Large Multimodal Models as General In-Context Classifiers

Marco Garosi[1] Matteo Farina[1] Alessandro Conti[1] Massimiliano Mancini[1] Elisa Ricci[1,2]
[1]University of Trento [2]Fondazione Bruno Kessler
https://circle-lmm.github.io

Abstract

Which multimodal model should we use for classification? Previous studies suggest that the answer lies in CLIP-like contrastive Vision-Language Models (VLMs), due to their remarkable performance in zero-shot classification. In contrast, Large Multimodal Models (LMM) are more suitable for complex tasks. In this work, we argue that this answer overlooks an important capability of LMMs: in-context learning. We benchmark state-of-the-art LMMs on diverse datasets for closed-world classification and find that, although their zero-shot performance is lower than CLIP's, LMMs with a few in-context examples can match or even surpass contrastive VLMs with cache-based adapters, their "in-context" equivalent. We extend this analysis to the open-world setting, where the generative nature of LMMs makes them more suitable for the task. In this challenging scenario, LMMs struggle whenever provided with imperfect context information. To address this issue, we propose CIRCLE, a simple training-free method that assigns pseudo-labels to in-context examples, iteratively refining them with the available context itself. Through extensive experiments, we show that CIRCLE establishes a robust baseline for open-world classification, surpassing VLM counterparts and highlighting the potential of LMMs to serve as unified classifiers, and a flexible alternative to specialized models.

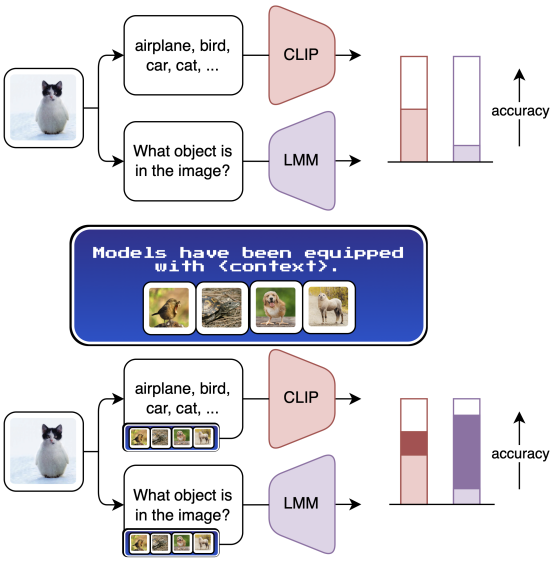

Figure 1. **The role of context in classification.** CLIP-like models outperform Large Multimodal Models (LMMs) in closed-world classification. However, we show that context dramatically unlocks LMM performance in closed-world classification while allowing them to surpass VLMs on open-world classification.

1. Introduction

Recent advances in large-scale contrastive vision-language models (VLMs) [30, 41] reshaped the landscape of image classification. In particular, these models can effectively tackle arbitrary classification tasks by measuring the similarity between a list of text labels and the input image in their multimodal representation space. This ability is closely tied to their strong zero-shot transfer performance [30], unlocked by their large pre-trained datasets. While VLMs strive for classification, the same cannot be affirmed for Large Multimodal Models (LMMs) [1, 2, 7, 20, 37].

A few studies [12, 21, 39, 46] have analyzed how well LMMs can recognize objects under both closed- and open-world settings. In the closed-world case, models are required to select a label from a predefined set of categories, whereas in open-world [12], they must answer open-ended queries such as *"What is the object in the image?"*. These investigations typically compare LMMs to VLMs [30, 41], and consistently find VLMs as highly competitive against their generative counterparts [12, 46]. This is surprising, as LMMs are usually benchmarked in much more complex scenarios [15, 22, 40, 42], and should find classification a much easier task. Thus, this finding prompts a fundamental question: *are LMMs worse than VLMs at classification, or are they not properly conditioned for the task?*

In this work, we address this question through the lens of *In-Context Learning* (ICL), investigating whether examples

provided at inference time can enhance the performance of LMMs on image classification tasks. ICL enables models to perform new tasks without parameter updates by conditioning generation on a few input-output examples. While ICL has been extensively studied in large language models [6, 8, 35], its application to visual tasks has only recently emerged [29, 34, 44, 45]. These works suggest that visual ICL can unlock latent discriminative and reasoning abilities by allowing models to access additional cues, potentially narrowing the gap with specialized discriminative systems.

We begin with an in-depth analysis of ICL applied to several LMMs in the closed-world classification setting, comparing their behavior with caching-based strategies for VLMs [43], where few-shot examples are encoded into a key–value memory. Our experiments show that LMMs, when conditioned on annotated examples, exhibit substantial performance gains and, in several configurations, close or even surpass the gap with contrastive VLMs. This finding challenges the conventional assumption that LMMs are inherently weaker at discriminative perception tasks.

Motivated by these results, we extend our investigation to the more challenging open-world classification scenario [12]. This setting introduces two major challenges: (1) the absence of fixed class labels, preventing balanced per-class sampling, and (2) the lack of human supervision, requiring automatic labeling of in-context examples. To address these issues, we propose a novel, annotation-free method that leverages ICL to refine pseudo-labels for unlabeled images iteratively. Our approach, termed ***CIRCLE Iteratively Refines Contextual Learning Examples*** (CIRCLE), employs a *circular iterative procedure* in which pseudo-labels assigned to in-context examples are progressively updated across multiple rounds, allowing the model to self-correct and dynamically infer the level of visual granularity required for the task. This mechanism enhances LMM performance in open-world scenarios without requiring external supervision, consistently outperforming VLMs and challenging the common assumption that VLMs are superior to LMMs at discriminative tasks.

Contributions. In summary, our main contributions are:

- We provide the first systematic analysis of ICL in LMMs for closed-world image classification, and compare their behavior with caching-based VLMs, showing that LMMs with ICL can match and even surpass VLMs.
- We present CIRCLE, a new approach that enhances LMMs for open-world classification using only unlabeled images as ICL examples, iteratively refining their pseudo-labels.
- With an extensive benchmark on open-world classification, we show that, while naïve ICL struggles in this setting, CIRCLE largely improves the performance of the base model, consistently surpassing those of VLMs, making a valid case for adopting LMMs for discriminative tasks.

