OPEN-WORLD ZERO-SHOT IMAGE CLASSIFICATION USING AN LLM-GUIDED DISCOVERY LOOP

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Vision-Language models such as CLIP have enabled zero-shot image classification using only the pretrained model and a complete list of category names, vastly lowering labeling costs compared to traditional supervised approaches. However, in practice the full class list may not be available, creating an obstacle to applying CLIP-style models. Vocabulary-free image classification (VIC) approaches this problem by trying to infer the class list from just the images, however VIC is fundamentally ambiguous as images may be classified according to different attributes and granularities, each of them equally valid. We therefore propose Open-World Zero-Shot Image Classification (OWIC) as a well-defined alternative setting in which some category names are given, and the task is to discover the remaining ones based on the pattern defined by the existing classes. To tackle this task, we propose Multimodal Agentic Novelty Detection (MAND), an LLM-driven discovery loop in which a multimodal LLM identifies novel-class images and proposes category names for them that are then used by a contrastive image-text model such as CLIP. A similarity-based stopping criterion helps minimize LLM usage to lower compute cost. Once all category names are discovered the LLM can be discarded, maintaining deployment efficiency. Experimental results show MAND outperforming strong baselines based on vocabulary-free methods adapted to this new task. In addition to leveraging pretrained models, we also show that finetuning can further improve the LLM's performance within the discovery loop, generalizing even to domains beyond the training data. Code will be published.

1 INTRODUCTION

Image classification in supervised settings with large-scale labeled training data has seen immense progress (Beyer et al., 2020), with less and less room for improvement. As a result, the research focus has shifted towards scenarios with reduced supervision, reducing labeling costs at the expense of increasing the difficulty of the task. These efforts culminate in abandoning sample labels entirely and moving to zero-shot settings. Contrastive image-text models such as CLIP (Radford et al., 2021) have shown strong results in this field, enabling accurate classification using just the category names. However, full knowledge of the categories present in the data can often not be assumed, the data may contain additional categories that have gone undetected due to the lack of labels. To overcome this limitation, vocabulary-free methods such as CaSED (Conti et al., 2023) and FiNDR (Demidov et al., 2026) attempt to discover the category list automatically.

In practice though, these approaches attempt to solve an ambiguous problem. The same data can be categorized in many, equally valid ways, e.g. it can be clustered on different granularity levels, or by different attributes (see Figure 1). By assuming no knowledge of the categories in the dataset, vocabulary-free methods lose any signal on what classification rule the undiscovered classes follow. In contrast, if some category names are given, the task has a clear and unique solution dictated by the classification rule implicit to the known classes. Assuming knowledge of some categories is also not unreasonable in practice, usually classification is done with a use-case in mind that dictates the classification rule, so some knowledge about the categories in the data is to be expected.

Based on this insight, we propose Open-World Zero-Shot Image Classification (OWIC), a zero-shot setting where only some category names are given, and the remaining ones need to be discovered. OWIC fills the gap between VIC and standard zero-shot settings, allowing for unknown categories

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: The classification rule and the label granularity cannot be identified from images alone, making Vocabulary-free Image Classification an ambiguous task in practice. However, if some category names are known the rule can be inferred, so Open-World Zero-Shot Image Classification is well-defined.

while remaining well-defined. To tackle this problem we propose Multimodal Agentic Novelty Detection (MAND), an LLM-driven discovery loop that processes images sequentially to detect novel categories and label them with a category name. To avoid processing every image in the dataset, MAND is equipped with a stopping criterion to terminate the discovery loop when no further progress is being made. The final classification step is performed using a contrastive image-text model, so deployment remains lightweight and does not require the LLM. While MAND works well with pretrained models, we also explore finetuning the LLM to further improve accuracy.

In summary, our contributions are:

- We define Open-World Zero-Shot Image Classification (OWIC), a realistic, well-posed zero-shot classification setting. Compared to vocabulary-free classification it is less ambiguous while also allowing unknown categories in the dataset.
- We propose MAND, a discovery method that utilizes a multimodal language model to iteratively identify novel categories. MAND outperforms several strong baselines, including vocabulary-free methods adapted to the new setting.
- We also develop a finetuning pipeline to further improve the LLM's performance as part of MAND. Despite large domain gaps between training and evaluation datasets the finetuned model outperforms the pretrained model.

2 Related Work

2.1 Vocabulary-Free Image Classification

Vocabulary-free Image Classification (VIC) was introduced by Conti et al. (2023) as a new zero-shot classification setting that assumes no knowledge of the category names. The same work also proposed CaSED (Conti et al., 2023), the first method to tackle VIC. CaSED relies on a large caption database sourced from multiple public datasets such as CC12M (Changpinyo et al., 2021) from which the most

relevant captions for each image are retrieved using CLIP similarity. From these captions, candidate classes are extracted and scored using a combination of text-to-image similarity and text-to-text similarity to the centroid of the captions. Later methods leverage the world knowledge and reasoning capabilities of large language models instead of relying on caption databases. FineR (Liu et al., 2024) combines a VQA model to extract relevant information from the images with an LLM to provide expert knowledge. First, the VQA model extracts the super category of the dataset, then the LLM provides a list of attributes that are of interest to distinguish subcategories. The VQA model then retrieves the attribute values along with a general description of the image, which are then fed to the LLM to generate candidate names. FiNDR (Demidov et al., 2026) replaces the interaction between VQA and LLM with a multimodal LLM that can both process the images and reason about the category names. First, three images are presented to the LLM to infer metadata about the dataset such as the taxonomic group present in the images and the target granularity. Equipped with this meta information, the LLM then generates candidate categories for each image, which are then filtered using CLIP similarity. Both LLM-based methods are dependent on correctly extracting metadata about the task, a step that suffers from ambiguity and often has multiple possible solutions. In contrast, OWIC is well-defined due to the known categories implicitly defining the desired classification rule.

2.2 Category Discovery

VIC assumes no knowledge of the target categories, other category discovery settings however function more similar to OWIC in that they assume that some target categories are given. These include Novel Class Discovery (NCD) and the more realistic Generalized Category Discovery (GCD). NCD (Han et al., 2019) is a task where the data is split into a labeled and an unlabeled portion, with the labeled samples belonging to a known set of classes while the unlabeled samples belong to a disjoint novel category set. The goal is to correctly classify all samples, which involves discovering the novel categories while also maintaining accuracy on the known classes. GCD (Vaze et al., 2022) builds on NCD and moves to a more realistic setting where the unlabeled data may contain both novel and known classes. While NCD and GCD are similarly to OWIC in that some categories are provided while others have to be discovered, they make use of sample labels and per-dataset training, whereas OWIC requires categorizing the images zero-shot.

