Decompose, Compare, and Decide: Multimodal LLMs are Implicit Few-Shot Learners

Yunhan Wang, Eshika Khandelwal, Edson Araujo, Walid Bousselham, Nina Shvetsova, and Hilde Kuehne

Tuebingen AI Center, University of Tuebingen, Germany
yunhan.wang@student.uni-tuebingen.de

Abstract. Multimodal Large Language Models (MLLMs) have demonstrated remarkable abilities when analyzing images, yet translating these capabilities to few-shot image classification remains challenging. To bridge this gap, we present DeCoDe, a simple yet effective technique that enables off-the-shelf MLLMs to act as strong few-shot classifiers without any additional training. Our approach builds on the idea of few-shot classification as a set of pairwise image comparisons, decomposing the task into a set of binary decisions. Given a query image and a support image from a candidate class, the MLLM is prompted to decide whether the two images depict the same class. The logit corresponding to an affirmative response is then used as a similarity score to assign the query image to the most likely class. While this already yields good results, we show that providing additional high-level information, such as the data domain, to the model further improves performance. Our evaluation provides an extensive analysis of various inference variants on a suite of twelve datasets, six established and six newly curated few-shot benchmarks spanning across diverse domains. The results show that the proposed simple decomposition technique can turn off-the-shelf MLLMs into powerful few-shot learners, significantly outperforming current state-of-the-art few-shot methods on both standard and novel domains. Code is available at https://github.com/yunhanwang1105/DeCoDe.

Keywords: Few-Shot Learning · Multimodal Large Language Models · Vision-Language Models · In-Context Learning

1 Introduction

Few-shot learning (FSL) enables models to generalize to new tasks and domains using only a small number of labeled examples per class. This can be especially valuable for real-world scenarios where large annotated datasets are unavailable or adaptation to novel tasks is required at inference time. Early few-shot methods required model training on a set of base classes in order to enable recognition of held-out novel classes [12, 21, 33, 38, 40, 43]. With the emergence of foundation models trained on large-scale web data via self- or weakly-supervised objectives, such as DINO [29] or CLIP [32], the focus shifted towards exploring

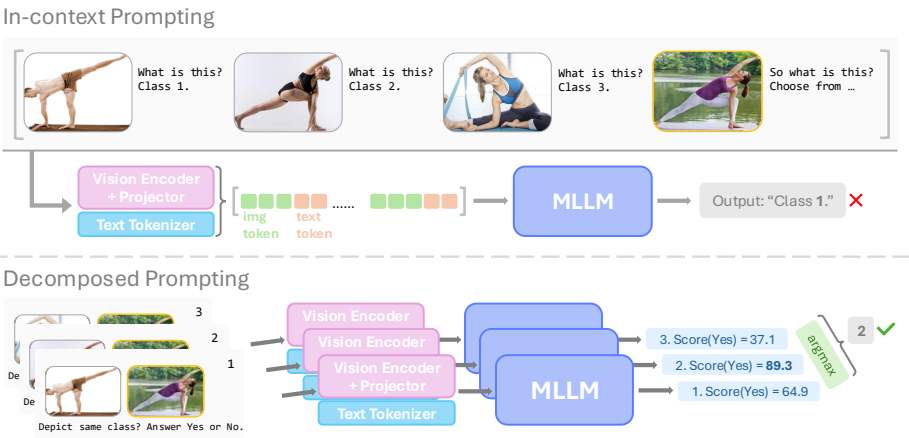


Fig. 1: We propose a decomposed prompting technique (DeCoDe) for few-shot classification with MLLMs. We **decompose** the task into pairwise support–query comparisons, asking whether two images belong to the same class. By ranking the model's affirmative responses across candidate pairs (**compare**) and selecting the highest-scoring logit as the predicted class (**decide**), MLLMs become strong few-shot classifiers without any training. Unlike standard in-context prompting, which relies heavily on semantic label names, our decomposed approach succeeds even with anonymized labels by forcing the model to perform direct visual-to-visual comparison.

how pretrained representations can be leveraged for improved few-shot classification with minimal or no additional training [6,8,14,41,49,51,52]. More recently, multimodal large language models (MLLMs), such as LLaVA-OneVision [22], InternVL3 [54], or Qwen3-VL [3], have demonstrated strong vision-language capabilities, raising the question of whether such models can also serve as effective few-shot learners [25,45].

So far, MLLMs have been explored for few-shot learning in three ways: by extracting discriminative embeddings from the internal representations of the MLLM [27], by verbalizing discriminative visual features and matching them to query images [45], or via in-context learning where support and query images are jointly provided in a single prompt [25]. However, in-context prompting has been shown to perform poorly for few-shot image classification when used off-the-shelf [25,27], often requiring additional fine-tuning to encourage stronger reliance on the support examples [25].

To address this limitation, the proposed approach reformulates few-shot classification as a set of pairwise comparisons. Instead of reasoning over all support examples jointly, the few-shot task is decomposed into binary prompts asking the MLLM whether a query image and a support image belong to the same class. Unlike existing methods, particularly [25,45], the proposed technique considers the logit of the `Yes` token to rank all candidate pairs and assign the query image to the most likely class. While this simple technique, when applied to standard

MLLMs, is already sufficient to outperform current state-of-the-art methods, it shows that the models can benefit from contextual information, such as the domain of the data. Namely, we found that the performance of the pairwise decomposed setting further improves when models are given the topic of the data, usually based on the dataset name, such as cars, yoga poses, etc. We thus coin the respective workflow as ***De****compose*, ***Co****mpare*, ***De****cide* (DeCoDe) and show that this technique can transform state-of-the-art MLLMs into high-performing few-shot learners.

However, evaluating the few-shot capabilities of web-scale pretrained models remains challenging. In particular, the performance of models on standard few-shot benchmarks may be influenced by potential overlap between benchmark datasets and the models' pretraining corpora. As a result, models may rely on memorized knowledge rather than adapting to the provided support examples [20]. To address this issue, we conduct an extensive analysis on two sets of datasets and in two evaluation modes. *First*, with respect to datasets, we consider six standard few-shot datasets, as well as six novel datasets, which can be considered as out-of-domain for web-scale trained models. We observe that most MLLMs already achieve near-perfect accuracy on the standard datasets in a 0-shot setting, often even exceeding 1-shot performance. In contrast, we consider novel, out-of-domain datasets as datasets that show substantially low 0-shot performance as well as a clear improvement from 0-shot to a simple 1-shot baseline. We consider this as an indication that the model might have seen less of the original data and needs to rely more on visual cues. *Second*, we consider two different few-shot settings for MLLMs: first, we consider a semantic setting, in which semantically meaningful class labels are provided together with the support images, and, second, we consider an anonymized setting, where support images are labeled only with numerical identifiers (e.g., `Class 1`, `Class 2`), preventing the model from relying directly on textual knowledge. Our analysis shows that the anonymization alone is enough to reduce the performance of MLLMs for in-context learning, even on standard benchmarks, whereas the proposed decomposition is able to significantly increase the performance in all cases, with and without class labels, as well as for standard and novel datasets.