2. Related work

Vision-Language Models (VLMs) as Classifiers. VLMs, such as CLIP [30] or SigLIP [41], align image and text representations in a shared, dual-encoder embedding space. These models, trained on web-crawled datasets (*e.g.*, 400M image-text pairs in [30]), enable zero-shot classification by computing the cosine similarity between an image embedding and the text embeddings of class names. Despite their impressive zero-shot capabilities, previous studies [30, 47, 48] have shown that VLMs exhibit limited generalization to fine-grained classification tasks (*e.g.*, distinguishing between types of flowers or cars) and specialized domains that are underrepresented in the training data (*e.g.*, satellite image classification). Consequently, significant research has focused on improving VLM performance without updating parameters. These training-free adaptation strategies typically involve incorporating additional domain-specific knowledge into textual prompts or leveraging a cache of labelled images [16, 36, 43]. For example, Tip-Adapter [43] uses similarities between input image and cached examples for prediction, SuS-X [36] automatically constructs its cache through generative models or retrieval, and ComCa [16] adapts the cache to the task by analyzing web-scale databases and querying LLMs. A key architectural constraint of dual-encoder VLMs is their inherent restriction to the closed-world setting, where the prediction space is explicitly defined by a user-provided list of admissible classes [30, 41]. However, recent works [10, 11, 23] have explored ways to adapt VLMs for open-world classification. For instance, CaSED [10] leverages an external vision-language database to allow CLIP to operate effectively in an open-world setting.

Large Multimodal Models as Classifiers. Several studies have focused on assessing the general capabilities of LMMs [15, 22, 40, 42]. Specifically, for image classification, prior works have examined performance in both closed-world settings [21, 46] and open-world scenarios [12, 39]. Notably, Zhang *et al.* [46] evaluated multiple LMMs across several datasets, finding that generative LMMs generally underperform compared to contrastive models. Furthermore, Conti *et al.* [12] showed that LMMs often struggle with label granularity, tending to predict generic terms (*e.g.*, "flower") rather than precise categories (*e.g.*, "water lily"). However, previous studies have not studied the role of in-context examples in this setting. In this paper, we challenge the assumption that contrastive VLMs outperform LMMs, showing that (i) context is crucial and (ii) LMMs equipped with ICL can achieve superior accuracy under the same conditions.

In-Context Learning. Since the advent of LMMs, one of the biggest challenges is adapting these models to specific tasks due to their large parameter size. In-Context Learning [6], a technique well-known in natural language processing but only recently explored in computer vision [29, 44, 45], has

emerged as a strategy to address this issue. Previous works [29, 44] on visual ICL have investigated how the selection of in-context examples affects downstream performance in several visual tasks, such as semantic segmentation or detection, revealing that different examples can lead to significantly different results. Other studies [45] focused on proposing strategies for selecting in-context examples, considering both unsupervised and supervised approaches. In this work, we investigate the connection between LMMs with ICL and CLIP-based caching methods, and introduce a novel ICL strategy that leverages unlabeled images as context. We show that CIRCLE significantly improves classification performance in the challenging open-world setting.

3. Closed-world classification

This section provides a comparative analysis of Few-Shot and In-Context Learning for Closed-World Classification, *i.e.*, the standard scenario where the set of classes is known *a priori*, with particular focus on comparing traditional contrastive VLMs with generative LMMs. We first formalize the classification setting for both model architectures (Sec. 3.1) and then detail their primary mechanisms for Few-Shot Learning: caching for VLMs and In-Context Learning (ICL) for LMMs (Sec. 3.2). Finally, we introduce our experimental setup and present a detailed analysis of the results.

3.1. Preliminaries

In a nutshell, classification aims to assign a semantic label to an input image. Traditionally, a classifier ϕ is a mapping $\phi : \mathcal{V} \rightarrow \mathcal{S}$, *i.e.*, from an image $v \in \mathcal{V}$ to a label $s \in \mathcal{S}$, with $\mathcal{V}$ denoting the image space and $\mathcal{S}$ denoting a set of semantic labels. In standard classification, $\mathcal{S} = \{s_1, \cdots, s_n\}$ is assumed to be a finite set with *known* elements. In other words, $\mathcal{S}$ is "closed", inspiring the naming of the established Closed-World Classification (CWC) setting.

CWC with Contrastive VLMs. Contrastive Vision-Language Models can be readily employed as zero-shot closed-world classifiers by leveraging their aligned encoders. We denote the visual encoder with $\phi_{\text{vis}}^{\text{VLM}} : \mathcal{V} \rightarrow \mathcal{Z}$ and the textual encoder with $\phi_{\text{text}}^{\text{VLM}} : \mathcal{T} \rightarrow \mathcal{Z}$, where, for an hidden size h, $\mathcal{Z}$ is the $h-1$ unit-hypersphere manifold, and $\mathcal{T}$ is the text space. When classes in $\mathcal{S}$ are expressed in natural language (*i.e.*, $\mathcal{S} \subset \mathcal{T}$), classification for an image $v \in \mathcal{V}$, is carried out by finding the class whose text embedding has the highest cosine similarity with the embedding of v. Formally, the predicted label $\hat{s}$ is:[1]

$$\hat{s} = \underset{s \in \mathcal{S}}{\text{argmax}} \ \langle \phi_{\text{vis}}^{\text{VLM}}(x), \phi_{\text{text}}^{\text{VLM}}(s) \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity.

[1] For ease of notation, we omit any template prepend to class names, *e.g.*, "A photo of a", "itap of a", etc.

Table 1. **Closed-world results.** Averaged accuracy over the ten datasets. Higher is better, **bold** indicates best. See the *Supp. Mat.* for detailed per-dataset results and different context sizes.

