Unlocking Few-Shot Capabilities in LVLMs via Prompt Conditioning and Head Selection

Adhemar de Senneville[1], Xavier Bou[2], Jérémy Anger[1], Rafael Grompone[1], and Gabriele Facciolo[1,3]

[1] Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, France
[2] École Polytechnique, AMIAD, France
[3] Institut Universitaire de France
adhemar.de_senneville@ens-paris-saclay.fr

Abstract. Current Large Vision Language Models (LVLMs) excel at many zero-shot tasks like image captioning, visual question answering and OCR. However, these same models suffer from poor performance at image classification tasks, underperforming against CLIP-based methods. Notably, this gap is surprising because many LVLMs use CLIP-pretrained vision encoders. Yet LVLMs are not inherently limited by CLIP's architecture with independent vision and text encoders. In CLIP, this separation biases classification toward class-name matching rather than joint visual–text reasoning. In this paper we show that, despite their poor raw performance, LVLMs can improve visual feature class separability at inference using prompt conditioning, and LVLMs' internal representations, especially attention heads, can outperform the model itself at zero-shot and few-shot classification. We introduce Head Ensemble Classifiers (HEC) to bridge the performance gap between CLIP-based and LVLM-based classification methods. Inspired by Gaussian Discriminant Analysis, HEC ranks the most discriminative vision and text heads and combines them into a training-free classifier. We show that HEC achieves state-of-the-art performance in few-shot and zero-shot classification across 12 datasets. Code: `github.com/AdhemarDeSenneville/HEC`

Keywords: Zero shot classification · Few shot classification · Large Vision-Language Model · CLIP

1 Introduction

Recent LVLMs show remarkable progress in a wide range of computer vision tasks, including image captioning [10, 36], text transcription [40], Visual Question Answering (VQA) [43] and grounding [52, 64]. In addition, they offer the versatility to address all of these problems with a single pre-trained model and no additional fine-tuning. However, LVLMs still lag behind the state of the art in few-shot image classification [42, 68], particularly compared to CLIP-based models [4, 39, 54, 61]. This is surprising as many LVLMs inherit from a CLIP [54] vision encoder, yet score below CLIP at zero-shot [68].

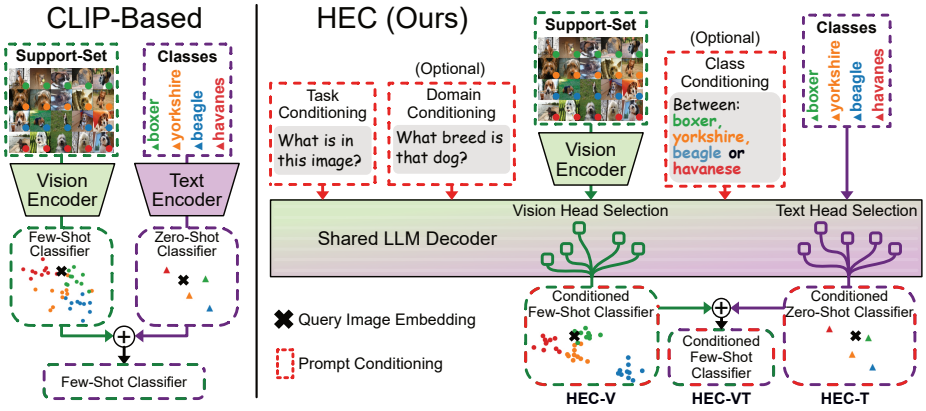


Fig. 1: CLIP-based vs HEC (Ours). CLIP-based methods encode class names and support set images independently to construct a zero-shot and a few-shot classifier respectively. Our method keeps the same two-classifier structure. However, both distributions go through a shared LLM decoder which can be conditioned by a text prompt to include guidance on the domain or the classes of the support set. The few-shot classifier (HEC-V) builds on the distribution of a sparse set of heads from the LLM decoder. This subset, which we refer to as *vision-heads*, is selected using Gaussian Discriminant Analysis [6] (Fig. 2). The zero-shot classifier (HEC-T) builds on the distribution of another sparse set of heads, which we refer to as *text-heads*. Similarly to CLIP-based methods, the two classifiers can be combined in a single one (HEC-VT) by adding their output probabilities.

Nonetheless, despite exhibiting strong performances, CLIP-based methods still have some limitations. First, CLIP mainly matches images to class names or descriptions, making it weaker when depending on domain text contextualization [8, 17, 53]. Secondly, CLIP vision and text encoders are independent, so image–text interaction is limited at decision time, as illustrated in Fig. 1 (left). In contrast, LVLMs use a vision encoder to convert an image into a sequence of vision tokens, or use a tokenizer to convert a text prompt into a sequence of text tokens. Then, these vision and text tokens are jointly processed by a shared LLM transformer decoder allowing image–text interactions during inference.

There is a mismatch between the rich internal representations that LVLMs should inherit from their CLIP encoder and their weak final output predictions [68]. Consequently, previous works either focus on using LVLMs to generate captions to improve CLIP performance [37, 45], or extensive finetuning on downstream classification tasks [23, 50]. This mismatch led us to investigate whether the LVLMs' internal representations can be leveraged for few-shot classification.

Inspired by Gaussian Discriminant Analysis (GDA) [6, 61] and recent works on training-free LVLM adaptation [29, 47], we propose two simple yet effective mechanisms to condition and extract LVLMs multimodal representations. First, we use text prompts to condition LVLMs feature distribution during inference. The prompt, concatenated with vision tokens, specifies contextual information

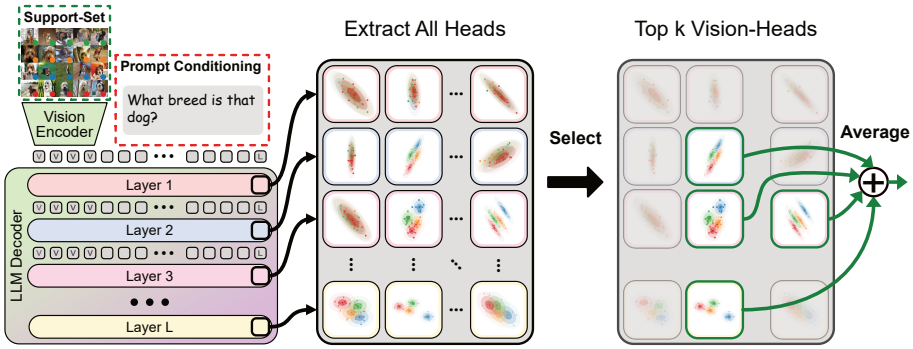


Fig. 2: Overview of HEC-V. Given a prompt, we first encode all the images from the support set with the LVLM. We then extract the distribution of attention vectors (3) for the last token across all heads in every layer . Then, based on a Gaussian Discriminant Analysis [6], we rank each head based on its class separability. Lastly, given a query image, we ensemble the predictions of the top k heads for that task by averaging their class probabilities.

of the few-shot task, as shown in Fig. 1. Then, we extract features by selecting a sparse set of attention heads that maximize few-shot and zero-shot classification performance. More specifically we select the best attention heads at few-shot (denoted *vision-heads*) using GDA, selecting heads that maximize performance given a set of labeled images (support set). Similarly, we introduce a mechanism to select the best attention heads at zero-shot (denoted *text-heads*).

