However, their open-ended outputs also require new evaluation protocols to assess prediction correctness. To this end, recent work [11] reformulates Open World Classification (OWC) task by treating LMMs as universal classifiers and proposes evaluation metrics based on language models and textual embedding similarity.

Despite achieving strong performance on coarse-grained OWC, LMMs often struggle with fine-grained categories, frequently generating overly generic predictions (e.g., "flower" instead of "lily"), as illustrated in Fig. 1 (a). To obtain more specialized predictions, one intuitive solution is to incorporate contextual information (e.g., domain priors) into the query prompt. As shown in Fig. 1 (b), such contextual cues can substantially improve the performance of LMMs. However, assuming access to such prior knowledge at test time is unrealistic in open-world scenarios, where the domain of incoming data is unknown and highly variable. Moreover, due to the large parameter scale of LMMs, repeatedly fine-tuning them for different testing domains is computationally expensive and typically requires large-scale labeled data, which is impractical for low-resource tasks. Therefore, it becomes crucial to design a lightweight method that can automatically inject contextual information during inference, particularly for fine-grained and low-resource scenarios [14, 46].

On the other hand, Retrieval-Augmented Test-Time Adaptation [12, 16, 22, 27] (Retrieval-Augmented TTA) has emerged as a promising strategy for adapting models to incoming test samples, by building a memory buffer and leveraging contextual information from the test stream, as illustrated in Fig. 1(c). However, directly applying it to fine-grained OWC faces two fundamental challenges: ***1) How to construct a memory that provides informative and representative context in open-world settings?*** Unlike the zero-shot classification

task in Retrieval-Augmented TTA [12, 16, 22, 27], OWC lacks ground-truth labels, which makes supervised or pseudo-label updates impractical. Meanwhile, without ground-truth guidance, a naive memory strategy may accumulate visually similar but non-discriminative samples, making it harder to retrieve useful fine-grained evidence for accurate recognition. Thus, the challenge is to automatically extract and retain key fine-grained attributes while keeping the memory diverse to avoid redundant samples. ***2) How to effectively transform accumulated context into structured prompts?*** Retrieved samples often contain redundant narratives or distracting noise. If directly used as prompts, these fragmented observations may not provide enough useful evidence for categorization. Therefore, a refinement step is needed to summarize representative features from the raw context and convert them into prompts that guide LMMs toward more reliable predictions.

To address these challenges, we propose a training-free framework, Adaptive Memory Context Injection (AMCI), following a multi-stage fashion. **Stage 1: Memory Initialization.** During the initial phase of the test stream, the model initializes a memory with visual indices and attribute-augmented descriptions generated through standard zero-shot LMM inference. This stage establishes a foundational memory that captures the initial semantic distribution of the test stream, providing the necessary comparative anchors for subsequent observations. **Stage 2: Retrieval and Diversity Maintenance.** To construct an informative and representative memory, this stage uses a dynamic similarity threshold coupled with a most visually similar replacement strategy. For each query, AMCI retrieves the k nearest neighbors from the memory to provide contextual information for the current observation. The inclusion of new samples is controlled by an EMA-updated threshold, which filters samples with higher novelty and reduces redundancy. If the memory reaches its capacity, AMCI replaces the existing entry most visually similar to the current query with the current entry, effectively preventing the memory from being overwhelmed by biased or redundant data. **Stage 3: Context-aware Refinement and Inference.** This stage converts accumulated context into structured prompts via a refinement LLM. This module augments the initial description of a query with fine-grained attributes extracted from neighboring instances. By aggregating fine-grained attributes from neighboring instances, the LLM complements the initial description with discriminative details that may be missing from the query alone, thereby improving prediction reliability without backpropagation.

Our main contributions are summarized as follows:

- We identify the lack of contextual guidance during inference as an important factor contributing to the degraded performance of LMMs in fine-grained OWC.
- We propose AMCI, a training-free and model-agnostic framework that leverages a dynamic memory to achieve test-time context injection and refine model outputs without labels or backpropagation.
- We extend the existing OWC benchmark to low-resource scenarios and introduce a new metric for more reasonable evaluation tailored to fine-grained

OWC. Extensive experiments across multiple datasets demonstrate that AMCI consistently improves the performance of various LMMs.

2 Related Work

Open-World Classification with LLMs. Image classification [6, 15, 24, 25] underpins the perceptual power of multimodal models. While contrastive VLMs like CLIP [39] excel in zero-shot classification, they remain constrained by predefined label sets. This has led to the emergence of Vocabulary-free Image Classification (VIC) [9, 10] for unconstrained label identification. Generative LMMs have achieved remarkable success in reasoning and scientific comprehension [21, 29, 31, 35]. However, their foundational classification performance, particularly in open-world settings, remains under-explored. Zhang et al. [45] analyze the image classification performance of LMMs, defining OWC in the context of LMMs as a setting characterized by a lack of constraints in the output space. Their findings suggest that generative LMMs underperform their contrastive counterparts. Conversely, Liu et al. [34] challenge this view in closed-world environments, demonstrating the latent classification potential of modern LMMs, though their analysis did not extend to open-world challenges. Conti et al. [11] further expand the evaluation in [45] by introducing four complementary metrics across 10 datasets, providing extensive studies on OWC and error analysis for LMMs. In parallel, CIRCLE [19] and SpeciaRL [2] also explore improving LMM-based OWC, via iterative multi-granular pseudo-label refinement and reinforcement learning with dynamic rewards, respectively. *In this paper, we delve into the performance bottlenecks of LMMs in extreme low-resource and fine-grained scenarios, uncovering a prevalent 'vagueness bias' stemming from the lack of contextual information. To address this, we propose AMCI, a dynamic memory-augmented adaptation framework designed to activate the expert-level discriminative potential of LMMs.*

