

Retrieval-enriched zero-shot image classification in low-resource domains

Nicola Dall’Asen^{1,2} Yiming Wang³ Enrico Fini^{1*} Elisa Ricci^{1,3}

¹University of Trento ²University of Pisa ³Fondazione Bruno Kessler

Correspondence: nicola.dallasen@unitn.it

Project website: <https://fodark.github.io/CoRE>

Abstract

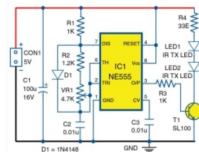
Low-resource domains, characterized by scarce data and annotations, present significant challenges for language and visual understanding tasks, with the latter much under-explored in the literature. Recent advancements in Vision-Language Models (VLM) have shown promising results in high-resource domains but fall short in low-resource concepts that are under-represented (*e.g.* only a handful of images per category) in the pre-training set. We tackle the challenging task of zero-shot low-resource image classification from a novel perspective. By leveraging a retrieval-based strategy, we achieve this in a training-free fashion. Specifically, our method, named CORE (Combination of Retrieval Enrichment), enriches the representation of both query images and class prototypes by retrieving relevant textual information from large web-crawled databases. This retrieval-based enrichment significantly boosts classification performance by incorporating the broader contextual information relevant to the specific class. We validate our method on a newly established benchmark covering diverse low-resource domains, including medical imaging, rare plants, and circuits. Our experiments demonstrate that CORE outperforms existing state-of-the-art methods that rely on synthetic data generation and model fine-tuning.

1 Introduction

Low-resource domains refer to those rare domains where the data or its annotation is truly scarce. Similarly, low-resource languages are those that have significantly less content available online (Maugeresse et al., 2020) with respect to other high-resource languages, like English. There exist abundant research on the topic in the context of natural language processing (Ranathunga et al., 2023; Adams et al., 2017; Fadaee et al., 2017; Pan et al.,

*Currently at Apple

Input



Retrieved captions

“**Electronics** that control **LED** patterns”
“The basic **circuit diagram** of **LED**”
“This clearly Image **Circuit or Diagram**”

“A photo of a melanoma”

“**Skin Cancer** The Dangers of **Melanoma**”
“Two types of **skin melanoma** on the neck”
“**Mole** on the human skin”



“The soft gradient of **Japanese Painted Fern**”
“This is a **plant** imported from **Japan**.”
“Image of a Tree **Fern** and **Japanese maple**”

Figure 1: Our retrieval-based solution enriches both images and textual descriptors with real-world captions which contain **domains** and **classes**. Even when the captions are generic (third row for each example), they can still restrict the focus to the correct domain.

2017). However, surprisingly, the vision counterpart, *i.e.* *low-resource visual domains*, is much under-explored despite the numerous practical applications. In this paper, we focus on classifying images in low resource domains, *i.e.* where we can find only a handful of images per category. The causes for such limited data can be various: for example, when only certain devices are capable of capturing the visual content, *e.g.* astronomy or medical imaging; the visual content itself is sensitive or private, *e.g.* due to privacy issues, or rarely appears in nature, *e.g.* deep ocean animals, or other long-tailed categories. Their associated annotations can also be limited due to the expertise requires, in particular for niche fields, *e.g.* electric design or phytology.

Recent large vision-language models (VLMs) have fostered a paradigm shift in image classification. Their flexibility and generalization, enabled by web-scale pre-training with text-image pairs,

makes them versatile tools in many sub-fields of computer vision. Numerous works have appeared, with the objective of tuning VLMs, *e.g.* CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023), to address image zero-shot (Jia et al., 2022) or few-shot (Chowdhury et al., 2023; da Costa et al., 2023) image classification. However, the images involved in those studies are mostly in *high-resource* image domains, where there exist thousands of images on the Internet for VLMs to learn from during the web-scale pre-training (Udandarao et al., 2024).

However, directly performing zero-shot classification in low-resource domains does not yield satisfactory performance due to the data scarcity in pre-training. Even supervised fine-tuning might fall short in learning the underlying data distribution due to the very limited amount of data and annotation. Among the techniques that have been explored in the pioneering work (Zhang et al., 2024), one prominent recipe is to fine-tune the VLMs on data augmented via synthetic generation (*e.g.* Stable Diffusion (Rombach et al., 2021)). Despite the performance improvements, by analyzing the generated images, we observe that image generation models are also affected by the low-resource nature of the task. The generation quality is largely dependent on the noise injected on the real samples: by injecting limited noise, the synthetic images appear very similar to the original samples, being correct but not diverse, while by injecting more noise, the synthetic images diversify in appearance, but are mostly semantically incorrect and exhibit domain-specific rule violations. This is because the data distribution of rare domains is not well-represented in the generative models latent space (Mokady et al., 2022; Trabucco et al., 2024).

Instead of generating synthetic images as data augmentation, we explore the possibility of retrieving relevant information from a textual corpus, crawled from the Internet, to enrich the data representation at inference time, as shown in Fig. 1. It turns out that retrieval is also non-trivial in the low-resource regime as (i) pre-trained models generally under-represent the low-resource domains, thus greatly limiting the retrieval efficacy; (ii) large web-crawled databases can contain noisy or incorrect content, a problem that is more severe in low-resource domains.

Thus, a careful design is required to leverage the retrieved data. In this work, we propose the first training-free and retrieval-based method, CoRE (Combination of Retrieval Enrichment), to

tackle low-resource image classification. Following a VLM-based zero-shot classification paradigm, we propose to enrich the representation for both the query image and the class prototypes with textual content retrieved with different encoder backbones from large web-crawled databases. Specifically, for the query image, we employ the pre-trained image encoder from a VLM as our vision retrieval backbone. We perform image-to-text retrieval, obtaining the most relevant captions with respect to the query image.

From our preliminary analysis, we observe that although the specific category (*e.g.* “LED”) appears sparsely in the retrieval, its broader category (“circuit”) does occur frequently. Previous studies have empirically demonstrated that enriching the prompt with the broader concept, together with noise, can significantly boost the zero-shot recognition performance (Roth et al., 2023). We thus enrich the image embedding by combining it with the textual embedding from image-to-text retrieval. Similarly, we construct the enriched class prototypes. For each class, we form its corresponding text prompt and embed it with a pre-trained text encoder to retrieve captions that are most relevant. Then, the retrieved captions are encoded with the VLM text encoder and aggregated together with the textual embedding of the original class prompt. The final categorization is obtained by computing the cosine similarity between the enriched visual representation against the enriched textual class prototypes.

To validate the effectiveness of our proposed method, we also collect a set of datasets that covers diverse low-resource domains, including medical imaging, rare plants, and circuits. CoRE can effectively improve the data representation in a training-free fashion, with a noticeable improvement in image classification performance on all the datasets, outperforming the state-of-the-art method that involves synthetic image generation and model fine-tuning.

To summarize, our contributions are:

- We propose the first training-free retrieval-based method CoRE for addressing zero-shot low-resource image classification;
- we propose a data representation enrichment strategy for both query image and class prototypes, using the textual content retrieved from the database;
- we establish a benchmark featuring zero-shot

low-resource image classification, composed of representative datasets and VLM-based baselines and state-of-the-art methods;

- our training-free method is effective in classifying low-resource images, outperforming competitors with training and other training-free baselines by a large margin.