Building on these two mechanisms, we introduce three distinct Head Ensemble Classifiers (HEC), see Fig. 1. When class names are unknown, we show that combining vision-heads produces a few-shot classifier (HEC-V) that improves its performance by conditioning the classification with a domain-specific prompt (e.g., `What breed is that dog?`). This type of prompt guidance domain adaptation is not possible with vision models and CLIP-based models. We also show that combining text-heads into a classifier (HEC-T) improves for free the zero-shot accuracy of a given LVLM. Lastly, we bridge the performance gap between training-free CLIP-based methods and LVLM-based methods by introducing a classifier that combines text-heads and vision-heads (HEC-VT).

In summary, our contributions are:

- We experimentally show that adding prompt conditioning to LVLMs improves the class separability of its internal distributions, especially within a sparse set of heads.
- We introduce a method to select and combine those top heads at test-time.
- We report state-of-the-art performance at few-shot and zero-shot classification, bridging the gap between CLIP-based and LVLM-based methods without the need of additional fine-tuning.

2 Previous Work

Training-Free CLIP Most research effort around multimodal few-shot image classification has been focused around CLIP-based architectures. CLIP in its original formulation can be applied easily to zero-shot image classification [54] using the known class names. Prior work shows that CLIP's zero-shot classification can be improved with parameter-free attention maps [21], similarly we propose a training-free adaptation of LVLMs for zero-shot classification. On the other hand, when class names are unknown but a small labeled support set is available, the training-free baselines include nearest-centroid classifier [13,57] and linear probing with closed-form solutions [3, 5, 61]. When both class names and the support set are available, a first line of work simply adds logits from a zero-shot and a few-shot classifier together [61, 67]. Instead, some improve upon this by combining text and visual frozen features to build a single classifier [4, 39, 69].

LVLMs in Few-Shot Learning A first application of LVLMs to few-shot learning was to use them to generate descriptions to guide a CLIP-based classifier [37, 44, 45], rather than using the LVLM as the classifier. To mitigate low classification performance, a line of work directly fine-tunes the model on fine-grained classification tasks. Finedefics [23] is trained using attribute descriptions and CLIP-like contrastive losses. In [41], meta-training improves in-context learning performance. Recent work explicitly fine-tunes LVLMs as CLIP encoders [50], training them with contrastive objectives aligning image and text embeddings [31, 65]. Similarly to us, VLM2Vec [31] uses instruction conditioning (e.g. `Instruction: Represent the given image and the related question`), which we refer to as Task conditioning. However, in our work, performance gains arise from adding domain and class conditioning. Our method can be applied to any LVLM, making it complementary to fine-tuning approaches.

Training-Free LVLM A straightforward training-free method to improve performance is in-context learning, where few example images per class are added to the prompt [2]. However, the performance decreases rapidly with the number of shots and classes present in the prompt [11, 30, 55]. Some lines of work mitigate this by introducing task vectors that compress many in-context tasks in a single prompt [25, 26]. MTV [29] improved on that furthermore by selecting task vectors inside a sparse set of attention heads. Recently SAVs [47] directly selects top heads from the support set using nearest centroid classifier without having to compute task vectors. Our work builds on SAVs by improving the head selection and ensemble mechanisms.

3 Preliminaries

In this section, we first formalize the few-shot and zero-shot image classification setup. We then conduct a preliminary investigation on where LVLMs build their representation during an image classification task. Particularly we observe that (1) prompt conditioning allows the last token of the LLM decoder to build multimodal representations that outperform the vision backbone. (2) A sparse set of

attention heads contain representations that outperform the prediction based on the last token.

3.1 Problem Formulation

We consider the N-way few-shot image classification over classes $\mathcal{C} = \{1, \dots, N\}$. The support set is composed of image–label pairs $\{(x_i, y_i)\}_{i=1}^{NK}$, with only a small number of K samples per class (K-shots). A class text t_c is associated to each label. The query set is the set of unlabeled images we aim to classify.

In the training-free paradigm, we use a frozen foundation model to represent both images and class texts in a shared embedding space [61, 67, 69]. We denote $z_i^{(\mathrm{v})}$ the embedding of a support image x_i, and $z_c^{(\mathrm{t})}$ the embedding of the class text t_c. In the **text-zero-shot** setting, the class logits are given by the dot product $z_q^{(\mathrm{v})\top} z_c^{(\mathrm{t})}$ where $z_q^{(\mathrm{v})}$ is the query image embedding. In the **vision-few-shot** setting, we define a classifier from the support set embeddings $\{z_i^{(\mathrm{v})}, y_i\}_{i=1}^{NK}$, to predict the class of the query embedding $z_q^{(\mathrm{v})}$. In the **vision-text-few-shot** setting, we are given $\{z_c^{(\mathrm{t})}\}_{c\in\mathcal{C}}$ and $\{z_i^{(\mathrm{v})}, y_i\}_{i=1}^{NK}$. A simple option to address this problem is to add the logits of the text-zero-shot and a vision-few-shot classifiers [61, 67]. However, some methods introduced vision–text coupling by treating the text-zero-shot classifier as a fixed prior and adjusting its logits using the support set [4, 39, 69].

CLIP encoder: Extracting text and image CLIP embeddings is straightforward. It uses two separate encoders, the vision encoder $z_i^{(\mathrm{v})} = f^v(x_i)[\mathrm{CLS}]$ and the text encoder $z_c^{(\mathrm{t})} = f^t(\texttt{"a photo of a \{}t_c\texttt{\}"})[\mathrm{CLS}]$, to map images and class prompts into a shared embedding space. Here, [CLS] denotes the extraction of the class token, meaning that the embedding is a single token.