The contributions of this work can be summarized as follows:

- We propose a simple, yet effective technique of *Decompose, Compare, Decide* (DeCoDe), to turn MLLMs into powerful few-shot learners by ranking the logits corresponding to an affirmative response to a binary prompt question.
- We show that decomposed MLLMs can further benefit from high-level cues such as domain information and propose a new few-shot evaluation scenario by providing the global dataset domain as context.
- We conduct a systematic evaluation of current MLLMs for few-shot learning across standard and novel out-of-domain datasets and under different modes, including anonymized class labels, revealing when models rely on memorized pretraining knowledge versus true few-shot adaptation.

2 Related Works

VLM Embeddings for Few-Shot Learning. Vision-Language Models (VLMs) such as CLIP provide a strong foundation for FSL by enabling classification based on embedding similarity. Prior work has adapted such embeddings for few-shot tasks via prompt learning [7, 18, 46, 51–53] or adapter-based methods [14, 42, 47, 49]. More recent works explore additional adaptation strategies. Two-Stage Few-Shot (2SFS) [10] performs two-stage PEFT-based representation learning; Logit Deconfusion [23] addresses class bias at the logit level; ProKeR [6] applies training-free kernel regression over pretrained CLIP features; LoRA Recycle [17] enables tuning-free adaptation via reuse of pre-trained LoRAs [16]. CAML [11] is a large-scale, meta-trained framework built on a ViT [9] backbone that enables in-context classification using frozen image embeddings. Unlike methods that modify pretrained encoders, our proposed technique does not adapt the underlying representation or update parameters. Instead, we investigate how frozen MLLMs can perform few-shot classification purely through structured inference.

Finally, Kravets et al. [20] discuss the problem that standard CLIP few-shot evaluations are partially transductive due to pretraining overlap, and introduce an unlearning-based pipeline to construct inductive benchmarks that reveal substantial performance drops in existing methods. We follow this idea, but instead of unlearning and thus changing the model parameters, we use alternative datasets that can be considered more out-of-domain for web-trained multimodal models.

MLLMs for Few-Shot Learning. Beyond adapting pretrained VLM encoders, SAVs [27] extracts representations from MLLMs for downstream few-shot classification. VRL [45] leverages MLLMs to automatically derive interpretable verbalized features that capture inter-class differences and intra-class commonalities. Other works explore generative few-shot prompting with MLLMs [25, 27], yet such models often remain less effective with respect to embedding-based few-shot methods, despite strong performance on open-ended vision–language tasks. A central challenge is their reliance on semantic language priors rather than visual evidence from the support set [5]. GFSL [25] addresses this in part through label anonymization, applied within a fine-tuning strategy. Across these works, evaluations vary considerably in scope: SAVs focus on standard benchmarks, VRL reports results on novel out-of-domain datasets. Building on this idea, we jointly examine zero-shot and few-shot regimes, the effect of semantic versus anonymized labels, and evaluate across both standard and novel datasets, providing a unified evaluation across these axes.

3 Method

In this section, we introduce our inference framework for few-shot classification. Fig. 1 illustrates the inference pipeline of our framework. Given multimodal context consisting of images and text, the MLLM is prompted to predict the

Variant	Paradigm	Label	Domain	Description
0-shot w. labels	In-context	Semantic	✗	No support images; only query image and labels given
1-shot w. labels	In-context	Semantic	✗	Support images with labels given
1-shot w/o labels	In-context	Anonymous	✗	Support images with anonymized labels e.g., `Class 1`, ..., `Class 5`
Ours (Decomposition)				
1-shot dec. w. labels	Decompose	Semantic	✗	Binary comparisons w. label added to binary question
1-shot dec.+D_{info} w. labels	Decompose	Semantic	✓	Binary comparisons w. label + Domain-adapted prompt wording
1-shot dec. w/o labels	Decompose	Anonymous	✗	Binary comparisons w/o label
1-shot dec.+D_{info} w/o labels	Decompose	Anonymous	✓	Binary comparisons w/o label + Domain-adapted prompt wording

Table 1: Overview of inference variants. Each variant is characterized by its task structure, label naming strategy, and distinguishing mechanism. *dec.* = decompose, D_{info} = domain information.

class of a query image based on the provided support examples. Depending on the prompting paradigm, the prediction is either obtained directly from the model's generated response or derived from the model's output logits for specific answer tokens (e.g., 'Yes').

Table 1 summarizes the inference variants evaluated in this work. The variants differ along three dimensions: (i) the prompting paradigm (in-context learning vs. decomposed pairwise prompting); (ii) the use of dataset-level domain information; and (iii) the anonymization of semantic class labels. In addition, we evaluate a 0-shot setting as a baseline to assess whether the models can classify a query based solely on its prior knowledge without visual support examples. In all experiments, we follow the prompt design of [25] and explore more prompt options in the supplementary material.

3.1 Problem Formulation

Classical few-shot learning protocols [12, 38, 43] divide a dataset into a set of base classes used for supervised pre-training and a set of novel classes reserved for few-shot evaluation. In contrast, training-free approaches leverage VLMs and MLLMs [6, 27, 42, 49], bypassing supervised base-class training, thanks to rich representations acquired during large-scale web pretraining. In this paper, we follow a training-free protocol in the standard *N-way K-shot* setting. The model's task is to classify a query image into one of N classes given a support set of K labeled examples per class. We define the support set $\mathcal{S}$ to be sampled from the test set as:

$$\mathcal{S} = \big\{(x^s_{n,k},\, c_n) \mid n = 1, \ldots, N,\; k = 1, \ldots, K\big\}, \tag{1}$$

where N is the number of classes sampled, K is the number of samples per class, and $\mathcal{C} = \{c_1, \ldots, c_N\}$ is the set of corresponding class names where each c_n is the semantic name of the n-th class. We formulate the query set $\mathcal{Q}$ as:

$$\mathcal{Q} = \{(x^q, c_y)\}, \tag{2}$$

where x^q is the query image and $y \in \{1, \ldots, N\}$ is its ground-truth class index with respect to the support classes $\mathcal{C}$.