2 Related work

High-resource data. The *de-facto* standard in Deep Learning has become to train larger foundation models with a high volume of data, *e.g.* ImageNet (Deng et al., 2009) or LAION-5B (Schuhmann et al., 2022) for vision, which contain up to 5 billion images, or FineWeb (Penedo et al., 2024) for text, which contains 15 trillion of tokens. For vision-related tasks, data focuses on natural images, which are plentiful online and can be easily obtained. However, when moving to more specific domains, *e.g.* medical (Irvin et al., 2019), satellite imaging (Helber et al., 2019), or long-tailed distributed data (Van Horn et al., 2018), data becomes less available. Although these rare domains have been understudied in favor of higher-available data, we follow the study of (Zhang et al., 2024) and investigate domains where the number of available data is in the order of hundreds, and training is an under-performing option.

We propose a novel way to address the challenge, and we exploit the knowledge coming from web-scale image-text pairs datasets through retrieval. Retrieval-based solutions have proved successful for image classification (Liu et al., 2023b; Conti et al., 2023) and NLP tasks (Lewis et al., 2020), and we propose to enrich the data representation in VLMs.

Multimodal foundation models. Multimodal Foundation Models, such as CLIP (Radford et al., 2021) or BLIP (Li et al., 2022, 2023) have gained a lot of popularity due to their outstanding zero-shot capabilities, derived from their weak-supervised training on web-crawled data. The most common paradigm is to train on image-text pairs, which can be easily obtained on the web (Changpinyo et al., 2021; Schuhmann et al., 2022), but recent approaches like ImageBind (Girdhar et al., 2023) bridge several modalities, *e.g.* images, videos, audio, text, and thermal. We focus in particular on Vision-Language Models (VLM).

Adaptation of VLM to downstream tasks is performed in several ways, *e.g.* full finetuning, or

Parameter-Efficient FineTuning (PEFT) methods, *e.g.* LoRA (Hu et al., 2022), Prompt Tuning (Jia et al., 2022), or Textual Prompt Tuning (Zhou et al., 2022). These approaches are not suitable for the rare domain setting as the amount of data is limited, domain-specific, and extremely different from the pre-training data of the original VLM.

VLM few-shot learning. Several works have proved the effectiveness of few-shot learning when adapting VLMs to downstream tasks. Some notable works include a combination of PEFT strategy with VLMs, such as APoLLO (Chowdhury et al., 2023) that synthetically augments both the visual and textual branch of CLIP, or DISEF (da Costa et al., 2023) that employs LoRA and synthetic images to fine-tune CLIP. We closely follow the work of (Zhang et al., 2024), where the authors fine-tune ImageBind (Girdhar et al., 2023) with an AdaptFormer (Chen et al., 2022) module on a combination of real and synthetic data for low-resource rare domains. Differently from these works, we do not employ synthetic data and we do not fine-tune, instead, we propose a training-free zero-shot solution through retrieval to adapt both the visual and textual representation in CLIP.

3 Background

To establish a foundation in understanding our method, we provide a brief introduction to Vision-Language Models and web-scale databases, the two essential elements in our method design.

3.1 Vision-Language Models

Vision-Language Models (VLM) (Radford et al., 2021; Jia et al., 2021) learn a function $f_{VLM} : \mathcal{V} \times \mathcal{L} \rightarrow \mathbb{R}$ with the goal to maximize their representation similarities \mathbb{R} in order to map images in visual space \mathcal{V} and texts in language space \mathcal{L} to the same latent space. In particular, a VLM is composed of a vision encoder $f_{VLM}^V : \mathcal{V} \rightarrow \mathbb{R}^d$ that maps images to a visual embedding and a language encoder $f_{VLM}^L : \mathcal{L} \rightarrow \mathbb{R}^d$ that maps the text in natural language (after converting into tokens) to an embedding. VLMs learn to project the two modalities to the same latent space \mathbb{R}^d via contrastive learning with millions of web-crawled image-text pairs, enabling image classification using text by evaluating their similarity in this shared space.

At inference, the VLM can be used to assess the similarity between an image sample and a set of prompt texts that are composed from a number of

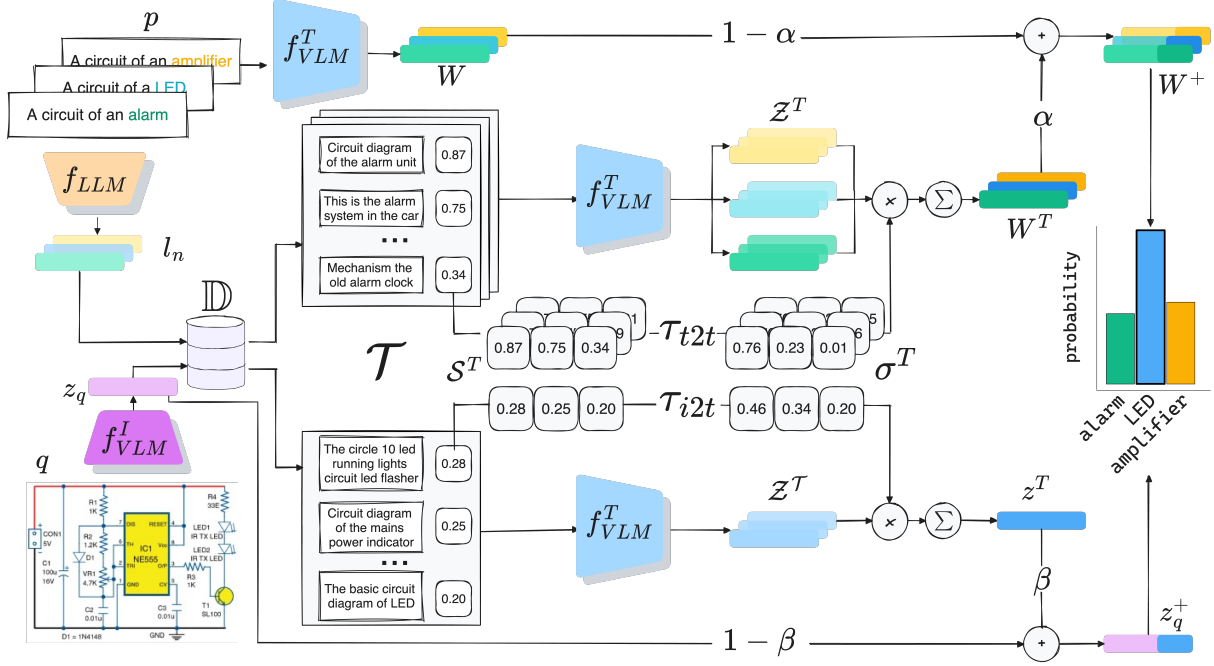


Figure 2: Our CORE enriches both the image embedding z_q and the class prompts p with retrieved captions from a large-scale web-crawled database \mathbb{D} . We weight the retrieved captions \mathcal{T} with their similarity scores S^T , which we skew with controllable temperatures τ_{i2t} and τ_{t2t} . By combining the retrieved captions embedding with the original representations W and q through α and β , we obtain enriched representations W^+ and z_q^+ which we employ for zero-shot classification.