Retrieval-Augmented Test-Time-Adaptation. Test-time adaptation (TTA) improves robustness by exploiting unlabeled test data at deployment time, either through optimization-based parameter updates or training-free prediction refinement [17,18,22,40]. More recently, Retrieval-Augmented TTA [12,16,22,27] incorporates retrieved evidence to enhance test-time inference. TT-RAA [16] maintains an internal memory that compresses the incoming test stream into summary representations such as feature-level prototypes or running distribution statistics, and uses this memory to correct VLM predictions online. Unlike TT-RAA, RA-TTA [27] relies on external retrieval from a large-scale image pool using fine-grained textual descriptions, and aggregates the retrieved evidence to refine predictions under distribution shift. Existing approaches are mostly studied under closed-world recognition with a fixed label set and mainly focus on output-side calibration or prediction correction. Some retrieval-based methods further rely on large external retrieval corpora. *In contrast, our framework constructs an internal memory from incoming test samples and injects instance-level comparative context with attribute-level evidence into inference, using diversity-aware updates to keep the memory informative with limited test-time overhead.*

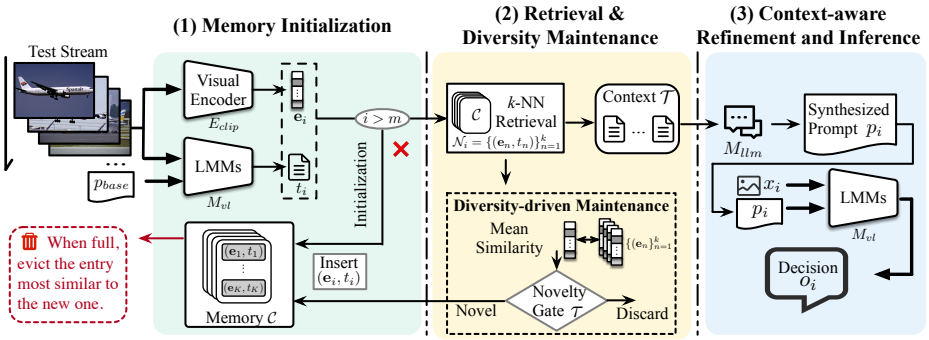

Fig. 2: Overview of the AMCI. (1) **Memory Initialization**: For each incoming image, AMCI builds a visual index e_i and generates an attribute-aware description t_i to initialize the memory $\mathcal{C}$. (2) **Retrieval and Diversity Maintenance**: AMCI retrieves the k-nearest neighbors as the context T. A novelty gate with an adaptive threshold τ and a replacement strategy of the most visually similar entry are utilized to keep memory diverse and avoid information redundancy. (3) **Context-aware Refinement and Inference**: The refinement LLM M_{llm} summarizes the retrieved context T and outputs a synthesized prompt p_i. The LMM M_{vl} predicts the final decision o_i based on the augmented prompt.

3 Method

Problem Definition. We denote an LMM as a function f_{LMM} that maps an input image $x \in \mathcal{X}$ and a textual prompt $p \in \mathcal{T}$ to a textual output $y = f_{LMM}(x, p) \in \mathcal{T}$. In the conventional closed-world setting, the output is restricted to a predefined, static label set $\mathcal{K} \subset \mathcal{T}$. In contrast, the OWC setting allows the model to operate within its original, unconstrained semantic space $\mathcal{T}$, where the output y is expected to accurately describe a specific semantic entity Y encountered in the open world. This setup requires the model to capture subtle discriminative cues to yield accurate semantic descriptions without relying on a fixed category list.

Framework Overview. The proposed AMCI framework unlocks the potential of LMMs for fine-grained OWC by transitioning the model from isolated, stateless inference to context-aware adaptation. By leveraging a dynamic memory, AMCI enables context-aware refinement to recover discriminative details often lost in standard zero-shot settings. The overall pipeline is illustrated in Fig. 2, and the detailed algorithm is provided in the Appendix.

3.1 Memory Initialization

Motivation. Initial empirical tests reveal that one limitation of LMMs is their limited ability to make use of contextual information when incoming instances are processed in isolation. Without sufficient context to serve as a referential anchor, the model fails to align specialized, domain-specific visual features with its pre-trained knowledge. To address the issue, we implement this stage to

provide the contextual foundation necessary for adaptation. We maintain a non-parametric memory $\mathcal{C} = \{(\mathbf{e}_i, t_i)\}_{i=1}^{M}$, where $\mathbf{e}_i \in \mathbb{R}^D$ denotes the visual indexing and $t_i \in \mathcal{T}$ represents the attribute-augmented description. By establishing this foundational memory, AMCI can effectively reuse fine-grained context, thereby unlocking the capability of LMMs to distinguish between highly similar entities under an unsupervised setting.

Visual Indexing Construction. To provide a robust mechanism for retrieval, we first compute a visual embedding $\mathbf{e}_i = E_{clip}(x_i)$ for each query x_i utilizing a frozen CLIP encoder. This process maps the raw pixel data into a structured embedding space where visually similar entities are positioned in close proximity. These embeddings serve as the search **keys** of the memory $\mathcal{C}$, allowing the framework to perform similarity-based retrieval during the inference stream.

Attribute-Augmented Description Generation. We focus on augmenting the memory with **values** that capture the discriminative attributes of perceived samples. For each instance x_i, we prompt the LMM to generate an attribute-augmented description $t_i = M_{vl}(x_i, p_{base})$ that is stored as the corresponding value in the **key-value pair** $(\mathbf{e}_i, t_i)$. The prompt p_{base} is defined as:

> *"What type of object is in this photo? You may include a brief mention of key visual attributes if it helps recognition."*

By explicitly encouraging the model to surface specific textures, shapes, or structural parts, we ensure that each entry contains rich semantic context. The resulting key-value pair $(\mathbf{e}_i, t_i)$ is then inserted into $\mathcal{C}$, forming the foundational repository for context-aware inference. This pairing of **visual keys** and **textual values** supports the later retrieval stage, enabling AMCI to exploit contextual information to align fine-grained visual features with LMMs' knowledge.