LVLM as a CLIP encoder: Unlike CLIP-based methods, encoding text and images using LVLMs is not straightforward. Inspired by [31, 50, 65], we propose a framework for using LVLMs as CLIP encoders. Let f^v be the vision encoder that maps an image x_i to a sequence of vision tokens, and let $\mathrm{LLM}(\cdot)$ be the LLM decoder that takes that sequence of vision tokens as well as text tokens as input. For instance, Qwen2-VL has a ViT visual encoder with 32 layers and an LLM decoder with 28 layers. We extract the output embedding from the summary token, i.e. $\mathrm{ST}(\cdot)$, which returns the last token of the last layer embedding of the LLM decoder [50]. Unlike CLIP, LVLM embeddings can be prompt-conditioned as the LLM decoder jointly processes vision tokens with text tokens from the prompt π. The LLM decoder input is the concatenation $[f^v(x_i); \pi]$ (see Fig. 2), and the embedding of an image is obtained as

$$z_i^{(\mathrm{v})} = \mathrm{ST}(\mathrm{LLM}([f^v(x_i); \pi])). \tag{1}$$

To obtain the class embedding, we replace image tokens by a textual class description:

$$z_c^{(\mathrm{t})} = \mathrm{ST}(\mathrm{LLM}([\texttt{"You are given an image of a \{}t_c\texttt{\}."}; \pi])). \tag{2}$$

The prompt π can incorporate incrementally different levels of guidance depending on available information. Thus, we propose three levels of conditioning:

- **Task Conditioning**: A task-specific prompt (e.g., `What is the object in the image?`) pushes the summary token toward discriminative representation. As shown in [50], adding constraints in the prompt such as `Answer in one word` can improve performance by encouraging the model to compress information in the next-token representation. It is important to note that, at this stage, by switching the prompt we can solve other classification tasks such as VQA, image–text pair classification or image retrieval (see supplementary **??**).
- **Domain Conditioning**: Similarly to the case of task guidance, in fine-grained settings, rephrasing the prompt to be domain-specific (e.g., `What breed is that dog?`) should additionally push the summary token toward domain-discriminative representation.
- **Class Conditioning**: Moreover, if the candidate classes are known, appending to the domain prompt the class list (e.g., `Between: boxer, yorkshire, beagle or havanese.`) should additionally push the summary token toward class-discriminative representation.

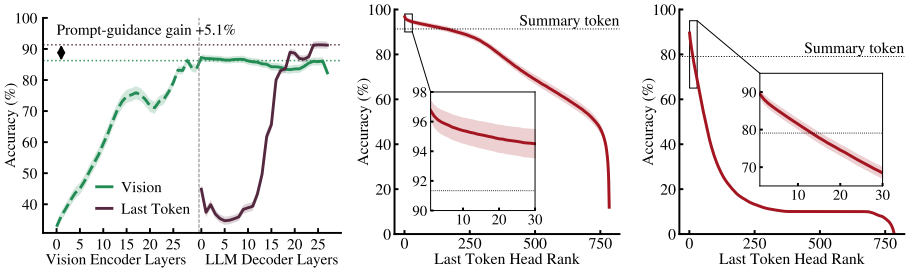
Head Extraction In the following we formalize how features from a given head are extracted. We index attention heads by $m \in \{1, \dots, M\}$, where the total number of heads is $M = L \cdot H$ with L the number of layers, and H the number of heads per layer. For each head m, we compute the corresponding head embedding as

$$\mathbf{h}_m = \text{softmax}\left(\frac{\mathbf{q}_m \mathbf{K}_m^\top}{\sqrt{D}}\right)\mathbf{V}_m, \tag{3}$$

where $\mathbf{q}_m$ is the query vector of only the last token in the input sequence, $\mathbf{K}_m$ and $\mathbf{V}_m$ are the key and value matrices of the current head at the current layer, and D is the head dimension. For the rest of the method, we L2-normalize each $\mathbf{h}_m$ making dot products equivalent to cosine similarities. Following [47] we denote $\mathbf{h}_m$ as an *attention vector* for head m. We denote $\mathbf{h}_{i,m}^{(\text{v})}$ the attention vector when encoding the image x_i and $\mathbf{h}_{c,m}^{(\text{t})}$ when encoding the text class t_c.

3.2 What happens inside LVLMs

To study the impact of prompt conditioning we conducted a series of experiments by randomly selecting 1000 10-way 4-shot tasks across 10 datasets. We processed the 40 images of the support set through Qwen2-VL using a prompt with Class conditioning. We then measure few-shot and zero-shot accuracy at different locations in the model. Given either a token-embedding support set $\{z_i, y_i\}_{i=1}^{NK}$ extracted from an intermediate LLM layer, or an attention vector support set $\{\mathbf{h}_{i,m}^{(\text{v})}, y_i\}_{i=1}^{NK}$, we fit a ridge linear classifier [6] to measure the few-shot accuracy of each distribution on the query set [1]. Given class attention vectors $\{\mathbf{h}_{c,m}^{(\text{t})}\}_{c\in\mathcal{C}}$, we measure the head zero-shot accuracy on the query set using as class logits



(a) Evolution of embedding few-shot accuracy across vision encoder layers and LLM decoder layers

(b) Few-shot accuracy of all last token heads, ranked from best to worst

(c) Zero-shot accuracy of all last token heads, ranked from best to worst

Fig. 3: Experiments to identify where the best classification representation lies in LVLMs. At different locations of Qwen2-VL, we compute linear probing accuracy averaged over a thousand 10-way 4-shot tasks across 10 datasets using Class conditioning. (a) Although vision tokens yield strong accuracy early on, inherited from the CLIP vision transformer, the last token builds better representations by integrating multimodal features from vision and text prompt tokens. (b) For each few-shot setup, a small number of top heads called *vision-heads* yield better performance than the summary token. (c) Similarly, for each zero-shot setup, a small number of top heads called *text-heads* yield better performance than the summary token.

$\mathbf{h}_{q,m}^{(\mathrm{v})\top}\mathbf{h}_{c,m}^{(\mathrm{t})}$. Results are shown in Fig. 3. We refer to the supplementary material **??** for additional details on these experiments.

Figure 3a shows the accuracy of the averaged vision tokens across the vision encoder and the LLM decoder. As expected, as images are processed by the vision encoder, representations shift from low-level to more discriminative high-level features, resulting in improved accuracy. Once visual tokens are processed by the LLM decoder, their classification accuracy stalls. Meanwhile, the last token representations refine from layer to layer by attending to the vision tokens and the prompt conditioning tokens, improving its accuracy. As a result, the summary token yields higher accuracy than vision tokens, indicating that LLM joint decoding of vision and prompt tokens steers representations during inference toward being more class-discriminative. **This demonstrates the LVLMs ability to refine visual features at inference using prompt conditioning.**

In Fig. 3b, we select the 784 heads of the 28 layers of the last token. For each task, we rank them from best to worst accuracy and show their average few-shot accuracy. We see that about 25% of heads have better classification power than the LLM output, and that, more specifically, a handful of heads yields an accuracy gain of more than 10% compared to the summary token. **This shows that for a given task, a sparse set of heads yields better performance at few-shot than the summary token itself.**

Similarly in Fig. 3c we rank each head by its zero-shot classification accuracy and find that about 10 heads yield better zero-shot performance than the summary

Prompt: `What type of bird is this?` Prompt: `What type of plane is this?`

Fig. 4: Top head attention map. We concatenate bird [59] and aircraft [46] datasets images horizontally in one support set. We then select the top vision-head for bird classification using the prompt `What type of bird is this?` and do the same for plane using the prompt `What type of plane is this?`. The attention map of the bird (left) and plane (right) top vision-head is overlaid on top of the image.

token. **This shows that for a given task, a sparse set of heads yields better performance at zero-shot than the summary token itself.**

Figure 4 shows attention maps of the top vision-head for different Domain prompts on a fixed support set. It shows that by combining prompt conditioning and top vision-head selection, we indeed retrieve domain-specific features.

