# Meta-Prompting for Automating Zero-shot Visual Recognition with LLMs Supplementary Material

**M. Jehanzeb Mirza**[†1,2] **Leonid Karlinsky**[3] **Wei Lin**[4]
**Sivan Doveh**[5,6] **Jakub Micorek**[1] **Mateusz Kozinski**[1]
**Hilde Kuhene**[3,7] **Horst Possegger**[1,2]

[1]ICG, TU Graz, Austria. [2]CDL-EML. [3]MIT-IBM Watson AI Lab, USA.
[4]JKU, Austria. [5]IBM Research, Israel. [6]Weizmann Institute of Science, Israel.
[7]University of Bonn, Germany.

As supplementary material for our MPVR: Meta Prompting for Visual Recognition, we first list additional implementation details (Section 1). Then, for additional insights, we provide an ablation on the use of the in-context dataset employed for meta-prompting (Section 2). Moving forward, we provide results with different strategies employed for prompting multimodal language models (MMLMs) for the task of object recognition (Section 3), demonstrating we used the best performing available strategy for the MMLM baseline in the main paper. Then, we provide results for ensembling (in probability space) the vision language model (VLM) prompts generated through our MPVR (Section 4). Later, we conclude with experiments performed during the rebuttal phase (Section 5) and detailed (dataset-wise) results (Section 6).

## 1 Implementation Details

All our experiments are performed on a single NVIDIA 3090 GPU. To obtain the results for the baselines, we use their official codebase and run the baselines locally with all their recommended parameters and settings. For CUPL [14] we only report the results on the datasets, for which the authors provided the category-level VLM prompts. Since CUPL uses hand-crafted dataset-specific LLM queries to generate the category-level VLM prompts, for some datasets these queries are not available, so we were not able to generate the VLM prompts for those datasets. We used the category-level VLM attributes provided by DCLIP [12] in their official repository[1]. For the datasets, not listed in their repository, we used their official code to generate the attributes and used them for obtaining the Waffle [17] results, following the official publication. In contrast to CUPL [14], the attributes can be generated for any dataset, only by providing the class names from the downstream datasets. Similarly, following the official publication and settings proposed in Waffle [17], the datasets for which the high-level concepts are not available (*i.e.,* ImageNet [5], ImageNetv2 [16], CIFAR10/100 [9]), their two variants, Waffle+Con and Waffle+Con+GPT, collapse to only the Waffle results, in all the tables.

[1] https://github.com/sachit-menon/classifybydescriptionrelease

## 2 Meta Prompt

In the main manuscript, we arbitrarily employed the Describable Textures Dataset (DTD) [4] as the in-context example dataset for all our experiments. However, when the target dataset is DTD, we switched the in-context example dataset to EuroSAT [6]. Here, we studied the effect of employing different in-context datasets. For example, when employing an alternative in-context dataset, such as Flowers [13] or CUBS [19] for DTD (as the target dataset), the variance in results is only ±0.71, considerably lower than the gains of 8.4% (50.8% *vs.* 42.4%) obtained over the baseline of CLIP + 'dataset-specific templates', for the ViT-B/32 backbone from CLIP [15].

Similarly, while using an alternative in-context dataset, Flowers or Cubs, for the target dataset EuroSAT, the variance in obtained results is only ±0.44, again considerably lower than the gains of 9.8% (55.6% *vs.* 45.8%) obtained over the baseline of CLIP + 'dataset-specific templates'. Furthermore, for completeness, we also provide 2 complete meta-prompt examples in Figure 1 while choosing different in-context demonstrators (*i.e.,* DTD [4] and Flowers [13]) and target datasets (*i.e.,* ImageNet-R [7] and DTD [4]).

## 3 Prompt Engineering for MMLM

To address the sensitivity of MMLMs to different prompting strategies, we extensively tested the following different prompting variations used for the task of category recognition for MMLMs. These prompting strategies are also illustrated in Figure 2 for the EuroSAT dataset.

*Categories as Numbered Options:* The prompt to the MMLM [11] contained the categories (the model needed to choose from) listed as numbered options.

*Categories as Alphabet Options:* The prompt to the MMLM [11] contained the categories (the model needed to choose from) listed as English alphabet options.

*Categories as List:* In this prompting strategy, we provided the category names as a list and the MMLM was prompted to output the exact name of the category for each test image.

In Table 1 we list the results for different prompting strategies and find that the best results were obtained when LLAVA-1.6 [11] was prompted with categories (to choose from) as numbered options. For the fairest comparison, the LLAVA-1.6 [11] baseline results reported in Table 4 of the main manuscript were obtained using this (top-performing) prompting option for all the tested datasets.

## 4 Ensembling Descriptions

In the main manuscript (Table 3) we provide results by constructing the zero-shot classifier by ensembling the VLM prompts in two different ways:

| | Numbered Options | Alphabet Options | List Option |
|---|---|---|---|
| EuroSAT | 41.3 | 38.7 | 34.4 |

**Table 1:** Top-1 accuracy (%) with different prompting strategies for LLAVA-1.6 [10].

| Ensemble in Embedding Space | | | | | |
|---|---|---|---|---|---|
| | EuroSAT | Flowers | DTD | Resisc | Mean |
| Top-1 (%) - ViT-B/32 | 55.6 | 73.9 | 50.8 | 64.0 | 61.1 |
| **Ensemble in Probability Space** | | | | | |
| | EuroSAT | Flowers | DTD | Resisc | Mean |
| Top-1 (%) - ViT-B/32 | 54.5 | 73.0 | 51.0 | 61.3 | 60.0 |

**Table 2:** Comparison of constructing the zero-shot classifier by ensembling the GPT MPVR prompts over the embedding or probability space.

*Embedding Space:* The zero-shot classifier is constructed as the mean of the embeddings (from the text encoder of CLIP [15]) from the different sources (*e.g.,* Mixtral [8] or GPT [3]) VLM prompts).

*Probability Space:* The zero-shot classifier is constructed as the mean of the probabilities (*e.g.,* from softmax) obtained by different VLM prompt sources (*e.g.,* Mixtral [8] or GPT [3]) for MPVR.