Fig. 2: Variants of prompt formulations used in our experiments. **(a)** standard in-context prompting with semantic labels. **(b)** standard in-context prompting with anonymized labels. **(c)** decomposed pairwise prompting with semantic labels. **(d)** decomposed pairwise prompting with anonymized labels. **(e)** decomposed pairwise prompting with domain information and semantic labels. **(f)** decomposed pairwise prompting with domain information and anonymized labels.

3.2 In-Context Few-Shot Classification with MLLMs

To assess the few-shot capabilities of current MLLMs, we consider few-shot learning as an in-context learning task. Given a support set $\mathcal{S}$ and a query image x^q, the model is prompted to predict the class of the query image via multiple-choice selection over the candidate classes. The prompt consists of the concatenation of all support images with their corresponding class labels, followed by the query image and the list of candidate class options. The resulting prompt p is structured as:

$$p = x^s_{1,1} \oplus \texttt{Str} \oplus c_1 \ldots x^s_{N,K} \oplus \texttt{Str} \oplus c_N \;\; \oplus x^q \oplus \texttt{Str} \oplus 1.\, c_1; 2.\, c_2; \ldots; \tag{3}$$

The model is expected to output a class index $\hat{y} \in \{1, \ldots, N\}$, which we extract from the generated text via regular expression matching. Full prompt can be found in the top-left of Figure 2 (a). We further instruct the model to output the expected answer format. Exact prompts and instruction-following errors are reported in the supplementary material.

3.3 Decomposed Few-Shot Classification with MLLMs

In contrast to in-context learning, we propose to decompose the few-shot classification problem into a set of pairwise comparisons. For each support example and the query image, the model is prompted to determine whether the two images belong to the same class. Given $(\mathcal{S}, x^q)$, we decompose the support set into a set of pairwise prompts:

$$\mathcal{P} = \big\{ p_{n,k} \mid n = 1, \ldots, N,\ k = 1, \ldots, K \big\}, \tag{4}$$

Each prompt compares the query image with a support example:

$$p_{n,k} = x^s_{n,k} \oplus x^q \oplus \texttt{Str} \oplus c_n \tag{5}$$

All prompts are independently processed by the model. The full prompt can be found in the mid-right of Figure 2 (c). For each prompt $p_{n,k}$, the model's output logit for the token `Yes` is used as the score indicating whether the pair belongs to the same class:

$$s^{\text{pair}}_{n,k} = \mathit{logit}(\texttt{Yes} \mid p_{n,k}), \qquad \hat{y} = \arg\max_n \ \frac{1}{K} \sum_{k=1}^{K} s^{\text{pair}}_{n,k}\,, \tag{6}$$

where $\hat{y} \in \{1, \ldots, N\}$ is the index with the highest averaged pairwise score as the predicted class. This choice is consistent with recent VQA-based scoring approaches such as VQAScore [24], which scores image-text alignment using the `Yes` token for evaluating text-to-visual generation; in contrast, we score support–query image-pair matches for few-shot image classification.

3.4 Decomposed Prompting with Domain Information

Decomposed prompting resembles the classical few-shot classification scenario, where a single representation is compared against a set of others. However, MLLMs also offer the ability to further refine this setting by encouraging the model to focus on dataset-specific concepts. To leverage those capabilities, we replace the generic term `class` with a domain-appropriate term D_{info} (e.g., `pose`, `texture`, `species`) that reflects the concept represented in a given dataset. The full prompt can be found in the bottom-right of Figure 2 (e). The descriptors D_{info} are inferred for each dataset as shown in Table 2. This grounds each comparison in both visual and textual semantics, allowing the MLLM to leverage its pretrained domain-level knowledge alongside the visual

Table 2: Mapping of datasets to their domain information text.

Dataset	D_{info} text
mini-ImageNet [43]	–
UCF101 [39]	action category
CUB [44]	bird species
Aircraft [26]	aircraft variant
Dogs [19]	dog breed
DomainNet [30]	–
Lego [15]	Lego brick type
Industrial [35]	industrial product
Yoga [34]	yoga pose
Egyptian hieroglyph [13]	Egyptian hieroglyph
Flying insects [31]	insect species
Arabic sign language [1]	sign language

evidence from the support image[1]. We further provide preliminary evidence in the supplementary material that D_{info} can be inferred automatically with competitive performance, and we also include example dataset labels to illustrate how D_{info} captures domain-level information distinct from class labels.

3.5 Prompting Without Class Labels

To isolate the model's ability to learn from visual examples alone, we replace all semantic class names c_n with abstract identifiers (e.g., `Class 1`, ..., `Class N`) in both the support demonstrations and the option list. All other formatting stays identical. This forces the MLLM to rely on visual similarity rather than leveraging any prior knowledge associated with class names. The full anonymous prompts can be found in the right column of Figure 2.

3.6 Episode Construction

We construct evaluation episodes using the standard N-way K-shot protocol [6, 27, 42, 49]. In each episode e, an N-way label set is sampled, $\mathcal{C} = \{c_1, \ldots, c_N\}$, with support set $\mathcal{S} = \{(x^s_{n,k}, c_n)\}^{N\times K}_{n=1,k=1}$ containing K examples per class, and a single query sample $\mathcal{Q} = \{(x^q, c_y)\}$. Given $(\mathcal{S}, x^q)$, the model predicts $\hat{y} \in \{1, \ldots, N\}$. Performance is measured as mean episodic accuracy over a set of evaluation episodes $\mathcal{E}$:

$$\text{Acc} = \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \mathbf{1}[\hat{y}_e = y_e].$$

For each dataset, we pre-sample a fixed set of 1000 evaluation episodes using a fixed random seed. We verify that cumulative accuracy stabilizes well within 1000 episodes. All evaluation runs are performed on the same set of episodes to ensure fair comparison across models and variants.

4 Evaluation

4.1 Datasets

We evaluate the proposed technique on a diverse collection of datasets spanning established few-shot benchmarks and novel specialized domains.