Model	Zero-Shot	Few-Shot & In-Context Learning					
		4-shot	Δ	8-shot	Δ	16-shot	Δ
Contrastive Vision-Language Models ("Few-Shot" = Tip-Adapter)							
CLIP ViT-B/32	62.6	66.0	+3.4	68.2	+5.6	70.1	+7.5
k-NN	N/A	54.2	-8.4	60.3	-2.3	64.8	+2.2
CLIP ViT-B/16	65.6	69.7	+4.1	71.4	+5.8	74.4	+8.8
k-NN	N/A	61.6	-4.0	66.8	+1.2	70.9	+5.3
CLIP ViT-L/14	72.9	75.6	+2.7	77.5	+4.6	79.8	+6.9
k-NN	N/A	67.2	-5.7	73.1	+0.2	76.8	+3.9
Large Multimodal Models ("Few-Shot" = Vanilla ICL)							
Qwen-2-VL 7B	61.3	68.1	+6.8	73.8	+12.5	79.0	+17.7
Qwen-2.5-VL 7B	61.2	61.3	+0.1	70.1	+8.9	76.5	+15.3
LLaVa OneVision 7B	55.8	60.4	+4.6	60.8	+5.0	60.8	+5.0
Phi-3.5-Vision	38.5	50.5	+12.0	59.1	+20.6	67.7	+29.2
Phi-4-MM	39.6	47.8	+8.2	51.6	+12.0	53.0	+13.4

CWC with LMMs. Unlike Contrastive VLMs, LMMs typically comprise a vision encoder $\phi_{\text{vis}}^{\text{LMM}}$, a connector $\phi_{\text{conn}}^{\text{LMM}}$, and an LLM decoder $\phi_{\text{text}}^{\text{LMM}}$. For an hidden size h, an input image v and a textual query q, the forward pass takes the form:

$$\mathbf{V} = \phi_{\text{conn}}^{\text{LMM}}(\phi_{\text{vis}}^{\text{LMM}}(v)) \in \mathbb{R}^{L_v \times h}; \tag{2}$$

$$y = \phi_{\text{text}}^{\text{LMM}}\Big([\mathbf{V}, \mathbf{Q}]\Big) \in \mathcal{T}, \tag{3}$$

where $[\cdot, \cdot]$ denotes first-axis concatenation, L_v is the number of image tokens, and $\mathbf{Q} \in \mathbb{R}^{L_q \times h}$ the tokenized representation of q with length L_q. The output $y \in \mathcal{T}$ is a free-form text as a result of autoregressive generation: this requires to reformulate CWC from textual responses. An established way is to format q as a Multiple-Choice Question (MCQ) followed by class options in natural language, *e.g.*, "A: water lily, B: sunflower, C: daffodil". In the following, we will denote with $q_{\mathcal{S}}$ the MCQ for the label set $\mathcal{S}$. For a given image v we obtain a textual output through Eq. (3). A prediction $\hat{s}$ is then typically parsed via string exact or fuzzy matching.

3.2. Few-Shot and In-Context Learning for CWC

Although Zero-Shot classification is arguably the most widely adopted protocol, both Contrastive VLMs and LMMs benefit from a support set of labeled examples: a *context*.

For Contrastive VLMs, a simple and effective way to exploit an available context is to refine the prediction for x according to its similarity with the context images. Thanks to its simplicity and influential impact on a variety of follow-up works (*e.g.*, [16, 36]), we consider the approach introduced by Tip-Adapter [43] as the central Few-Shot Learning method for contrastive VLMs in this work. In a nutshell, with $\mathcal{C}$ being a uniform $k \times |\mathcal{S}|$-sized context comprised of $k \in \mathbb{N}$ samples for each semantic category in $\mathcal{S}$, Few-Shot Learning boils down to logit refinement through visual similarity between a query image v and the context images:

$$\underset{s\in\mathcal{S}}{\text{argmax}}\left[\underbrace{\langle\phi_{\text{vis}}^{\text{VLM}}(v),\phi_{\text{text}}^{\text{VLM}}(s)\rangle}_{\text{Zero-shot score}}+\underbrace{\omega\sum_{x\in C_s}\langle\phi_{\text{vis}}^{\text{VLM}}(v),\phi_{\text{vis}}^{\text{VLM}}(x)\rangle}_{\text{In-Context Refinement}}\right]. \tag{4}$$

Here, ω is a weighting factor and $\mathcal{C}_s \subset \mathcal{C}$ is the subset of k context examples belonging to class s. Importantly, the logit for any s can only be refined if $\mathcal{C}_s$ is *not* empty, which implies that $\mathcal{C}$ must contain samples for *all* pre-defined categories.

For LMMs, a natural way to exploit $\mathcal{C}$ is to plug samples within the so-called "context window", which enables implicit adaptation through attention. For n in-context examples, let $\mathbf{X}_i = \phi_{\text{conn}}^{\text{LMM}}(\phi_{\text{vis}}^{\text{LMM}}(x_i))$ be the encoded image patches of the i-th context image, and $\mathbf{T}_i$ be the tokenized representation of its class name. Among many possible orderings, we consider a "Vanilla ICL" setup of the form:

$$y = \phi_{\text{text}}^{\text{LMM}}\Big([\underbrace{\mathbf{X}_1,\mathbf{T}_1,\ldots,\mathbf{X}_n,\mathbf{T}_n}_{\text{Context }\mathcal{C}},\ \underbrace{\mathbf{V},\mathbf{Q}_{\mathcal{S}}}_{\text{Img and MCQ}}\ ]\Big), \tag{5}$$

with $\mathbf{Q}_{\mathcal{S}}$ being the tokenized representation of the MCQ $q_{\mathcal{S}}$. In the next Section, we try to answer the following research question: *Are Contrastive VLMs truly better than LMMs for discriminative tasks?*

3.3. Are CLIP-like VLMs really better than LMMs?

Prior work [12] conveyed a strong message: despite their strengths in generative tasks, LMMs are largely outperformed by Contrastive VLMs when it comes to discriminative ones. In this Section, we examine whether such a claim holds when a shared context is available. To this end, we compare Tip-Adapter [43] as a representative for VLMs, and Vanilla ICL as the nearest equivalent for LMMs.

Datasets. We use the established Few-Shot Classification suite [10, 12, 32, 48], including Caltech101 [14] SUN397 [38], Flowers102 [27], Food101 [5], Oxford Pets [28], FGVC Aircraft [24] Stanford Cars [19], DTD [9], UCF101 [33], and EuroSAT [17]. To ensure an exact comparison with [12], we borrow the benchmark categorization: ① *Prototypical Datasets*, including Caltech101 and SUN397; ② *Non-Prototypical Datasets*, encompassing DTD, UCF101, and EuroSAT; ③ *Fine-grained Datasets*, with Oxford Pets, Food101, and Flowers102; and ④ *Very Fine-grained Datasets*, with Stanford Cars and FGVC Aircraft.