N classes $\{c_n\}_{n=1}^N$ (Radford et al., 2021). Given a query image $q \in \mathcal{V}$, we can obtain its visual embedding via the image encoder as $z_q = f_{VLM}^V(q)$. We can build the class prototypes $W \in \mathbb{R}^{N \times d}$ as the textual embeddings using the VLM text encoder. Specifically, for each class c_n , we build the text prompt $p_n = \{\text{prefix}\} [\text{CLS}]_n$, where $\{\text{prefix}\}$ usually is “a photo of a” and $[\text{CLS}]_n$ is the name of c_n in text. Then the class prototypes can be formed as $W = f_{VLM}^L(\{p\}_n)$. Finally, we can compute the cosine similarities between the image embedding z_q and the class prototypes W in order to predict the class \hat{c} for the query image q :

$$\hat{c} = \arg \max_c (z_q \times W^T). \quad (1)$$

3.2 Retrieval databases

Together with the VLMs, there is also the emergence of large vision-language databases that can be leveraged for VLM pre-training and retrieval. The community has collected and released several web-scale image-text pairs datasets, such as LAION (Schuhmann et al., 2022), CC12M (Changpinyo et al., 2021), and COYO (Byeon et al., 2022), with respectively 400M/5B, 12M and 700M image-text pairs. Formally, let $\mathbb{D} = \{(i, t)_m\}_{m=1}^M$ be the

retrieval database that consists of M items, where each item is paired with an image i_m and its associated textual description t_m . We are mostly interested in the textual content of \mathbb{D} as it contains rich language-induced semantics (Conti et al., 2023) that might contribute to enriching the knowledge that is specific low-resource domain.

With millions and billions of data items, it is critical performing retrieval efficiently. To this purpose, we focus on embedding-based retrieval, whose goal is to retrieve the most similar elements from a database from a query embedding. Such retrieval is supported by off-the-shelf tools, e.g. Faiss (Douze et al., 2024), that enable similarity search and clustering of dense vectors at scale. In particular, we prepare the text-image database \mathbb{D} by encoding data items into embeddings using encoders of VLMs. Given an embedding z corresponds to either visual or textual modality, we can retrieve a set of k textual descriptions \mathcal{T} that are most similar to the z from the database \mathbb{D} :

$$\mathcal{T} = \text{top-k}_t(\langle z, f_{VLM}^L(t) \rangle), \forall t \in \mathbb{D}, \quad (2)$$

where $\langle \cdot \rangle$ computes the cosine similarity between two embeddings, and $\text{top-k}(\cdot)$ returns k textual descriptions with the highest cosine similarities.

4 CORE

We focus on zero-shot low-resource image classification, following the paradigm of VLMs as described in Sec 3.1. Low-resource domains are generally not well represented by VLMs given their limited availability in the pre-training dataset (Udandarao et al., 2024). There is a need to steer the original data representation towards more specialized low-resource domains represented by the query image q .

To this end, our training-free method CORE extracts domain-relevant information from a large text-image database \mathbb{D} , to enrich the representation of both class prototypes and the query image.

For the class prototypes, *i.e.* textual representation of the classes $\{c_n\}_{n=1}^N$, we retrieve semantically close captions for each class with the embedding of prompt text regarding the c . On one hand, the retrieved captions contain rich domain-level information while being less specific to the exact class. On the other hand, the class prompt is only specific to the class of interest without much prior of the domain information. Therefore, we further join the retrieved captions with the prompt class text to obtain the textual class prototypes W^+ enriched with the domain context.

A similar rationale is applied to enrich the query image representation via image-to-text retrieval, implemented by the image encoder f_{VLM}^V of a pre-trained VLM. With the visual embedding z_q of the query image q , we retrieve the set of captions that are the most aligned to z_q in the shared latent space, and use the retrieved captions to obtain an enriched image representation z_q^+ . The final class is predicted using z_q^+ and W^+ similar to Eq. 1:

$$\hat{c} = \arg \max_c (z_q^+ \times W^{+\top}). \quad (3)$$

We describe each retrieval branch in detail in the following sections and we show an overview of our CORE in Fig. 2.

4.1 Class representation enrichment

We leverage text-to-text retrieval to enrich the class prototypes of the set of predefined classes $\{c_n\}_{n=1}^N$. For each class c_n , we first build the text prompt p_n following the prompt format as described in Sec. 3.1. Different prompt templates are also experimented in Sec. 5.2.

For the text-to-text retrieval, we leverage the encoder of a LLM to obtain the textual embeddings for both the per-class text prompt, *i.e.* $l_n =$

$f_{LLM}(p_n)$ and all textual content in the database \mathbb{D} . For each l_n , we then retrieve from \mathbb{D} a set of k most similar textual descriptions \mathcal{T}_n with respect to class c_n . Each retrieved text $t_i \in \mathcal{T}_n$ has an associated cosine similarity in the range of $[-1, 1]$ *w.r.t.* the embedding of the prompt text l_n , forming a set of k scores \mathcal{S}_n^T .

As the retrieved texts contain rich domain-level context, we further embed them and merge their textual embeddings to obtain a domain-specialized embedding. Since the eventual classification is achieved in the latent space of VLMs, we leverage the text encoder of the VLM to embed the retrieved texts, *i.e.* $\mathcal{Z}_n^T = f_{VLM}^L(\mathcal{T}_n)$. As the retrieved captions are associated with different similarity scores \mathcal{S}_n^T , we weigh their contribution to form the domain-specialized embedding accordingly. Specifically, we propose to build a probability distribution out of the similarities scores as:

$$\sigma_n^T = \text{softmax} \left(\frac{\mathcal{S}_n^T}{\tau_{t2t}} \right), \quad (4)$$

where τ_{t2t} is the temperature parameter that controls the skewness of the distribution.

For the class c_n , the embedding from retrieved texts \mathcal{Z}_n^T are then combined as a weighted sum with the weight being σ_n^T , forming the domain-specialized embeddings for all classes W^T .

Finally, we build the retrieval-enriched class prototypes by linearly interpolating prompt-text class prototypes W as described in Sec 3.1 and the retrieved domain-specialized ones W^T , with a interpolation factor α as:

$$W^+ = \alpha W^T + (1 - \alpha)W, \quad (5)$$

where α is a hyperparameter and we shows its impact in Section 5.2.

4.2 Image query representation enrichment

Symmetrically, we want to enrich the query image representation and exploit the rich web-scale semantics of captions to build a more representative image embedding.