Standard benchmarks. We evaluate on standard few-shot classification benchmarks following [11, 25, 27], including mini-ImageNet [43], a widely used general-purpose few-shot benchmark; CUB [44], Aircraft [26], and Dogs [19], which test fine-grained visual discrimination within natural categories; UCF101 [39] (middle frame of video clip), representing action recognition under a video-to-image transfer. We further remove action images with strong object bias [37]. In DomainNet [30], we sample episodes within a single domain at a time, excluding the *real* domain to avoid overlap with the mini-ImageNet domain.

[1] For mini-ImageNet and DomainNet, we use a generic instruction asking the model to focus on the depicted concept rather than domain or texture. See supplement.

Novel datasets. We introduce a benchmark suite of six visually distinctive datasets to probe model generalization under stronger domain shift. Namely, we consider the Lego bricks dataset (Lego) [15], industrial parts (Ind.) [35], yoga poses (Yoga) [34], ancient Egyptian hieroglyphs (Hiero.) [13], flying insects (Insect) [31] and Arabic alphabet sign language (Sign) [1]. These cover highly specialized concepts as shown by the particularly low 0-shot recognition performance of pretrained MLLMs. We thus assume that those concepts are only sparsely represented in standard text-image pretraining, making them suitable test cases for evaluating the visual few-shot capabilities of pretrained MLLMs.

4.2 Experimental Details

Models. We analyze three open-source state-of-the-art MLLMs of comparable scale (7–8B parameters): Qwen2.5-VL-7B-Instruct [4], Qwen3-VL-8B-Instruct [3], and InternVL3-8B [54]. All models are used in inference-only mode, with token-level scores extracted directly from output logits as described in Eq. (6).

SFT Baseline. We use supervised fine-tuning (SFT) for Qwen3-VL-8B-Instruct with LoRA [16] on the semantic in-context few-shot formulation. The training data consists of 1000 5-way 1-shot episodes sampled from the miniImageNet base classes and the model is trained to output the correct query class. We fine-tune for 3 epochs with LoRA rank 16 and alpha 16, applying LoRA to all non-visual linear modules while keeping the visual modules frozen. Optimization uses a learning rate of (5×10^{-5}), warmup ratio 0.03. The model is then evaluated directly on all target datasets without any further tuning.

FSL Baselines. We compare against a diverse set of few-shot learning baselines. CLIP [32] and SigLIP [48] are vision-language models that support 0-shot classification via text-image cosine similarity, and few-shot classification via K-nearest neighbour (KNN) search over image embeddings. CLIP uses a ViT-B/32 backbone, while the SigLIP model variant is `base-patch16-224`. DINOv2 `base` model [29] and I-JEPA (ViT-H/14) [2] are vision-only self-supervised models, evaluated via KNN over frozen embeddings. Tip-Adapter [49] builds a training-free cache model from CLIP features (ViT-B/32). CAML [11] is a large-scale meta-learning method with a ViT-H encoder pre-trained on Laion-2B [36] that performs in-context classification from frozen embeddings. ProKeR [6] implements training-free kernel regression over pretrained CLIP features (ViT-B/32). Finally, SAVs [27] directly leverage internal representations of MLLMs for training-free few-shot classification. We implement SAVs on top of Qwen2.5-VL-7B-Instruct, making it the closest baseline to our approach.

4.3 Comparison to State-of-the-art

Table 3 compares our method against few-shot baselines and recent MLLM-based approaches across standard and novel benchmarks, evaluated under both semantic and anonymized settings.

Panel A: Main comparison

Method		Standard Datasets							Novel Datasets							
		mini	UCF	CUB	Air.	Dogs	Dom.	Avg	Lego	Ind.	Yoga	Hiero.	Insect	Sign	Avg	Total
With labels (Semantic)																
Reported results																
SAVs [27]	1-shot	–	–	98.7	–	–	–	–	–	–	–	–	–	–	–	–
CAML [11]	1-shot	96.2	–	91.8	63.3	–	–	–	–	–	–	–	–	–	–	–
GFSL [25]	1-shot	98.2	–	96.6	96.6	96.7	–	–	–	–	–	–	–	–	–	–
Reproduced baselines																
CLIP [32]	0-shot	95.5	81.2	94.0	72.9	90.4	86.2	86.7	30.3	31.7	31.3	39.9	57.7	20.0	35.2	60.9
ProKeR [6]	1-shot	98.1	89.0	94.7	73.4	92.5	**90.3**	89.3	49.7	51.7	43.2	68.1	70.3	38.2	53.5	71.4
SigLip [48]	0-shot	98.3	85.4	**97.5**	79.0	**98.7**	88.2	91.5	63.0	49.4	**91.8**	47.2	72.9	20.2	57.4	74.5
SAVs [27]	1-shot	98.4	**94.4**	96.0	**89.4**	93.1	86.2	**92.9**	**68.2**	57.8	62.4	64.4	80.4	28.8	60.3	76.6
CAML [11]	1-shot	**98.7**	90.7	95.3	86.9	94.9	81.2	91.3	62.3	**78.0**	25.2	**84.4**	**95.4**	**52.3**	**66.3**	**78.8**
InternVL3																
	0-shot	**99.1**	**95.7**	89.6	73.8	**93.7**	**92.0**	90.7	53.1	60.8	58.5	45.2	68.2	18.8	50.8	70.7
	1-shot	76.2	94.8	83.2	30.3	86.6	61.8	72.2	49.0	64.9	45.2	35.6	82.4	20.0	49.5	60.8
	1-shot dec.	98.4	94.5	**94.1**	**75.7**	92.7	89.8	**90.9**	**62.8**	**75.0**	**73.6**	**75.7**	**89.0**	**58.7**	**72.5**	**81.7**
Qwen2.5-VL																
	0-shot	97.2	93.0	96.4	89.5	91.8	85.3	92.2	50.1	42.9	50.7	43.1	80.0	20.7	47.9	70.1
	1-shot	88.1	84.7	66.6	61.9	81.0	73.6	76.0	54.9	49.9	67.9	66.0	88.8	59.3	64.5	70.2
	1-shot dec.	**98.5**	**93.7**	**96.9**	**94.6**	**95.8**	**90.5**	**95.0**	**65.5**	**73.5**	**80.6**	**78.1**	**93.3**	**67.7**	**76.5**	**85.7**
Qwen3-VL																
	0-shot	**99.1**	96.6	96.5	89.1	**96.4**	**92.6**	95.1	63.4	62.8	66.0	48.1	80.1	21.3	57.0	76.0
	1-shot	85.3	92.7	85.0	75.3	86.2	84.6	84.9	66.9	80.3	74.5	82.4	**89.1**	68.4	76.9	80.9
	1-shot dec.	98.9	**97.3**	**96.8**	**92.2**	96.1	91.9	**95.5**	**70.6**	**81.4**	**83.4**	**89.5**	**89.1**	**70.6**	**79.1**	**87.3**
	1-shot SFT	98.9	97.6	97.6	90.1	96.9	92.9	95.7	77.9	77.9	81.7	81.6	95.5	70.1	80.8	88.2
Without labels (Anonymous)																
Reproduced baselines																
I-JEPA [2]	1-shot	76.8	75.5	53.8	32.5	79.1	39.8	59.6	45.1	42.1	60.0	68.4	45.1	31.3	48.7	54.1
Tip-A. [49]	1-shot	**97.7**	82.2	93.9	70.9	90.2	**87.0**	**87.0**	31.2	39.5	33.2	42.2	62.1	21.5	38.3	62.6
CLIP [32]	1-shot	86.5	83.0	80.6	61.0	66.7	63.9	73.6	51.3	51.5	35.3	80.4	76.8	41.5	56.1	64.9
DINO [29]	1-shot	93.0	**95.2**	**97.8**	64.6	**95.5**	74.1	86.7	65.0	73.6	52.1	85.1	66.0	40.2	63.7	75.2
SigLip [48]	1-shot	91.3	88.6	90.4	**85.2**	84.7	70.1	85.1	**70.3**	**75.7**	**62.4**	**87.9**	**91.0**	**48.2**	**72.6**	**78.8**
InternVL3																
	1-shot	22.9	36.0	20.4	21.0	21.0	19.9	23.5	23.0	27.5	23.4	21.0	23.3	21.9	23.4	23.4
	1-shot dec.	**94.8**	**92.8**	**90.6**	**70.8**	**91.4**	**78.8**	**86.5**	**55.4**	**81.2**	**51.9**	**92.1**	**86.0**	**76.2**	**73.8**	**80.2**
Qwen2.5-VL																
	1-shot	23.0	70.4	38.4	47.9	43.7	30.9	42.4	41.3	49.4	57.2	57.1	**88.8**	68.8	60.4	51.4
	1-shot dec.	**97.0**	**94.9**	**96.2**	**92.8**	**92.7**	**82.9**	**92.8**	**60.8**	**73.9**	**74.6**	**86.2**	72.1	**76.5**	**74.0**	**83.4**
Qwen3-VL																
	1-shot	21.4	26.6	38.0	39.4	27.7	20.5	28.9	25.6	31.1	21.3	30.0	23.6	28.1	26.6	27.8
	1-shot dec.	**94.9**	**93.0**	**87.0**	**91.1**	**88.1**	**79.1**	**88.9**	**63.7**	**81.2**	**67.5**	**94.5**	**33.9**	**72.6**	**68.9**	**78.9**
	1-shot SFT	51.4	51.8	80.3	65.1	62.3	26.8	56.3	44.2	51.1	51.4	58.4	68.2	54.4	54.6	55.5