In Table 3 (main manuscript) we observed different behaviors (in terms of the obtained results) from these two sources of ensembling. In theory, an ensemble over the probability space can also be obtained for the individual category-specific VLM prompts (from stage 2) of the MPVR. However, for datasets with a larger number of classes (*e.g.,* ImageNet [5] with 1000 classes), such an ensemble is prohibitively expensive (as also noted in [15]). Nevertheless, for completeness, in Table 2, we provide results for the two ensembling methods for datasets with a smaller number of classes. From these results, we observe that the two different ensembling methods do not result in a huge deviation in performance. Note, to obtain all the MPVR results in all our experiments reported in the main paper, we always construct the zero-shot classifier as the mean of the embeddings from the VLM prompts for each category.

## 5 Additional Insights and Experiments

This section provides additional insights and experiments the reviewers requested during the review process. First, we examine the role of two-stage prompting in MPVR, then study the concerns of data leakage (due to LLM already knowing the downstream datasets) and also look into detail why adding the class information in the meta prompt hurt the MPVR performance, later provide a few

| CLIP | CLIP+MPVR | GEM | GEM+MPVR |
|---|---|---|---|
| 11.2 | **15.0** | 46.2 | **51.3** |

| | base | novel |
|---|---|---|
| OVD | 56.6 | 36.9 |
| OVD+MPVR | **57.1** | **40.6** |

**Table 3: Left:** Semantic Segmentation mIOU (CLIP ViT-B/16) on Pascal VOC. **Right:** Object Detection mAP@50 on MS-COCO.

| datasets | direct-replace | MPVR |
|---|---|---|
| flowers | 66.9 | 75.2 |
| sun | 63.4 | 67.0 |
| food | 78.5 | 81.3 |
| eurosat | 50.3 | 55.6 |
| mean | 64.8 | 69.8 |

**Table 4:** Top-1 accuracy by directly replacing the task-specific information in the in-context prompts and the 2-stage MPVR.

| | **EuroSAT** | **INR** | **Flowers** | **INS** | **DTD** | **FGVCAircraft** | **Food** | **kinetics400** | **Caltech101** | **places365** |
|---|---|---|---|---|---|---|---|---|---|---|
| OpenCLIP | 42.9 | 74.4 | 69.8 | 53.0 | 49.2 | 23.0 | 78.2 | 38.6 | 95.9 | 42.1 |
| MPVR (GPT) | 57.6 | 77.6 | 74.3 | 54.9 | 61.7 | 26.0 | 78.7 | 42.3 | 94.8 | 43.7 |
| SigLip | 41.3 | 89.3 | 84.3 | 67.1 | 62.1 | 40.7 | 89.1 | 46.1 | 97.8 | 41.4 |
| MPVR (GPT) | 46.4 | 90.3 | 88.7 | 68.3 | 66.7 | 45.9 | 88.5 | 48.4 | 97.2 | 43.6 |
| | **CUBS200** | **ImageNet** | **Cars** | **SUN397** | **ImageNetV2** | **CIFAR10** | **CIFAR100** | **OxfordPets** | **UCF101** | **RESISC45** |
| OpenCLIP | 65.2 | 66.1 | 88.3 | 68.2 | 57.9 | 93.7 | 75.8 | 87.3 | 63.4 | 55.9 |
| MPVR (GPT) | 67.0 | 67.0 | 88.2 | 69.6 | 59.0 | 93.9 | 75.8 | 91.4 | 66.9 | 66.6 |
| SigLip | 65.5 | 75.7 | 90.7 | 69.6 | 68.4 | 92.5 | 70.9 | 93.2 | 70.8 | 60.3 |
| MPVR (GPT) | 66.3 | 76.2 | 90.3 | 70.9 | 69.0 | 92.6 | 71.1 | 93.9 | 69.6 | 64.3 |

**Table 5:** Top-1 Accuracy (%) for OpenCLIP (ViT-B/32) and SigLIP (ViT-B/16).

qualitative examples and finally conclude with results for additional downstream tasks and comparison with OpenCLIP [18] and SigLIP [21] backbones.

*Effect of 2-Stages in MPVR:* The role of in-context examples provided to the LLM is only to specify the desired output format to the LLM (*i.e.,* a `python` code with category name placeholders). To further analyze if 'LLM merely replaces the corresponding part of in-context prompts', we manually replaced the downstream task specification (*e.g.,* `texture` → `flower`, for Oxford Flowers dataset) in the in-context prompts (provided in the supplementary Fig. 1) and generated the VLM prompts directly from stage 2 (circumventing stage 1). Results in Table 4

show that our proposed meta-prompting allows for more diverse task-adaptive LLM knowledge extraction, not possible through simple heuristic replacement.

*Class names in meta prompt and data leakage* We manually verified the LLM queries generated from meta-prompts with and without additional class name information and found: 1) with class names, the generated queries are less diverse and focus on specifics, thus restricting the diversity of the VLM prompts; 2) some queries are not syntactically correct, resulting in less effective VLM prompts. To test against leakage of dataset information, we build a new dataset dubbed as `Mixture Dataset`, by combining Flowers, Textures, and EuroSAT datasets. Our MPVR-generated VLM prompts improve the CLIP 0-shot accuracy from 46.7% $\rightarrow$ 51.7%, suggesting that the performance gains with MPVR are not due to data leakage.

*Qualitative Examples* In the two randomly chosen prompts from the best performing *Clematis* (left) and worst performing *Ball moss* (right) category in the Oxford Flowers dataset (Figure 3) we observe that instead of describing the *Ball moss* flower in the UK, the prompt is about the Spanish moss of the same name.

*Additional VLMs and downstream tasks* MPVR also scales to VLMs trained on different sources of data. We see an average improvement of 3.3% (64.4% $\rightarrow$ 67.8%) and 1.6% (70.8% $\rightarrow$ 72.4%) for **OpenCLIP** and **SigLIP**. Dataset-wise results are provided in Table 5. Furthermore, in Table 3 we see that our MPVR improves upon vanilla CLIP and the training-free SOTA method GEM [2] for semantic segmentation, and also improves Open Vocabulary Object Detector OVD [1].