In conclusion:

1. At inference time, the summary token gains in class separability thanks to prompt conditioning. This arises from the LLM decoder multimodal ability to jointly process vision tokens with text prompt tokens.
2. For each classification task, a sparse set of last token heads yield better performance than the summary token at zero-shot and few-shot.

Therefore, identifying those top heads at test time has the potential to greatly increase few-shot and zero-shot performance of LVLMs.

4 Method

In this section, we propose a test-time ranking procedure to select the top vision and text heads. We then show how to combine predictions from those heads into a single classifier. An overview of the method is illustrated in Fig. 2.

4.1 Vision-Heads Ranking

Let us see how to rank the top vision-heads from a labeled support set. Given the attention vector support set $\{\mathbf{h}_{i,m}^{(\mathrm{v})}, y_i\}_{i=1}^{NK}$ in a head m, the classifier and its ranking score are both derived from a classical Gaussian Discriminant Analysis

(GDA) [6,61]. We assume a class-conditional generative model for the observed feature distribution. Specifically, each attention vector follows a Gaussian distribution, centered at the class mean $\boldsymbol{\mu}_{m,c} \in \mathbb{R}^D$, with a covariance matrix $\boldsymbol{\Sigma}_m \in \mathbb{R}^{D\times D}$ that is shared across classes i.e.

$$(\mathbf{H}_m^{(\mathrm{v})} \mid Y = c) \sim \mathcal{N}\big(\boldsymbol{\mu}_{m,c}, \boldsymbol{\Sigma}_m\big)\,. \tag{4}$$

A key reason this generative model performs well is that, in fine-grained few-shot classification, features often lie in a thin, anisotropic region of the embedding space. The shared covariance among all classes captures principal directions common across classes, which improves discrimination. In addition, with very few samples per class, class-specific covariance estimates are underconstrained, one shared covariance gives a more stable estimate from the full support set.

Given the observed features in a head, we estimate the mean $\hat{\boldsymbol{\mu}}_{m,c}$ and the precision matrix $\widehat{\boldsymbol{\Sigma}}_m^{-1}$ using the unbiased sample mean estimator and the empirical Bayes ridge-type estimator [33] respectively:

$$\hat{\boldsymbol{\mu}}_{m,c} = \frac{1}{K}\sum_{i:\, y_i=c} \mathbf{h}_{i,m}^{(\mathrm{v})}, \qquad \widehat{\boldsymbol{\Sigma}}_m^{-1} = D\Big((KN-1)\widehat{\boldsymbol{\Sigma}}_m + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_m)\mathbf{I}_D\Big)^{-1}, \tag{5}$$

where $\mathbf{I}_D$ is the identity matrix and $\widehat{\boldsymbol{\Sigma}}_m$ is the empirical covariance. The class logits $\ell_{i,m,c}$ of the model are the log-probabilities of observing the features and the class label

$$\begin{aligned} \ell_{i,m,c} &= \log p\Big(\mathbf{h}_{i,m}^{(\mathrm{v})}, y = c\Big) = \log p\Big(\mathbf{h}_{i,m}^{(\mathrm{v})} \mid y = c\Big) + \log p(y = c) \\ &= -\frac{1}{2}\big(\mathbf{h}_{i,m}^{(\mathrm{v})} - \hat{\boldsymbol{\mu}}_{m,c}\big)^\top \widehat{\boldsymbol{\Sigma}}_m^{-1}\big(\mathbf{h}_{i,m}^{(\mathrm{v})} - \hat{\boldsymbol{\mu}}_{m,c}\big) + C, \end{aligned} \tag{6}$$

where the constant C cancels out in the softmax. We compute class probabilities $p_{i,m,c}^{(\mathrm{v})}$ via a temperature-scaled softmax, and define the vision-head score as

$$s_m^{(\mathrm{v})} = \frac{1}{KN}\sum_{i=1}^{KN} p_{i,m,y_i}^{(\mathrm{v})}, \qquad \text{with} \qquad p_{i,m,c}^{(\mathrm{v})} = \frac{\exp\big(\ell_{i,m,c}/\tau\big)}{\sum_{j=1}^{N}\exp\big(\ell_{i,m,j}/\tau\big)}. \tag{7}$$

This score can be interpreted as a soft accuracy on the support set, since in the limit $\tau \to 0$, $s_m^{(\mathrm{v})}$ is the support set accuracy. In the few-shot setting, the Gaussian model often overfits the support set. Hence, the support set accuracy saturates at 100% for many heads, reducing the ability to distinguish top heads from the others. In such cases, using softmax probabilities prevents $s_m^{(\mathrm{v})}$ from saturating. This score is inspired by prior work on neural checkpoint ranking [38,62], which aims to rank model transferability instead. We define $\mathcal{H}^V$ as the set of top k vision-heads according to the ranking score.

4.2 Text-Heads Ranking

Given the class attention vectors for the m-th text-head $\{\mathbf{h}_{c,m}^{(\mathrm{t})}\}_{c\in\mathcal{C}}$, ranking the heads is straightforward as we simply evaluate the zero-shot soft accuracy of

each head on the support set. More specifically, for each support, we compute class logits as the dot product $\mathbf{h}_{i,m}^{(\mathrm{v})\top}\mathbf{h}_{c,m}^{(\mathrm{t})}$ and compute class probabilities by applying a softmax. The head score $s_m^{(\mathrm{t})}$ is the average probability assigned to the ground-truth label over the support set:

$$s_m^{(\mathrm{t})} = \frac{1}{KN}\sum_{i=1}^{KN} p_{i,m,y_i}^{(\mathrm{t})}, \qquad p_{i,m,c}^{(\mathrm{t})} = \frac{\exp\left(\mathbf{h}_{i,m}^{(\mathrm{v})\top}\mathbf{h}_{c,m}^{(\mathrm{t})}\right)}{\sum_{j=1}^{N}\exp\left(\mathbf{h}_{i,m}^{(\mathrm{v})\top}\mathbf{h}_{j,m}^{(\mathrm{t})}\right)}. \tag{8}$$

The set of top k text-heads according to the ranking score $s_m^{(\mathrm{t})}$ is $\mathcal{H}^T$. It is important to note that this ranking needs labels, which are not available in pure zero-shot scenarios. However, in the vision-text-few-shot setup, a labeled set is available, making our method capable of training-free zero-shot adaptation. Moreover, experiments show that text-heads are shared across tasks and domains, hence a fixed $\mathcal{H}^T$, determined once per model, transfers across tasks.

