Making Large Vision Language Models to Be Good Few-Shot Learners

Fan Liu[1] *, **Wenwen Cai**[1], **Jian Huo**[1], **Chuanyi Zhang**[1] *, **Delong Chen**[2], **Jun Zhou**[3]

[1]Hohai University, China
[2]Hong Kong University of Science and Technology, China
[3]Griffith University, Australia
fanliu@hhu.edu.cn, caiwenwen@hhu.edu.cn, huojian@hhu.edu.cn, zhangchuanyi@hhu.edu.cn, delong.chen@connect.ust.hk, jun.zhou@griffith.edu.au

Abstract

Few-shot classification (FSC) is a fundamental yet challenging task in computer vision that involves recognizing novel classes from limited data. While previous methods have focused on enhancing visual features or incorporating additional modalities, Large Vision Language Models (LVLMs) offer a promising alternative due to their rich knowledge and strong visual perception. However, LVLMs risk learning specific response formats rather than effectively extracting useful information from support data in FSC. In this paper, we investigate LVLMs' performance in FSC and identify key issues such as insufficient learning and the presence of severe position biases. To tackle above challenges, we adopt the meta-learning strategy to teach models "learn to learn". By constructing a rich set of meta-tasks for instruction fine-tuning, LVLMs enhance the ability to extract information from few-shot support data for classification. Additionally, we further boost LVLM's few-shot learning capabilities through label augmentation (LA) and candidate selection (CS) in the fine-tuning and inference stages, respectively. LA is implemented via a character perturbation strategy to ensure the model focuses on support information. CS leverages attribute descriptions to filter out unreliable candidates and simplify the task. Extensive experiments demonstrate that our approach achieves superior performance on both general and fine-grained datasets. Furthermore, our candidate selection strategy has been proven beneficial for training-free LVLMs.

Code — https://github.com/HUOUO7/MLVLM-FSL

1 Introduction

Few-shot classification (FSC), a specific application of few-shot learning (FSL) (Wang et al. 2020), draws inspiration from human learning capabilities. It enables models to classify even previously unseen classes using minimal labeled data. Typical research focused on training robust visual embedding networks (Vinyals et al. 2016) or leveraging additional attributes (Reed et al. 2016) to mitigate the lack of supervision. Nevertheless, the small amount of data tends to result in unsatisfying generalization.

Recently, Large Vision Language Models (LVLMs) like GPT-4V (Achiam et al. 2023) and Qwen-VL (Bai et al. 2023) integrate powerful language models with advanced visual encoders. They obtain rich perceptual capabilities and comprehensive knowledge via training on extensive multimodal data. They also have in-cotext learning ability that learns from demonstrations and becomes potentially suitable for FSC. However, researchers (Tai et al. 2024) have found that current LVLMs can hardly link unseen image-text pairs and recognize novel categories from support samples. It results from that models tend to focus on specific answer formats instead of grasping the provided information.

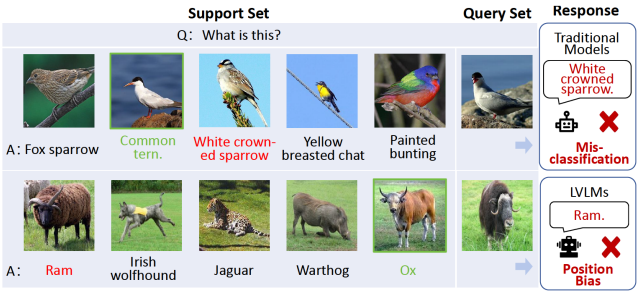

Figure 1: The challenges of FSL. Typical models often suffer from poor generalization, leading to incorrect classifications. Directly applying LVLMs to FSL also encounters position bias that models favor the first option they encounter.