3.2 Retrieval and Diversity Maintenance

Motivation. As the memory $\mathcal{C}$ begins to accumulate data from the test stream, its utility is primarily determined by the **relevance** of the retrieved context and the **diversity** of the stored information. We observe that the performance of the framework degrades significantly when these two factors are neglected by naive policies. Specifically, the performance decline is caused by two factors associated with unconstrained memory updates. First, the lack of similarity constraints during retrieval, typically seen in random retrieval, introduces irrelevant samples that act as semantic noise, which distracts the inference of the LMMs (analyzed in Sec. 5.3). Second, a simple chronological update strategy, such as a first-in-first-out (FIFO) policy, can make the memory dominated by redundant samples from majority classes, thereby weakening its ability to preserve discriminative cues for rare categories (shown in Sec. 5.3). To resolve these issues, we implement a process that implements relevance-based retrieval and prioritizes informational novelty during memory updates, effectively balancing the quality and diversity of the stored context.

Relevance-based Retrieval. To ensure that the context is relevant to the current instance, AMCI performs a k-nearest neighbor search within the memory $\mathcal{C}$ based on visual similarity. Utilizing the visual index $\mathbf{e}_i$ as a query key, we

identify the set of neighbors $\mathcal{N}_i = \{(\mathbf{e}_n, t_n)\}_{n=1}^{k}$ that exhibit the highest cosine similarity to the current query. These retrieved descriptions $\{t_n\}_{n=1}^{k}$ serve as contextual references for the query, facilitating the transition from isolated zero-shot prediction to context-aware inference. By anchoring the current sample in visually correlated context, the LMMs can more effectively distinguish specialized fine-grained categories.

Diversity-driven Maintenance. To prevent the memory from being dominated by redundant samples, we adopt a maintenance strategy that prioritizes informational novelty. The novelty of a new sample is quantified by the mean similarity s_i to the retrieved neighbors:

$$s_i = \frac{1}{k} \sum_{(\mathbf{e}_n, t_n) \in \mathcal{N}_i} \cos(\mathbf{e}_i, \mathbf{e}_n) \tag{1}$$

We employ an adaptive threshold τ, updated via Exponential Moving Average (EMA) as $\tau = (1 - \lambda)\tau + \lambda s_i$, to act as a gatekeeper: a sample is only admitted to the memory if $s_i < \tau$. Furthermore, when the capacity K of the memory is reached, we maintain the memory by replacing the existing entry j^* that is most visually similar to the current query:

$$j^* = \underset{e_j \in \mathcal{C}}{\operatorname{argmax}} \cos(\mathbf{e}_i, \mathbf{e}_j) \tag{2}$$

This strategy helps the memory retain discriminative fine-grained visual cues while avoiding redundancy, which is important for OWC inference.

3.3 Context-aware Refinement and Inference

Motivation. The design of this stage is motivated by the observation that the accumulated context, although relevant, often lacks the structure necessary for final prediction. As shown in the ablation studies of Sec. 5.3, a naive concatenation of retrieved descriptions fails to provide a structured prompt, as the LMMs struggle to capture the relationships between multiple descriptions. Consequently, we propose to leverage a refinement LLM to transform the accumulated context into a structured prompt. This stage allows the framework to exploit the context while keeping all parameters frozen, and helps the LMMs to make predictions without relying on predefined labels.

Context-aware Prompt Refinement. Following the retrieval of k neighbors within the memory $\mathcal{C}$, we obtain a set of textual descriptions $T = \{t_n\}_{n=1}^{k}$. Instead of directly using the retrieved context, we use a refinement LLM M_{llm} to reorganize T into a synthesized, attribute-aware prompt p_i. This prompt emphasizes visual attributes and domain-specific cues relevant to the current query x_i. Formally, the synthesized prompt is generated as:

$$p_i = M_{llm}(\text{RefinePrompt}, T) \tag{3}$$

where *RefinePrompt* denotes the instruction used to reorganize the retrieved textual descriptions.

Comprehensive Inference. The synthesized prompt p_i is then used as the textual input, together with the query image x_i, for the final prediction:

$$o_i = M_{vl}(x_i, p_i) \tag{4}$$

AMCI helps the LMMs to distinguish visually similar instances by converting memory entries into a targeted, attribute-aware prompt, while keeping all model parameters frozen.

4 Fine-Grained OWC Benchmark and Evaluation Metrics

Dataset Composition and Scenarios. Our benchmark comprises seven datasets categorized into two distinct groups based on semantic granularity and data scarcity. (i) **Fine-grained:** We utilize *Flowers102* [37], *Food101* [5], *Oxford Pets* [38], *Stanford Cars* [23], and *FGVC Aircraft* [36], all of which demand fine-grained discrimination across closely related visual categories. (ii) **Low-resource:** To evaluate the framework under extreme distribution shifts and data scarcity, we incorporate two specialized datasets: *iNaturalist 2021* [42, 46], featuring rare biological data with an extreme long-tail distribution, and *HAM10000* [14, 41], consisting of specialized dermatoscopic images for skin lesion analysis. These benchmarks represent the critical challenges of open-world understanding where fine-grained knowledge is essential and visual priors are inherently sparse. Statistics are provided in the Appendix.

Evaluation Metrics. Standard lexical matching is inadequate for OWC, as LMMs often generate verbose responses that encompass the correct concept yet fail strict string matching (e.g., "a majestic sofa" vs. "sofa"). To ensure a rigorous assessment, we adopt a multi-dimensional metric suite: (i) **Text Inclusion (TI)** checks for strict string containment of the ground-truth label within the prediction; (ii) **Categorical Fidelity Inclusion (CFI)** utilizes an LLM-judge to evaluate whether the prediction achieves categorical alignment while penalizing generic or vague responses; (iii) **Semantic Similarity (SS)** measures global alignment by calculating the cosine similarity between the Sentence-BERT embeddings of the prediction and the label; and (iv) **Concept Similarity (CS)** captures maximum local alignment by identifying the specific phrase within the response that best matches the label embedding, thereby mitigating the impact of redundant descriptions.