4.3 Head Ensemble Classifiers

Given $\mathcal{H}^V$ and $\mathcal{H}^T$, we introduce 3 classifiers. HEC-V averages class probabilities of vision-heads to produce a vision-few-shot classifier and HEC-T averages class probabilities of text-heads to produce a text-zero-shot classifier:

$$\bar{p}_{q,c}^{(\text{HEC-V})} = \frac{1}{|\mathcal{H}^V|}\sum_{m\in\mathcal{H}^V} p_{q,m,c}^{(\mathrm{v})}, \qquad \bar{p}_{q,c}^{(\text{HEC-T})} = \frac{1}{|\mathcal{H}^T|}\sum_{m\in\mathcal{H}^T} p_{q,m,c}^{(\mathrm{t})}. \tag{9}$$

Lastly, HEC-VT adds HEC-V and HEC-T class probabilities to produce a vision-text-few-shot classifier

$$\bar{p}_{q,c}^{(\text{HEC-VT})} = \frac{\alpha\,\bar{p}_{q,c}^{(\text{HEC-V})} + \bar{p}_{q,c}^{(\text{HEC-T})}}{\alpha+1}, \tag{10}$$

where α is a hyper-parameter. Despite its simplicity, we find that ensembling heads by averaging class probabilities yields strong performance. It is robust to poorly ranked heads without adding extra hyperparameters or computations.

5 Experiments

In this section we evaluate our methods on 12 datasets in three different setups. We first benchmark HEC-V on **vision-few-shot** with Domain conditioning, when class names are unknown. We then benchmark HEC-T on **text-zero-shot** with Class conditioning. Lastly, we benchmark HEC-VT on **vision-text-few-shot** using Domain conditioning. We also include additional experiments on prompt conditioning and head selection. Additional image-text benchmark [12, 20, 22, 28, 34, 70] results beyond image classification are reported in the supplementary **??**.

Dataset. Following previous works [4, 54, 61], we use 10 publicly available image classification datasets across different domains, covering a diverse range of visual recognition problems: EuroSAT (ESAT) [24], UCF101 (UCF) [58], DTD [15], Caltech101 (CAL) [19], SUN397 (SUN) [63], OxfordPets (PETS) [51], StanfordCars (CARS) [32], Flowers102(FLWR) [48], Food101(FOOD) [7], and FGVCAircraft (FGVC) [46]. The last 5 are fine-grained image classification benchmarks. We include in our experiments two additional fine-grained image classification datasets CUB-200 (BIRD) [59] and Traffic-Signs (SIGN) [27]. We treat ImageNet [16] as a general classification dataset, non domain-specific, and use it to select the top vision and text-heads shared across domains.

Protocol. We compare against state-of-the-art training-free CLIP-based baselines: closed-form linear probing [6] (Probing), CLIP [54], TipAdapter [67], GDA [61], and ProKeR [4]. Prior work reports results using the original OpenAI CLIP [54]. Our method, paired with recent LVLMs, significantly outperforms these baselines, in part because recent LVLMs inherit recent and stronger CLIP backbones than OpenAI CLIP. For a fair comparison, we try to disentangle the backbone performance from the contribution of our method. We therefore evaluate on two LVLMs where the pretrained CLIP backbone they inherit from is known. More specifically Qwen2-VL (7B) [60] and LLaVA-OV (7B) [35] use respectively DFN [18] and SigLIP [66] before instruction tuning. For LVLM-based training-free adaptation, we compare our work to state-of-the-art SAVs [47].

All CLIP-based methods are evaluated using the same prompts originally introduced by TipAdapter [67]. All LVLM-based methods are evaluated with the same prompts (see details in the supplementary **??**). All results are averaged over 5 random seeds. For every dataset, we select the optimal set of hyperparameters using the original hyperparameter sweep used by each method. For linear probing, we use a ridge classifier and sweep the regularization coefficient with values ranging from 0.001 to 10. To show the robustness of our method, we set $\tau = 10$ and set to 20 the number of top vision-heads and the number of top text-heads we select across all models and benchmarks. Only for HEC-VT, we sweep α from 0.1 to 10. All experiments are done on a single NVIDIA V100 GPU.

5.1 Vision-Few-Shot Classification

We benchmark our method in the vision-few-shot setting, where class names are unknown. We evaluate in N-way 4-shot with N equal to the total number of classes in each dataset. The results are reported in Tab. 1. For LVLM-based methods, we use Domain prompt conditioning (DC) e.g., `What breed is that dog?`, except for linear probing where we additionally test Task prompt conditioning (TC) i.e., `What object is in the image?`. This allows to extract more domain-specific features, improving class separability, which is impossible for vision models and CLIP models. Hence, we also evaluate several vision models and CLIP models using linear probing. We compare DFN [18], OpenAI CLIP [54], OpenCLIP [14] for CLIP models. We also compare against the DINO series of vision models DINOv1 [9], DINOv2 [49], and DINOv3 [56].

Table 1: Vision-Few-Shot classification accuracy (%) on 4-shot across 12 datasets without knowing class names. DC indicates Domain prompt conditioning and TC indicates Task prompt conditioning. HEC-V achieves state-of-the-art performance on average and across all datasets except FLWR, BIRD and ESAT. Underline denotes the best LVLM method. **Bold** denotes the best overall. Methods marked with † do not use hyperparameter tuning.

Model	Method	PETS	ESAT	UCF	SUN	CAL	DTD	AIR	FOOD	FLWR	CARS	BIRD	SIGN	AVG
DINOv1	Probing	81.9	**81.6**	71.8	50.8	86.9	49.6	25.5	37.9	88.1	28.3	54.2	46.3	58.6
DINOv2	Probing	78.5	73.6	67.6	63.4	87.0	51.6	29.9	43.0	97.0	35.9	65.5	35.6	60.7
DINOv3	Probing	88.0	77.2	82.8	72.7	95.4	63.6	56.9	74.3	**99.5**	79.4	**77.3**	53.1	76.7
OpenAI CLIP	Probing	72.9	73.6	81.5	73.2	90.2	55.8	28.2	73.4	89.5	56.9	53.6	56.5	67.1
OpenCLIP	Probing	73.0	75.6	79.5	72.1	90.4	60.3	29.6	62.9	89.0	75.0	51.2	62.7	68.4
DFN	Probing	84.5	80.8	79.8	73.9	94.4	62.8	40.0	77.5	96.8	85.6	66.9	68.2	75.9
DFN+LLM (Qwen2-VL)	Probing(TC)	84.9	75.6	81.4	77.2	93.4	58.2	40.9	81.1	98.1	79.8	65.1	71.3	75.6
	Probing(DC)	92.0	74.0	82.3	79.8	94.2	66.9	60.7	82.7	98.2	89.2	69.5	69.2	79.9
	SAVs†(DC) [47]	91.0	72.0	80.5	81.7	94.4	70.5	59.3	84.9	97.8	89.5	69.8	68.1	80.0
	HEC-V†(DC)	**92.2**	78.8	**85.0**	**82.4**	**95.5**	**71.8**	**62.2**	**85.3**	98.5	**89.8**	72.0	**75.5**	**82.4**

HEC-V achieves the best average accuracy 82.4%, surpassing all LVLM-based methods including SAVs [47] across all datasets. Compared to the strongest non-LVLM baselines HEC-V is best on 9/12 datasets, outperforming its vision backbone DFN on all datasets except EuroSAT. The strong results of HEC-V against DINOv3, despite using no hyperparameter tuning and relying on a less recent backbone, signals a promising new direction for training-free vision-few-shot classification. Qwen2-VL Probing(TC) is already competitive with strong visual backbones. However, Probing(DC) adds 4% in accuracy, confirming the hypothesis that domain conditioning helps retrieve domain-specific features.