## 6 Detailed Results

For completeness, here we provide dataset-wise results for two experiments in the main manuscript: ensembling different text sources (Table 3) and mean results (over 20 datasets) for different backbones, listed in Table 2. To this end, in Table 6 and Table 7 we provide the dataset-wise results by mixing different text sources and ensembling these sources either in the embedding space or the probability space. In Tables 8 & 9 we provide the dataset-wise results for ViT-B/16 and ViT-L/14 from CLIP [15]. Furthermore, in Tables 10, 11 & 12 we list the detailed results for 3 different backbones (ViT-B/32, ViT-B/16 and ViT-L/14) from MetaCLIP [20]. The detailed (dataset-wise) results also highlight that our MPVR performs favorably on most datasets when compared to the state-of-the-art methods.

| ViT-B/32 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ImageNet** | **ImageNetv2** | **C10** | **C100** | **Caltech101** | **Flowers** | **Stanford Cars** | **Cubs** | **Pets** | **DTD** |
| GPT+TEMP | 45.9 | 40.2 | 87.6 | 59.0 | 88.1 | 72.6 | 40.8 | 54.6 | 86.2 | 48.3 |
| MIXTRAL+TEMP | 46.7 | 41.2 | 87.6 | 59.0 | 88.8 | 74.4 | 40.0 | 55.4 | 86.5 | 47.5 |
| GPT+MIXTRAL | **64.9** | **57.4** | **89.8** | **66.0** | **92.8** | **76.0** | **59.2** | **55.8** | **88.6** | **51.1** |
| GPT+MIXTRAL+TEMP | 46.1 | 40.8 | 86.7 | 60.7 | 89.9 | **76.0** | 40.7 | 55.7 | 87.9 | 49.2 |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| GPT+TEMP | 80.1 | 20.9 | 42.3 | 66.1 | 62.8 | 36.6 | 58.8 | 29.3 | **56.6** | 62.2 |
| MIXTRAL+TEMP | 80.8 | 22.2 | 41.9 | 66.0 | 59.8 | 35.7 | 61.1 | 15.5 | 50.6 | 61.6 |
| GPT+MIXTRAL | **81.1** | **22.3** | **42.7** | **67.1** | **67.1** | **43.4** | **70.3** | **44.0** | 56.4 | **65.0** |
| GPT+MIXTRAL+TEMP | 80.2 | 21.1 | 42.6 | 66.2 | 63.0 | 37.7 | 61.9 | 29.5 | 55.3 | 63.3 |
| **ViT-B/16** | | | | | | | | | | |
| | **ImageNet** | **ImageNetv2** | **C10** | **C100** | **Caltech101** | **Flowers** | **Stanford Cars** | **Cubs** | **Pets** | **DTD** |
| GPT+TEMP | 49.9 | 44.8 | 84.6 | 63.7 | 89.2 | 76.1 | 45.8 | 58.6 | 87.6 | 53.2 |
| MIXTRAL+TEMP | 50.5 | 45.5 | 86.9 | 62.6 | 89.8 | 78.3 | 44.5 | 59.7 | 89.6 | 50.0 |
| GPT+MIXTRAL | **69.7** | **63.4** | **91.2** | **69.6** | **94.4** | **79.2** | **65.6** | 59.8 | **90.7** | **55.4** |
| GPT+MIXTRAL+TEMP | 50.2 | 45.1 | 85.8 | 64.3 | 91.5 | 78.9 | 46.4 | **60.1** | 90.1 | 53.4 |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| GPT+TEMP | 85.8 | 27.8 | 43.1 | 67.9 | 67.0 | 40.8 | 64.2 | 36.0 | 58.5 | 65.4 |
| MIXTRAL+TEMP | 86.2 | **29.3** | 42.7 | 68.2 | 65.3 | 39.8 | 66.4 | 17.1 | 55.0 | 63.8 |
| GPT+MIXTRAL | **86.5** | 28.3 | **43.6** | **68.8** | **70.1** | **48.0** | **78.4** | **50.6** | **60.2** | **67.2** |
| GPT+MIXTRAL+TEMP | 86.0 | 28.1 | **43.6** | 68.1 | 67.4 | 42.1 | 67.8 | 36.0 | 59.1 | 66.1 |
| **ViT-L/14** | | | | | | | | | | |
| | **ImageNet** | **ImageNetv2** | **C10** | **C100** | **Caltech101** | **Flowers** | **Stanford Cars** | **Cubs** | **Pets** | **DTD** |
| GPT+TEMP | 62.2 | 57.3 | 90.1 | 75.5 | 94.3 | 81.9 | 58.9 | 64.8 | 92.9 | 62.0 |
| MIXTRAL+TEMP | 62.4 | 56.9 | 93.3 | 76.7 | 93.6 | 82.0 | 54.9 | 66.6 | 92.2 | 60.4 |
| GPT+MIXTRAL | **76.9** | **71.0** | **96.2** | **79.4** | **95.5** | **83.9** | **78.1** | **67.3** | **93.8** | **62.9** |
| GPT+MIXTRAL+TEMP | 62.4 | 57.3 | 91.9 | 76.8 | 94.7 | 82.5 | 59.2 | 67.2 | 93.2 | 62.8 |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| GPT+TEMP | 91.2 | 33.1 | 43.4 | 72.6 | 73.2 | 50.7 | 81.4 | 49.5 | **67.2** | 70.0 |
| MIXTRAL+TEMP | 91.3 | **36.1** | 42.7 | 72.3 | 73.7 | 49.8 | 82.9 | 28.2 | 60.4 | 69.1 |
| GPT+MIXTRAL | **91.6** | 34.3 | **43.8** | **72.9** | **77.4** | **55.5** | **88.6** | **61.2** | 65.6 | **71.5** |
| GPT+MIXTRAL+TEMP | 91.4 | 33.5 | **43.8** | 72.7 | 74.4 | 51.7 | 83.5 | 49.4 | 65.4 | 70.2 |

**Table 6:** Top-1 accuracy (%) while ensembling different text sources in the embedding space. Here, the zero-shot classifier is constructed by taking the mean of the embeddings from the different individual text sources.