Motivation of the new CFI metrics. We find that the Llama Inclusion (LI) metric proposed in OWC [11] tends to reward overly broad descriptions rather than fine-grained precision (e.g., predicting "airplane" for "Boeing 707-320"). To address this issue, we introduce CFI, which penalizes generic predictions and favors informative, subtype-specific outputs. CFI is defined by four rules. 1) **Anti-ambiguous gating** assigns a score of 0 if the prediction only provides a parent category without subtype evidence. 2) **Evidence-based verification** uses an LLM-judge (Llama-3.2-3B) to check whether discriminative visual traits are present. 3) **Synonym awareness** validates scientific names, professional aliases, and domain-specific synonyms. 4) **Contradiction veto** assigns a score of

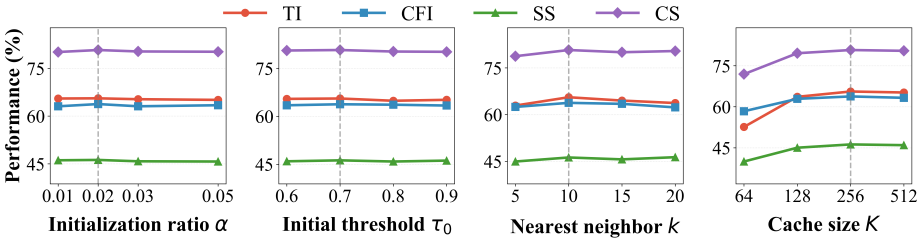

Fig. 3: Sensitivity analysis of key hyper-parameters.

0 if the prediction contradicts the target category or its defining characteristics. By prioritizing categorical fidelity over generic correctness, CFI provides a more reliable metric for fine-grained OWC.

5 Experiments

5.1 Experimental Setup

We follow the same setting as [11] and evaluate the performance of 13 publicly available LMMs across diverse architectures and parameter scales: Idefics2 [26], InstructBLIP [13], InternVL2 (at 2B, 4B, and 8B scales) [7, 8], LLaVA-1.5 [33], LLaVA-NeXT [32] (with Mistral-7B and Vicuna-7B backbones), LLaVA-OV [28] (with Qwen2-0.5B and Qwen2-7B variants), Phi-3-Vision [1], and Qwen2-VL [43] (available in 2B and 7B variants). Unless otherwise stated, all models utilize the same prompt described in Sec. 3.1. We utilize CLIP-ViT-L/14 for visual feature extraction and retrieval, while Llama-3-8B-Instruct serves as the refinement LLM for contextual prompt synthesis. Hyperparameter settings are kept uniform across all experiments: initialization ratio $\alpha = 0.02$, maximum cache size $K = 256$, initial threshold $\tau = 0.7$, and nearest neighbor count $k = 10$. AMCI introduces no model training or fine-tuning, operating purely through the three-stage inference process introduced in Sec. 3.

5.2 Comparison with State-of-the-Art OWC Methods

As shown in Tab. 1, integrating the AMCI framework consistently improves performance across all evaluated LMMs with model scales ranging from 0.5B to 8B parameters. On average across 13 configurations, AMCI yields notable gains in Fine-grained tasks, including 8.2% in TI, 16.3% in CFI, 3.5% in SS, and 10.2% in CS. In Low-resource scenarios, AMCI also achieves stable improvements of 0.7% in TI, 7.9% in CFI, 2.3% in SS, and 5.4% in CS. Also, there is **significant improvements on CFI metric**. For example, LLAVA-OV (Qwen2 7B) exhibits a CFI increase from 13.9% to 47.6%, demonstrating AMCI's effectiveness in mitigating the generic prediction bias of vanilla LMMs. By injecting streaming contextual evidence, AMCI enables models to shift from overly general descriptions to more discriminative predictions. Notably, AMCI is especially **effective for smaller models**. For instance, LLAVA-OV (0.5B) enhanced with AMCI

Table 1: Overall performance on Fine-grained and Low-resource scenarios. We report the average results across five Fine-grained datasets (Flowers102, Food101, Oxford Pets, Stanford Cars, FGVC Aircraft) and two Low-resource datasets (HAM10000, iNaturalist).