3 Methodology

3.1 Problem Formulation

In Vocabulary-free Image Classification (VIC), we are given an image dataset $\mathbb{D} = \{x_i\}_{i=1}^{N}$ consisting of N images. Each sample x_i has a label $y_i \in \mathbb{C}$, where $\mathbb{C}$ is the set of ground-truth class names, however neither the sample labels nor $\mathbb{C}$ are known. Consequently, the goal of VIC is to infer both the class list and the sample labels from just the given images. This task is ambiguous as there are often multiple valid category lists for the same dataset, grouping the images by different attributes or at different hierarchy levels. To remove this ambiguity, we define Open-World Zero-Shot Image Classification (OWIC) as an alternative setting, where some category names $\mathbb{C}_{\text{known}} \subset \mathbb{C}$ are given, and the task is to infer the remaining names $\mathbb{C}_{\text{novel}} = \mathbb{C} \setminus \mathbb{C}_{\text{known}}$. Here, $\mathbb{C}_{\text{known}}$ constrains the possible candidates for $\mathbb{C}_{\text{novel}}$ through the classification rule implicit to the known classes. Note that the sample labels remain unavailable for both known and novel categories.

3.2 Discovery Loop

Our proposed Multimodal Agentic Novelty Detection (MAND) utilizes a multimodal large language model to detect and name novel categories, and a contrastive vision-language model such as CLIP (Radford et al., 2021) for detecting duplicate classes and determining when to stop the discovery. An overview of MAND's discovery loop can be seen in Figure 2. First, an image is sampled uniformly at random, and the LLM is queried using the image and the current class list $\mathbb{C}_{\text{disc}}$, initialized as $\mathbb{C}_{\text{disc}} = \mathbb{C}_{\text{known}}$, to determine whether the image belongs to a known category or not (refer to Appendix A.1 for the full prompts used for MAND). The query is posed as a yes/no question, so only a single token is generated, and the answer is extracted using constrained decoding, effectively making this a binary classification task. One significant challenge for this decision stage is the shift in the share of novel classes in the unseen data. As more categories are discovered, this

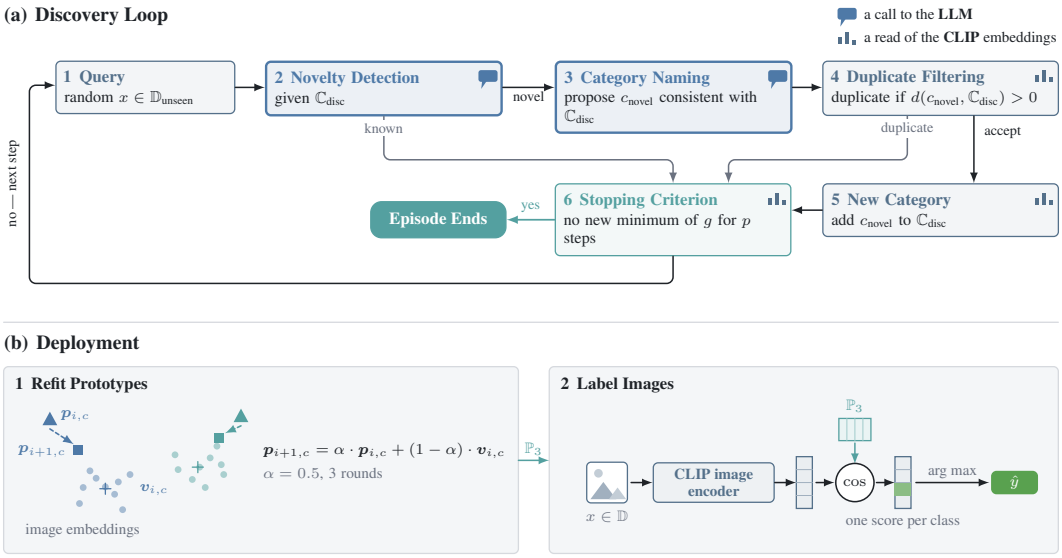

Figure 2: Overview of our proposed MAND discovery loop. The loop consists of two LLM queries for novelty detection and naming, followed by a filtering step. The stopping criterion helps minimize the amount of LLM queries needed. During deployment, only the CLIP model is required, so inference remains efficient.

share declines throughout the episode, resulting in a prior shift for the novelty detection task. The LLM however has an inbuilt prior π_{train} that remains unchanged throughout the episode, making it prone to underdiscovery when a lot of samples are novel, and to overdiscovery when few novel categories remain to be discovered. To mitigate this, we adjust the LLM logits during the decision stage following Saerens et al. (2002) as

$$\begin{aligned} l_{\text{adj,novel}} &= l_{\text{LLM,novel}} - \log \pi_{\text{train}} + \log \hat{\pi}_n \text{ , and} \\ l_{\text{adj,known}} &= l_{\text{LLM,known}} - \log(1 - \pi_{\text{train}}) + \log(1 - \hat{\pi}_n) \text{ ,} \end{aligned} \tag{1}$$

where $l_{\text{LLM,novel}}$ is the LLM logit for the 'novel' decision, $l_{\text{adj,novel}}$ is the corresponding adjusted logit, $l_{\text{LLM,known}}$ and $l_{\text{adj,known}}$ are the corresponding logits for the 'known' decision, and $\hat{\pi}_n$ is our estimate of the true novel class prior π_n at the current step n. As π_{train} is unknown for the pretrained model, we set it to 0.5, which means the prior adjustment may be off by an additive constant. Despite this, the trajectory of the prior over time remains valid, and it remains effective against overdiscovery.