Starting from a query image q . For image-to-text retrieval, we use the visual encoder of the VLM to obtain the image embedding $z_q = f_{VLM}^V(q)$, while we use the text encoder f_{VLM}^L to embed the database \mathbb{D} , in this way image and text are mapped in the same space. We retrieve from \mathbb{D} the k most similar captions \mathcal{T} , with their associated cosine similarities \mathcal{S}^T . Similar to the class prototype




Method	Circuits		iNaturalist2021 (LT100)		HAM10000	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
ImageBind  (Zhang et al., 2024)	24.10	49.30	31.60 [†]	60.50 [†]	54.60 [†]	96.56 [†]
SigLIP@384px  (Zhai et al., 2023)	19.53 [†]	30.61 [†]	34.50 [†]	63.50 [†]	54.60 [†]	95.90 [†]
CLIP ViT-L (Radford et al., 2021)	7.98	29.13	8.00	22.60	45.27	90.80
CLIP ViT-L@336px (Radford et al., 2021)	9.09	30.33	7.60	22.70	40.97	90.27
BLIP2-EVA (Li et al., 2023)	17.63	N/A	1.40	N/A	2.91	N/A
LlaVA 1.6 34B (Liu et al., 2023a)	29.59	N/A	0.60	N/A	10.59	N/A
ImageBind (Girdhar et al., 2023)	22.36	51.02	6.70	23.90	14.43	84.25
SigLIP@384px (Zhai et al., 2023)	35.81	58.63	19.10	45.70	57.64	96.16
CoRE (Ours — CC12M)	<u>42.94</u> _{7.13}	<u>67.71</u> _{9.08}	<u>21.40</u> _{2.30}	42.59 _{3.11}	<u>61.54</u> _{3.90}	<u>95.70</u> _{0.46}
CoRE (Ours — COYO-700M)	43.88 _{8.07}	71.99 _{13.36}	22.10 _{3.00}	<u>44.10</u> _{1.60}	62.21 _{4.57}	94.51 _{1.65}

Table 1: Top-1 and top-5 accuracy on the proposed benchmark. [†] indicates our re-implementation.  denotes fine-tuning. We highlight **best** and second best results. We report **gain** and **loss** *w.r.t.* the best training-free solution.

branch, we use the scores to build a probability distribution:

$$\sigma^T = \text{softmax} \left(\frac{\mathcal{S}^T}{\tau_{i2t}} \right), \quad (6)$$

where τ_{i2t} controls the final distribution skewness.

We then embed the retrieved captions with the VLM text encoder $\mathcal{Z}^T = f_{VLM}^L(\mathcal{T})$, and combine them as a weighted sum using σ^T , obtaining z^T .

Finally, the original query image embedding z_q and the retrieved captions embedding z^T are linearly interpolated to build the final query representation as:

$$z_q^+ = \beta z^T + (1 - \beta) z_q, \quad (7)$$

where the hyperparameter β controls the importance of the two components and we study the effect of this parameter in Section 5.2.

5 Experiments

We evaluate our proposed method CORE in comparison with state-of-the-art Vision-Language approaches using three challenging low-resource datasets. We describe the dataset used and the evaluation protocol we follow. We discuss the results *w.r.t.* the baseline methods, and we ablate the proposed components and architectural choices.

Datasets. We consider three datasets covering different low-resource scenarios for our analysis: we employ the Circuit Diagram Classification dataset (Zhang et al., 2024)¹ that comprises 1,332 circuit diagram images covering 32 different classes, the images are scraped from the web

and textbooks. The authors split the data into 154 images for training and the remaining for testing, resulting in an average of ~ 5 samples per class available at training time.

The second dataset we consider is iNaturalist 2021 (Van Horn et al., 2018) which contains 10,000 species with a fine-grained classification. The dataset features many visually similar species, captured in a wide variety of situations. In order to remain in the rare domain setting, we restrict our analysis to the rarest 100 species in terms of available training samples and set the maximum amount of training shots to 5. We test on 10 images for each class, therefore the test set contains 1,000 samples.

The last dataset we employ is HAM10000 (Tschandl et al., 2018), which comprises dermatoscopic images from different populations. The dataset has 7 classes that include a representative collection of all important diagnostic categories in the realm of pigmented lesions. The test set includes 1,511 images.

Database(s). We use CC12M (Changpinyo et al., 2021) as our source dataset to build the database \mathbb{D} , following previous works (Conti et al., 2023). We also use a subset (10%) of COYO-700M to test the scaling laws of the retrieval database. We focus on their textual corpus as we are interested in retrieving the relevant pieces of text. The retrieval databases are implemented using Faiss (Johnson et al., 2019) on pre-extracted embeddings. We use the SigLIP (Zhai et al., 2023) text encoder for the image-to-text retrieval, while we employ a text-only encoder to obtain a stronger textual semantic in text-to-text retrieval. In particular, we use SFR-Embedding-Mistral (Meng et al., 2024) as f_{LLM} .

Evaluation protocol. As a common practice in image classification tasks, we report the perfor-

¹The other released datasets relate to retrieval tasks.

mance as the top-1 and top-5 accuracy. For VLMs, we compare the index of the highest logit to the ground truth, while for Large Multimodal Models (LMM) we parse the LMM output to extract the predicted class name and match it with the dataset class names.

Implementation details. CORE leverages an LLM encoder f_{LLM} , implemented as SFR-Embedding-Mistral (Meng et al., 2024), an image encoder f_{VLM}^V , implemented as the SigLIP (Zhai et al., 2023) vision encoder, and a text encoder f_{VLM}^L , implemented as the SigLIP text encoder. The number k of retrieved captions is set to 10 following previous works (Conti et al., 2023), nevertheless, the temperature parameters τ_{t2t} and τ_{i2t} allow restricting the effect of the lower-ranked retrieved captions by skewing the distribution towards the most similar samples.

For our CORE, we empirically find that a high temperature τ_{i2t} for image-retrieved captions leads to more favorable results, therefore we set it to 100. For text-retrieved captions, a lower temperature is more beneficial and we set it to 1, skewing the distribution more towards the high-confidence samples. For the effect of α and β we refer to Sec. 5.2.

For the zero-shot and retrieval prompting strategies, we find that having a domain-specific prompt for zero-shot and a generic prompt for retrieval leads to favorably good results across all the datasets, and we provide a complete study of this in Sec. 5.2. For implementation details of the baselines we refer to Appendix A

5.1 Comparisons

Baselines. We compare against state-of-the-art VLMs and LMMs in the training-free zero-shot scenario. In particular, we compare with CLIP (Radford et al., 2021), using the ViT-L/14 and ViT-L/14@336px vision encoder variants, BLIP-2 (Li et al., 2023) using the ViT-g/14 vision encoder from EVA-CLIP (Fang et al., 2023), LLaVA-NeXT (Liu et al., 2023a) with a 34B LLM, ImageBind (Girdhar et al., 2023) as in (Zhang et al., 2024), and SigLIP (Zhai et al., 2023). For the last two models, we also implement a training-based variant as in (Zhang et al., 2024), which trains an AdaptFormer (Chen et al., 2022) and a linear classification head.

Discussion. We present a quantitative evaluation of CORE and the baselines in Tab. 1. We can see that among the training-free approaches, CORE out-

performs the others by a substantial margin (up to 8.07% in top-1 accuracy on Circuits). When comparing to fine-tuned solutions (🔥), CORE still outperforms them, even given the substantial gain between zero-shot and fine-tuned ImageBind (up to 40% in HAM10000).

Fine-tuning SigLIP seems detrimental to performance, except for iNaturalist, and we deem the lower performance *w.r.t.* fine-tuned ImageBind to the higher number of trainable parameters, as SigLIP embeddings are bigger than ImageBind ones (1152 vs 1024). We study this behavior further in Section 5.2.