	Fine-grained				Low-resource			
Model	**TI**	**CFI**	**SS**	**CS**	**TI**	**CFI**	**SS**	**CS**
IDEFICS2 [26] 8B	1.8	8.0	34.6	38.5	0.0	21.5	30.2	33.1
+ **AMCI (Ours)**	4.5	26.2	36.6	40.5	0.1	28.4	31.9	35.4
INSTRUCTBLIP [13] Vicuna 7B	6.2	18.4	33.4	42.0	0.0	20.2	29.2	31.0
+ **AMCI (Ours)**	15.3	47.8	35.5	54.8	0.3	32.2	29.5	38.4
PHI-3-VISION [1]	8.1	31.3	30.6	42.7	0.0	27.8	29.0	34.8
+ **AMCI (Ours)**	18.0	51.3	35.8	55.6	0.3	37.0	28.9	39.6
INTERNVL2 [7, 8] 2B	9.2	37.7	32.2	48.0	0.6	33.4	33.3	43.8
+ **AMCI (Ours)**	11.2	46.2	33.1	53.0	0.8	42.5	34.2	46.3
INTERNVL2 [7, 8] 4B	10.4	37.5	32.7	48.9	0.1	31.5	33.4	43.1
+ **AMCI (Ours)**	15.5	48.9	34.9	55.3	0.6	40.3	34.6	48.4
INTERNVL2 [7, 8] 8B	14.3	44.1	35.3	53.7	3.9	37.7	33.7	47.8
+ **AMCI (Ours)**	19.1	51.4	36.5	58.8	4.1	**44.1**	34.9	50.1
LLAVA-1.5 [33] 7B	5.1	40.7	28.3	41.9	0.1	29.6	26.7	37.4
+ **AMCI (Ours)**	9.9	47.2	32.2	50.8	0.1	36.0	27.3	39.1
LLAVA-NEXT [32] (Mistral 7B)	16.6	50.2	34.8	54.8	0.1	34.0	29.4	39.9
+ **AMCI (Ours)**	17.2	51.4	35.8	56.2	0.1	34.7	29.6	40.7
LLAVA-NEXT [32] (Vicuna 7B)	10.7	42.9	33.2	50.3	0.5	33.6	28.9	38.5
+ **AMCI (Ours)**	14.9	51.2	35.9	56.7	0.4	38.5	31.2	42.8
LLAVA-OV [28] (Qwen2 0.5B)	3.9	17.2	35.3	40.8	0.3	25.1	27.8	35.5
+ **AMCI (Ours)**	13.8	42.4	39.1	54.7	0.9	35.1	29.6	41.4
LLAVA-OV [28] (Qwen2 7B)	3.9	13.9	36.2	39.2	0.1	25.1	24.0	30.6
+ **AMCI (Ours)**	18.6	47.6	39.6	60.0	0.2	37.0	31.9	43.1
QWEN2VL [43] 2B	26.6	47.6	42.5	63.0	1.0	30.8	29.9	40.5
+ **AMCI (Ours)**	**45.0**	**63.1**	**50.3**	**79.2**	5.1	38.5	36.4	52.1
QWEN2VL [43] 7B	21.0	36.6	37.3	55.1	1.8	31.7	32.1	43.5
+ **AMCI (Ours)**	41.5	63.0	47.2	75.7	**5.2**	40.2	**37.2**	**52.9**
Open-world baselines								
CaSED [9]	16.7	21.3	51.8	52.4	1.8	15.1	30.0	31.9
CLIP RETRIEVAL	22.3	24.7	41.6	61.7	3.2	17.6	33.6	41.1
Closed-world baselines								
CLIP [39]	71.7	54.5	83.2	83.2	38.1	42.8	73.5	73.5
SigLIP [44]	83.2	67.6	92.7	92.7	51.8	56.9	86.4	86.4

achieves 13.8% TI, surpassing several larger vanilla models such as InstructBLIP (6.2%) and Phi-3-Vision (8.1%), indicating that context-aware adaptation can partially compensate for limited model capacity in fine-grained open-world scenarios. Compared with **retrieval-based baselines**, AMCI provides stronger semantic refinement by synthesizing candidate evidence into attribute-aware predictions, leading to higher CS and SS scores. In low-resource settings, where vanilla LMMs often fail, AMCI enables meaningful prediction capability (e.g., Qwen2VL 7B reaching 5.2% TI), highlighting the importance of leveraging local distribution cues when global priors are unavailable. **Summary:** AMCI consistently improves fine-grained open-world classification across model scales and scenarios, with particularly strong gains in low-resource robustness.

Table 2: Ablation study based on Qwen2-VL 7B. (a) Core modules. (b) Memory update policy. (c) Retrieval method. (d) Prompt and querying.

		Oxford Pets				FGVC Aircraft				HAM10000			
	Configuration	**TI**	**CFI**	**SS**	**CS**	**TI**	**CFI**	**SS**	**CS**	**TI**	**CFI**	**SS**	**CS**
(a)	w/o Memory	52.6	58.4	40.0	71.9	18.2	45.6	28.4	54.2	1.2	22.4	26.8	35.1
	w/o LLM Refinement	16.0	39.5	26.0	48.9	1.3	36.8	21.4	33.5	0.3	18.5	24.2	30.6
	w/o Memory & w/o LLM	12.3	14.8	25.2	44.6	1.7	5.1	21.9	30.9	3.4	31.2	36.1	45.8
(b)	w/ FIFO Replacement	61.2	60.8	43.1	77.4	30.5	52.1	33.5	63.8	6.4	29.5	39.8	51.2
	w/ LRU Replacement	60.8	61.4	43.8	76.5	32.8	51.5	34.2	62.4	7.1	29.2	37.4	53.4
	w/ Fixed Threshold	64.1	63.2	45.6	78.4	34.5	56.4	35.1	65.9	8.4	33.1	44.0	56.2
(c)	w/ Random Retrieval	53.1	58.2	40.3	72.3	28.5	48.1	30.2	56.4	4.2	27.5	35.2	46.8
	w/ Top-1 Retrieval	60.4	59.5	42.8	75.5	31.2	54.2	32.1	61.4	6.8	31.4	38.5	50.8
	w/ CLIP Textual	64.8	62.5	44.8	79.8	35.1	56.8	36.4	64.5	7.8	33.9	42.5	55.4
(d)	w/ Minimal Prompt	48.2	52.1	39.7	67.1	26.5	46.2	31.4	55.2	5.1	28.4	36.2	48.5
	w/ Decomposed Querying	63.8	62.9	45.2	79.5	34.2	56.1	35.8	64.2	8.2	33.6	43.1	56.2
	Full AMCI (Ours)	**65.6**	**63.8**	**46.2**	**80.7**	**36.0**	**57.7**	**37.0**	**66.7**	**8.7**	**34.2**	**44.6**	**57.7**

Table 3: Comparison of AMCI with CoT prompting methods across three datasets.