Models. For Contrastive VLMs, we evaluate the three most common CLIP [30] variants using ViT-B/32, ViT-B/16, and ViT-L/14 [13]. For LMMs, we consider a broad spectrum of publicly available models, including the Qwen series (Qwen2-VL [37], Qwen2.5-VL [3]), LLaVA OneVision [20], and the Phi series (Phi-3.5.Vision [1] and Phi-4-MultiModal [25]), ensuring coverage across several design choices for multimodal pretraining, *e.g.*, LLM decoders, pretraining data, vision encoders, and alignment tasks.

Methods. Following established literature, we evaluate at $k \in \{4, 8, 16\}$ shot availabilities, which implies a $k\times|\mathcal{S}|$-sized *shared* context for both VLMs and LMMs. For the latter, to avoid overly long sequence lengths leading to intractable memory usage, we reduce the effective context size to exactly k (image, label) pairs by drawing from $\mathcal{C}$ in two flavors: *random* (see the *Supp. Mat.*) and *similarity-based* sampling, for which we use the smaller CLIP ViT-B/32 to retrieve the k most similar samples to the given query. This might naturally raise a critical confounder, *i.e.*, whether any performance gain observed with LMMs *solely* stems from correctly labeled retrieved examples, making ICL collapse to a majority vote within the available context [4]. Note that the pipeline ① unlabeled query $\rightarrow$ ② k retrieved samples $\rightarrow$ ③ majority vote exactly corresponds to the k-NN algorithm. Hence, to avoid such a confounder, we report k-NN results within the available context for all different CLIP models.

Metrics. Since $\mathcal{S}$ is closed, we use the *Textual Inclusion* (`TI`) metric from [12], which simply resorts to substring matching. For CLIP models, `TI` is equivalent to top-1 accuracy.

Experimental results, averaged across all datasets, are given in Tab. 1[2] and convey a strong message: *despite starting from significantly lower zero-shot performance, LMMs can fill the gap with Contrastive VLMs as the context size increases*. Specifically, we observe that at lower shot counts (*e.g.*, 4), LMMs often underperform even a simple CLIP k-NN majority vote, but the picture changes dramatically with higher shot counts (*e.g.*, 16). For the strongest LMMs, we observe performance boosts relative to zero-shot inference up to $+29.2\%$ for Phi-3.5-Vision and $+17.7\%$ for Qwen2-VL-7B on average. Importantly, *the strongest LMMs can even match the strongest VLMs when provided with sufficient context*. For example, Qwen2-VL-7B matches CLIP-ViT-L/14, the strongest contrastive model, when k=16 in-context examples are given. These results show, for the first time, that generative models might serve as a drop-in replacement to VLMs when facing discriminative tasks, and further allows to see contrastive VLMs through different lenses: instead of directly solving a discriminative task, in the future they might serve better as context builders for LMMs. The most significant advantage relates to **sample efficiency**, as shown in Fig. 2. Using only k=16 shots, LMMs can improve up to $> +50\%$, w.r.t. their zero-shot behavior, while CLIP ViT-B/32 (which shows greater relative improvements than the stronger ViT-L/14 variant), peaks at $\approx +25\%$, which entails $2\times$ sample efficiency in terms of relative gains. Through these results, we speculate that with well-engineered context curation, LMMs might surpass VLMs in future research, even in a closed-world setup that naturally suits the latter.

[2]Please refer to the *Supp. Mat.* for the detailed results.

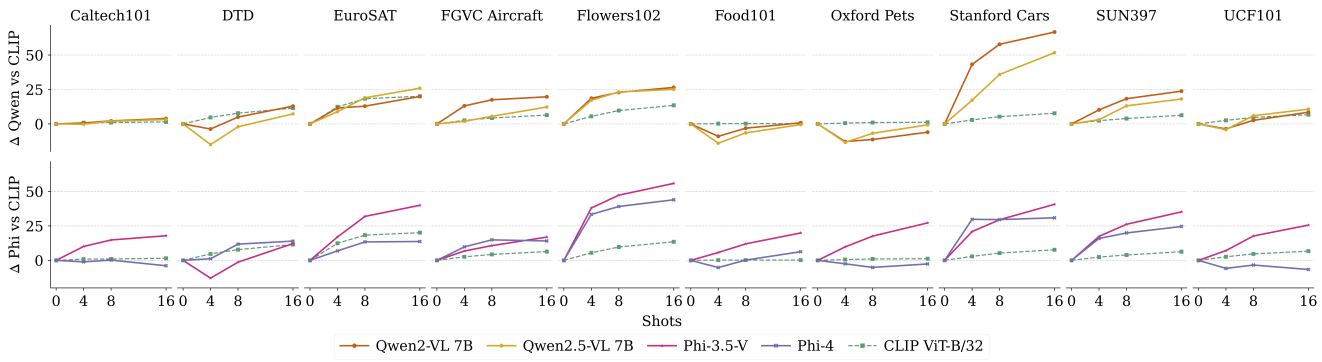

Figure 2. **Sample efficiency.** We visualize the *relative improvement* of a k-shot context w.r.t. the corresponding zero-shot model. For contrastive VLMs (dashed lines), we use Tip-Adapter [43]. For LMMs (solid lines), a simple Vanilla ICL setup. We report both the Qwen family (top row) and the Phi series (bottom row). LMMs benefit much more from additional context than VLMs on most datasets, with peaks of up to $> +50\%$ (*e.g.*, Qwen2-VL-7B on Stanford Cars, Phi3.5-V on Flowers102). In contrast, CLIP-ViT-B/32 peaks at $\approx +25\%$.

Algorithm 1 PyTorch-style code for CIRCLE

```
# image_shots = given unlabeled images (m,3,H,W)
# iters = number of iterations
def classify(model, images):
    context = build_context(model)
    response = model(images, context)
    return response

def build_context(model):
    # step1: pseudo-label images
    labels = forward(model, image_shots)
    context = zip(image_shots, labels)
    # step2: recursive steps
    for iter in iters:
        labels = forward(model, image_shots,
            context)
        context = zip(image_shots, labels)
    return context

def forward(model, images, context=None):
    responses = []
    for i, image in enumerate(images):
        # context contains the other m-1 images
        cur_context = None
        if context:
            cur_context = context.copy()
            cur_context.remove(i)
        responses[i] = model(image, cur_context)
    return responses
```

4. Open-world classification

In this section, we expand our analysis to Open-World Classification (OWC), where the goal is to classify an image *without* a predefined set of classes. In other words, $\mathcal{S}$ is "open" and its elements are not known a priori.