| Incontext Dataset: DTD \| Target Dataset: ImageNet-R | Incontext Dataset: Flowers-102 \| Target Dataset: DTD |
|---|---|
| You are provided with prompt template examples for a dataset, which are provided to the LLM to generate descriptions for the categories in these datasets. Your task is to generate 30 diverse prompts for another dataset for which you are also provided the dataset name and the description. Format it correctly for use in a Python script, and do not repeat the prompts.<br>Example<br>Dataset Name: Describable Textures Dataset (DTD)<br>Description: The Describable Textures Dataset (DTD) is an evolving collection of textural images in the wild, annotated with a series of human-centric attributes, inspired by the perceptual properties of textures. This data is made available to the computer vision community for research purposes.<br>Prompts:<br>prompts.append("Describe how does the " + category + " texture looks like.")<br>prompts.append("How can you recognize the texture of " + category + "?")<br>prompts.append("What does the texture of " + category + " look like?")<br>prompts.append("Describe an image from the internet of the " + category + " texture.")<br>prompts.append("How can you identify the texture of " + category + "?")<br><br>Dataset Name: ImageNet-R(endition)<br>Description: ImageNet-R(endition) contains art, cartoons, deviantart, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video game renditions of ImageNet classes.<br><br>Prompts:<br>prompts.append("Describe the artistic representation of the " + category + ".")<br>prompts.append("How would you visually recognize the " + category + " class in art or cartoons?")<br>prompts.append("Provide a detailed description of the graffiti or street art related to " + category + ".")<br>prompts.append("What are the distinctive visual features of the embroidery depicting the " + category + "?")<br>prompts.append("Describe the graphics that represent the visual essence of the " + category + " class.")<br>prompts.append("Illustrate the origami models inspired by the " + category + ".")<br>prompts.append("Explain the characteristics of paintings that portray the " + category + ".")<br>prompts.append("Detail the visual patterns associated with the " + category + " class.")<br>prompts.append("Describe the visual appearance of plastic objects related to the " + category + ".") | You are provided with prompt template examples for a dataset, which are provided to the LLM to generate descriptions for the categories in these datasets. Your task is to generate 30 diverse prompts for another dataset for which you are also provided the dataset name and the description. Format it correctly for use in a Python script, and do not repeat the prompts.<br>Example<br>Dataset Name: Oxford Flowers Dataset<br>Description: Oxford Flowers consists of 102 flower categories. The flowers chosen to be flowers commonly occur in the United Kingdom.<br>Prompts:<br>prompts.append("Describe how does the flower type " + category + " looks like.")<br>prompts.append("How can you recognize the flower type " + category + "?")<br>prompts.append("What does the flower type " + category + " look like?")<br>prompts.append("Describe an image from the internet of the flower type " + category + ".")<br>prompts.append("How can you identify the flower type of " + category + "?")<br><br>Dataset Name: Describable Textures Dataset (DTD)<br>Description: The Describable Textures Dataset (DTD) is an evolving collection of textural images in the wild, annotated with a series of human-centric attributes, inspired by the perceptual properties of textures. This data is made available to the computer vision community for research purposes.<br><br>Prompts:<br>prompts.append("Describe the visual characteristics of the texture labeled as " + category + ".")<br>prompts.append("How would you recognize the texture labeled as " + category + "?")<br>prompts.append("What are the key features of the texture labeled as " + category + "?")<br>prompts.append("Provide a detailed description of the appearance of the texture labeled as " + category + ".")<br>prompts.append("If you see an image with the texture labeled as " + category + ", what would stand out to you?")<br>prompts.append("Imagine you encounter a surface with the texture labeled as " + category + ". How would you describe it?")<br>prompts.append("What visual attributes define the texture category " + category + "?")<br>prompts.append("Describe an image featuring the texture labeled as " + category + ".")<br>prompts.append("Create a caption for an image showcasing the texture labeled as " + category + ".") |

**Fig. 1:** Exemplary meta-prompts (and a few LLM generated responses) for MPVR using different in-context (left: DTD [4], right: Flowers [13]) and target (left: ImageNet-R [7], right: DTD [4]) datasets.

| Categories as Numbers | Categories as English Alphabets |
|---|---|
| Identify the category of this satellite image from the following options:<br><br>0. Annual Crop Land<br>1. Forest<br>2. Herbaceous Vegetation Land<br>3. Highway or Road<br>4. Industrial Buildings<br>5. Pasture Land<br>6. Permanent Crop Land<br>7. Residential Buildings<br>8. River<br>9. Sea or Lake<br><br>Answer with the option's number from the given choices directly. | Identify the category of this satellite image from the following options:<br><br>A. Annual Crop Land<br>B. Forest<br>C. Herbaceous Vegetation Land<br>D. Highway or Road<br>E. Industrial Buildings<br>F. Pasture Land<br>G. Permanent Crop Land<br>H. Residential Buildings<br>I. River<br>J. Sea or Lake<br><br>Answer with the option's letter from the given choices directly |
| **Categories as List** | |
| Identify the category of this satellite image from the following list:<br><br>[Annual Crop Land, Forest, Herbaceous Vegetation Land, Highway or Road, Industrial Buildings, Pasture Land, Permanent Crop Land, Residential Buildings, River, Sea or Lake]<br><br>Answer with the exact category name from the given list of categories. | |

**Fig. 2:** Example of different prompting options explored for LLAVA for EuroSAT [6].

| | |
|---|---|
| 1. Unique Shape: The clematis flower has an interesting shape that sets it apart from other flowers. It has 6-8 pointed petals that radiate outwards, giving it a star-like appearance.<br>2. Petals: The clematis flower has 4-8 sepals that resemble petals and come in a wide range of colors, including white, pink, purple, blue, and red. | 1. Ball moss, also known as Spanish moss, is a epiphytic plant that grows in spherical clumps on branches and limbs of trees.<br>2. The flower of the ball moss, also known as Tillandsia recurvata, is a small, delicate, and intricate structure. It is a greenish-yellow color and is composed of multiple individual flowers, each about 1 cm in length. |