Panel B: With dataset-specific domain information

Method		mini	UCF	CUB	Air.	Dogs	Dom.	Avg	Lego	Ind.	Yoga	Hiero.	Insect	Sign	Avg	Total
With labels (Semantic)																
InternVL3																
	1-shot dec.	**98.4**	94.5	94.1	75.7	92.7	**89.8**	90.9	62.8	75.0	73.6	75.7	89.0	58.7	72.5	81.7
	1-shot dec. $+D_{\text{info}}$	98.0	**94.9**	**95.1**	**76.6**	**93.0**	**89.8**	**91.2**	**64.7**	**75.3**	**80.1**	**77.3**	**91.8**	**65.1**	**75.7**	**83.5**
Qwen2.5-VL																
	1-shot dec.	**98.5**	93.7	96.9	94.6	95.8	90.5	95.0	65.5	73.5	80.6	78.1	93.3	67.7	76.5	85.7
	1-shot dec. $+D_{\text{info}}$	**98.5**	**94.3**	**97.6**	**95.8**	**96.5**	**91.0**	**95.6**	**66.4**	**74.2**	**84.3**	**88.1**	**97.0**	**80.9**	**81.8**	**88.7**
Qwen3-VL																
	1-shot dec.	98.9	**97.3**	96.8	92.2	96.1	91.9	95.5	70.6	**81.4**	83.4	89.5	89.1	70.6	79.1	87.3
	1-shot dec. $+D_{\text{info}}$	**99.0**	**97.3**	**97.9**	**93.6**	**96.6**	**92.0**	**96.1**	**72.9**	80.5	**88.3**	**90.2**	**97.4**	**82.0**	**85.2**	**90.6**
Without labels (Anonymous)																
InternVL3																
	1-shot dec.	94.8	92.8	90.6	70.8	**91.4**	78.8	86.5	55.4	**81.2**	51.9	**92.1**	86.0	76.2	73.8	80.2
	1-shot dec. $+D_{\text{info}}$	**95.1**	**95.0**	**95.3**	**77.0**	90.7	**80.6**	**89.0**	**61.2**	78.2	**84.8**	90.3	**93.8**	**86.9**	**82.5**	**85.7**
Qwen2.5-VL																
	1-shot dec.	97.0	**94.9**	96.2	92.8	92.7	82.9	92.8	60.8	**73.9**	74.6	**86.2**	72.1	76.5	74.0	83.4
	1-shot dec. $+D_{\text{info}}$	**97.2**	92.4	**97.4**	**95.2**	**94.7**	**84.3**	**93.5**	**63.0**	72.1	**84.3**	83.0	**97.5**	**82.0**	**80.3**	**86.9**
Qwen3-VL																
	1-shot dec.	94.9	93.0	87.0	91.1	88.1	**79.1**	88.9	63.7	**81.2**	67.5	**94.5**	33.9	72.6	68.9	78.9
	1-shot dec. $+D_{\text{info}}$	**95.2**	**96.6**	**96.8**	**92.4**	**95.4**	78.7	**92.5**	**68.9**	80.3	**89.6**	91.2	**98.2**	**86.6**	**85.8**	**89.2**

Table 3: Comparison with state-of-the-art on standard/novel benchmarks for 5-way 1-shot, resp. 0-shot if indicated. Panel A reports the main comparison; Panel B reports the effect of dataset-specific domain information D_{info}. *dec.* = decompose and *SFT* = supervised fine-tuning. Best results are highlighted **bold** per subsection. Colored rows mark our method variants.