5.2 Text-Zero-Shot Classification

We benchmark our method in the text-zero-shot setting. The results are reported in Tab. 2. For the baseline, we follow the standard LVLM zero-shot protocol [23], framing classification as next-token prediction with a prompt that associates a letter with each class (e.g., `A: boxer, B: yorkshire terrier, C: golden retriever, ...`). As a stronger baseline, we also report the summary token (ST) zero-shot accuracy (1)(2) following [50]. Lastly, we evaluate HEC-T in the zero-shot setup. HEC-T requires a labeled support set to select text-heads so we perform the head selection only once using the average ranking score over 100 randomly selected ImageNet tasks. We use that fixed set of 20 heads for all datasets. Note that, because the head set is selected using labeled ImageNet tasks, this protocol is not purely zero-shot in the strict CLIP sense. Downstream datasets have less than 13% class overlap with ImageNet. HEC-T and ST use a prompt with Class conditioning. For each dataset, we report the average over 100 10-way 0-shot tasks, to fit all classes into the prompt without degrading the performance of the baseline. To verify the generality of the method, we test on two LVLMs: Qwen2-VL and LLaVA-OV.

On average, HEC-T improves Qwen2-VL zero-shot by +10.1% surpassing its backbone (DFN) by 1.5% on average. For both Qwen2-VL and LLaVA-OV,

Table 2: Zero-Shot classification accuracy (%) on 10-way 0-shot across 12 datasets. Our method HEC-T provides a training-free zero-shot adaptation that outperforms previous baselines and is competitive with CLIP backbones that LVLMs inherit from. HEC-T provides meaningful gain while enabling to zero-shot an unlimited number of classes. Underline denotes the best LVLM method. **Bold** denotes the best including its CLIP backbone. Green denotes the absolute gain of HEC-T over the LVLM Baseline.

Model	Method	PETS	ESAT	UCF	SUN	CAL	DTD	AIR	FOOD	FLWR	CARS	BIRD	SIGN	AVG
DFN	Zero-Shot	**97.6**	53.0	87.8	**97.4**	**99.3**	78.4	70.7	**96.7**	**93.8**	**99.6**	**96.5**	45.3	84.7
DFN+LLM (Qwen2-VL)	Baseline	84.1	33.6	85.6	94.2	98.4	70.9	62.1	91.2	78.3	91.1	69.7	54.5	76.1
	ST [50]	90.1	41.6	90.5	94.5	98.8	77.3	66.4	92.3	90.0	96.0	78.5	56.8	81.1
	HEC-T	95.2	**54.0**	**92.6**	97.0	99.3	**84.2**	**78.7**	95.2	91.1	97.7	85.9	**63.8**	**86.2**
		+11.1	+20.4	+7.0	+2.8	+0.9	+13.3	+16.6	+4.0	+12.9	+6.7	+16.2	+9.3	+10.1
SigLIP	Zero-Shot	**97.0**	**41.1**	87.1	96.8	**99.5**	84.1	**79.6**	**97.1**	**95.8**	**99.5**	**95.5**	47.1	**85.0**
SigLIP+LLM (LLaVA-OV)	Baseline	79.6	37.2	92.1	96.0	98.6	77.8	61.9	94.8	66.7	92.5	63.7	72.6	79.7
	ST [50]	69.8	32.1	91.7	52.4	95.9	44.2	64.7	69.3	50.0	93.6	59.1	59.5	65.2
	HEC-T	83.8	40.0	**94.2**	**97.2**	99.1	**84.5**	73.0	95.4	72.2	96.8	67.2	**73.7**	82.1
		+4.2	+2.8	+2.1	+1.3	+0.5	+6.7	+11.1	+0.6	+5.5	+4.3	+3.5	+1.1	+2.4

HEC-T outperforms ST and the baseline on every dataset. For LLaVA-OV, HEC-T yields a smaller gain of +2.4% over the baseline, but improves ST by +16.9%. Notably, ST and HEC-T are the only methods that can scale with the number of classes as the baseline is limited by the context window. HEC-T's consistent gains across 12 heterogeneous benchmarks support that top text-heads transfer across domains. However, the CLIP backbones still win on more datasets overall (DFN beats Qwen2-VL HEC-T on 7/12 datasets; SigLIP beats LLaVA-OV HEC-T on 8/12). While HEC-T bridged the gap between LVLMs and CLIPs in zero-shot scenarios, CLIP still yielded strong performance. We notice that in general, CLIP wins on saturated benchmarks. For Qwen2-VL, HEC-T wins only when the performance is below 90%. This hints that LVLMs with HEC-T are more robust to domains under-represented in pretraining data.

5.3 Vision-Text-Few-Shot Classification

We benchmark our method in the vision-text-few-shot setting. We evaluate in N-way 4-shot with N equal to the total number of classes in the dataset. The results are reported in Tab. 3. We evaluate all baselines using CLIP and LVLM as an encoder (1) (2). For LVLM-based methods, we use Domain prompt conditioning. We do not include classes in the prompt, as most datasets have $N \gg 20$ classes, which would cause a drop in performance.

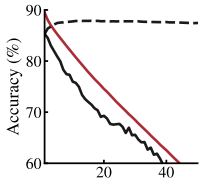
HEC-VT outperforms all LVLM-based baselines by more than 3% on average. Combining HEC-T and HEC-V improves performance on every dataset except UCF that already yields strong results with HEC-V. Averaged over all datasets, HEC is the only LVLM method that surpasses the best CLIP-based baseline. However, CLIP-based methods still achieve higher accuracy on 5 out of the 12 datasets. We hypothesize that part of that performance gap could be linked to the post-training of Qwen2-VL. Similarly to zero-shot, we notice that HEC-VT wins on less saturated benchmarks. Additionally, HEC-VT consistently outperforms

Table 3: Vision-Text-Few-shot classification accuracy (%) on 4-shot across 12 datasets. We report results for CLIP-based and LVLM-based baselines. Combining HEC-T and HEC-V in a single classifier gives state-of-the-art performance, outperforming previous CLIP-based or LVLM-based baselines. Underline denotes the best LVLM-based method. **Bold** denotes the best overall. Methods marked with † do not use hyperparameter tuning.