Nevertheless, previous works did not explore the base-to-novel experimental setup in FSC, where inference classes do not overlap with training classes. In this paper, we attempt to evaluate the performance of LVLMs in a base-to-novel few-shot setting. We observed that LVLMs could hardly utilize the information from support samples and achieved suboptimal performance. Through the examination of their outputs, we noticed a position bias: LVLMs tend to favor the first few candidate answers they encounter. Position bias has proven to be a prevalent issue of modern Large Language Models (LLMs) which likely stems from a combination of training data patterns (Hsieh et al. 2024), the left-to-right structure of causal transformers (Zheng et al. 2023b), specific models and tasks (Zheng et al. 2023a). LVLMs may inherit this problem from their core LLM component and cannot effectively leverage support information in the few-shot task. Given the challenges LVLMs face in FSC, we

*corresponding author

adopt a meta-learning (Hospedales et al. 2021) strategy to teach models to effectively learn from support samples. It is a process by which individuals increasingly manage their internalized perception, exploration, learning, and growth habits (Saunders and Wong 2020). Meta-learning has since evolved into the concept of "learn to learn" and become a classic paradigm in FSL.

To this end, we explore the challenges and opportunities of applying LVLM to FSC. Initially, we construct a rich set of instruction-following meta-tasks sourced from diverse domains. This operation enables the model to learn how to extract information from support data. To further improve the few-shot learning capabilities of LVLMs, we design label augmentation and candidate selection methods in the fine-tuning and inference stages, respectively. Specifically, LVLMs' overconfidence in existing knowledge may lead to neglecting support data. Considering the autoregressive token modeling strategy of LVLMs, we adopt a straightforward yet effective character perturbation strategy in class names as label augmentation during fine-tuning. This strategy enhances the model's focus on the candidate classes in the support samples instead of knowledge from pretraining. Moreover, LVLM's strong chain-of-thought (Wei et al. 2022) and image captioning capabilities motivate us to develop an adaptive attribute description generator to provide additional information for candidate selection. Directly entering these descriptions into LVLM may complicate the context and degrade performance. Instead, we use aggregated text similarity scores on these descriptions to select candidates. This approach filters out unreliable options and simplifies the classification task. Finally, we conduct a thorough analysis of the initial suboptimal performance of LVLMs in FSC and why our method effectively addresses these issues and enhances performance.

Our contributions can be summarized as follows:

- We investigate the initial challenges LVLMs face in FSC and propose a meta-learning-based instruction fine-tuning approach. This method enhances the LVLM's ability to learn from support data in few-shot scenarios.
- We develop two strategies to optimize LVLMs across different stages: 1) During instruction fine-tuning, we introduce label augmentation to ensure the model focuses on support data; 2) At inference, we implement an adaptive pipeline that generates and utilizes auxiliary attribute descriptions for candidate selection.
- Our approach achieves state-of-the-art performance on eight FSC benchmarks, demonstrating the feasibility of applying LVLMs to both general and fine-grained image classification tasks. This success highlights the effectiveness of our meta-learning-based fine-tuning and semantic augmentation strategies.

2 Related Work

2.1 Few-Shot Learning

Few-shot learning, which aims to recognize new categories with limited labeled data, can be divided into visual-based methods and semantic-based methods. Visual-based methods focus on extracting category-related features from images for classification. These methods can be generally grouped into three types: optimization-based methods, memory-based methods, and metric-based methods. The first aims to learn a set of initial model parameters that can quickly adapt to new categories (Finn, Abbeel, and Levine 2017; Ravi and Larochelle 2016; Elsken et al. 2020). The second expedites new-task learning via integrating external memory components for storing and using prior-task knowledge (Santoro et al. 2016; Cai et al. 2018; Gidaris and Komodakis 2018). The last aims to learn a metric space where inter-class distances are maximized while intra-class distances are minimized (Snell, Swersky, and Zemel 2017; Sung et al. 2018; Vinyals et al. 2016). Semantic-based methods attempt to enhance visual recognition performance by fusing the complementary information of visual and textual modalities (Chen et al. 2023b; Xu and Le 2022). Xing et al. (2019) proposed an adaptive fusion mechanism to combine visual prototypes with semantic prototypes obtained through word embeddings of class labels. Peng et al. (2019) utilized graph convolutional networks to incorporate additional knowledge from knowledge graphs. Yan et al. (2022) proposed a word embedding-guided attention mechanism to obtain label prototypes for multi-label few-shot learning problems. These methods usually introduce complex network frameworks to effectively utilize textual information. In contrast, our method eliminates the necessity for manual annotation to gather elaborate textual knowledge or devise complicated network architectures. Instead, it takes full advantage of the extensive knowledge within LVLM and its strong image-text alignment to classify new classes.