**Fig. 3:** Qualitative Examples

| ViT-B/32 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ImageNet** | **ImageNetv2** | **C10** | **C100** | **Caltech101** | **Flowers** | **Stanford Cars** | **Cubs** | **Pets** | **DTD** |
| GPT+TEMP | **64.8** | **57.2** | **89.9** | **66.3** | 92.7 | 73.7 | **59.4** | 55.8 | 88.3 | 51.3 |
| MIXTRAL+TEMP | 63.8 | 56.3 | 89.5 | 65.5 | **93.0** | 75.2 | 58.2 | 55.5 | 88.4 | 50.3 |
| GPT+MIXTRAL | 62.9 | 55.9 | 89.7 | 66.1 | 92.8 | 76.1 | 52.4 | **55.9** | 88.8 | 51.2 |
| GPT+MIXTRAL+TEMP | 62.9 | 55.8 | 89.8 | 66.1 | 92.9 | **76.2** | 52.4 | 55.8 | **89.0** | **51.4** |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| GPT+TEMP | 81.1 | 21.9 | 42.1 | **67.0** | **68.0** | **43.7** | 70.1 | **44.2** | 55.0 | 63.9 |
| MIXTRAL+TEMP | **81.2** | **22.3** | 42.1 | 66.5 | 66.4 | 42.2 | 70.1 | 42.7 | 53.3 | 64.5 |
| GPT+MIXTRAL | 79.2 | 22.2 | 42.6 | 66.2 | 67.2 | 43.4 | **70.3** | 43.8 | **56.4** | **65.0** |
| GPT+MIXTRAL+TEMP | 79.2 | **22.3** | **42.7** | 66.2 | 67.2 | 43.4 | **70.3** | 43.9 | 56.1 | **65.0** |

| ViT-B/16 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ImageNet** | **ImageNetv2** | **C10** | **C100** | **Caltech101** | **Flowers** | **Stanford Cars** | **Cubs** | **Pets** | **DTD** |
| GPT+TEMP | **69.8** | **63.2** | 90.8 | 69.7 | 94.0 | 76.8 | **65.4** | 58.8 | 89.8 | **56.6** |
| MIXTRAL+TEMP | 68.8 | 62.2 | 91.1 | 69.2 | 94.0 | 78.2 | 62.2 | **60.3** | 90.4 | 54.1 |
| GPT+MIXTRAL | 67.7 | 61.6 | 91.1 | 69.7 | **94.4** | 79.4 | 57.0 | 60.1 | 91.0 | 55.6 |
| GPT+MIXTRAL+TEMP | 67.7 | 61.5 | **91.2** | **69.8** | **94.4** | **79.7** | 57.0 | 60.1 | **91.1** | 55.9 |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| GPT+TEMP | 86.4 | 27.5 | 43.0 | **68.9** | **70.7** | **48.0** | 78.3 | 50.7 | 59.1 | 66.1 |
| MIXTRAL+TEMP | **86.6** | **29.8** | 42.6 | 68.7 | 68.8 | 46.9 | 78.4 | 49.7 | 59.0 | 66.7 |
| GPT+MIXTRAL | 84.7 | 28.1 | **43.5** | 68.4 | 70.0 | **48.0** | 78.4 | 50.8 | **60.3** | **67.1** |
| GPT+MIXTRAL+TEMP | 84.6 | 28.0 | **43.5** | 68.4 | 69.8 | 47.9 | **78.5** | **50.9** | 60.0 | 67.0 |

| ViT-L/14 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ImageNet** | **ImageNetv2** | **C10** | **C100** | **Caltech101** | **Flowers** | **Stanford Cars** | **Cubs** | **Pets** | **DTD** |
| GPT+TEMP | **76.8** | **70.9** | 95.8 | 79.1 | **96.2** | 83.7 | **78.4** | 65.2 | 93.6 | **62.9** |
| MIXTRAL+TEMP | 75.9 | 69.6 | 96.1 | 79.3 | 95.3 | 83.6 | 70.9 | **67.5** | 93.0 | 61.6 |
| GPT+MIXTRAL | 75.2 | 69.7 | **96.2** | **79.5** | 95.8 | **84.4** | 65.7 | 67.3 | **93.9** | 62.7 |
| GPT+MIXTRAL+TEMP | 75.2 | 69.7 | **96.2** | **79.5** | 95.7 | 84.3 | 65.8 | 67.4 | **93.9** | **62.9** |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| GPT+TEMP | **91.5** | 34.3 | 43.5 | **72.9** | **77.9** | **55.7** | 88.5 | 61.1 | **67.6** | 71.0 |
| MIXTRAL+TEMP | 91.4 | **37.5** | 42.5 | 72.4 | 75.9 | 54.6 | **88.6** | 60.0 | 61.9 | 71.1 |
| GPT+MIXTRAL | 90.6 | 35.6 | **43.8** | 72.1 | 77.5 | 55.6 | **88.6** | 61.1 | 65.2 | **71.7** |
| GPT+MIXTRAL+TEMP | 90.6 | 35.5 | **43.8** | 72.2 | 77.5 | 55.6 | **88.6** | **61.2** | 65.1 | 71.6 |

**Table 7:** Top-1 accuracy (%) while ensembling different text sources in the probability space. Here, the zero-shot classifier is constructed by taking the mean of the softmax probabilities from the different individual classifiers.

| | ImageNet | ImageNetv2 | C10 | C100 | Caltech101 | Flowers | Stanford Cars | Cubs | Pets | DTD |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 66.7 | 60.9 | 90.1 | 68.4 | 93.3 | 67.5 | 65.5 | 55.1 | 88.2 | 43.3 |
| DS-TEMP | 68.3 | 61.9 | 90.8 | 68.2 | 92.9 | 70.7 | **66.2** | 56.1 | 89.1 | 43.2 |
| CUPL | **69.7** | **63.4** | 90.3 | 69.0 | 94.4 | 70.9 | 60.0 | 56.0 | **91.2** | 53.3 |
| D-CLIP | 68.6 | 62.2 | 89.6 | 68.4 | **94.5** | 72.1 | 63.7 | 56.7 | 90.3 | 42.8 |
| Waffle | 68.3 | 62.3 | 90.8 | 68.8 | 93.7 | 72.2 | 64.0 | 57.0 | 89.2 | 41.9 |
| Waffle+Con | 68.3 | 62.3 | 90.8 | 68.8 | 90.7 | 69.0 | 63.9 | 56.5 | 89.4 | 42.7 |
| Waffle+Con+GPT | 68.3 | 62.3 | 90.8 | 68.8 | 94.4 | 72.3 | 63.8 | 56.8 | 89.7 | 42.8 |
| MPVR (Mixtral) | 68.8 | 62.2 | **91.1** | 69.1 | 94.2 | **78.4** | 62.2 | **60.4** | 90.3 | 53.7 |
| MPVR (GPT) | **69.7** | **63.4** | 90.8 | **69.5** | 94.1 | 76.9 | 65.4 | 59.0 | 89.9 | **56.1** |