With semantic labels. On standard datasets, 0-shot MLLMs already achieve remarkably strong performance. For example, Qwen3-VL and InternVL3 reach above 90% average accuracy without any support examples. Interestingly, 1-shot in-context inference often performs worse than 0-shot (e.g., InternVL3 drops from 90.7% to 72.2% average; Qwen2.5-VL from 92.2% to 76.0%). This indicates that

adding support images does not guarantee effective adaptation and may even introduce interference when strong semantic priors dominate.

In contrast, on novel datasets, the trend reverses: 1-shot generally improves over 0-shot. For instance, Qwen3-VL increases from 57.0% (0-shot avg) to 76.9% (1-shot avg), suggesting that support examples become more beneficial as pretraining overlap decreases. This highlights the importance of evaluating beyond saturated standard benchmarks.

Across both dataset groups, our decompositional approach consistently outperforms 1-shot inference and matches or surpasses 0-shot performance on standard benchmarks, while substantially outperforming 0-shot performance on novel datasets. Adding dataset-specific domain information further improves results, yielding the best overall averages (e.g., 90.6% for Qwen3-VL, 88.7% for Qwen2.5-VL) and establishing new state-of-the-art few-shot performance.

Compared to the FSL baseline methods, mainly Qwen3-VL models outperform current methods in the simple decomposed setting, while the other two models perform on par. This is expected for standard datasets, given the large-scale pretraining data, but the difference becomes more pronounced for novel datasets, where MLLMs significantly outperform our considered FSL baselines in most cases while remaining on par with the SFT baseline.

Without semantic labels (Anonymous). When semantic class names are removed, the performance of in-context 1-shot MLLMs drops dramatically. InternVL3 and Qwen3-VL collapse to near-random prediction (around 23–28% total average in 5-way classification), revealing a heavy reliance on semantic priors rather than true support–query alignment. Qwen2.5-VL shows slightly more robustness but still degrades substantially.

In stark contrast, the decomposition approach remains highly effective in the anonymized regime. For all three MLLMs, our method restores strong few-shot performance. With Qwen3-VL, it surpasses the SFT baseline by 23.4 points. This shows that structured pairwise inference explicitly enforces visual correspondence and mitigates the semantic dominance.

Interestingly, also in this setting, where MLLMs cannot rely on their textual knowledge, they are able to outperform few-shot baseline methods. We again observe that, also here, the performance difference to the baseline models is smaller on the standard benchmarks, but widens for the novel benchmark setting. This shows that MLLMs might have learned general visual patterns that help them to generalize beyond the training distribution.

Moreover, in Panel B of Table 3, incorporating **domain information** leads to consistent overall gains across models and evaluation settings. The improvement is particularly clear on the novel datasets, suggesting that D_{info} provides useful high-level context about the label space when the classes are less familiar. These results indicate that lightweight domain descriptions can complement decision-based prediction and improve recognition without additional training.

Overall, the results confirm that the proposed decomposed inference framework consistently enables robust few-shot adaptation across both semantic and anonymized settings, achieving state-of-the-art performance.

Method	Standard Datasets				Novel Datasets				
	mini	CUB	Dogs	Avg	Lego	Yoga	Hiero.	Avg	Total
With labels (Semantic)									
InternVL3									
0-shot	**99.1**	89.6	**93.7**	94.1	53.1	58.5	45.2	52.3	73.2
0-shot dec.	97.0	88.2	91.5	92.2	49.4	53.9	42.4	48.6	70.4
1-shot (baseline)	76.2	83.2	86.6	82.0	49.0	45.2	35.6	43.3	62.6
1-shot calibrated	94.0	**95.7**	90.9	93.5	**65.4**	56.6	47.2	56.4	75.0
1-shot dec. (ours)	98.4	94.1	92.7	95.1	62.8	73.6	75.7	70.7	82.9
1-shot dec.+D_{info} (ours)	98.0	95.1	93.0	**95.4**	64.7	**80.1**	**77.3**	**74.0**	**84.7**
Qwen2.5-VL									
0-shot	97.2	96.4	91.8	95.1	50.1	50.7	43.1	48.0	71.5
0-shot dec.	97.6	96.4	96.0	96.7	65.5	56.1	45.7	55.8	76.2
1-shot (baseline)	88.1	66.6	81.0	78.6	54.9	67.9	66.0	62.9	70.8
1-shot calibrated	94.8	96.1	92.7	94.5	64.1	70.7	72.7	69.2	81.8
1-shot dec. (ours)	**98.5**	96.9	95.8	97.1	65.5	80.6	78.1	74.7	85.9
1-shot dec.+D_{info} (ours)	**98.5**	**97.6**	**96.5**	**97.5**	**66.4**	**84.3**	**88.1**	**79.6**	**88.5**
Qwen3-VL									
0-shot	**99.1**	96.5	96.4	97.3	63.4	66.0	48.1	59.2	78.2
0-shot dec.	96.2	96.0	95.9	96.0	63.4	62.7	46.7	57.6	76.8
1-shot (baseline)	85.3	85.0	86.2	85.5	66.9	74.5	82.4	74.6	80.0
1-shot calibrated	94.5	93.2	92.4	93.4	70.9	80.0	85.8	78.9	86.2
1-shot dec. (ours)	98.9	96.8	96.1	97.3	70.6	83.4	89.5	81.2	89.2
1-shot dec.+D_{info} (ours)	99.0	**97.9**	**96.6**	**97.8**	**72.9**	**88.3**	**90.2**	**83.8**	**90.8**

Table 4: Ablation on decomposition and context across standard and novel benchmarks (5-way). Top-1 accuracy (%). *dec.* = decompose, D_{info} = dataset specific domain info. Best and second-best results are highlighted **bold** and underlined per section.

4.4 Ablation

We ablate three design choices that determine the effectiveness of our framework.

Ablation of Decomposed Prompting. First, we analyze whether the performance gains stem from the binary decomposition itself or from the support–query comparison it introduces. To this end, we consider a 0-shot decompose variant where no support image is given; the model answers: ‘‘`Is this image depicting` c_n`? Yes or No.`’’ In Table 4, 0-shot decomposition does not improve over standard 0-shot inference and can even perform slightly worse (e.g., Qwen3-VL: 78.2% → 76.8%). This indicates that the gains of 1-shot decomposition do not arise from the binary prompt structure, but from the support-query comparisons.