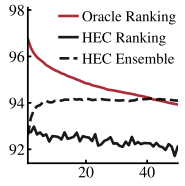
Model	Method	PETS	ESAT	UCF	SUN	CAL	DTD	AIR	FOOD	FLWR	CARS	BIRD	SIGN	AVG
DFN	Zero-Shot† [54]	92.0	51.6	63.4	79.5	95.6	51.1	29.6	87.2	82.0	92.1	78.0	28.1	69.2
	Probing [6]	84.9	81.6	79.9	73.7	94.4	61.4	38.0	77.0	97.3	85.2	66.7	65.9	75.5
	TipAdapter [67]	92.3	69.2	77.4	80.5	95.8	64.4	40.2	87.2	97.0	92.7	78.7	50.0	77.1
	GDA [61]	**92.8**	78.0	84.0	83.2	96.5	70.0	46.3	87.2	98.5	**93.6**	80.3	67.8	81.5
	ProKeR [4]	91.2	**82.4**	83.9	82.2	**97.0**	66.2	43.1	**87.8**	98.3	**93.6**	**81.5**	64.6	81.0
DFN+LLM (Qwen2-VL)	Zero-Shot† [50]	55.0	48.8	30.8	65.5	73.3	32.0	30.1	72.4	8.2	45.5	6.1	30.9	41.5
	Probing [6]	92.0	74.0	82.3	79.8	94.2	66.9	60.7	82.7	98.2	89.2	69.5	69.2	79.9
	TipAdapter [67]	79.0	50.4	57.6	76.1	84.9	58.5	51.8	79.0	74.7	74.6	54.0	49.0	65.8
	GDA [61]	92.3	69.2	79.5	82.1	94.0	68.7	60.4	83.9	96.4	87.9	69.4	64.5	79.0
	ProKeR [4]	86.7	74.4	77.4	81.4	94.1	67.2	55.8	84.4	94.4	86.9	65.7	63.3	77.6
	SAVs† [47]	91.0	72.0	80.5	81.7	94.4	70.5	59.3	84.9	97.8	89.5	69.8	68.1	80.0
	HEC-T†	85.2	55.6	69.3	75.1	92.4	62.6	31.6	83.9	50.7	72.6	47.1	47.9	64.5
	HEC-V†	92.2	78.8	**85.0**	82.4	95.5	71.8	62.2	85.3	98.5	89.8	72.0	75.5	82.4
	HEC-VT	**92.8**	82.0	**85.0**	**83.3**	95.6	**72.7**	**62.3**	85.7	**98.6**	90.1	72.1	**76.2**	**83.0**

Conditioning	HEC-T Acc.	HEC-T ER	HEC-V Acc.	HEC-V ER
None	$82.34_{0.8}$	-	$90.43_{0.5}$	-
Task	$84.09_{0.7}$	↓ 9.89%	$91.42_{0.5}$	↓ 10.31%
Domain	$87.08_{0.6}$	↓ 18.81%	$92.78_{0.4}$	↓ 15.85%
Class	$88.45_{0.6}$	↓ 10.63%	$94.14_{0.4}$	↓ 18.84%

(a) 10-way 4-shot performance under different conditioning. ER stands for Error Reduction in percentage.



(b) text-head zero-shot accuracy



(c) vision-head few-shot accuracy

Fig. 5: Ablation studies. Prompt Conditioning (left) and Head Ranking (right).

CLIP-based methods on less object-centric datasets, such as textures (DTD), scenes (SUN), and human actions (UCF).

5.4 Ablation Studies: Prompt Conditioning and Head Ranking

Figure 5a reports performance given four types of prompt conditioning: None, Task, Domain, and Class. Incrementally adding conditioning results in better zero-shot and few-shot performance. Given the setup of Sec. 3, Figs. 5b and 5c report the performance of our ranking and ensemble method on the top 50 heads, showing HEC robustness.

Ablations and Comparison with SAVs. To better isolate our algorithmic contributions, Tab. 4 shows how HEC improves over SAVs by isolating five contributions in the low-shot and high-shot regimes. Our head ranking (SAR+GDA) is the main contribution. In the low-shot regime, most gains come from SAR due to overfitting, while in high-shot, the less noisy GDA estimate boosts performance. Domain Conditioning (DC) does not explain most of the gain in this setup. In

Table 4: Ablation from SAVs to HEC-VT. DC: Domain Conditioning. GDA: Gaussian discriminant classifier Eq. (6). SAR: Soft-Accuracy Ranking Eq. (7). PE: Probability Ensembling Eq. (9).

Method (10-way)	DC	SAR	GDA	PE	HEC-T	4-shot	16-shot
In-Context	-	-	-	-	-	84.98	NA
SAVs						87.03	94.45
+ DC	✓					88.13	95.23
+ SAR	✓	✓				93.28	94.99
+ GDA	✓		✓			83.91	95.06
Our ranking	✓	✓	✓			93.99	96.03
Ensemble	✓	✓		✓		93.28	95.12
HEC-V w/o DC		✓	✓	✓		93.23	95.70
HEC-V	✓	✓	✓	✓		94.09	96.17
HEC-VT	✓	✓	✓	✓	✓	**94.42**	**96.23**

the 4-shot setup, we also compare HEC against the in-context learning baseline, where support examples are directly provided in the LVLM prompt. We do not evaluate this baseline in the 10-way 16-shot setup, since the support set exceeds the model context window.

Further ablations on prompt sensitivity, and shot number are provided in the supplementary material **??**.

6 Conclusion

We've seen that HEC improves few-shot classification across a variety of setups, notably showcasing prompt-guided domain adaptation. In addition, it closes the performance gap between LVLM-based and CLIP-based methods without the need for fine-tuning. Thus we think that combining prompt conditioning with top head selection has the potential to generalize to other setups beyond few-shot and zero-shot classification. Future work will focus on adding more complex prompts and in-context examples.

Limitations We acknowledge that the need for an intermediate representation (i.e., $\mathbf{h}_m$) for classification is a limitation, especially for API-based usage. Also, class conditioning, while promising, is limited to a small number of classes. Furthermore, LVLM inference is more computationally intensive than CLIP models.