| | Food101 | Aircraft | Places365 | SUN397 | UCF101 | K400 | IN-R | IN-S | EuroSAT | Resisc45 |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 85.2 | 23.8 | 39.3 | 62.5 | 65.1 | 43.7 | 74.0 | 46.2 | 42.3 | 56.5 |
| DS-TEMP | 85.9 | 24.3 | 40.9 | 65.3 | 68.5 | 47.4 | 77.7 | 48.8 | 48.9 | 60.1 |
| CUPL | 86.1 | 26.6 | – | **69.0** | 68.9 | 46.0 | – | – | – | 66.2 |
| D-CLIP | 86.1 | 24.0 | 42.0 | 66.1 | 67.5 | 45.2 | 76.5 | 48.9 | 58.5 | 64.8 |
| Waffle | **86.9** | 24.9 | 42.0 | 65.4 | 67.1 | 46.1 | 77.0 | 49.1 | 49.6 | 64.8 |
| Waffle+Con | 86.5 | 24.2 | 39.8 | 62.9 | 66.5 | 45.1 | 76.3 | 48.2 | 48.1 | 61.7 |
| Waffle+Con+GPT | 86.7 | 24.9 | 42.4 | 66.4 | 68.4 | 46.0 | 77.0 | 49.5 | 55.6 | 65.2 |
| MPVR (Mixtral) | 86.6 | **29.9** | 42.7 | 68.9 | 68.9 | 46.9 | **78.4** | 49.7 | 59.2 | **66.7** |
| MPVR (GPT) | 86.4 | 28.0 | **43.1** | 68.8 | **70.9** | **48.0** | 78.2 | **50.6** | **59.6** | 66.2 |

**Table 8:** Top-1 accuracy (%) for 20 datasets obtained by employing the ViT-B/16 backbone from OpenAI CLIP [15]. S-TEMP refers to the results obtained by using the default template (`a photo of a <class name>`), while DS-TEMP refers to the results obtained by using the ensemble of dataset-specific prompts. An empty placeholder for CUPL [34] indicates that the respective baseline did not provide the handcrafted prompts for the dataset. For Waffle [17], mean results from 7 random runs are reported, following the original publication.

| | ImageNet | ImageNetv2 | C10 | C100 | Caltech101 | Flowers | Stanford Cars | Cubs | Pets | DTD |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 73.5 | 67.8 | 95.2 | 77.2 | 94.3 | 76.2 | 76.9 | 62.1 | 93.1 | 52.5 |
| DS-TEMP | 75.5 | 69.9 | 95.7 | 78.3 | 93.7 | 79.5 | 78.1 | 61.8 | 93.5 | 54.8 |
| CUPL | 76.7 | 70.8 | 95.8 | 78.6 | 96.1 | 79.6 | 64.2 | 60.3 | **94.3** | 61.1 |
| D-CLIP | 75.1 | 69.0 | 95.2 | 78.4 | **97.0** | 79.5 | 75.1 | 61.7 | 93.0 | 56.1 |
| Waffle | 75.1 | 68.9 | 96.0 | 78.4 | 96.2 | 78.3 | 76.5 | 62.3 | 93.2 | 55.3 |
| Waffle+Con | 75.1 | 68.9 | 96.0 | 78.4 | 93.9 | 77.3 | 76.7 | 63.1 | 93.4 | 53.7 |
| Waffle+Con+GPT | 75.1 | 68.9 | 96.0 | 78.4 | 96.9 | 79.0 | 75.9 | 62.0 | 93.1 | 56.1 |
| MPVR (Mixtral) | 75.9 | 69.6 | **96.1** | **79.3** | 95.4 | **83.8** | 70.6 | **67.7** | 93.1 | 61.6 |
| MPVR (GPT) | **76.8** | **70.9** | 96.0 | 79.2 | 96.1 | 83.6 | **78.3** | 65.5 | 93.7 | **62.9** |

| | Food101 | Aircraft | Places365 | SUN397 | UCF101 | K400 | IN-R | IN-S | EuroSAT | Resisc45 |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 90.3 | 30.0 | 40.1 | 67.6 | 73.8 | 51.3 | 85.4 | 58.3 | 55.1 | 63.2 |
| DS-TEMP | 90.9 | 31.8 | 41.2 | 69.0 | 76.2 | 55.0 | 87.8 | 59.8 | 63.2 | 68.0 |
| CUPL | 91.4 | 35.1 | – | 72.8 | 75.8 | 54.4 | – | – | – | **71.8** |
| D-CLIP | 91.1 | 31.8 | 42.3 | 69.6 | 76.2 | 52.5 | 86.8 | 59.0 | 54.6 | 70.7 |
| Waffle | **91.5** | 32.5 | 42.6 | 69.4 | 76.0 | 53.4 | 87.4 | 59.1 | 50.4 | 71.4 |
| Waffle+Con | 91.2 | 31.3 | 41.1 | 66.2 | 74.2 | 52.0 | 86.2 | 58.6 | 44.2 | 66.7 |
| Waffle+Con+GPT | 91.4 | 32.1 | 42.9 | 70.1 | 76.4 | 53.5 | 87.3 | 59.3 | 53.7 | 71.1 |
| MPVR (Mixtral) | 91.4 | **37.6** | 42.5 | 72.5 | 75.8 | 54.6 | **88.5** | 60.0 | 62.2 | 71.2 |
| MPVR (GPT) | **91.5** | 34.4 | **43.5** | **73.0** | **78.1** | **55.7** | 88.4 | **61.0** | **67.3** | 71.1 |