LMMs reach satisfactory results on Circuits, while the low performance on iNaturalist and HAM10000 is due to these models answering consistently with the same class name for almost all the samples.

On iNaturalist, supervised fine-tuning represents a stronger solution *w.r.t.* our CORE, and we deem this results to two factors: the nature of the dataset and the availability of the concepts in the retrieval database. Being the images of different plants and animals, the visual overlap is less prominent than in the other two datasets, therefore supervised fine-tuning can effectively separate different classes with a small amount of training data. Secondly, as in iNaturalist class names are provided with their Greek or Latin scientific name, this limits the amount of relevant data that can be retrieved from the database as even in large-scale image-text datasets, such as LAION 400M, this type of data is under-represented (Parashar et al., 2023). We mitigate this effect by using both scientific and common names when retrieving and building the zero-shot weights.

This benchmark allows us to show the training-free strength of our method, but also showcase the viability of training-based solutions given the right assumptions. Nevertheless, CORE remains the strongest training-free solution. We can also observe that employing a bigger retrieval database improves the top-1 accuracy across all the datasets.

5.2 Ablation

Various embedding fusion strategies. We ablate the contribution of each of our proposed components in Tab. 2, starting from the zero-shot only and reaching the full CORE. We first add the weighting between zero-shot weights and retrieved weights without the temperature on the similarities, effectively having a naïve average of retrieved embed-

α	τ_{i2t}	β	τ_{t2t}	Circuits		iNaturalist		HAM10000	
				Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
✓	✗	✗	✗	35.81	58.63	19.10	45.70	57.64	96.16
✓	✗	✗	✗	36.46	64.01	20.90	43.99	60.75	95.76
✓	✓	✗	✗	37.66	65.88	21.09	43.90	61.02	95.57
✓	✓	✓	✗	42.39	69.85	20.80	42.39	61.55	95.43
✓	✓	✓	✓	42.94	67.71	21.40	42.59	61.55	95.70

Table 2: Ablation of proposed components of CoRE CC12M, starting from the zero-shot only to the full CoRE. As shown in the table, each of the components can bring performance improvement across all the datasets, proving the effectiveness of the designed retrieval strategy.

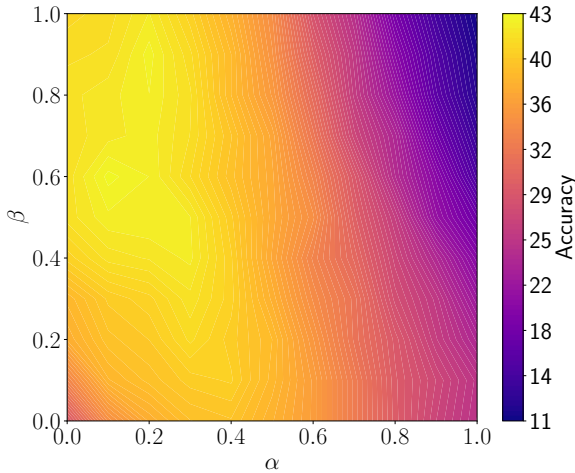


Figure 3: Top-1 accuracy of CoRE CC12M on Circuits with varying α and β . CoRE achieves the best performance with a balanced merge of image-retrieved captions ($\beta \sim 0.5$), while for class-relevant captions the best weighting is slightly lower ($\alpha \sim 0.2$).

dings. We then introduce τ_{i2t} to skew the distribution. We thus move to the image retrieval part, where we first add a naïve average embedding weighted by β , and finally the weighting on the similarities τ_{i2t} . We can see how each of the components brings an improvement in performance across all the datasets, proving the effectiveness of our retrieval strategies.

We further isolate the effect of the two merging weights α and β on the Circuits dataset in Fig. 3, where we can see that CoRE achieves the best performance with a balanced merge of image-retrieved captions ($\beta \sim 0.5$), while for class-relevant captions the best weighting is slightly lower ($\alpha \sim 0.2$). This indicates that the branch that obtains the most improvement from retrieval is the image one, while the zero-shot weights obtained from the classes are already representative, but they can still benefit from retrieved captions.

On the use of LLM for embedding. We ablate the choice of using an LLM to encode the text for

Encoder	Circuits		iNaturalist		HAM10000	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
SigLIP	41.46	64.56	20.90	43.20	60.02	90.53
Mistral	42.94	67.71	21.40	42.59	61.54	95.70

Table 3: Accuracy of CoRE CC12M using different models for text-to-text retrieval. The LLM (Mistral) embedding is stronger than vision-language aligned embedding in terms of the unimodal text-to-text retrieval.

the textual branch and compare this choice to using the VLM textual encoder f_{VLM}^L . In particular, we use the SigLIP@384px text encoder and perform both image-to-text and text-to-text retrieval in the SigLIP latent space. From Tab. 3, we see that LLM embeddings are stronger in the unimodal text-to-text retrieval. While SigLIP embedding is less performing than the LLM embeddings at text-to-text retrieval, the final classification performance is still better than all the other training-free baselines. Therefore SigLIP text embeddings remain a viable solution in case deploying the LLM (7B parameters) is infeasible due to computational constraints.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Zeroshot	30.80	28.48	25.32	31.26	27.83	<u>32.47</u>	35.81
Retrieved	26.07	23.75	27.64	18.00	13.27	19.76	22.73

Table 4: Zeroshot results on Circuits with different prefixes. Numbers denote: (1) a circuit diagram of. (2) a circuit of. (3) an electronic schematic of. (4) a photo of a circuit diagram: a. (5) a picture of a { } circuit. (6) a photo of an electronic circuit: a. (7) a photo of a.

Prompting strategy. We study the role of the “prefix” part of the queries when performing retrieval and when building the zero-shot weights. In Tab. 4 we show the effect of using only the zero-shot weights W or only the retrieved captions weights W^T for zero-shot classification. We can see that the best prefix to build zero-shot weights is not the best prefix to retrieve information from the database, where constraining the domain in the prompt becomes fundamental to obtaining good performance. Additionally, using only retrieved captions as class prototypes leads to unsatisfactory results. We then restrict the study to two styles of the prefix: a generic “a photo of a” and a domain-specific “a {domain} of a”.

We study the effect of these two prefixes in Table 5. For iNaturalist, since the images cover very different natural domains, *e.g.* animals, plants, fungi, etc. we cannot design a domain-specific prompt, and we can only use the domain-agnostic

Zeroshot	Retrieval	Circuits	iNaturalist	HAM10000
Domain	Generic	42.94	N/A	61.55
Domain	Domain	41.84	N/A	54.40
Generic	Domain	41.75	N/A	56.78
Generic	Generic	39.05	21.40	59.30

Table 5: Accuracy of our CORE CC12M with different prompting strategies for zero-shot weights and text-to-text retrieval. Employing a domain-specific prefix for zero-shot and a domain-agnostic prefix for retrieval leads to generally better results across all the benchmarks.

one. We can observe that employing a domain-specific prefix for zero-shot and a domain-agnostic prefix for retrieval leads to generally better results across all the benchmarks.