To estimate the novel class prior, we utilize the maximum CLIP similarity between an image and the known categories, defined as $\text{smax}(x, \mathbb{C}_{\text{disc}}) = \max_{c \in \mathbb{C}_{\text{disc}}} (\text{sim}_{it}(x, c))$, where sim_{it} is the image-text CLIP similarity. The maximum similarity of an image to any known category will usually be high if it is from a known class, and low if it is not, so we expect a bimodal distribution of the maximum similarities. At the start of the episode, we fit a Gaussian mixture model (GMM) on smax and find the threshold τ where both components have equal probability. Based on this, we estimate $\hat{\pi}_n$ at each step as the fraction of the unseen data where $\text{smax} < \tau$. This prior estimation depends on the quality of the CLIP embeddings, if the CLIP similarities are noisy then there will not be a clean bimodal distribution of the smax values, so prior estimation cannot work and should not be used in such cases. To detect such cases, we run a leave-one-out evaluation at the start of the episode, where we investigate the higher-mean component of the GMM with its samples $\mathbb{D}_{\text{high}}$ and delete one category at a time from $\mathbb{C}_{\text{known}}$. For each category $c \in \mathbb{C}_{\text{known}}$, we define its CLIP recall as

$$r(c, \mathbb{C}_{\text{known}}, \mathbb{D}_{\text{high},c}, \tau) = \frac{1}{|\mathbb{D}_{\text{high},c}|} \sum_{x \in \mathbb{D}_{\text{high},c}} \mathbb{1}\{ \max_{c' \in \mathbb{C}_{\text{known}} \setminus c} \text{sim}_{it}(x, c') < \tau \} \text{ ,} \tag{2}$$

where $\mathbb{D}_{\text{high},c}$ is the subset of $\mathbb{D}_{\text{high}}$ that has c as its most similar category. This recall signifies how many of its associated samples in the high GMM component drop out of this component if c is removed from the known classes. For a well-aligned CLIP model, r should be close to 1, lower values signify worse alignment between images and category names. We compute the mean CLIP recall over all categories and only apply prior correction for a dataset when the mean is larger than a threshold θ.

After the decision stage, if the sample is found to belong to a known class, the loop step ends, otherwise the LLM is queried again for a second turn to determine a suitable name for the newly discovered category, taking into account the known class names to guide the model towards a name that is consistent with them. To detect failures where a known-class image is falsely identified as novel, the generated names are filtered for duplicates. This is done by querying the CLIP-similarity between the novel category c_{novel} and the other category embeddings and comparing the maximum similarity to the maximum similarity between the known categories. The duplicate score is defined as

$$d(c_{\text{novel}}, \mathbb{C}_{\text{disc}}) = \max_{c \in \mathbb{C}_{\text{disc}}} \left(\text{sim}_t(c_{\text{novel}}, c)\right) - \max_{\substack{c_1, c_2 \in \mathbb{C}_{\text{disc}} \\ c_1 \neq c_2}} \left(\text{sim}_t(c_1, c_2)\right) , \tag{3}$$

where sim_t is the cosine similarity between the CLIP text embeddings of the two categories. The novel category is filtered out if $d(c_{\text{novel}}, \mathbb{C}_{\text{disc}}) > 0$. By conditioning the removal on the dataset-dependent maximum similarity of the known categories to each other instead of a fixed threshold, we create a robust, granularity-aware deduplication. This filter does not catch cases where the LLM hallucinates a name that is completely unrelated to the image, but such cases are less harmful than genuine duplicates, as these hallucinations will not get many samples assigned to them during inference, whereas duplicates may cause a ground-truth category to be split into multiple predicted categories.

3.3 Stopping Criterion

While this discovery loop could be run for every sample, doing so would require at least N LLM inference calls, making the method very expensive. In addition, continuing discovery after every category has been detected increases the chance of the LLM hallucinating categories. To mitigate this, we add a stopping criterion to terminate the discovery loop early. While our estimated novel rate $\hat{\pi}_n$ could in theory also serve as a criterion for stopping the episode, in practice we find that the estimate is too inaccurate, it is sufficient for providing an approximate guidance to the decision stage, but not suitable as a stopping criterion. Instead, we focus on the difference in smax between the samples $\mathbb{D}_{\text{seen}}$ which were already processed by the LLM, and $\mathbb{D}_{\text{unseen}}$, consisting of the unseen samples. By definition, the samples in $\mathbb{D}_{\text{seen}}$ should belong to known categories, either because they already belonged to a known class at the start of the step they were processed in, or because their category was added during the discovery step. In contrast, the samples in $\mathbb{D}_{\text{unseen}}$ may belong to either known or novel categories. As a result, there should be a substantial gap in the maximum image-text CLIP similarities between $\mathbb{D}_{\text{seen}}$ and $\mathbb{D}_{\text{unseen}}$. Formally, we define the gap g as

$$g(\mathbb{D}_{\text{seen}}, \mathbb{D}_{\text{unseen}}, \mathbb{C}_{\text{disc}}) = \frac{1}{|\mathbb{D}_{\text{seen}}|} \sum_{x \in \mathbb{D}_{\text{seen}}} \text{smax}(x, \mathbb{C}_{\text{disc}}) - \frac{1}{|\mathbb{D}_{\text{unseen}}|} \sum_{x \in \mathbb{D}_{\text{unseen}}} \text{smax}(x, \mathbb{C}_{\text{disc}}) . \tag{4}$$

To motivate the use of g as a stopping criterion, we show that under some assumptions, g is directly related to π_n and reaches 0 as π_n does.

Proposition 1 (Gap and novel rate)**.** *Write $y(x_i) = y_i$, let $a(x) = \text{sim}_{it}(x, y(x))$ be the similarity of x to its own class name, with sim_{it} fixed during the loop, and write $\bar{f}(\cdot)$ for the mean of a function f on $\mathbb{D}$ over a set of images. After n steps ($1 \leq n < N$), write g_n for the gap, let π_n be the share of images in $\mathbb{D}_{\text{unseen}}$ whose class name is not in $\mathbb{C}_{\text{disc}}$, and let $g_n^\star = \bar{a}(\mathbb{D}_{\text{seen}}) - \bar{a}(\mathbb{D}_{\text{unseen}})$. Assume that (i) on every run, the exact class name of every image in $\mathbb{D}_{\text{seen}}$ is in $\mathbb{C}_{\text{disc}}$ (correct LLM); (ii) for some $\Delta > 0$, $\text{sim}_{it}(x, c) \leq a(x) - \Delta$ for every $x \in \mathbb{D}$ and every name $c \neq y(x)$ that can belong to $\mathbb{C}_{\text{disc}}$ (CLIP margin); and (iii) each step draws the next image uniformly at random from $\mathbb{D}_{\text{unseen}}$, irrespective of the previous LLM outputs. Then, with $\Delta_{\max} = \max_{x \in \mathbb{D}} \big(a(x) - \text{smax}(x, \mathbb{C}_{\text{known}})\big)$, on every run $\Delta\, \pi_n \leq g_n - g_n^\star \leq \Delta_{\max}\, \pi_n$, and for every fixed n (ignoring the stopping rule) $\mathbb{E}[g_n^\star] = 0$, so that*

$$\Delta\, \mathbb{E}[\pi_n] \;\leq\; \mathbb{E}[g_n] \;\leq\; \Delta_{\max}\, \mathbb{E}[\pi_n]. \tag{5}$$