	Oxford Pets				FGVC Aircraft				HAM10000			
Model & Method	**TI**	**CFI**	**SS**	**CS**	**TI**	**CFI**	**SS**	**CS**	**TI**	**CFI**	**SS**	**CS**
Qwen2-VL 7B	12.1	13.5	25.1	43.1	1.4	4.6	20.6	30.8	3.3	29.4	35.3	44.5
w/ Zero-shot CoT	19.2	39.8	21.3	50.9	4.1	49.9	20.3	38.0	1.9	33.0	29.9	43.6
w/ LLaVA CoT	21.3	41.5	21.4	53.1	12.8	52.9	21.5	47.2	0.7	31.1	21.8	37.9
+ **AMCI (Ours)**	**65.6**	**63.8**	**46.2**	**80.7**	**36.0**	**57.7**	**37.0**	**66.7**	**8.7**	**34.2**	**44.6**	**57.7**
LLaVA-1.5 7B	1.1	29.2	21.5	35.9	0.1	43.7	19.0	29.0	0.1	26.3	28.0	33.5
w/ Zero-shot CoT	4.8	29.2	20.6	42.4	0.1	43.8	17.0	31.7	0.1	26.9	27.6	34.4
w/ LLaVA CoT	4.8	29.2	20.6	42.4	0.1	43.8	17.0	31.7	0.1	26.9	27.6	34.4
+ **AMCI (Ours)**	**18.2**	**47.1**	**32.1**	**53.8**	**0.5**	**48.6**	**20.7**	**33.6**	**0.1**	**31.4**	**28.5**	**36.2**
InternVL2 8B	13.8	20.7	27.3	46.0	4.6	47.4	25.5	42.5	7.7	37.1	40.1	52.4
w/ Zero-shot CoT	16.4	39.2	23.7	50.3	2.5	50.6	19.3	38.0	6.3	35.8	39.0	54.9
w/ LLaVA CoT	10.9	36.7	18.0	47.3	3.0	**50.7**	18.4	39.6	3.6	35.0	31.3	49.5
+ **AMCI (Ours)**	**37.8**	**53.9**	**36.9**	**64.5**	**5.6**	50.2	**27.0**	**46.8**	**8.0**	**47.4**	**42.8**	**56.5**

5.3 Ablation Study

Component Ablation. To evaluate the contribution of each component in AMCI, we conduct an ablation study as shown in Tab. 2 (a). Removing the memory module (w/o Memory) results in an average performance drop of 12.0% across the three datasets, highlighting the importance of contextual reference during inference. After removing the LLM refinement module (w/o LLM Refinement), the performance further decreases by 25.2%, indicating that refinement is essential for mitigating semantic noise in raw descriptions and improving fine-grained recognition. When both components are removed (w/o Memory & w/o LLM), the model suffers the largest decline, with an average performance drop of 27.2%. **Summary:** Both the memory module and LLM refinement play critical and complementary roles in AMCI, jointly enabling effective contextual reasoning and fine-grained prediction.

Table 4: Comparison with Retrieval-Augmented TTA methods. **CL** (Class Label) indicates whether the method requires pre-defined class names during inference (✓) or operates in a label-free open-world setting (×).

		Oxford Pets				FGVC Aircraft				HAM10000			
Model & Method	**CL**	**TI**	**CFI**	**SS**	**CS**	**TI**	**CFI**	**SS**	**CS**	**TI**	**CFI**	**SS**	**CS**
FreeTTA [12]	✓	**66.2**	**64.1**	45.8	80.4	**36.5**	**58.2**	**37.4**	**67.2**	**8.9**	**34.5**	**44.9**	**58.2**
TT-RAA [16]	✓	65.8	63.9	**46.5**	79.9	36.2	57.9	37.1	66.9	8.8	34.3	44.7	57.9
TDA [22]	✓	58.1	57.5	41.2	72.8	30.4	52.1	32.8	59.5	5.4	28.7	38.1	48.6
RA-TTA [27]	✓	62.4	61.2	44.1	78.2	33.1	54.2	34.5	62.4	7.2	31.5	40.2	52.1
Qwen2-VL 7B	×	12.1	13.5	25.1	43.1	1.4	4.6	20.6	30.8	3.3	29.4	35.3	44.5
+ **AMCI (Ours)**	×	65.6	63.8	46.2	**80.7**	36.0	57.7	37.0	66.7	8.7	34.2	44.6	57.7
FreeTTA [12]	✓	**38.5**	**54.1**	**37.5**	**65.2**	**6.2**	**51.2**	**28.1**	**48.5**	**8.5**	**48.2**	**43.5**	**57.4**
TT-RAA [16]	✓	38.2	53.8	37.2	64.9	5.9	50.8	27.5	47.6	8.2	47.8	43.1	56.9
TDA [22]	✓	29.8	47.2	31.5	55.4	3.5	47.8	23.2	41.5	6.5	42.4	38.5	50.1
RA-TTA [27]	✓	32.1	50.4	33.2	58.7	4.1	48.5	24.8	43.1	7.1	44.1	40.2	53.2
InternVL2 8B	×	13.8	20.7	27.3	46.0	4.6	47.4	25.5	42.5	7.7	37.1	40.1	52.4
+ **AMCI (Ours)**	×	37.8	53.9	36.9	64.5	5.6	50.2	27.0	46.8	8.0	47.4	42.8	56.5

Table 5: Inference latency and performance comparison between LlamaV-o1 and our AMCI. Time represents the total hours (h) required for inference on the full test set using a single H100 GPU.

	Oxford Pets					FGVC Aircraft				
Model & Method	**TI**	**CFI**	**SS**	**CS**	**Time**	**TI**	**CFI**	**SS**	**CS**	**Time**
Qwen2-VL 7B										
w/ LlamaV-o1	35.2	45.7	31.4	59.5	6.6	15.5	53.4	28.0	52.0	6.9
+ **AMCI (Ours)**	**65.6**	**63.8**	**46.2**	**80.7**	**1.4**	**36.0**	**57.7**	**37.0**	**66.7**	**2.1**
LLaVA-1.5 7B										
w/ LlamaV-o1	4.8	29.2	20.6	42.4	5.9	0.1	43.8	17.0	31.7	6.9
+ **AMCI (Ours)**	**18.2**	**47.1**	**32.1**	**53.8**	**2.0**	**0.5**	**48.6**	**20.7**	**33.6**	**1.9**

Memory Management Strategy. As shown in Tab. 2 (b), we compare different memory update policies. Traditional FIFO and LRU algorithms led to average performance decreases of 4.1% and 4.0%, respectively, demonstrating that in fine-grained scenarios, simple replacement mechanisms randomly discard visual anchors of high discriminative value. Furthermore, using a fixed similarity threshold (*Fixed Threshold*) resulted in an average performance drop of 1.2%. By utilizing an EMA-based dynamic threshold, Full AMCI adaptively regulates the memory update process based on the distribution of incoming samples, effectively accumulating fine-grained visual evidence that supports more precise and robust differentiation.