6 Conclusion

We presented CORE, a training-free retrieval-based zero-shot solution for low-resource image classification. Through the retrieval of semantically relevant textual information for both image representations and class prototypes, CORE is able to enhance the richness of feature representations and achieves state-of-the-art performance. Remarkably, it does so without the need for additional training or labeled data, outperforming existing training-based methods under extremely low-resource conditions. Moreover, we established a comprehensive benchmark using representative datasets and baseline models, providing a robust testing ground for the low-resource image classification task. The results demonstrate the efficacy and generalization capability of CORE across various low-resource domains, representing a significant step forward in low-resource image classification.

7 Limitations

While we prove text retrieval to be a strong training-free solution on the proposed benchmark, the performance is still limited by the representation of a domain in the external database. An example of this low coverage is provided by (Liu et al., 2023b), where Patch-Camelyon, a medical dataset, has a limited presence even in LAION 400M. We face the same problem when trying to apply CORE to Parasitic Egg Classification (Anantrasirichai et al., 2022), where retrieved captions from CC12M only contain the most common egg parasite name, and where supervised fine-tuning becomes the strongest way to adapt VLMs for the setting.

8 Ethical considerations

We employ large-scale web-crawled data to enrich our representations, and this type of data is by definition mostly uncurated, and it may reflect bias from the real world. This data might contain harmful or Not Safe For Work (NSFW) content, as demonstrated in previous studies (Thiel, 2023).

The scientific community has acted to mitigate these risks, *e.g.* employing NSFW filters before releasing the image-text pairs (Byeon et al., 2022). Nevertheless, we do not employ the image content of these datasets, and we use the textual part only at the semantic level to enhance the image classification performance. Therefore we never expose users to harmful or undesired content.

Acknowledgments

We thank CINECA and the ISCRA initiative for the availability of high-performance computing resources. This work was supported by the EU Horizon ELIAS (No. 101120237) and AI4TRUST (No.101070190) projects, and the FAIR - Future AI Research (PE00000013), funded by NextGeneration EU, and the PRIN LEGO-AI (Prot. 2020TA3K9N) project. This work was carried out in the Vision and Learning joint laboratory of FBK and UNITN.

References

- Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 1, Long Papers*.
- Nantheera Anantrasirichai, Thanarat H Chalidabhongse, Duangdao Palasuwan, Korranat Naruenatthanaset, Thananop Kobchaisawat, Nuntiporn Nunthanasup, Kanyarat Boonpeng, Xudong Ma, and Alin Achim. 2022. Icip 2022 challenge on parasitic egg detection and classification in microscopic images: Dataset, methods and results. In *IEEE International Conference on Image Processing (ICIP)*.
- Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhersch, Matthias Reso,

- Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*.
- Minwoo Byeon, Beomhee Park, Haechon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022. Adapterformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Sanjoy Chowdhury, Sayan Nag, and Dinesh Manocha. 2023. APoLLo: Unified adapter and prompt learning for vision language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. 2023. Vocabulary-free image classification. In *Advances in neural information processing systems (NeurIPS)*.
- Victor G Turrissi da Costa, Nicola Dall'Asen, Yiming Wang, Nicu Sebe, and Elisa Ricci. 2023. Diversified in-domain synthesis with efficient fine-tuning for few-shot classification. *arXiv preprint arXiv:2312.03046*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.
- Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Manan Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning (ICML)*.
- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning (ICML)*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning (ICML)*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

- Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. 2023b. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. [Sfr-embedding-mistral:enhance text retrieval with transfer learning](#). Salesforce AI Research Blog.
- Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Null-text inversion for editing real images using guided diffusion models. 2023 ieee. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th annual meeting of the association for computational linguistics (ACL): Volume 1, Long Papers*.
- Shubham Parashar, Zhiqiu Lin, Yanan Li, and Shu Kong. 2023. Prompting scientific names for zero-shot species recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Guilherme Penedo, Hyněk Kydlíček, Leandro von Werra, and Thomas Wolf. 2024. [Fineweb](#).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*.
- Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. 2023. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- David Thiel. 2023. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical Report. Stanford University, Palo Alto, CA. <https://purl.stanford...>
- Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. 2024. Effective data augmentation with diffusion models. In *International Conference on Learning Representations (ICLR)*.
- Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*.
- Vishaal Udandarao, Ameeya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip HS Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2024. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125*.
- Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Yunhua Zhang, Hazel Doughty, and Cees GM Snoek. 2024. Low-resource vision challenges for foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*.

A Baselines implementation details

For VLMs (*i.e.* CLIP, ImageBind, and SigLIP) we report the zero-shot results with zero-shot weights built starting from the prompt “{prefix} [CLS]”. As {prefix} we test both domain-specific (*e.g.* “a circuit diagram of a” for Circuits dataset) and domain-agnostic sentences (“a photo of a” as standard practice in CLIP), and we report the best result for each model. For iNaturalist we follow the insights of (Parashar et al., 2023) and try using common names of animals and plants instead of their scientific names. We find that merging the zero-shot weights of common and scientific names improves the overall accuracy, therefore we report these results for this dataset.

For LMM (*i.e.* BLIP2 and LLaVA) we feed the query image with the textual prompt “Question: what’s the name of the object in the image out of [class names]? Answer with the name only. Answer: ”, then parse the answer and match the name with the dataset classes.

For ImageBind🔥 and SigLIP🔥 we follow (Zhang et al., 2024) recipe for fine-tuning, and generate synthetic images using their pipeline. The generation involves using a Stable Diffusion model (Rombach et al., 2021) on top of noised images, and feeding the model with a domain-specific prompt, *e.g.* “a circuit diagram of a [CLS].” to re-generate the missing part. The amount of noise added from the diffusion schedule depends on the role of the synthetic image, *i.e.* it is set to 30% of the schedule for samples used as positive in the contrastive loss, and 60% for the negative samples. We maintain the original training hyperparameters (including batch size and learning rate), and we train an AdaptFormer (Chen et al., 2022) module with rank 2 and a linear classification head.

B On synthetic data augmentation

We extend the discussion from Sec. 1 on using synthetic data augmentation to address low-data scenarios. We argue that while synthetic data generation can effectively enhance recognition performance, it introduces training images that are either overly similar to the original samples or incorrect, violating domain rules, one example being the case of (Zhang et al., 2024). We showcase examples of both “positive” and “negative” types of synthetic augmentation in Fig. 7. We can see that “positive” samples do not differ significantly from the original sample and do not bring meaningful variation to the

original sample. On the other hand, the “negative” samples differ to the point of changing semantics (first row) or breaking the reference domain rules (third row).

C On having an image-to-image retrieval

Initially, we also implemented an image-to-image retrieval system leveraging DINOv2 embeddings. We hypothesized that enhancing the original image embedding with features from retrieved images would improve its capability to retrieve relevant textual information. However, this approach did not yield the desired results. Although the retrieved images were from the correct domain, they often lacked the appropriate semantic class. This mismatch caused the embeddings to deviate significantly from the target class, resulting in a text-retrieval performance decline. Quantitative results supporting this observation are presented in Tab. 8.