4.1. Preliminaries

OWC with Contrastive VLMs. Thanks to their architectural design, Contrastive VLMs are ultimately image-text retrieval systems. Therefore, a popular approach to enable OWC is to equip them with a large natural language corpus and consider the most similar caption to the input image as their response. An extension of this approach is CaSED [10], which extracts candidate class names after retrieving captions from a massive pre-built textual index.

OWC with LMMs is a natural scenario, which does not require to reformat a query q into an MCQ. Hence, the design of LMMs is naturally suited for this task.

4.2. Few-shot and In-Context Learning for OWC

In this setting, we assume access to an *unlabeled* set of m in-context images $\mathcal{C} = \{x_1, \dots, x_m\}$, and consider using pseudo-labeling to account for the lack of supervision.

For Contrastive VLMs, the advantage of having a cache for open-world is dubious, since abundant support data (*i.e.*, the retrieval index) are assumed to be available already. Moreover, as stated in Sec. 3.2, in-context examples for VLMs primarily serve for logit refinement, which is ill-posed in OWC since the notion of "logit" is inherently attached to a pre-defined category, which is absent in this scenario.

For LMMs, on the other hand, the context can still induce better responses not by steering the predictions towards a specific label set, but by conveying the task to the generative model [26], or by narrowing generation down to the visual domain depicted by the context.

4.3. What can a Context tell about the Open World?

From Sec. 3, we have empirical evidence that labeled in-context examples significantly benefit CWC. However, there is no exact notion of "label" in OWC, and we assume an unlabeled context comprised of m images only. An intuitive way to mimic the setting of Sec. 3 is to let the LMM generate pseudo labels $\hat{y}_i$ for each of the context images x_i, and to use their tokenized representations $\hat{\mathbf{T}}_i$ as a context. With this

strategy, the Vanilla ICL setup of Eq. (5) now reads:

$$y = \phi_{\text{text}}^{\text{LMM}}\Big([\mathbf{X}_1, \hat{\mathbf{T}}_1, \ldots, \mathbf{X}_m, \hat{\mathbf{T}}_m, \mathbf{V}, \mathbf{Q}]\Big). \quad (6)$$

We call such a setup *Pseudo In-Context Learning*.

The key limitation of this naïve approach is that it does not capture the inter-sample dependencies within $\mathcal{C}$, which is crucial for disambiguating user intent and converging on a consistent semantic granularity. In particular, we have seen how in-context examples improve classification performance. Thus, we can apply the same principle to improve the pseudo-labels of our in-context examples. We therefore introduce a new approach, ***CIRCLE*** *Iteratively* ***Refines*** *Contextual* ***L****earning* ***E****xamples* (CIRCLE), with a simple key idea: using the context $\mathcal{C}$ as a context for its own examples, recursively. CIRCLE treats $\mathcal{C}$ itself as a malleable resource, where each pseudo label is refined according to the state of all other in-context examples. Formally, let $\hat{\mathcal{C}}^t$ denote the state of the context at time t, with images x_i (immutable), and pseudo labels $\hat{y}_i^t$ evolving over time. At time $t = 0$, the LMM provides independent per-sample pseudo labels $\hat{y}_i^{t=0}$. From time $t = 1$ onward, the context for the j-th (in-context) example is the concatenation of all but the j-th sample itself, $\{(x_i, \hat{y}_i^{t-1}) : i \neq j,\ \forall\, i \in [1, \ldots, m]\}$, which has a tokenized representation $\mathbf{C}_{i\neq j}^t$ of the form:

$$\mathbf{C}_{i\neq j}^t = [\{\mathbf{X}_i, \hat{\mathbf{T}}_i^{t-1} : i \neq j, \forall\, i \in [1, \ldots, m]\}]. \quad (7)$$

Here, $\hat{\mathbf{T}}_i^{t-1}$ denotes the tokenized representation of the running pseudo label $\hat{y}_i^{t-1}$. Intuitively, $\mathbf{C}_{i\neq j}^t$ is a "leave-one-out" context with all but the j-th sample, which we can use to obtain a contextual pseudo label $\hat{y}_j^t$ for sample x_j as follows:

$$\hat{y}_j^t = \phi_{\text{text}}^{\text{LMM}}\Big([\mathbf{C}_{i\neq j}^t, \mathbf{X}_j, \mathbf{Q}]\Big) \quad (8)$$

This operation is parallelized across the context samples, obtaining new contextual pseudo labels $\{\hat{y}_i^t\}_{i=1}^m$ that capture inter-sample dependencies within the context. From the hints in Sec. 3, we expect that leveraging auxiliary information makes $\hat{y}_i^t$ more accurate than $\hat{y}_i^{t-1}$. Repeating this cyclical procedure T times yields a final context $\hat{\mathcal{C}}^T = \{x_i, \hat{y}_i^T\}_{i=1}^m$, which serves as context for the query image.

We report the pseudo-code for CIRCLE in Algorithm 1. We highlight that there are three main steps: pseudo-labeling the context images, recursively refining their labels using the others as context, and classifying an input image.

4.4. Evaluation protocol

Here, we examine OWC using contrastive VLMs and LMMs, along with their ICL extensions when applicable. The protocol maintains the outline from the closed-world setting, using the same ten datasets and models.

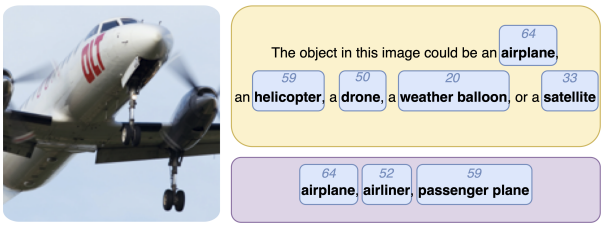

Figure 3. **Concept similarity issues.** `bCS` can be misleading, for instance by favoring comprehensive yet ungrounded lists of candidates (64 for *Vanilla* and CIRCLE). `mCS` instead rewards more coherent and precise answers (50 for *Vanilla*, 59 for CIRCLE).