Retrieval Strategy Ablation. As shown in Tab. 2 (c), we examine whether retrieval is necessary. Replacing retrieval with random sampling (*w/ Random Retrieval*) results in an average performance drop of 8.2%. We further study the retrieval granularity and find that using only the single most similar neighbor (*w/ Top-1 Retrieval*) leads to a 4.5% average decrease. Finally, we compare retrieval modalities: retrieving with CLIP textual features (*w/ CLIP Textual*) performs slightly worse than visual-feature retrieval, with a 1.2% average drop.

This can be attributed to the fact that images in open-world classification often preserve richer fine-grained structural cues than coarse text descriptions, making visual-embedding similarity search a more reliable reference for M_{vl}.

Prompt Design and Interaction Ablation. As shown in Tab. 2(d), using a minimal prompt (*w/ Minimal Prompt*) causes a 9.5% average performance drop, indicating that prompt design serves as a crucial quality controller for the streaming reference system. Weak prompts produce generic descriptions that contaminate the memory bank and hinder long-term adaptation. We also evaluate a decomposed querying strategy (*w/ Decomposed Querying*), which splits the attribute prompt into two sequential questions and concatenates the answers. This results in a 1.3% average decline, suggesting that simple answer stacking introduces redundancy and weakens discriminative cues.

5.4 Analysis and Visualization

Hyper-Parameters Analysis We conduct a sensitivity analysis on the key hyper-parameters of our AMCI, including the initialization ratio α, initial threshold τ_0, nearest neighbor count k, and cache size K, as illustrated in Fig. 3. All experiments are performed using Qwen2-VL-7B on the Oxford Pets dataset. The experimental results indicate that α, τ_0, and k exhibit high robustness across a broad range, with minimal fluctuations in the four metrics. Consequently, we set $\alpha = 0.02$, $\tau_0 = 0.7$, and $k = 10$. Regarding the cache size K, the performance generally improves as K increases. Considering the balanced performance across all four metrics, we select $K = 256$. These hyper-parameters are fixed and applied to all other models and datasets.

Comparison with Chain-of-Thought (CoT) Prompting As demonstrated in Tab. 3, our AMCI consistently outperforms mainstream CoT methods across all experimental settings. While Zero-shot CoT and LLaVA CoT improve classification performance to some extent by introducing reasoning chains, their gains on TI and SS remain overall limited. More importantly, CoT methods exhibit stagnant performance under the CFI metric. This indicates that while CoT increases the length of reasoning, it fails to capture the essential discriminative evidence required for fine-grained recognition. In contrast, AMCI achieves significant performance improvements. Across the three datasets and three base models, AMCI achieves average improvements of 11.7% and 12.3% over Zero-shot CoT and LLaVA CoT on the four metrics, respectively. Unlike CoT, which relies on the model's reasoning paths, AMCI leverages a discriminative feedback mechanism and a dynamic cache to provide reliable, domain-specific contextual information.

Comparison with Retrieval-Augmented TTA. We compare AMCI with TTA-based approaches [12, 16, 22, 27] in Tab. 4. Since they are not directly comparable, we adapt them to our framework by **introducing extra ground-truth labels.** Implementation details are provided in the Appendix. AMCI achieves a mean score of 44.8 across the four metrics, which is comparable to FreeTTA (45.4) and TT-RAA (45.1). Notably, methods such as RA-TTA and TT-RAA

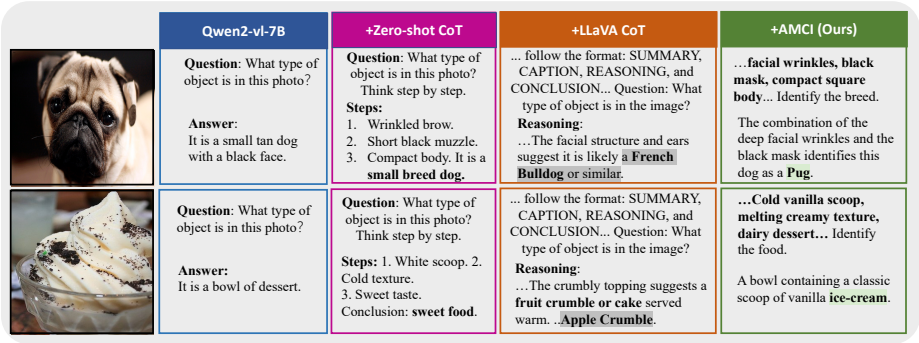

Fig. 4: Qualitative comparison. AMCI reduces the generic vagueness of baselines and the incorrect fine-grained predictions of LLaVA-CoT (e.g., misidentifying ice-cream as apple crumble) by injecting retrieved context.

rely on class-name vocabulary to form a CLIP closed-set distribution that provides confidence or soft-assignment signals. In contrast, AMCI attains nearly the same performance as these methods without requiring any label vocabulary or additional external retrieval corpus.

Performance and Computational Efficiency. As shown in Tab. 5, we compare AMCI with the long-chain reasoning baseline LlamaV-o1. Averaged over two base models and two fine-grained datasets, LlamaV-o1 takes 6.58h per evaluation, while AMCI takes 1.85h, yielding a 3.5× speedup. Meanwhile, AMCI achieves a higher mean score across the four metrics (44.3 vs. 31.9), showing a better trade-off between effectiveness and efficiency. We additionally provide a latency ablation in the supplementary material, which reports the runtime of retrieval, memory update, refinement, and LMM inference in AMCI.