D Interesting failure cases

We extend the discussion started in Section 7 regarding the failure cases of retrieval-based solutions *w.r.t.* to training-based ones. We include the quantitative results on the Parasitic Egg Recognition challenge (Anantrasirichai et al., 2022) in Tab. 9. We can see that in this case, the training-based solutions outperform the training-free solutions by a large margin. We deem this for two reasons: i) the classes differ visually, making a training approach powerful in telling the different classes, and ii) this type of data is under-represented in large-scale datasets (Liu et al., 2023b), making the retrieval less effective which, nevertheless, outperforms the other training-free solutions.

E Additional analyses

Dependency on sample quantity. The experiments in Tab. 1 are conducted with ~ 5 annotated samples per class as in the Circuits dataset of (Zhang et al., 2024), where only 154 images are available for training on 32 classes. We complement this analysis by showing in Tab. 6 the effect of having access to more training data for training-based solutions on HAM10000.

We can see that the model performance scales with the amount of annotated samples per class. SigLIP trained with 15 samples achieves better results than CORE. The low amount of training and




Number of shots per class	1		5		10		15		20	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
ImageBind  (Zhang et al., 2024)	10.85	68.70	54.60	96.56	55.59	96.10	46.72	97.88	57.84	98.81
SigLIP  (Zhai et al., 2023)	25.74	90.73	54.60	95.90	46.72	95.90	62.21	98.15	61.42	98.21

Table 6: Accuracy of training-based  solutions with increasing number of training shots per class on HAM10000.

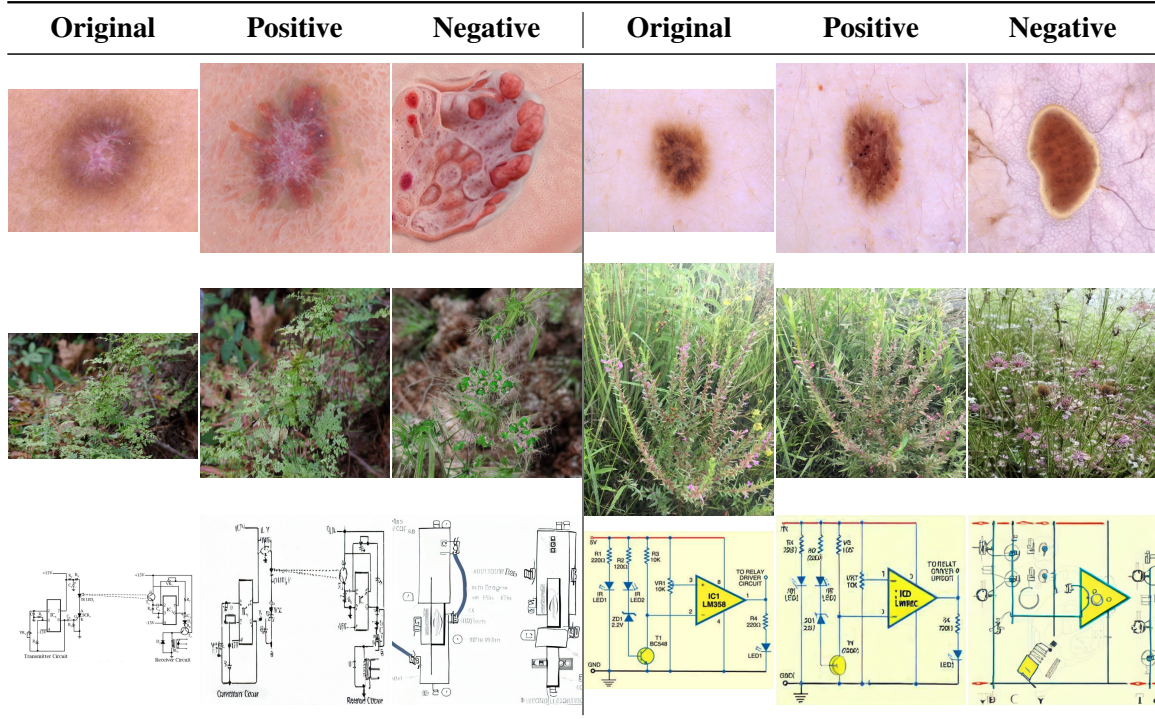


Table 7: Synthetic image from the baseline of (Zhang et al., 2024). We show original samples, the “positive” augmentation and the “negative” augmentation.

Method	Circuits	iNaturalist	HAM10000
DINOv2 img2img	23.28	7.78	59.96
CoRE (Ours — CC12M)	42.94	21.40	61.54

Table 8: Top-1 accuracy on the proposed benchmark using DINOv2 features to retrieve relevant images compared to our CoRECC12M.

validation data, on the other hand, makes the selection of the best-trained model noisy, therefore the improvement with more samples is not straightforward to assess.

F Accuracy@1 vs Accuracy@5 tradeoff

In our investigation, we observed that modifying the hyperparameters α , β , and τ resulted in a trade-off between Accuracy@1 and Accuracy@5. We prioritized Accuracy@1, which is widely regarded as the primary classification metric. This sensitivity was particularly pronounced in the HAM10000 dataset, where we identified several data points that constitute the Pareto frontier when plotting Accu-



Method	Acc@1
ImageBind  (Zhang et al., 2024)	33.27
SigLIP  (Zhai et al., 2023)	17.59
ImageBind (Girdhar et al., 2023)	9.18
SigLIP (Zhai et al., 2023)	12.72
CoRE (Ours — CC12M)	15.14

Table 9: Top-1 accuracy on the Parasitic Egg Classification.

racy@1 against Accuracy@5. Some notable examples from this frontier are presented in Table 10.

G Retrieval Efficiency

Our method CoRE leverages an efficient indexing and retrieval mechanism, implemented with FAISS (Douze et al., 2024), achieving 57.54 ms per image-text retrieval on a CPU. As text-to-text retrieval is performed offline, we do not include it in the runtime analysis. As shown in Table 11,

Acc@1	Acc@5
27.73	96.10
39.58	95.96
54.80	95.63
57.97	95.50
61.95	95.30
62.21	94.51

Table 10: Acc@1 vs Acc@5 tradeoff on HAM10000.

per image inference at runtime, CoRE requires on average 110 ms (the sum of SigLIP inference time and retrieval time), while the competitor with the best performance (SigLIP@384px) requires 53 ms. CoRE requires 110 ms, which is about one-third of other VLMs (e.g. CLIP ViT-L@336px, BLIP2-EVA) and is of one magnitude less than LLaVA 1.6 (34B).

In terms of memory/storage, in addition to the models, CoRE requires the storage of both the metadata (i.e., the filenames and original captions), and the indexes (i.e., the embeddings). For CC12M, the metadata is of 2.7 GB and the index is of 3.4 GB by SigLIP (3.1 GB by SFR-Embedding-Mistral). For our COYO-700M subset, the metadata is of 6.0 GB, and the indices are of 22 GB by SigLIP (23 GB by SFR-Embedding-Mistral). In comparison, other competitors do not require the storage of a database, but the models themselves could be storage-demanding, e.g. LLaVA 1.6 (34B) occupies about 65GB of storage, while CoRE (COYO-700M) in total occupies as little as 9.4GB, as the LLM is optional as shown in the ablation “On the use of LLM for embedding”.