Methods. For contrastive VLMs, we report the open-world baselines CaSED [10] and CLIP-Retrieval. For LMMs, we start by reporting a *Random* baseline for which we assume manually annotated ground truths (which, in practice, can easily be human-annotated given the modest context size). We then report *Pseudo* ICL and CIRCLE, assuming an unlabeled context with fully automated refinement. Importantly, all ICL variants use the *same* m=16 in-context images. Results at different shot availabilities are in the *Supp. Mat.*

Metrics. Following [12], we report performance using four distinct metrics designed to assess the correctness and relevance of generative outputs. For correctness, we evaluate the presence of the label in the output with *Llama Inclusion* (**`LI`**), which uses LLM-as-a-judge to compare the generated outputs with the ground truth[3]. For relevance, we estimate the similarity of the generation to the ground truth on a sentence-level (*i.e.*, *Semantic Similarity*, **`SS`**), and on a concept-level. For the latter, we extract all concepts from the free-form output using spaCy and compute the maximum Sentence-BERT [31] similarity with the ground truth. We denote this metric as *Best Concept Similarity*, **`bCS`**. Additionally, since **`bCS`** might lead to inflated and misleading scores (see Fig. 3), we further report the *Median Concept Similarity*, **`mCS`**.

Experimental results. We present OWC results in Tab. 2 and visualize some qualitative examples of the predictions of these methods in Fig. 4. There are three clear observations from the table. The first is the low performance of LMMs zero-shot w.r.t. VLMs models designed for OWC on semantic similarity metrics (*e.g.*, +15 `mCS` and +7 `SS` for CaSED in *Prototypical* vs. Qwen2-VL), they lag behind in `LI` (*e.g.*, -32 for the same models), as shown also in [12]. The second is that naïve ICL may degrade the performance of the zero-shot baseline across both correctness (*e.g.*, -20 `LI` for LLaVa OV Random on *Non-prototypical*) and relevance metrics (*e.g.*, -27 for `SS` for the same model). Additionally, pseudo-labeling alone does not mitigate this issue, showing similar performance degradation (*e.g.*, -33.5 `LI` and -35 `SS`

[3] As CIRCLE generates a list of comma-separated labels, we wrap its output in a fixed template to allow the LLM to process it correctly. More details available in the *Supp. Mat.*

Table 2. **Open-world results.** We report results for *Llama Inclusion* (LI), *Semantic Similarity* (SS), *Concept Similarity* (bCS), and *Median Concept Similarity* (mCS). Purple indicates our CIRCLE. Higher is better on all metrics. For each LMM, **bold** indicates the best result. See the *Supp. Mat.* for an extended version of the table.

Model	Method	Prototypical				Non-prototypical				Fine-grained				Very fine-grained			
		LI	SS	bCS	mCS	LI	SS	bCS	mCS	LI	SS	bCS	mCS	LI	SS	bCS	mCS
Contrastive Vision-Language Models																	
CaSED [10] (ViT-L/14)		46.3	58.9	59.8	58.8	18.6	41.8	42.4	41.8	46.6	60.7	61.6	60.7	47.1	38.5	38.5	38.5
CLIP retrieval (ViT-L/14)		42.9	40.2	60.6	31.1	28.1	43.4	32.4	23.8	45.4	42.9	65.4	31.8	18.1	39.7	56.1	29.5
Large Multimodal Models																	
Qwen2-VL 7B [2]	*Zero-Shot*	78.7	51.9	76.0	43.7	42.6	30.8	49.8	29.2	64.0	39.2	62.9	31.9	63.0	34.5	43.4	33.1
	Random Ctx	24.4	41.4	52.7	39.7	17.1	23.4	41.3	23.0	31.7	34.4	44.8	34.6	31.1	29.2	34.1	27.9
	Pseudo ICL	81.1	53.4	**76.2**	44.4	42.8	31.2	**50.1**	26.9	53.1	40.2	64.4	30.7	49.1	38.9	49.1	38.6
	CIRCLE	**91.5**	**65.6**	74.3	**63.5**	**61.6**	**41.9**	49.4	**40.5**	**87.3**	**61.1**	**72.0**	**57.3**	**91.5**	**42.5**	**50.2**	**39.8**
Qwen2.5-VL 7B [3]	*Zero-Shot*	82.9	47.9	**79.9**	31.1	45.9	30.5	54.0	24.8	73.8	47.0	**78.9**	29.5	69.0	45.8	68.6	27.1
	Random Ctx	82.5	53.5	78.2	43.1	48.2	35.7	**58.9**	27.3	70.6	49.0	76.7	36.3	34.7	**52.4**	**70.2**	**42.0**
	Pseudo ICL	80.6	49.3	78.6	31.0	42.7	31.6	53.2	24.2	63.8	45.1	74.8	28.4	39.7	46.0	63.4	25.5
	CIRCLE	**94.9**	**67.7**	68.1	**67.2**	**67.6**	**42.6**	45.1	**42.3**	**86.3**	**60.1**	60.9	**59.7**	**93.6**	36.4	36.6	36.5
LLaVa OV 7B [20]	*Zero-Shot*	53.2	56.2	62.0	53.4	28.1	31.6	43.8	30.2	40.4	39.0	43.9	37.2	**76.7**	31.8	32.3	30.9
	Random Ctx	14.0	29.3	36.7	29.6	8.6	26.0	39.3	24.1	21.0	33.2	35.8	33.4	75.8	30.8	30.8	30.8
	Pseudo ICL	19.7	31.2	41.1	30.5	3.3	18.3	31.5	19.7	20.3	35.5	40.0	34.7	70.4	30.6	31.1	29.8
	CIRCLE	**72.2**	**74.0**	**74.0**	**74.0**	**61.7**	**55.3**	**55.3**	**55.3**	**55.1**	**46.0**	**46.0**	**46.0**	74.2	**32.9**	**32.8**	**32.9**
Phi-3.5-V [1]	*Zero-Shot*	60.7	48.2	**65.6**	46.1	28.7	24.9	**36.7**	24.1	50.7	32.1	47.2	31.3	54.2	29.5	36.3	29.8
	Random Ctx	41.1	48.3	54.4	48.4	10.7	26.3	36.4	27.4	25.2	26.3	36.4	27.4	59.2	27.1	31.7	26.3
	Pseudo ICL	54.1	44.1	61.7	40.1	23.7	22.8	35.1	21.5	43.1	33.0	**48.6**	29.7	24.9	32.4	**41.8**	32.5
	CIRCLE	**92.1**	**59.7**	63.2	**60.3**	**58.3**	**30.0**	35.2	**31.7**	**88.1**	**39.2**	45.7	**42.1**	**99.6**	**33.0**	33.6	**33.1**
Phi-4-MM [25]	*Zero-Shot*	49.8	57.4	58.7	57.2	21.2	29.2	32.7	29.2	37.7	39.2	39.2	39.1	73.6	31.6	31.7	31.6
	Random Ctx	11.6	32.8	32.8	32.8	5.9	28.5	28.6	28.5	31.3	37.9	37.9	37.9	74.6	30.8	30.7	30.6
	Pseudo ICL	51.9	61.5	62.0	61.5	15.1	26.6	31.2	26.7	37.7	41.6	41.7	41.5	72.8	31.9	31.9	31.9
	CIRCLE	**91.5**	**65.5**	**70.1**	**66.4**	**67.6**	**43.2**	**46.1**	**43.4**	**79.1**	**53.3**	**55.5**	**53.0**	**75.2**	**40.2**	**42.5**	**37.9**