Visualization As shown in Fig. 4, the baseline often produces overly generic predictions (e.g., identifying a "pug" as "a small dog"), leading to imprecise recognition. Even complex reasoning prompts such as LLaVA-CoT may introduce incorrect fine-grained predictions, such as misclassifying "ice-cream" as "apple crumble" due to distracting textures. In contrast, AMCI alleviates this issue by leveraging discriminative attributes (e.g., "facial wrinkles" or "creamy texture"). This demonstrates that structured context injection is more effective for fine-grained OWC than increasing prompt complexity or reasoning steps.

6 Conclusion

We empirically find that LMMs' fine-grained OWC performance is often limited due to the lack of contextual information, resulting in overly generic predictions. To address this issue, we propose AMCI, which builds an internal memory from test streams and injects contextual information during inference, enabling more specific predictions in a training-free fashion. We propose the CFI metric and construct a low-resource fine-grained benchmark. Extensive experiments demonstrate the effectiveness of AMCI, highlighting the potential of context-driven, training-free adaptation for enhancing LMM specificity.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Nos. 62572166 and 62402157), the Fundamental Research Funds for the Central Universities (No. JZ2025HGTB0219), the EU Horizon Europe projects ELIAS (No. 101120237) and ELLIOT (No. 101214398), and the FIS project GUIDANCE (No. FIS2023-03251). The computations were performed on the HPC Platform of Hefei University of Technology.

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.C., Chen, Y.L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A.D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R.J., Hu, W., Huynh, J., Iter, D., Jacobs, S.A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y.J., Kurilenko, L., Lee, J.R., Lee, Y.T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C.C.T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L.L., Zhang, Y., Zhang, Y., Zhang, Y., Zhou, X.: Phi-3 technical report: A highly capable language model locally on your phone (2024), `https://arxiv.org/abs/2404.14219`
2. Angheben, S., Berasi, D., Conti, A., Ricci, E., Wang, Y.: Specificity-aware reinforcement learning for fine-grained open-world classification. In: CVPR. pp. 41467–41477 (2026)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), `https://arxiv.org/abs/2308.12966`
4. Bendale, A., Boult, T.: Towards open world recognition. In: CVPR. pp. 1893–1902 (2015)
5. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: ECCV. pp. 446–461. Springer (2014)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PmLR (2020)
7. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
9. Conti, A., Fini, E., Mancini, M., Rota, P., Wang, Y., Ricci, E.: Vocabulary-free image classification. NeurIPS **36**, 30662–30680 (2023)
10. Conti, A., Fini, E., Mancini, M., Rota, P., Wang, Y., Ricci, E.: Vocabulary-free image classification and semantic segmentation. IEEE TPAMI (2026)
11. Conti, A., Mancini, M., Fini, E., Wang, Y., Rota, P., Ricci, E.: On large multimodal models as open-world image classifiers. In: ICCV. pp. 16388–16398 (2025)
12. Dai, Q., Yang, S.: Free on the fly: Enhancing flexibility in test-time adaptation with online em. In: CVPR. pp. 9538–9548 (2025)

13. 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. NeurIPS **36**, 49250–49267 (2023)
14. Dall'Asen, N., Wang, Y., Fini, E., Ricci, E.: Retrieval-enriched zero-shot image classification in low-resource domains. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21287–21302 (2024)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
16. Fan, X., Chen, X., Yang, L., Yap, C.H., Qureshi, R., Dou, Q., Yap, M.H., Shah, M.: Test-time retrieval-augmented adaptation for vision-language models. In: ICCV. pp. 8810–8819 (2025)
17. Feng, C.M., Yu, K., Liu, Y., Khan, S., Zuo, W.: Diverse data augmentation with diffusions for effective test-time prompt tuning. In: ICCV. pp. 2704–2714 (2023)
18. Gao, Y., Shi, X., Zhu, Y., Wang, H., Tang, Z., Zhou, X., Li, M., Metaxas, D.N.: Visual prompt tuning for test-time domain adaptation. arXiv preprint arXiv:2210.04831 (2022)
19. Garosi, M., Farina, M., Conti, A., Mancini, M., Ricci, E.: Large multimodal models as general in-context classifiers. arXiv preprint arXiv:2602.23229 (2026)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
21. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. NeurIPS **36**, 31096–31116 (2023)
22. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. In: CVPR. pp. 14162–14171 (2024)
23. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
24. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS **25** (2012)
26. Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? NeurIPS **37**, 87874–87907 (2024)
27. Lee, Y., Kim, D., Kang, J., Bang, J., Song, H., Lee, J.G.: Ra-tta: Retrieval-augmented test-time adaptation for vision-language models. In: ICLR (2025)
28. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
29. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: CVPR. pp. 13299–13308 (2024)
30. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742. PMLR (2023)
31. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR. pp. 22195–22206 (2024)
32. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`

33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
34. Liu, H., Xiao, L., Liu, J., Li, X., Feng, Z., Yang, S., Wang, J.: Revisiting mllms: An in-depth analysis of image classification abilities. arXiv preprint arXiv:2412.16418 (2024)
35. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: ECCV. pp. 216–233. Springer (2024)
36. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
37. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
38. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR. pp. 3498–3505. IEEE (2012)
39. 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: ICML. pp. 8748–8763. PmLR (2021)
40. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. NeurIPS **35**, 14274–14289 (2022)
41. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 180161 (2018)
42. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: CVPR. pp. 8769–8778 (2018)
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
44. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
45. Zhang, Y., Unell, A., Wang, X., Ghosh, D., Su, Y., Schmidt, L., Yeung-Levy, S.: Why are visually-grounded language models bad at image classification? NeurIPS **37**, 51727–51753 (2024)
46. Zhang, Y., Doughty, H., Snoek, C.G.M.: Low-resource vision challenges for foundation models. In: CVPR. pp. 21956–21966 (June 2024)