Proof. Let $s(x) = \mathrm{smax}(x, \mathbb{C}_{\mathrm{disc}})$ and let $u(x) = a(x) - s(x)$ be the deficit of x; substituting $s = a - u$ in the definition of g_n gives $g_n - g_n^\star = \bar{u}(\mathbb{D}_{\mathrm{unseen}}) - \bar{u}(\mathbb{D}_{\mathrm{seen}})$. By (ii), the names in $\mathbb{C}_{\mathrm{disc}}$ score at most $a(x)$ on x, and at most $a(x) - \Delta$ except for $y(x)$ itself, while $\mathbb{C}_{\mathrm{known}} \subseteq \mathbb{C}_{\mathrm{disc}}$ because names are only added during the loop; hence $u(x) = 0$ if $y(x) \in \mathbb{C}_{\mathrm{disc}}$, and $\Delta \le u(x) \le a(x) - \mathrm{smax}(x, \mathbb{C}_{\mathrm{known}}) \le \Delta_{\max}$ otherwise. Hence $\Delta\, \pi_n \le \bar{u}(\mathbb{D}_{\mathrm{unseen}}) \le \Delta_{\max}\, \pi_n$, and $\bar{u}(\mathbb{D}_{\mathrm{seen}}) = 0$ by (i), which is the bound on every run. By (iii), $\mathbb{D}_{\mathrm{seen}}$ is a uniformly random n-subset of $\mathbb{D}$, so $\mathbb{E}[\bar{a}(\mathbb{D}_{\mathrm{seen}})] = \bar{a}(\mathbb{D}) = \mathbb{E}[\bar{a}(\mathbb{D}_{\mathrm{unseen}})]$ and $\mathbb{E}[g_n^\star] = 0$; as $\mathbb{D}$ is finite and $\mathrm{smax}(x, \mathbb{C}_{\mathrm{known}}) \le s(x) \le a(x)$, every term is bounded, and taking expectations gives equation 5. □

Thus $\mathbb{E}[g_n]$ is controlled from both sides by the expected share $\mathbb{E}[\pi_n]$ of unseen images whose class is still undiscovered: it is positive whenever $\mathbb{E}[\pi_n] > 0$, and zero once $\pi_n = 0$ almost surely; this state persists, as $\mathbb{C}_{\mathrm{disc}}$ only grows and $\mathbb{D}_{\mathrm{unseen}}$ only shrinks. The bound on $g_n - g_n^\star$ holds on every run and at every step, including the one at which the criterion stops the loop; only $\mathbb{E}[g_n^\star] = 0$ requires a fixed n. In practice assumptions (i) and (ii) will likely be violated to some degree as the LLM is bound to make mistakes and the CLIP model will misclassify some samples even given the ground-truth category names. Such violations will narrow the gap, but it remains a useful criterion even under practical error rates. In addition, other loop components also rely on the quality of the LLM and the CLIP model, so no additional points of failure are added to the method. Based on these observations, the most intuitive stopping criterion would be to stop once $g = 0$, however the LLM may not be able to discover all categories, resulting in g plateauing above 0. Instead, we adopt a progress-based criterion, where we define a patience p with $p = \max(c\sqrt{N}, 20)$, where c is a hyperparameter that controls the trade-off between earlier stopping and more complete discovery. We stop if g has not set a new minimum value for p steps. To avoid unstable gap values due to a small $\mathbb{D}_{\text{seen}}$ at the beginning of the episode, we do not compute the gap for the first 100 steps.

3.4 Finetuning Pipeline

MAND can be executed using a pretrained model, but to further improve accuracy we also investigate finetuning the LLM to enhance its ability to function as part of MAND. For this, we utilize ImageNet (Deng et al., 2009) as our training dataset, specifically we use the validation set with ImageNet-ReAL (Beyer et al., 2020) labels to benefit from its improved label quality. To ensure a diverse set of training episodes, we leverage the modified BREEDS (Santurkar et al., 2021) ImageNet hierarchy which, unlike the default WordNet (Miller, 1995) hierarchy, utilizes less abstract groupings and has a tree structure with no node having multiple parents. Based on this hierarchy, we sample tasks by first selecting a root node, then a target hierarchy level. The root node's children at the selected level are then randomly assigned to either the known categories, the novel categories, or they are excluded from the episode entirely. Lastly, the samples belonging to the selected categories are randomly included or excluded. The full task sampling procedure can be found in Appendix A.2.

After sampling the training task we run the MAND discovery loop, updating the LLM during the process using supervised finetuning. During training, the discovery loop differs in several ways from the deployed version. First, we sample images not uniformly at random, instead we apply balanced sampling between known and novel categories as controlling the training prior π_{train} by setting it to a fixed value makes the prior correction more accurate. Second, we utilize ground-truth stopping during training, i.e., we stop once all categories have been discovered, and we truncate episodes after 100 steps. These measures help increase task diversity by allowing us to process many tasks even with a small training budget, so the LLM is exposed to a variety of tasks at different granularity levels. During the finetuning, we also further enrich the LLM inputs by including the top-3 nearest neighbors of the query image, which are usually from the same class, as additional references in the prompt. This enables the model to base its decision on more information, enabling better decisions. We do not include reference images for the pretrained model because we found that they do not deliver the same improvement in that case, it appears that the ability to correctly interpret the additional information must be explicitly trained for the lightweight LLMs used in this study.

3.5 Deployment

After the discovery loop has terminated, we remove any category from $\mathbb{C}_{\text{disc}}$ whose text prototype is not the most similar prototype for any image. The other text prototypes form our initial prototype set $\mathbb{P}_0 = \{\boldsymbol{p}_{0,c}\}_{c \in \mathbb{C}_{\text{disc}}}$. Inspired by the modality coupling employed by FiNDR (Demidov et al.,

Table 1: Comparison to baseline methods on ten datasets. Both cluster accuracy and semantic accuracy are measured on all categories, known and novel. Best in **bold**, second underlined.