Ablation on Calibration Effects. Second, given that MLLMs exhibit strong language priors, we examine whether the performance of pairwise decomposition just stems from an implicit calibration. To this end, we consider an in-context 1-shot inference with calibration. Namely, we apply pointwise mutual information (PMI)-style calibration [50] to reduce option bias introduced by the prompt context. For each candidate class, we compute the standard prediction given the query image x^q and support set $\mathcal{S}$, and then subtract the score obtained when the query image is replaced by a placeholder text ‘‘`query image is omitted`’’

	Standard Datasets				Novel Datasets				
Scoring rule	mini	CUB	Dogs	Avg	Lego	Yoga	Hiero.	Avg	Total
(1) arg max[Confidence(`Yes`)]	73.3	84.6	86.6	81.5	46.9	71.2	68.2	62.1	71.8
(2) arg min[Score(`No`)]	97.0	96.1	92.9	95.3	64.1	74.4	85.7	74.7	85.0
(3) arg max[Score(`Yes`) - Score(`No`)]	97.1	96.2	92.6	95.3	63.9	74.4	86.1	74.8	85.1
(4) arg max[Score(`Yes`)]	97.1	96.1	92.7	95.3	64.1	74.6	86.2	75.0	85.1

Table 5: Comparison of pairwise scoring rules for Qwen2.5-VL in the 5-way 1-shot decomposed setting without semantic labels. All logit-style scoring variants perform nearly identically, confirming the scoring choice (4) of our method.

while keeping the rest of the prompt unchanged: $P(y|x^q, \mathcal{S}) - P(y|\mathcal{S})$. As shown in Table 4, PMI-style calibration mitigates option bias in in-context inference and improves total accuracy (e.g., InternVL3: 62.6% → 75.0%; Qwen3-VL: 80.0% → 86.2%). However, it remains inferior to decomposition. Replacing N-way inference with pairwise decomposition consistently yields larger gains (e.g., Qwen3-VL: 86.2% → 89.2%), with particularly great improvements on novel datasets.

Evaluation of Logit scoring. Having established decomposition as the key component, we verify that its performance is robust to the choice of logit scoring rule used for ranking. Our pairwise inference ranks each support–query pair using the logit score of the `Yes` token (Eq. (6)). Table 5 compares alternative scoring strategies on Qwen2.5-VL, including: (1) using the model's generated confidence score output for `Yes`; (2) minimizing the `No` logit score: Score(`No`); (3) a PMI-style correction Score(`Yes`) - Score(`No`); (4) the direct `Yes` logit score.

The exact prompt used to obtain the confidence score is provided in the supplementary material. The three logit-based variants yield nearly identical results across all datasets, confirming that the method is robust to the choice of ranking rule. In contrast, using the model's explicit confidence score leads to a substantial drop in accuracy (e.g., 73.3 on mini-ImageNet, 46.9 on Lego), indicating that generative confidence signals are less reliable for fine-grained pairwise matching. We provide further details on scoring in the supplement.

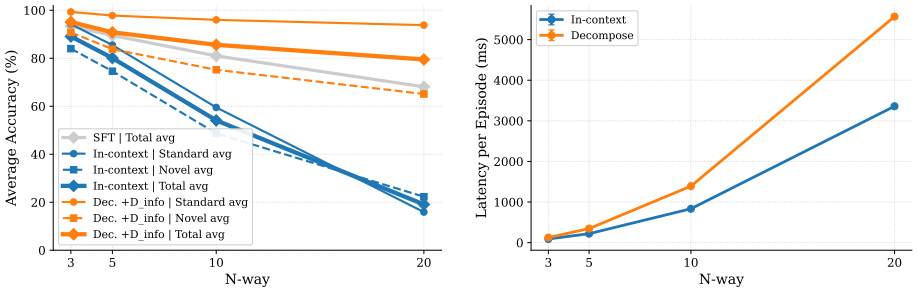
4.5 5-way 5-shot Analysis

We analyze the effect of increasing the number of shots under a fixed 5-way setting. The 5-way 5-shot episodes are constructed by extending the corresponding 5-way 1-shot episodes with four additional support samples per class, allowing for a controlled comparison (see also Equation 6). Results are reported in Table 6 for Yoga, Lego, and Industrial datasets.

While in-context learning benefits from a single support example, its performance deteriorates when moving to 5-shot, suggesting that longer in-context prompts introduce interference rather than improved adaptation. In contrast, decomposition with D_{info} consistently improves from 1-shot to 5-shot, achieving the best average accuracy (85.8%), outperforming the SFT baseline. This behavior highlights a key advantage of decomposition: additional support images contribute positively through independent pairwise aggregation.

Method	Novel Datasets							
	Yoga		Lego		Industrial		Avg	
	5w-1s	5w-5s	5w-1s	5w-5s	5w-1s	5w-5s	5w-1s	5w-5s
In-context	74.5	45.9	66.9	43.9	80.3	60.9	73.9	50.2
Decompose + D_{info}	**88.3**	**91.8**	**72.9**	**80.7**	**80.5**	**85.0**	**80.6**	**85.8**
SFT baseline	81.5	88.0	77.6	77.3	78.2	82.4	79.1	82.6

Table 6: Few-shot classification accuracy (%) on Yoga, Lego, and Industrial datasets under semantic 5-way 1-shot and 5-way 5-shot settings using Qwen3-VL.



(a) N-way-1-shot accuracy for comparing In-context and Decompose + domain info performance. We report the average accuracy for standard and novel datasets, as well as for all.

(b) N-way-1-shot runtime analysis comparing In-context and Decompose + domain info. We report the mean episode latency for single-pass and decomposed batch inference (batch size 64).

Fig. 3: Scaling with N-way using Qwen3-VL. (a) Accuracy comparison between In-context (with and without SFT) and Decompose + domain info. (b) Corresponding runtime analysis under identical decoding and batching settings. $N \in \{3, 5, 10, 20\}$.

4.6 N-way 1-shot Analysis

We further analyze the scalability with respect to the number of classes $N \in \{3, 5, 10, 20\}$ under the 1-shot setting. In Fig. 3a, In-context inference degrades sharply as N increases. While performance is competitive at small N, accuracy drops substantially at 10- and 20-way episodes. In contrast, our decomposed method remains robust across all N, outperforming the SFT baseline. The performance gap widens as the task becomes more challenging, demonstrating that structured pairwise comparison scales more reliably than standard N-way prompting. This trend holds consistently across standard, novel, and averaged results, confirming that the improved scalability is not dataset-specific but inherent to the inference structure.