Acknowledgements

This work was partially funded by AID-DGA (l'Agence de l'Innovation de Défense a la Direction Générale de l'Armement, Minitère des Armees), and was also partly funded by the ANR-DFG project BOFOR ANR-24-CE92-0048. This work was granted access to the HPC resources of IDRIS under the allocations 2025-AD011016525 made by GENCI.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. In: 5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings (2017)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Balestriero, R., LeCun, Y.: Lejepa: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544 (2025)
4. Bendou, Y., Ouasfi, A., Gripon, V., Boukhayma, A.: Proker: A kernel perspective on few-shot adaptation of large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25092–25102 (2025)
5. Bertinetto, L., Torr, P.H., Henriques, J., Vedaldi, A.: Meta-learning with differentiable closed-form solvers. In: 7th International Conference on Learning Representations, ICLR 2019 (2019)
6. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
7. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision (2014)
8. Cao, Q., Xu, Z., Chen, Y., Ma, C., Yang, X.: Domain prompt learning with quaternion networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26637–26646 (2024)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
10. Chen, G., Shen, L., Shao, R., Deng, X., Nie, L.: Lion: Empowering multimodal large language model with dual-level visual knowledge. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
11. Chen, S., Han, Z., He, B., Liu, J., Buckley, M., Qin, Y., Torr, P., Tresp, V., Gu, J.: Can multimodal large language models truly perform multimodal in-context learning? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
12. Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., Shen, Y., Liang, L., Gu, J., Chen, H.: Unified hallucination detection for multimodal large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3235–3252 (2024)
13. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: Exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9062–9071 (2021)
14. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2818–2829 (2023)
15. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

17. Fahes, M., Vu, T.H., Bursuc, A., Pérez, P., De Charette, R.: Poda: Prompt-driven zero-shot domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 18623–18633 (2023)
18. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A.T., Shankar, V.: Data filtering networks. In: The Twelfth International Conference on Learning Representations. OpenReview.net (2024)
19. Fei-Fei, L., Fergus, R., Perona, P.: 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)
20. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
21. Guo, Z., Zhang, R., Qiu, L., Ma, X., Miao, X., He, X., Cui, B.: Calip: Zero-shot enhancement of clip with parameter-free attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 746–754 (2023)
22. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)
23. He, H., Li, G., Geng, Z., Xu, J., Peng, Y.: Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In: The Thirteenth International Conference on Learning Representations (2025)
24. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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)
25. Hendel, R., Geva, M., Globerson, A.: In-context learning creates task vectors. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 9318–9333 (2023)
26. Hojel, A., Bai, Y., Darrell, T., Globerson, A., Bar, A.: Finding visual task vectors. In: European Conference on Computer Vision. pp. 257–273. Springer (2024)
27. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In: The 2013 international joint conference on neural networks (IJCNN). pp. 1–8. Ieee (2013)
28. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in neural information processing systems **36**, 31096–31116 (2023)
29. Huang, B., Mitra, C., Arbelle, A., Karlinsky, L., Darrell, T., Herzig, R.: Multimodal task vectors enable many-shot multimodal in-context learning. Advances in Neural Information Processing Systems **37**, 22124–22153 (2024)
30. Huang, C., Zhu, Y., Zhu, S., Xiao, J., Andrade, M., Chopra, S., Kira, Z.: Mimicking or reasoning: Rethinking multi-modal in-context learning in vision-language models. arXiv preprint arXiv:2506.07936 (2025)
31. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In: The Thirteenth International Conference on Learning Representations (2025)
32. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)

33. Kubokawa, T., Srivastava, M.S.: Estimation of the precision matrix of a singular wishart distribution and its application in high-dimensional data. Journal of multivariate Analysis **99**(9), 1906–1928 (2008)
34. Li, B., Lin, Z., Peng, W., Nyandwi, J.d.D., Jiang, D., Ma, Z., Khanuja, S., Krishna, R., Neubig, G., Ramanan, D.: Naturalbench: Evaluating vision-language models on natural adversarial samples. Advances in Neural Information Processing Systems **37**, 17044–17068 (2024)
35. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2024)
36. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
37. Li, W., Wang, Q., Meng, X., Wu, Z., Yin, Y.: Vt-fsl: Bridging vision and text with llms for few-shot learning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
38. Li, Y., Jia, X., Sang, R., Zhu, Y., Green, B., Wang, L., Gong, B.: Ranking neural checkpoints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2663–2673 (2021)
39. Li, Y., Guo, J., Qi, L., Li, W., Shi, Y.: Text and image are mutually beneficial: Enhancing training-free few-shot classification with clip. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5039–5047 (2025)
40. Liao, W., Wang, J., Li, H., Wang, C., Huang, J., Jin, L.: Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4038–4049 (June 2025)
41. Liu, F., Cai, W., Huo, J., Zhang, C., Chen, D., Zhou, J.: Making large vision language models to be good few-shot learners. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5415–5423 (2025)
42. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
43. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)
44. Liu, M., Roy, S., Wenjing, L., Zhong, Z., Sebe, N., Ricci, E., et al.: Democratizing fine-grained visual recognition with large language models. In: Proceedings of 2024 International Conference on Learning Representations (2024)
45. Liu, M., Wu, F., Li, B., Lu, Z., Yu, Y., Li, X.: Envisioning class entity reasoning by large language models for few-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 18906–18914 (2025)
46. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
47. Mitra, C., Huang, B., Chai, T., Lin, Z., Arbelle, A., Feris, R., Karlinsky, L., Darrell, T., Ramanan, D., Herzig, R.: Enhancing few-shot vision-language classification with large multimodal model features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2760–2772 (2025)
48. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)

49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (2024)
50. Ouali, Y., Bulat, A., Xenos, A., Zaganidis, A., Metaxas, I.M., Martinez, B., Tzimiropoulos, G.: Vladva: Discriminative fine-tuning of lvlms. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4101–4111 (2025)
51. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
52. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
53. Qu, X., Gou, G., Zhuang, J., Yu, J., Song, K., Wang, Q., Li, Y., Xiong, G.: Proapo: Progressively automatic prompt optimization for visual classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25145–25155 (2025)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
55. Santos, G.O.d., Colombini, E., Avila, S.: What do vision-language models see in the context? investigating multimodal in-context learning. arXiv preprint arXiv:2510.24331 (2025)
56. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
57. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
58. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
59. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., et al.: The caltech-ucsd birds-200-2011 dataset. Tech. rep., California Institute of Technology (2011)
60. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
61. Wang, Z., Liang, J., Sheng, L., He, R., Wang, Z., Tan, T.: A hard-to-beat baseline for training-free clip-based adaptation. In: The Twelfth International Conference on Learning Representations (ICLR) (2024)
62. Wang, Z., Luo, Y., Zheng, L., Huang, Z., Baktashmotlagh, M.: How far pre-trained models are from neural collapse on the target dataset informs their transferability. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5549–5558 (2023)
63. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: 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)
64. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. In:

The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
65. Yu, H., Zhao, Z., Yan, S., Korycki, L., Wang, J., He, B., Liu, J., Zhang, L., Fan, X., Yu, H.: Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6286–6297 (2025)
66. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
67. Zhang, R., Fang, R., Gao, P., Zhang, W., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
68. Zhang, Y., Unell, A., Wang, X., Ghosh, D., Su, Y., Schmidt, L., Yeung-Levy, S.: Why are visually-grounded language models bad at image classification? Advances in Neural Information Processing Systems **37**, 51727–51753 (2024)
69. Zhu, X., Zhang, R., He, B., Zhou, A., Wang, D., Zhao, B., Gao, P.: Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2605–2615 (October 2023)
70. Zong, Y., Bohdal, O., Yu, T., Yang, Y., Hospedales, T.: Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In: International Conference on Machine Learning. pp. 62867–62891. PMLR (2024)