	Aircraft		Caltech101		DTD		EuroSAT		Flowers102	
Method	cAcc	sAcc	cAcc	sAcc	cAcc	sAcc	cAcc	sAcc	cAcc	sAcc
CaSED	29.2	35.8	56.1	69.4	32.4	43.2	24.0	28.8	76.4	57.5
FiNDR	58.4	68.2	53.7	45.8	40.1	40.3	39.4	27.5	81.2	56.9
CaSED adap.	43.8	64.4	58.4	73.6	33.4	46.6	26.6	42.9	79.3	68.2
FiNDR adap.	60.4	71.7	73.3	82.2	52.2	**64.8**	51.3	59.9	82.6	70.3
MAND pre	55.5	76.0	**88.5**	**91.3**	**57.0**	64.6	**63.8**	61.5	81.2	**76.8**
MAND ft	**63.4**	**81.1**	88.1	87.6	55.7	63.1	62.7	**63.7**	**86.7**	74.7

	Food101		Pets		SCars		SUN397		UCF101	
Method	cAcc	sAcc	cAcc	sAcc	cAcc	sAcc	cAcc	sAcc	cAcc	sAcc
CaSED	67.1	71.1	63.4	66.7	42.5	55.9	42.2	57.7	47.8	57.8
FiNDR	60.5	70.1	78.8	85.4	76.3	80.7	31.8	39.4	51.7	43.2
CaSED adap.	74.8	81.5	71.0	80.1	64.0	82.2	46.7	65.9	56.1	70.9
FiNDR adap.	78.3	88.2	82.3	87.4	74.9	90.2	57.9	74.3	78.1	80.9
MAND pre	86.8	87.1	87.2	89.6	**82.8**	**94.3**	67.2	**78.6**	80.5	82.1
MAND ft	**88.9**	**88.2**	**90.8**	**90.1**	78.8	93.3	**68.4**	78.1	**85.1**	**83.2**

2026), the remaining prototypes are refitted using their associated image embeddings to reduce the modality gap between text prototypes and images. During each refit round $i \in \{0, 1, 2\}$, we select those images for each category c for which $\boldsymbol{p}_{i,c}$ is the most similar category prototype in terms of cosine similarity and calculate their mean image embedding $\boldsymbol{v}_{i,c}$. We then refit the prototype as

$$\boldsymbol{p}_{i+1,c} = \alpha \cdot \boldsymbol{p}_{i,c} + (1 - \alpha) \cdot \boldsymbol{v}_{i,c}\,, \tag{6}$$

where α is an update coefficient. This process is conducted in multiple rounds to take changes in the category assignments caused by the refit into account. Afterwards, any prototypes that lost all samples during the refit are removed. Once the prototypes are fitted, the final deployment requires only the CLIP image encoder, neither the text encoder nor the LLM are needed.

4 Experiments

4.1 Experimental Setup

Following CaSED, we evaluate on ten datasets, Caltech-101 (Fei-Fei et al., 2004), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), FGVC-Aircraft (Maji et al., 2013), Flowers-102 (Nilsback & Zisserman, 2008), Food-101 (Bossard et al., 2014), Oxford Pets (Parkhi et al., 2012), Stanford Cars (Krause et al., 2013), SUN397 (Xiao et al., 2010), and UCF101 (Soomro et al., 2012). Each dataset's categories are evenly split into known and novel classes. We evaluate on two target metrics, cluster accuracy (cAcc), and semantic accuracy (sAcc). Cluster accuracy uses Hungarian Matching (Kuhn, 1955) to find the optimal cluster match between the predicted and the ground-truth categories and then reports the accuracy under that assignment. Semantic accuracy evaluates how well the semantics of the predicted class names match the ground truth. It uses a pretrained sentence transformer to compute text embeddings for each sample based on its assigned category, then computes the semantic similarity to the ground truth label. We report the accuracy across all categories, detailed numbers for the known or novel split can be found in Appendix C.

We choose Qwen3.5-4B (Qwen Team, 2026) as the multimodal LLM and Perception Encoder PE-Core-L14-336 (Bolya et al., 2025) as the CLIP model. To reduce inference time, we use the model in non-thinking mode. We use greedy decoding to minimize randomness, but different image sampling orders still introduce variance, so we report the mean of 10 evaluation runs for each result. By default, we set the stopping criterion patience coefficient c to 7 and the threshold θ that governs when to

Table 2: MAND component analysis. Starting from a baseline we incrementally add additional components. We report the mean over all ten datasets.

	Pretrained				Finetuned			
Method	cAcc	sAcc	Steps	Cat. error	cAcc	sAcc	Steps	Cat. error
Baseline	67.4	77.6	2532	157.2	70.4	78.8	2117	127.4
+ duplicate filter	71.0	78.9	2500	70.6	72.0	79.3	2058	76.2
+ prototype refit	73.6	79.4	2500	70.6	74.0	78.2	2058	76.1
+ prior correction	74.2	80.2	2329	20.7	73.5	78.8	1950	44.9
+ CLIP recall gate	75.1	80.2	2341	19.1	75.0	79.2	1957	41.6
+ reference images	70.5	79.6	1969	26.6	76.9	80.3	1995	23.5

apply prior correction to 0.5. The refit update coefficient α is set to 0.5. For the finetuned version of MAND, we train the LLM for $10\,000$ steps with a batch size of 8 across two RTX Pro 6000 Blackwell GPUs. The model is optimized using the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of $1\mathrm{e}{-5}$ and a 500-step warmup followed by a linear decay to 0. To minimize memory usage during training, we finetune the model using Low-Rank Adaptation (Hu et al., 2022).

We compare against two baseline methods, CaSED (Conti et al., 2023) and FiNDR (Demidov et al., 2026). CaSED is a retrieval-based method that fetches candidate category names from a database, while FiNDR utilizes a multi-modal LLM to generate candidates. As both methods were designed for VIC instead of OWIC, we adapt them by giving them access to the list of known classes. In addition, we adapt both FiNDR versions to work without a separate balanced discovery set, enabling a fair comparison to MAND which does not require such a set. For more details on the adaptation of the baselines to OWIC, we refer to Appendix B. All methods utilize the same CLIP model and LLM.

4.2 Results

Table 1 presents our main results compared to the baselines. They show that pretrained MAND clearly outperforms both adapted baselines, achieving higher cluster accuracy across all datasets compared to CaSED, and across eight datasets compared to FiNDR. Finetuning the LLM further improves cAcc, achieving substantial improvements on FGVC-Aircraft, Flowers-102, Oxford Pets, and UCF101, with the only large regression being Stanford Cars. However, sAcc stagnates, showing that the quality of the proposed names does not increase. The results for CaSED and FiNDR show that our adaptation was successful, with both methods showing improvements. FiNDR is particularly interesting as it gains very little on some datasets while showing massive improvements on others, likely due to a failure to identify the classification rule without guidance from the known classes. This highlights that access to some known category names can be essential to enabling discovery, in line with the motivation for our new OWIC setting.