**Table 9:** Top-1 accuracy (%) for 20 datasets obtained by employing the ViT-L/14 backbone from OpenAI CLIP [15].

| | ImageNet | ImageNetv2 | C10 | C100 | Caltech101 | Flowers | Stanford Cars | Cubs | Pets | DTD |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 64.1 | 56.3 | 91.2 | 66.8 | **95.5** | 69.8 | 71.7 | 61.8 | 86.9 | 47.6 |
| DS-TEMP | 65.6 | 57.5 | 91.3 | **70.2** | 93.8 | 71.3 | **72.1** | 62.5 | **88.7** | 51.8 |
| CUPL | **66.0** | 57.5 | 90.3 | 68.4 | **95.5** | 68.3 | 61.3 | 61.5 | 88.5 | 58.4 |
| D-CLIP | 64.0 | 55.0 | 90.9 | 68.4 | 94.9 | 67.6 | 66.6 | 62.1 | 87.9 | 50.0 |
| Waffle | 63.5 | 55.5 | 90.9 | 67.2 | 93.9 | 69.7 | 68.8 | 61.7 | 88.6 | 50.0 |
| Waffle+Con | 63.5 | 55.5 | 90.9 | 67.2 | 88.2 | 68.6 | 69.1 | 61.8 | **88.7** | 48.5 |
| Waffle+Con+GPT | 63.5 | 55.5 | 90.9 | 67.2 | 94.8 | 68.7 | 68.1 | 62.1 | 88.1 | 51.0 |
| MPVR (Mixtral) | 64.8 | 57.4 | 91.2 | 68.9 | 94.3 | **78.4** | 68.3 | **65.2** | 88.1 | **61.5** |
| MPVR (GPT) | **66.0** | **57.6** | **91.4** | 69.2 | 94.5 | 74.8 | 71.2 | 64.6 | 88.0 | **61.5** |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| S-TEMP | 76.7 | 24.3 | 39.6 | 64.8 | 64.4 | 36.9 | 71.5 | 52.3 | 49.4 | 56.4 |
| DS-TEMP | **77.3** | 26.9 | 40.1 | 65.3 | 66.1 | 39.1 | 74.8 | 53.9 | 50.4 | 60.6 |
| CUPL | 77.0 | 32.3 | _ | **67.7** | 64.2 | 39.3 | _ | _ | _ | 63.9 |
| D-CLIP | 76.7 | 25.3 | 42.0 | 64.3 | 65.6 | 37.6 | 73.2 | 52.3 | 49.0 | 62.4 |
| Waffle | 77.2 | 26.0 | 42.1 | 65.8 | 64.1 | 38.1 | 73.9 | 52.9 | 42.3 | **64.4** |
| Waffle+Con | 77.1 | 25.4 | 41.4 | 66.0 | 63.5 | 37.4 | 72.8 | 52.8 | 37.8 | 63.6 |
| Waffle+Con+GPT | 77.2 | 25.7 | **42.4** | 65.6 | 65.8 | 38.4 | 73.8 | 52.9 | 46.7 | 63.4 |
| MPVR (Mixtral) | 77.0 | **35.4** | 41.4 | 67.3 | 65.4 | 39.9 | **75.7** | 53.1 | 56.3 | 61.4 |
| MPVR (GPT) | 77.1 | 31.8 | **42.4** | 65.8 | **67.3** | **40.6** | 75.6 | **54.1** | **58.7** | 63.6 |

**Table 10:** Top-1 accuracy (%) for 20 datasets obtained by employing the ViT-B/32 backbone from MetaCLIP [20].

| | ImageNet | ImageNetv2 | C10 | C100 | Caltech101 | Flowers | Stanford Cars | Cubs | Pets | DTD |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 70.0 | 61.8 | 89.9 | 64.9 | 95.7 | 71.7 | 74.7 | 69.5 | 88.5 | 53.0 |
| DS-TEMP | 70.8 | 62.6 | **90.1** | 66.5 | 95.6 | 73.8 | 75.8 | 69.7 | 90.5 | 56.3 |
| CUPL | 70.9 | 62.5 | 89.2 | 65.5 | 95.5 | 70.8 | 0.5 | 68.9 | 89.8 | 62.2 |
| D-CLIP | 69.0 | 60.7 | 88.6 | 64.6 | 95.7 | 72.7 | 71.9 | 68.4 | 90.1 | 53.5 |
| Waffle | 69.1 | 61.0 | 87.9 | 64.9 | 95.0 | 73.3 | 73.1 | 68.2 | 90.8 | 53.5 |
| Waffle+Con | 69.1 | 61.0 | 87.9 | 64.9 | 94.1 | 72.1 | 72.3 | 68.5 | **91.0** | 52.5 |
| Waffle+Con+GPT | 69.1 | 61.0 | 87.9 | 64.9 | **95.8** | 72.9 | 73.0 | 68.1 | 90.7 | 55.1 |
| MPVR (Mixtral) | 69.8 | 62.0 | 89.8 | 65.6 | 95.5 | **80.6** | 74.0 | 71.2 | 90.4 | 64.1 |
| MPVR (GPT) | **71.2** | **62.9** | 89.8 | **66.6** | 94.8 | 75.9 | **75.9** | **71.4** | 89.9 | **64.4** |
| | **Food101** | **Aircraft** | **Places365** | **SUN397** | **UCF101** | **K400** | **IN-R** | **IN-S** | **EuroSAT** | **Resisc45** |
| S-TEMP | 83.8 | 26.3 | 41.6 | 68.8 | 67.0 | 39.9 | 80.1 | 56.5 | 50.9 | 63.5 |
| DS-TEMP | **84.1** | 28.3 | 41.7 | 68.4 | **69.0** | 43.2 | 81.8 | **58.7** | 55.2 | 63.9 |
| CUPL | 84.0 | 34.4 | _ | **69.5** | 66.6 | 43.3 | _ | _ | _ | 67.0 |
| D-CLIP | 83.7 | 30.1 | 42.5 | 66.8 | 67.2 | 41.6 | 79.5 | 57.1 | 56.1 | 67.3 |
| Waffle | 83.9 | 30.5 | 42.3 | 67.7 | 68.4 | 42.4 | 80.0 | 56.9 | 53.3 | 67.8 |
| Waffle+Con | 83.9 | 30.2 | 41.5 | 68.2 | 66.3 | 41.7 | 79.4 | 57.5 | 49.7 | 67.8 |
| Waffle+Con+GPT | 84.0 | 30.4 | 42.8 | 68.0 | 68.5 | 42.4 | 80.1 | 57.2 | 55.9 | 68.3 |
| MPVR (Mixtral) | 84.0 | **37.8** | 41.4 | 69.4 | 67.9 | 44.2 | **82.2** | 57.2 | **59.7** | 67.0 |
| MPVR (GPT) | 83.6 | 34.0 | **43.0** | 69.4 | 68.8 | **44.9** | 82.1 | 58.2 | 57.8 | **69.1** |