for the same settings). Figure 4 shows some examples where the baselines fail to steer the model to the correct generation.

The third observation is that CIRCLE inverts these trends, providing results consistently higher than any ICL counterpart and VLM, across both correctness and relevance metrics and for all dataset categories. For instance, on the *Prototypical* tasks, CIRCLE improves the LI score of Qwen2-VL to 91.5, significantly outperforming the second-best baseline, *Pseudo ICL* (81.1). Similarly, it improves the performance of *Zero-Shot* Phi-3.5-V (60.7) by +31.4% on *Prototypical* tasks. This trend holds across all models and task groups.

Crucially, this notable increase in LI is not achieved at the expense of semantic quality, a common trade-off for baselines. For example, *Zero-Shot* Qwen2.5-VL achieves high LI (82.9) but comparatively low SS (47.9) and mCS (31.1), suggesting verbose or semantically imprecise outputs. In contrast, CIRCLE shows strong performance on *both* inclusion and semantic relevance metrics simultaneously.

Through these results, we can draw two key conclusions: (i) ICL does not guarantee improvements of OWC performance of LMMs; (ii) CIRCLE, a simple yet carefully designed strategy that exploits ICL itself to refine the context, overcomes these limitations, and consistently outperforms both the base model and VLM counterparts. Thus, with the right strategy, LMMs are more suitable than VLMs for open-world classification, We refer to the *Supp. Mat.* for a comprehensive, metric-by-metric analysis of the results.

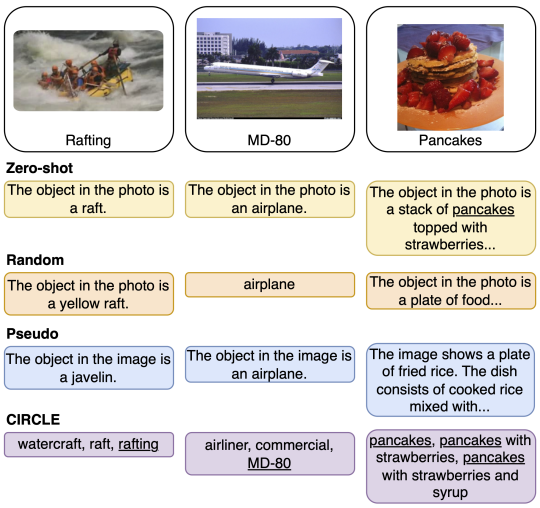

Figure 4. **Qualitative results.** We visualize some qualitative results on examples from UCF101, FGVC Aircraft, and Food101. Underline indicates a correct prediction of the ground truth.

4.5. Ablation study

In this section, we analyze CIRCLE's components, studying the impact of context size and test-time compute available.

Context size. In Fig. 5(a), we analyze the effect of adding more samples to the context. We observe that adding examples has a positive effect on semantic-related metrics (SS,

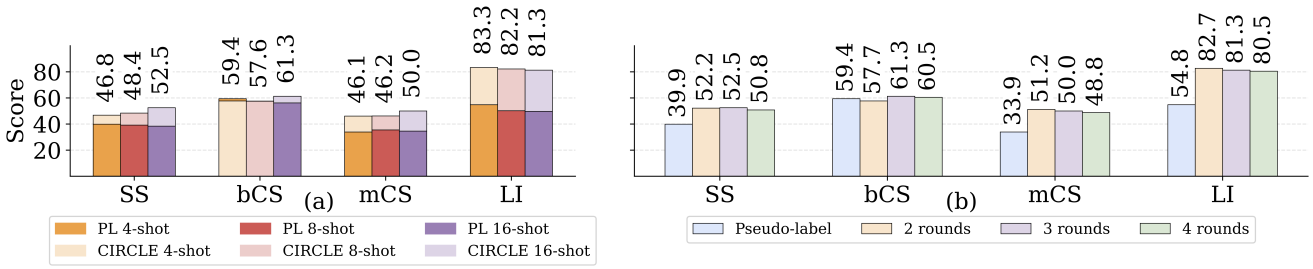

Figure 5. Purple denotes our default configuration, applied to Qwen2-VL 7B [37]. We report *semantic similarity* (SS), *Concept Similarity* (bCS), *Median Concept Similarity* (mCS), and *Llama Inclusion* (LI). **(a) Ablation on context size.** Results are reported for *Pseudo-labeling* (PL) and CIRCLE, with 4, 8, and 16 shots. **(b) Ablation on CIRCLE rounds.** We report results comparing *Pseudo-labeling* against increasing numbers of CIRCLE refinement rounds. See the *Supp. Mat.* for an extended version of these results.

bCS, mCS), while LI remains stable. As the LMM has to find the correct level of information to solve the task, providing more examples enhances the model's overall perspective.