2.2 LVLM Instruction Tuning

LVLM marks a major leap forward in vision-language modeling, which is designed to process and interpret cross-modal information, capable of proficiently handling complex tasks that require a deep understanding of context. Inspired by the remarkable success of LLM instruction fine-tuning (Ouyang et al. 2022; Wang et al. 2022), the LVLM community has increasingly focused on incorporating instruction-following data to further enhance the model's understanding of downstream tasks. Recently, LLaVA-1.5 (Liu et al. 2024) improved its instructions based on LLaVA's framework and obtained performance on a wider range of VQA tasks. InstructBLIP (Dai et al. 2023) enhanced zero-shot capabilities by performing instruction fine-tuning on multiple datasets. Shikra (Chen et al. 2023a) and Kosmos-2 (Peng et al. 2023) extended LVLM to visual ground truth tasks using instructions with bounding box coordinates. Qwen-VL-Chat (Bai et al. 2023) improved the model's multi-round dialogue interaction capabilities through instruction fine-tuning. Therefore, in this context, we introduce the meta-learning paradigm, focusing on organizing Meta-task instruction-following datasets and fine-tuning LVLM. The purpose of this approach is to fully tap the potential of LVLM in few-shot classification tasks. By leveraging meta-learning, we expect the model to better adapt to downstream tasks.

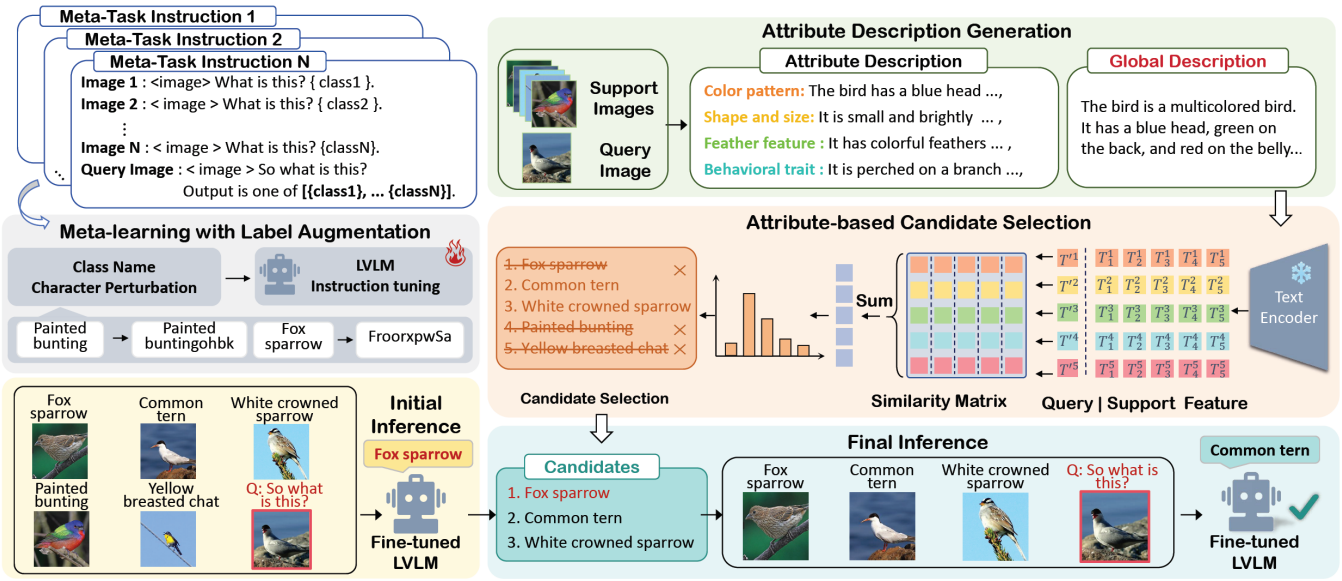

Figure 2: Overview of our approach. We construct meta-task instructions and apply character perturbation as label augmentation during fine-tuning. In the inference phase, we generate attribute descriptions for each image in the instructions through an adaptive framework. Then, we leverage these descriptions to select candidate classes. If the model's initial inference does not match any of the candidates, we reorganize the meta-task instructions and query the model again for the final inference.