We analyze the impact of the individual components of MAND in Table 2. We show the mean cAcc, sAcc, as well as the mean step count and the category count error in percent. The results indicate that the duplicate filter and the prototype refit are both instrumental to MAND's accuracy, while also improving the final category number. The prior correction appears to have no clear advantage on accuracy, however detailed examination reveals that most datasets benefit, while accuracy on FGVC-Aircraft collapses due to MAND discovering almost no novel categories. This is consistent with the dataset's low CLIP recall, indicating that our prior estimation cannot work here. Adding the recall gate reverts the regression on Aircraft, while maintaining the gains on other datasets. For a more detailed discussion of the recall gate we refer to Appendix C. It is also clear that many optimizations benefit the pretrained model more than the finetuned model, with the gap between the two gradually narrowing as they are added. Adding additional reference images however only has a positive effect when finetuning the LLM, justifying our choice to not include them for the pretrained model. This shows that one key benefit of finetuning is the ability to leverage information the pretrained model cannot process.

To verify the effectiveness of our similarity-based stopping criterion, we perform an ablation study in which we compare the criterion to two methods based on category count statistics. The first criterion is based on the RichnEst estimator (Daley & Smith, 2013) originally developed for predicting the

Table 3: Ablation comparing different stopping criteria. We report the mean over all ten datasets.

	Pretrained				Finetuned			
Stopping Criterion	cAcc	sAcc	Steps	Cat. error	cAcc	sAcc	Steps	Cat. error
Similarity gap (default)	75.1	80.2	2341	19.1	76.9	80.3	1995	23.5
RichnEst	74.7	79.8	7889	31.0	76.3	80.0	6628	34.8
Coverage	74.9	80.8	2507	21.2	76.6	80.5	2381	23.7

number of distinct molecules in a sequencing library using the count frequencies of a sample. We apply it by first running discovery for 10% of the dataset, then estimating the total category number. Discovery is then stopped when that number has been reached. The second criterion is based on sample coverage, defined as the share of the population belonging to the currently discovered categories. The coverage is estimated from the number of singletons and doubletons in the seen data following Chao & Jost (2012). After a sweep, we set the coverage target at 0.997 and use this target value to stop the episode. The results of the comparison are displayed in Table 3. They show that RichnEst requires more steps, while also lagging in accuracy due to a tendency to overdiscover. Coverage on the other hand is more competitive, achieving accuracy similar to our criterion but requiring more steps, especially for the finetuned model.

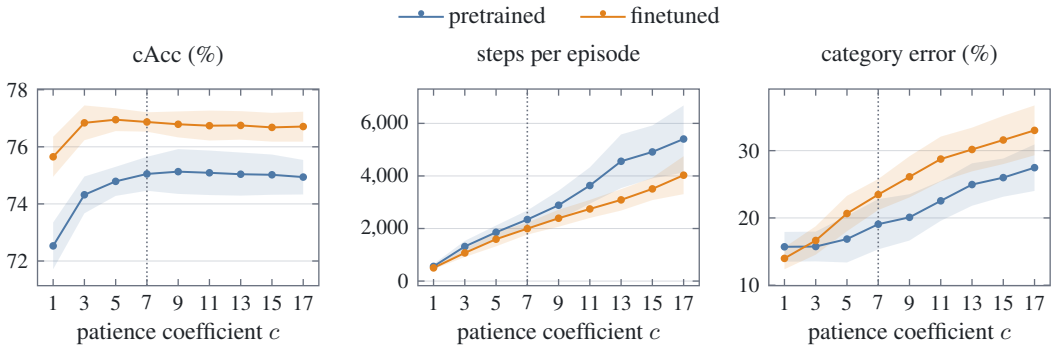

Figure 3: Sensitivity analysis of the patience coefficient c. We measure the impact on cluster accuracy, episode length, and category count error.

To investigate the impact of varying the patience coefficient c, we conduct a sensitivity study in Figure 3. The results show a quick ramp-up in cluster accuracy followed by a long plateau, demonstrating the criterion's robustness. However, we observe increased episode lengths as c increases, highlighting that a smaller c remains desirable for efficiency reasons. The category error increases as well for larger c, despite the plateau in cAcc, indicating that the categories being added are spurious, with few samples assigned to them. The comparison between the two models shows the finetuned model requiring a lower c and fewer steps to achieve its peak accuracy. This might be a result of the prior correction being more accurate on the finetuned model due to the known π_{train}.

5 Conclusion

In this work, we propose Open-World Zero-Shot Image Classification (OWIC), a new task that fills the gap between classical zero-shot classification with all categories given, and Vocabulary-free Image Classification (VIC). Compared to VIC, OWIC is less ambiguous due to the known categories providing implicit guidance on the underlying classification rule. We further propose Multimodal Agentic Novelty Detection (MAND), a method that approaches OWIC using an LLM-based discovery loop. The results show that MAND outperforms the baselines considerably while also demonstrating the value of its individual components. Future work in this field could investigate whether it is possible to move away from name-based classification, which always requires the model to have encountered the category in its training data, towards a more general framework that is able to detect previously entirely unseen categories.

AI USE STATEMENT

In this work, we used generative AI tools to help implement the method, as well as to run exploratory experiments and analyze and interpret their results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, and design or provide feedback on research methodology or experiments. Other required disclosure tasks (generate synthetic data sets, assist with translation, clean and reformat datasets) are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, identify relevant literature, and summarize or analyze existing literature. We have reviewed all AI-assisted work, meaning we verified LLM-generated code and always followed up on exploratory AI-led experiments with manual experiments. All experimental results reported in this work are the result of manually decided and launched experiments. AI-supported literature search was always backed up by manual literature search to ensure completeness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We enable reproduction of our results through a detailed description of our method in section 3, with additional implementation details given in section 4, such as the exact models used. More details are given in Appendix A, including the exact prompts used to query the LLM. To reproduce our evaluation setup, we state the datasets and metrics used in section 4, as well as the adaptations we made to the baselines in Appendix B. In addition, we will publish our source code upon acceptance to further increase the ease of reproducing our results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60884–60937, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021.

Anne Chao and Lou Jost. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93(12):2533–2547, 2012.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 3606–3613, 2014.

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. *Advances in Neural Information Processing Systems*, 36: 30662–30680, 2023.

Timothy Daley and Andrew D Smith. Predicting the molecular complexity of sequencing libraries. *Nature methods*, 10(4):325–327, 2013.