Test-time compute. We measure the impact of the amount of test-time compute in Fig. 5(b), showing how performance varies as the number of recursive rounds for CIRCLE increases. While allowing the model to spend more compute resources (*i.e.*, iterating more times) has a strongly beneficial effect over the *pseudo-labeling* setting (no refinement rounds), we find that it leads to diminishing returns.

4.6. Streaming ICL

CIRCLE's simple design enables flexible adaptation to multiple scenarios. We show this concept in the context of online ICL, taking inspiration from online cache building in VLM classification [18]. We extend our method to work online, *i.e.*, on a stream of samples. To avoid designing ad-hoc filtering protocols, we explore a simple strategy: selecting m random unlabeled examples from the history of test data. From there, naïve pseudo-labeling or CIRCLE can be employed to assign labels to the in-context samples. We report the results for this experiment in Fig. 6 for Qwen2-VL, comparing zero-shot predictions with pseudo-labeling and CIRCLE. Even in this setting, CIRCLE consistently improves over the Pseudo-labeling, with the latter sometimes even decreasing the performance of the base model (*e.g.*, -18 LI on fine-grained). In particular, it consistently achieves the best LI (*e.g.*, +16 on non-prototypical), and it is either best or second-best in all settings and for all relevance metrics. These results confirm the robustness of our approach even in a difficult streaming environment. A comprehensive analysis for all models and tasks is in the *Supp. Mat.*

5. Conclusions

In this paper, we demonstrated that LMMs show strong data efficiency in closed-world image classification through ICL. Still, their predictions are fragile and highly sensitive to context noise. These issues become more severe in open-world settings, where standard ICL fails to form consistent

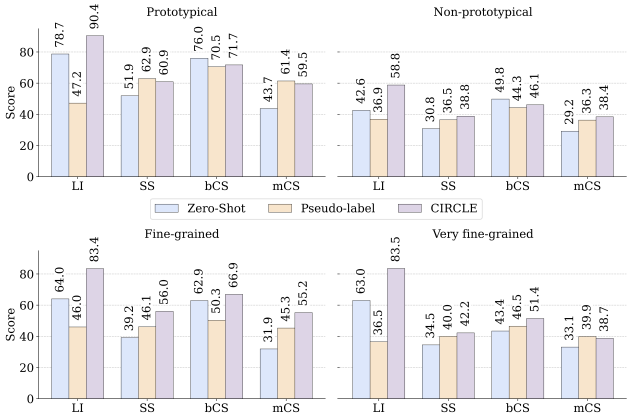

Figure 6. Purple indicates our CIRCLE, applied to Qwen2-VL 7B [37]. We report *semantic similarity* (SS), *Concept Similarity* (bCS), *Median Concept Similarity* (mCS), and *Llama Inclusion* (LI). See the *Supp. Mat.* for an extended version of these results.

semantic interpretations. To address this, we introduced ***CIRCLE Iteratively Refines Contextual Learning Examples***, a training-free self-refinement mechanism that models dependencies across unlabeled in-context examples, enabling the LMM to construct a coherent, task-aligned structure. Our experiments show that this refinement is essential: CIRCLE consistently stabilizes LMM outputs and yields high-precision predictions that outperform all baselines in the OW setting, demonstrating that targeted refinement is necessary to unlock LMMs' discriminative capabilities.

Limitations. Although CIRCLE requires no human annotations, this lack of supervision may make the refinement converge to semantically coherent but task-misaligned label interpretations. In addition, the streaming variant's dynamic memory updates can introduce non-trivial computational overhead when processing large or continuous data streams.

Future work. Two promising directions exist: (i) incorporating lightweight training or parameter-efficient tuning to stabilize refinement and potentially recover the task structure with ambiguous unlabeled data; (ii) improving the streaming mechanism's efficiency (*e.g.*, via memory compression, selective updating, scalable retrieval strategies), enabling broader deployment in resource-constrained scenarios.

Acknowledgments. We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). This work was supported by the EU Horizon ELIAS (No. 101120237), ELLIOT (No. 101214398), and TURING (No. 101215032) projects. This work was supported by Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007) and European Union (Next Generation EU).

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*, 2024. 1, 4, 7

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv*, 2023. 1, 7

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4, 7

[4] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *CVPR-WS*, 2024. 4

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 4

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv*, 2020. 2

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 2

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 4

[10] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. *NeurIPS*, 2023. 2, 4, 5, 6, 7

[11] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification and semantic segmentation. *arXiv*, 2024. 2

[12] Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. *ICCV*, 2025. 1, 2, 4, 6

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 4

[14] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004. 4

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv*, 2023. 1, 2

[16] Marco Garosi, Alessandro Conti, Gaowen Liu, Elisa Ricci, and Massimiliano Mancini. Compositional caching for training-free open-vocabulary attribute detection. In *CVPR*, 2025. 2, 3

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 4

[18] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, 2024. 8

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV-WS*, 2013. 4

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv*, 2024. 1, 4, 7

[21] Huan Liu, Lingyu Xiao, Jiangjiang Liu, Xiaofan Li, Ze Feng, Sen Yang, and Jingdong Wang. Revisiting mllms: An in-depth analysis of image classification abilities. *arXiv*, 2024. 1, 2

[22] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024. 1, 2

[23] Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Rar: Retrieving and ranking augmented mllms for visual recognition. *arXiv*, 2024. 2

[24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv*, 2013. 4

[25] Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. 4, 7

[26] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv*, 2022. 5

[27] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian conference on computer vision, graphics & image processing*, 2008. 4

[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 4

[29] Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. What factors affect multi-modal in-context learning? an in-depth exploration. *NeurIPS*, 2024. 2, 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 4

[31] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 2019. 6

[32] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 2022. 4

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012. 4

[34] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 2

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 2

[36] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *ICCV*, 2023. 2, 3

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv*, 2024. 1, 4, 8

[38] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 4

[39] Kaiyu Yue, Bor-Chun Chen, Jonas Geiping, Hengduo Li, Tom Goldstein, and Ser-Nam Lim. Object recognition as next token prediction. In *CVPR*, 2024. 1, 2

[40] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 1, 2

[41] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1, 2

[42] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *NAACL*, 2025. 1, 2

[43] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv*, 2021. 2, 3, 4, 5

[44] Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv*, 2024. 2, 3

[45] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *NeurIPS*, 2023. 2, 3

[46] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *NeurIPS*, 2024. 1, 2

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2, 4