**Table 11:** Top-1 accuracy (%) for 20 datasets obtained by employing the ViT-B/16 backbone from MetaCLIP [20].

| | ImageNet | ImageNetv2 | C10 | C100 | Caltech101 | Flowers | Stanford Cars | Cubs | Pets | DTD |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 75.1 | 68.5 | 94.9 | 74.4 | 96.8 | 76.7 | 84.5 | 76.0 | 88.7 | 58.8 |
| DS-TEMP | 76.2 | 69.9 | 95.7 | **77.4** | 96.3 | 77.4 | **84.9** | 75.2 | **93.7** | 60.5 |
| CUPL | 76.5 | 69.9 | 95.0 | 76.3 | 97.0 | 75.8 | 81.1 | 74.4 | 92.7 | 64.5 |
| D-CLIP | 74.4 | 67.8 | 95.7 | 75.9 | 97.0 | 76.7 | 82.9 | 74.7 | 93.0 | 58.0 |
| Waffle | 74.3 | 67.9 | 95.6 | 76.6 | 96.2 | 78.3 | 83.0 | 74.5 | 92.9 | 59.6 |
| Waffle+Con | 74.3 | 67.9 | 95.6 | 76.6 | 95.3 | 78.6 | 83.8 | 75.0 | 93.2 | 57.0 |
| Waffle+Con+GPT | 74.3 | 67.9 | 95.6 | 76.6 | **97.4** | 77.5 | 83.2 | 74.5 | 93.0 | 60.2 |
| MPVR (Mixtral) | 75.5 | 68.6 | **95.9** | 76.5 | 96.6 | **85.5** | 82.2 | **77.9** | 92.6 | **67.3** |
| MPVR (GPT) | **76.6** | **70.1** | 95.1 | 76.0 | 96.0 | 84.9 | 83.7 | 77.6 | 93.0 | 65.8 |

| | Food101 | Aircraft | Places365 | SUN397 | UCF101 | K400 | IN-R | IN-S | EuroSAT | Resisc45 |
|---|---|---|---|---|---|---|---|---|---|---|
| S-TEMP | 88.6 | 35.6 | 42.2 | 72.1 | 75.2 | 48.6 | 87.7 | 63.9 | 49.7 | 61.6 |
| DS-TEMP | 88.5 | 40.0 | 42.0 | 72.0 | 75.9 | 51.0 | 88.9 | 65.3 | 56.8 | 69.1 |
| CUPL | 89.0 | 41.2 | – | 71.9 | 75.0 | 51.1 | – | – | – | 71.2 |
| D-CLIP | 88.4 | 39.5 | 43.5 | 71.1 | 75.9 | 49.5 | 87.7 | 64.2 | **61.3** | 67.9 |
| Waffle | 88.7 | 39.0 | 43.3 | 71.7 | 75.9 | 49.8 | 87.5 | 64.0 | 59.6 | 69.5 |
| Waffle+Con | 88.9 | 38.8 | 41.4 | 71.4 | 75.1 | 49.2 | 87.2 | 64.1 | 59.3 | 65.2 |
| Waffle+Con+GPT | 88.7 | 39.7 | 43.0 | 72.3 | 76.3 | 50.2 | 87.9 | 64.3 | **61.3** | 69.0 |
| MPVR (Mixtral) | **89.1** | **49.5** | 40.2 | **73.1** | 74.8 | 51.8 | **89.4** | 65.1 | 56.3 | 70.5 |
| MPVR (GPT) | 88.8 | 46.7 | **43.8** | 72.5 | **77.5** | **52.3** | 89.2 | **65.5** | 58.0 | **72.8** |

**Table 12:** Top-1 accuracy (%) for 20 datasets obtained by employing the ViT-L/14 backbone from MetaCLIP [20].

## References


1. Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F.: Bridging the Gap between Object and Image-level Representations for Open-Vocabulary Detection. NeurIPS (2022) 5
2. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding Everything: Emerging Localization Properties in Vision-language Transformers. In: Proc. CVPR (2024) 5
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. arXiv:2005.14165 (2020) 3
4. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing Textures in the Wild. In: Proc. CVPR (2014) 2, 7
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. CVPR (2009) 1, 3
6. Helber, P., Bischke, B., Dengel, A., Borth, D.: EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. In: Proc. IGARSS (2018) 2, 8
7. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In: Proc. ICCV (2021) 2, 7
8. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of Experts. arXiv preprint arXiv:2401.04088 (2024) 3
9. Krizhevsky, A., Hinton, G.: Learning Multiple Layers of Features from Tiny Images. Tech. rep., Department of Computer Science, University of Toronto (2009) 1
10. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 (2023) 3
11. Liu, H., Li, C., Li, Y., Lee, Y.J.: LLaVA-Next (LLaVA 1.6). arXiv:2310.03744 (2023), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 2
12. Menon, S., Vondrick, C.: Visual Classification via Description from Large Language Models. Proc. ICLR (2023) 1
13. Nilsback, M.E., Zisserman, A.: Automated Flower Classification Over a Large Number of Classes. In: Proc. ICVGIP (2008) 2, 7
14. Pratt, S., Liu, R., Farhadi, A.: What does a platypus look like? Generating customized prompts for zero-shot image classification. arXiv:2209.03320 (2022) 1
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models from Natural Language Supervision. In: Proc. ICML (2021) 2, 3, 5, 10
16. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet Classifiers Generalize to ImageNet? In: Proc. ICML. pp. 5389–5400. PMLR (2019) 1
17. Roth, K., Kim, J.M., Koepke, A., Vinyals, O., Schmid, C., Akata, Z.: Waffling around for Performance: Visual Classification with Random Words and Broad Concepts. arXiv preprint arXiv:2306.07282 (2023) 1, 10
18. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b:

An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022) 4
19. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. rep., California Institute of Technology (2011) 2
20. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying CLIP Data. In: Proc. ICLR (2023) 5, 11, 12
21. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-training. In: Proc. ICCV (2023) 4