Dmitry Demidov, Muhammad Zaigham Zaheer, Zongyan Han, Omkar Thawakar, and Rao Anwer. Thinking beyond labels: Vocabulary-free fine-grained recognition using reasoning-augmented LMMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16855–16864, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

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 *2004 Conference on Computer Vision and Pattern Recognition workshop*, pp. 178–178. IEEE, 2004.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8400–8408. IEEE, 2019.

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*, 12(7):2217–2226, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. OpenReview.net, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *International Conference on Learning Representations*. OpenReview.net, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*. OpenReview.net, 2019.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

George A Miller. WordNet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Maria-Elena Nilsback and Andrew Zisserman. 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.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen Blog, 2026. URL `https://qwen.ai/blog?id=qwen3.5`. Accessed September 22, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*. OpenReview.net, 2021.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2022.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492. IEEE, 2010.

A Additional Details on MAND

A.1 Prompts used during the Discovery Loop

The decision stage prompt is

> {image}{ref images} Does the image belong to one of these categories? The categories are {category prompt}. After the first image, the most visually similar images to it are shown as references; they are usually related but may not always be of the same category. Respond only with 'yes' or 'no'.

Here {image} corresponds to the query image, {ref images} to the reference images, and {category prompt} is a comma separated list of the category names. When not using reference images, the images and the corresponding explanatory text are omitted. During the naming stage, the conversation is continued with

> Find a suitable category name for the {image str} that does not appear in the list and that follows the same classification granularity and pattern as the given categories. Respond with the category name only.

Here, {image str} is "first image" when using reference images, and "image" otherwise.

A.2 Task Sampling Procedure during Finetuning

Listing 1: Simplified Python-style pseudocode for the task sampling procedure. In practice we set the minimum number of known and novel categories to 5, and the minimum number of samples per class to 100.

```
def get_random_task(
        hierarchy,
        imagenet_dataset,
        min_known_classes,
        min_novel_classes,
        min_samples_per_class,
        rng,
):
    # Weight candidate nodes by how many leaves are reachable from them
    candidate_roots, candidate_weights = get_candidates(hierarchy)
    root_node = rng.choice(candidate_roots, p=candidate_weights)

    # Weight the levels by the number of reachable leaf nodes
    # Otherwise some levels would be overrepresented
    # compared to how much data they contain
    levels, level_weights = get_level_weights(root_node, hierarchy)
    target_level = rng.choice(levels, p=level_weights)
    nodes_at_level = get_nodes_at_level(root_node, target_level, hierarchy)

    # Randomly select the number of total / known classes
    min_num_classes = min_known_classes + min_novel_classes
    max_num_classes = len(nodes_at_level) + 1
    num_classes = rng.integers(
        low=min_num_classes,
        high=max_num_classes,
    ).item()
    max_known_classes = num_classes - min_novel_classes + 1


```

```
    num_known_classes = rng.integers(
        low=min_known_classes,
        high=max_known_classes,
    ).item()

    # Randomly select the classes for the episode
    all_classes = rng.choice(
        nodes_at_level,
        size=num_classes,
        replace=False,
    )
    known_classes = rng.choice(
        all_classes,
        size=num_known_classes,
        replace=False,
    )

    # Identify the samples we are going to use in this episode
    imagenet_labels = imagenet_dataset.targets
    transformed_labels = transform_to_level(imagenet_labels, target_level)
    sample_mask = np.zeros_like(transformed_labels, dtype=bool)

    # Randomly select a subset of samples from the selected classes
    for category in all_classes:
        category_samples = np.where(transformed_labels == category)[0]

        min_samples = min(min_samples_per_class, len(category_samples))
        max_samples = len(category_samples) + 1
        num_samples = rng.integers(
            min_samples,
            max_samples,
        ).item()
        selected_samples = rng.choice(
            category_samples,
            size=num_samples,
            replace=False,
        )
        sample_mask[selected_samples] = True

    return all_classes, known_classes, sample_mask
```

B ADAPTING CASED AND FINDR TO OWIC

To enable a fair comparison, we adapt the two VIC baselines to OWIC by enabling them to leverage the known categories as additional information, thus putting them on an even footing with MAND. We adapt CaSED to OWIC by simply injecting the known classes into the candidate set for every image. For FiNDR, we include the known classes in the meta prompt, allowing the model to infer the meta information more reliably by considering the known categories. In addition, we also include them in the main prompt to enable the model to recognize known-category images more easily. We also change the operating mode of FiNDR regarding the discovery set. In the original paper, FiNDR operates on a class-balanced discovery set with three images per class, which is used to perform discovery. We believe that this requirement would often be impossible to fulfill in practical applications, as a balanced sample cannot be guaranteed without labels. In addition, choosing the discovery set size dependent on the number of categories leaks the category count and circumvents the trade-off between class coverage and potential oversplitting. To enable a fair comparison, we

Table 4: Sweep over FiNDR discovery set size factor k. We report the mean values across all datasets.

	FiNDR		FiNDR adap.	
k	cAcc	sAcc	cAcc	sAcc
25	50.5	50.3	64.0	74.6
50	55.9	53.9	69.1	76.9
75	57.2	55.7	69.1	77.0
100	56.9	56.8	68.9	76.8
150	55.7	57.9	65.9	76.3
200	54.2	58.5	63.8	75.3

instead randomly select a discovery set from the evaluation data with a size of $k\log(N)$, where k is a constant factor. We sweep over k to find the optimal value and set it to 75 according to the results in Table 4.

C Additional Results

Table 5: Full main results for cluster accuracy, split by known/novel. Best in **bold**, second underlined.