4.7 Runtime

The main limitation of decomposition is that it requires access to token logits and incurs $O(N)$ comparisons for an N-way problem. Thus, to investigate the

computational cost difference, we measure runtime as end-to-end episode latency (ms) in the 1-shot N-way setting using Qwen3-VL ($N \in \{3, 5, 10, 20\}$), averaged over 50 episodes per N. We compare in-context prompting, which requires one forward pass per episode, and with the proposed pairwise decomposition, which requires N comparisons per episode. All experiments were conducted on a single NVIDIA H100 GPU (80GB) [28] using `Bfloat16` precision, deterministic decoding, left padding, CUDA synchronization, and a batch size of 64 for both episodes and generation. We report the mean latency per episode.

Results in Figure 3b show approximately linear scaling for both single inference and decomposition. The runtime gap is modest at small N but widens at 10–20 way (approximately 1.7× slower at 20-way), reflecting the computational trade-off introduced by structured decomposition. While those numbers reflect the single GPU setting, we argue that, especially for real-world scenarios, settings with larger N or K would most likely aim for a parallelization across multiple GPUs, as all forward paths can be processed independently, which would allow to mitigate longer runtime in those cases if needed.

5 Conclusion

We presented a simple yet effective framework for few-shot image classification with multimodal large language models based on *Decompose, Compare, Decide* (DeCoDe). Instead of prompting the model to directly select among class names, our method reformulates classification as binary support–query matching, using the logit of the `Yes` response as a similarity score aggregated across support examples. This structured inference encourages direct visual comparison and reduces reliance on semantic priors from class labels.

Across both standard and novel benchmarks, the proposed decomposition consistently improves performance over conventional in-context prompting and competitive training-free baselines, achieving state-of-the-art results. The gains are particularly pronounced in the anonymized setting where semantic labels are removed, revealing that structured pairwise comparison more effectively leverages the support set.

These findings suggest a practical inference-time adaptation strategy: users can define new visual categories using a few support images, while the MLLM classifies queries by explicitly comparing them against the provided references without parameter updates.

Beyond the inference method, we introduce a controlled evaluation protocol that disentangles semantic priors from visual adaptation by comparing zero-shot and few-shot regimes with and without semantic labels across diverse datasets. Our results show that standard in-context prompting relies heavily on semantic cues, while the proposed decomposition better captures support adaptation.

References

1. Al-Barham, M., Alsharkawi, A., Al-Yaman, M., Al-Fetyani, M., Elnagar, A., SaAleek, A.A., Al-Odat, M.: Rgb arabic alphabets sign language dataset. arXiv preprint arXiv:2301.11932 (2023)
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: CVPR (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Baldassini, F.B., Shukor, M., Cord, M., Soulier, L., Piwowarski, B.: What makes multimodal in-context learning work? In: CVPR (2024)
6. Bendou, Y., Ouasfi, A., Gripon, V., Boukhayma, A.: Proker: A kernel perspective on few-shot adaptation of large vision-language models. In: CVPR (2025)
7. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Plot: Prompt learning with optimal transport for vision-language models. In: ICLR (2023)
8. Chi, Z., Gu, L., Liu, H., Wang, Z., Wu, Y., Wang, Y., Plataniotis, K.N.: Learning to adapt frozen CLIP for few-shot test-time domain adaptation. In: ICLR (2025)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Farina, M., Mancini, M., Iacca, G., Ricci, E.: Rethinking few-shot adaptation of vision-language models in two stages. In: CVPR (2025)
11. Fifty, C., Duan, D., Junkins, R.G., Amid, E., Leskovec, J., Re, C., Thrun, S.: Context-aware meta-learning. In: ICLR (2024)
12. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
13. Franken, M., van Gemert, J.C.: Automatic egyptian hieroglyph recognition by retrieving images as texts. In: ACM MM (2013)
14. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. IJCV (2024)
15. Garciam, P.: Lego brick sorting image recognition (2019), kaggle
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
17. Hu, Z., Wei, Y., Shen, L., Yuan, C., Tao, D.: Unlocking tuning-free few-shot adaptability in visual foundation models by recycling pre-tuned loras. In: CVPR (2025)
18. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR (2023)
19. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs
20. Kravets, A., Chen, D., Namboodiri, V.P.: Rethinking few shot clip benchmarks: A critical analysis in the inductive setting. In: ICCV (2025)
21. Kukleva, A., Kuehne, H., Schiele, B.: Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In: ICCV (2021)

22. 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. arXiv preprint arXiv:2408.03326 (2024)
23. Li, S., Liu, F., Hao, Z., Wang, X., Li, L., Liu, X., Chen, P., Ma, W.: Logits deconfusion with clip for few-shot learning. In: CVPR (2025)
24. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: ECCV (2024)
25. Liu, F., Cai, W., Huo, J., Zhang, C., Chen, D., Zhou, J.: Making large vision language models to be good few-shot learners. In: AAAI (2025)
26. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
27. 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: ICCV (2025)
28. NVIDIA: NVIDIA H100 Tensor Core GPU Architecture (2022), whitepaper
29. 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. arXiv preprint arXiv:2304.07193 (2023)
30. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)
31. Piosenka, G.: Butterfly and moths image classification 100 species (2023), kaggle
32. 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: ICML (2021)
33. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
34. Saxena, S.: Yoga pose image classification dataset (2021), kaggle
35. Schuerrle, B., Sankarappan, V.: Industrial classification dataset (2023), kaggle
36. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
37. Shvetsova, N., Nagrani, A., Schiele, B., Kuehne, H., Rupprecht, C.: Unbiasing through textual descriptions: Mitigating representation bias in video benchmarks. In: CVPR (2025)
38. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
39. 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)
40. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018)
41. Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J.B., Isola, P.: Rethinking few-shot image classification: a good embedding is all you need? In: ECCV (2020)
42. Udandarao, V., Gupta, A., Albanie, S.: Sus-x: Training-free name-only transfer of vision-language models. In: ICCV (2023)
43. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016)
44. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., et al.: The caltech-ucsd birds-200-2011 dataset. Tech. rep.

45. Yang, C.F., Yin, D., Hu, W., Ji, H., Peng, N., Zhou, B., Chang, K.W.: Verbalized representation learning for interpretable few-shot generalization. In: ICCV (2025)
46. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR (2023)
47. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: CVPR (2023)
48. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
49. Zhang, R., Wei, Z., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: ECCV (2022)
50. Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: ICML (2021)
51. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)
52. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV (2022)
53. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: ICCV (2023)
54. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)