H Examples of retrieved captions

We provide a qualitative showcase of retrieved captions by our CoRE for both text-to-text and image-to-text in Tab. 12 and Tab. 13. We use the 10% subset of COYO-700M as the database, and we also show some failure cases (e.g. sixth row of Tab. 13).

I Computational requirements

Synthetic data generation for trained baselines, as the amount of training data is low, has an upper bound of 12 GPU/hours on an NVIDIA A100 for iNaturalist. The subsequent model training has an

upper bound of 4 GPU/hours on an NVIDIA A100 for SigLIP.

For our CoRE, the most time-consuming task is represented by the external database embedding. The upper bound is 16 GPU/days for the COYO-700M subset. CC12M has been embedded in 3 GPU/days. Retrieval can then be performed without access to any GPU and on any consumer hardware in ~ 58 ms.

J Licenses

Most of the datasets used (Circuits (Zhang et al., 2024), iNaturalist (Van Horn et al., 2018), HAM10000 (Tschandl et al., 2018), and Parasitic Egg Classification (Anantrasirichai et al., 2022)) are released under Creative Commons Attribution Non-Commercial 4.0. COYO-700M (Byeon et al., 2022) is released under Creative Commons Attribution 4.0, while CC12M (Changpinoy et al., 2021) is released “as is”.

LLaVA (Liu et al., 2023a) and SigLIP (Zhai et al., 2023) are released under Apache 2.0. ImageBind (Girdhar et al., 2023) is released under Creative Commons Attribution Non-Commercial ShareAlike 4.0. CLIP (Radford et al., 2021) is released under MIT. BLIP (Li et al., 2023) is released under BSD-3-Clause. SFR-Embedding-Mistral (Meng et al., 2024) is released under Creative Commons Attribution Non-Commercial 4.0.

PyTorch (Ansel et al., 2024) is released “as is”. Faiss (Douze et al., 2024) is released under MIT.

Method	Inference (ms)	Params	Storage (GB)
CLIP ViT-L	55	123M f_{VLM}^T + 303M f_{VLM}^I	1.6
CLIP ViT-L@336px	368	123M f_{VLM}^T + 304M f_{VLM}^I	1.6
BLIP2-EVA	397	2.9B f_{LLM} + 986M f_{VLM}^I	2.3
LLaVA 1.6 (34B)	4970	34B f_{LLM} + 303M f_{VLM}^I	65
ImageBind	52	302M f_{VLM}^T + 632M f_{VLM}^I	3.7
SigLIP@384px	53	449M f_{VLM}^T + 428M f_{VLM}^I	3.3
CoRE	110	449M f_{VLM}^T + 428M f_{VLM}^I + 7B f_{LLM}^*	17.3 + 6.1 (CC12M)

Table 11: Retrieval efficiency of our CoRE and all the baselines in terms of inference time, parameter count, and storage requirements.

Input	Retrieved captions		
A circuit diagram of an amplifier.	Electrical Diagram Amplifier	Audio amplifier circuit diagram with layout	Circuit diagram for transistor as audio amplifier
A circuit diagram of a LED.	Technical Drawing of an LED	The symbol for a light emitting diode.	One LED with leads all bent out
A circuit diagram of an audio mixer.	Photo of an audio mixer board	8-Channel Audio Mixer picture 1	Mixer with volume faders and pan knobs
A skin lesion of Bowen’s disease.	Pre-Cancerous Actinic Keratosis	A biopsy specimen showing hyperkeratosis, papillom	Looking for premalignant skin cancer
A skin lesion of melanoma.	Melanoma of the Skin, Cut-section	This picture shows a melanoma lesion with varying colors.	A mole that turned out to be melanoma skin cancer
A skin lesion of melanocytic nevi.	This picture shows a melanoma lesion with varying colors.	Recurrence of melanocytic naevus	Graphic of a melanoma
A photo of a Lygodium japonicum.	A picture of a Japanese holly fern	japanese painted fern has silver metallic fronds	Pyrrosia Ferns on moss covered trunk
A photo of a Salvinia minima.	12 Water Spangles(Salvinia Minima), Live Aquarium/Aquatic	Salvinia Minima (Water Spangles) floating aquarium plant	water drop salvinia sp trichomes stock image
A photo of a Azolla filiculoides.	The Azolla fern has leaves floating on the water surface	Image of Azolla filiculoides(). Click to enlarge parts of image.	Picture of Fern

Table 12: Examples of retrieved captions in text-to-text using the COYO-700M subset.

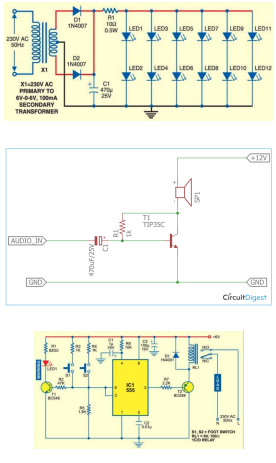
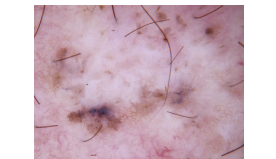
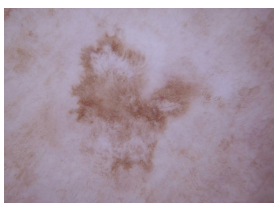


Input	Retrieved captions		
	<p>Hobby Electronics Circuits: AC Powered 220V Led Circuit</p>	<p>Transformerless Led Lighting Led Lamp Circuit Electronics</p>	<p>Astonishing Christmas Lights</p>
	<p>PIEZO SOUNDER WITH BUILT-IN CIRCUIT</p>	<p>How To Build A Speaker Circuit With Adjustable Volume</p>	<p>What all do I need for a simple speaker circuit?</p>
	<p>IC 555 dry run protection</p>	<p>Wireless Remote Control Switch</p>	<p>230V AC Mains Over Voltage Protection Circuit</p>
	<p>Skin coloured papules centred around hair follicles.</p>	<p>Dermoscopic image of a porokeratosis of Mibelli lesion</p>	<p>Pigmented basal cell carcinoma dermoscopy</p>
	<p>Dermoscopy. Brown and blue-grey dots/clods.</p>	<p>Dermoscopic image of a porokeratosis of Mibelli lesion</p>	<p>Dermoscopy. Chaos and clues</p>
	<p>brown blotches are formed</p>	<p>Oil Red Staining.</p>	<p>pigment used as normal pigment pattern</p>
	<p>Notogrammitis billardierei (Finger fern) at Wingecarribee</p>	<p>Asplenium polyodon (East Maui) This image is licensed</p>	<p>Not sure what this fern is I thought maybe Buckler fern. Any ideas?</p>
	<p>A seed fern frond is prepared for analysis.</p>	<p>A tiny plant on a tree fern's trunk</p>	<p>Asplenium polyodon (West Maui)</p>
	<p>Astrolepis cochisensis Cochise Scaly Cloakfern</p>	<p>Notogrammitis billardierei (Finger fern)</p>	<p>Ferns emerging from charred earth</p>

Table 13: Examples of retrieved captions in image-to-text using the COYO-700M subset.