	Aircraft		Caltech101		DTD		EuroSAT		Flowers102	
Method	known	novel	known	novel	known	novel	known	novel	known	novel
CaSED	28.9	29.4	45.1	76.3	29.1	35.5	28.3	19.3	75.6	77.3
FiNDR	57.6	59.3	52.9	55.0	39.7	40.4	42.3	36.3	83.6	78.8
CaSED adap.	54.4	33.1	48.6	76.3	32.1	34.7	28.0	25.1	83.0	75.5
FiNDR adap.	61.2	59.6	68.1	82.8	**65.8**	39.2	54.3	48.1	86.3	78.8
MAND pre	63.5	47.6	**89.1**	87.6	60.0	**54.2**	72.0	**54.9**	**92.3**	70.1
MAND ft	**63.9**	**62.9**	85.9	**92.3**	57.5	54.1	**75.2**	49.3	90.9	**82.6**

	Food101		Pets		SCars		SUN397		UCF101	
Method	known	novel	known	novel	known	novel	known	novel	known	novel
CaSED	66.5	67.7	55.0	72.1	48.5	36.6	42.7	41.6	49.1	46.5
FiNDR	58.2	62.8	77.8	79.7	78.2	74.4	32.9	30.6	54.5	48.6
CaSED adap.	82.4	67.0	76.4	65.3	**86.3**	42.5	53.3	39.1	65.2	46.6
FiNDR adap.	82.6	74.0	81.9	82.7	80.3	69.7	66.8	47.8	81.6	74.4
MAND pre	90.0	83.4	87.3	87.1	86.0	**79.8**	69.1	65.0	88.0	72.7
MAND ft	**90.7**	**87.1**	**89.4**	**92.2**	83.2	74.5	**70.8**	**65.8**	**89.3**	**80.6**

Here, we present additional results that could not be included in the main paper due to space constraints. First, we show a detailed view of our main results in Table 5 and Table 6, showing the cAcc and sAcc separately for known and novel categories. As expected, all OWIC methods on average perform better on the known categories than on the novel ones. However, it is notable that the adapted version of FiNDR also outperforms the VIC version on the novel classes, showing that the additional task information contained in the known categories can guide the discovery of novel classes. CaSED does not show the same gains, most likely due to not using an LLM that would be able to leverage the additional context into better classification.

In addition, we examine the CLIP recall metric that we use to decide whether to apply prior correction. In Table 7, we show the CLIP recall, the estimated novel rate, and the true novel rate at the start of the episode for each dataset. We observe that for most datasets the CLIP recall is high, only FGVC-Aircraft stands out at 0.23, making it the only dataset for which prior correction is disabled at $\theta = 0.5$. This coincides with the highest estimation error, showing that CLIP recall is a useful metric for predicting the utility of prior correction. For the other datasets the estimation error is much lower, consistent with their higher CLIP recall. UCF101 has the second-lowest recall at 0.6, and also the second-highest error at -0.14, while all other datasets remain below an absolute error of 0.1. These

Table 6: Full main results for semantic accuracy, split by known/novel. Best in **bold**, second underlined.

	Aircraft		Caltech101		DTD		EuroSAT		Flowers102	
Method	known	novel	known	novel	known	novel	known	novel	known	novel
CaSED	34.8	36.7	63.3	80.7	40.6	45.8	24.7	33.3	52.8	62.1
FiNDR	69.3	67.0	44.9	47.6	38.1	42.4	24.5	30.8	54.1	59.7
CaSED adap.	76.4	52.4	69.7	80.7	47.0	46.2	42.4	43.3	74.5	61.9
FiNDR adap.	74.4	69.0	83.2	80.4	**81.1**	49.2	60.1	**59.6**	76.5	64.0
MAND pre	**91.2**	60.8	**97.0**	**80.8**	75.5	54.2	80.7	40.9	**87.9**	65.7
MAND ft	87.4	**74.9**	94.7	74.7	72.2	**54.3**	**81.0**	45.1	79.8	**69.7**

	Food101		Pets		SCars		SUN397		UCF101	
Method	known	novel	known	novel	known	novel	known	novel	known	novel
CaSED	69.8	72.5	62.3	71.3	54.2	57.6	57.7	57.8	59.1	56.6
FiNDR	68.2	72.2	83.3	**87.5**	79.0	82.3	40.4	38.2	44.8	41.5
CaSED adap.	90.3	72.5	88.5	71.4	97.6	67.2	72.7	58.1	83.5	57.7
FiNDR adap.	90.7	**85.6**	87.5	87.3	95.9	84.6	84.1	63.3	91.3	70.0
MAND pre	92.3	81.8	92.2	87.0	98.7	**90.1**	84.8	**71.5**	94.0	69.6
MAND ft	**92.8**	83.6	**92.9**	87.2	**99.1**	87.7	**84.9**	70.3	**94.5**	**71.2**

Table 7: Overview of the different CLIP recall values for each dataset, as well as their initial novel rate estimate at the start of the episode, compared to the true novel rate.

Dataset	CLIP recall	Estimated Novel Rate	True Novel Rate	Error
Aircraft	0.23	0.15	0.50	-0.35
Caltech101	0.77	0.30	0.35	-0.05
DTD	0.76	0.57	0.51	+0.06
EuroSAT	0.88	0.53	0.48	+0.05
Flowers102	0.89	0.57	0.50	+0.07
Food101	0.93	0.48	0.50	-0.01
Pets	0.95	0.54	0.49	+0.05
SCars	0.94	0.49	0.51	-0.02
SUN397	0.66	0.50	0.47	+0.03
UCF101	0.60	0.34	0.49	-0.14

results show that the recall gate only disables prior correction in cases where the novel rate estimate is genuinely unusable.

Table 8: Comparison of different pretrained LLMs as part of MAND. The results report the mean across all datasets.

LLM	cAcc	sAcc	Steps	Cat. error
Qwen3.5-4B (default)	75.1	80.2	2341	19.1
Qwen3-VL-4B	72.6	78.5	2541	30.1
Gemma4-E4B	69.1	74.0	2017	26.7

Lastly, we investigate the performance of MAND on different LLM backbones. We replace Qwen3.5-4B with Qwen3-VL-4B (Bai et al., 2025) and Gemma4-E4B (Team et al., 2026), respectively. The results in Table 8 show that Qwen3.5 remains the strongest model, with Qwen3-VL placing second and Gemma 4 last.

D Discussion of Limitations

The most severe limitation of our approach is that the ability to accurately classify a concept depends on both the LLM and the CLIP models having had prior exposure to the concept during their pretraining. This means that, for example, if we were to extend the Stanford Cars dataset with a new car model that was not in the training data for either model, our discovery loop would be unable to accurately name it. This reduces the utility of our method to identifying categories that are novel in the sense that their presence in the data was not previously known, whereas categories that are novel in the sense that they have never been encountered before cannot be handled effectively. In practice, both LLMs and CLIP models are trained on enormous datasets that are updated for each new model generation to ensure that the models have access to the most recent knowledge, which greatly reduces the chance of encountering concepts that are genuinely unknown to the models. In addition, data contamination remains an issue when benchmarking LLMs and CLIP models. Given that all image classification benchmarks used in this paper are multiple years old, it cannot be ruled out that the images were part of the training data for the models we are using. These limitations are not unique to MAND, instead they also apply to previous VIC methods, such as FiNDR. Overcoming them remains a promising research direction that could enable truly generalist discovery.