3 Method

3.1 Problem Definition

The dataset of FSL is generally divided into two parts: a base set $D_{base} = \{(x, y)|x \in \mathcal{X}_{base}, y \in \mathcal{C}_{base}\}$ for pre-training to initialize the model and a novel set $D_{novel} = \{(x, y)|x \in \mathcal{X}_{novel}, y \in \mathcal{C}_{novel}\}$ for testing, where x denotes the image and y represents the label. The label space for both sets is disjoint, meaning that $C_{base} \cap C_{novel} = \varnothing$. During testing, the support set $S = \{(x_i, y_i)_{i=0}^{N \times K}\}$ is randomly selected from D_{novel}, which includes N classes, each containing K samples. The model must then accurately classify the images in the query set $Q = \{(x_i, y_i)_{i=0}^{N \times M}\}$ into one of the N classes present in the support set S, where M is the number of query samples per class. This classification task is generally referred to as an N-way K-shot task.

3.2 Instruction Tuning With Label Augmentation

To explore the direct application of LVLM on FSL tasks, we first organize the commonly used FSL evaluation datasets into N-way K-shot format. Specifically, we design Meta-Task Instructions to prompt LVLM to generate responses, as shown in Figure 2. However, we find that directly applying LVLM on N-way K-shot setting does not yield satisfactory performance. To enhance LVLM's performance, we adopt an instruction tuning in a meta-learning manner. Specifically, we collect various datasets from areas such as scene recognition, general object recognition, sentiment analysis, fine-grained recognition, and remote sensing. These datasets are then organized into meta-task instructions.

Following the meta-task instruction tuning, LVLM evaluates the similarity between query and support samples and aligns query samples with candidate answers. Consequently, the fine-tuned model can make more accurate predictions from the limited examples provided by the meta-task instructions. However, LVLM can sometimes be overly confident, relying on categories seen in the pre-training data and overlooking the support information during query sample classification. To avoid this problem, we propose a label augmentation (LA) method via a character perturbation strategy to enhance the model's focus on the support data.

Before introducing the character perturbation strategy, we should first take an inside look at the model's output process. The token embeddings W of the LVLM are trained to represent the entire textual space. When given an image embedding X_v, the LVLM identifies the image and outputs the predicted token in the following manner (Yue et al. 2024):

$$P(w|X_v) = argmax\ \sigma(W f(X_v)^T), \tag{1}$$

where σ is the softmax function, f is to transform X_v for aligning with W, and w represents the most probable single token for X_v.

Based on our meta-task instructions, we use interleaved $M(M = N \times K)$ image-text pairs where image features are represented as $X_v = \{X_v^i|i \in [1, M]\}$ and the prompts "What is this? {classname}." are tokenized as $X_p = \{X_p^i|i \in [1, M]\}$. Query image feature is X_v^q, the query prompt "So what is this? Output is one of [candidate class list] " is X_p^q. Then we can derive the complete formal expression of the input:

$$\begin{aligned} X = X_v^1 \oplus [IMG] \oplus X_p^1 \oplus ... \\ \oplus X_v^M \oplus [IMG] \oplus X_p^M \\ \oplus X_v^q \oplus [IMG] \oplus X_p^q, \end{aligned} \tag{2}$$

where $\oplus$ is the concatenation operation and [IMG] is a special token to indicate the boundary. Assuming a class

C has T tokens. Now predicting C is equivalent to auto-regressively predicting its tokens:

$$P(C) = P(w_1, ..., w_T | X_b, X_p) = \prod_{t=1}^{T} P(w_t | w_{<t}, X), \quad (3)$$

where w_t is the t-th token of C, and $w_{<t}$ is the sequence of tokens before the t-th token.

Through the above analysis, to prevent LVLM from becoming overly confident, a straightforward approach is to disrupt common token sequences. For instance, during pre-training, the token sequence 'yel'-'low' for the word 'yellow' is typical. However, by perturbing the original word 'yellow' to 'yelowla' during fine-tuning, it results in a new token sequence 'yel'-'ow'-'la'—which is strange to the LVLM. Therefore, LVLM is enforced to focus on support data instructions to learn the task paradigm. Here is how we implemented the perturbation:

- **Split and combined:** split the class name according to a specific symbol such as '_', and then re-combine the class name after splitting. *e.g.*'A330-300' to '300-A330'.
- **Reversed:** take the last few characters of the class name and place them at the beginning of the other perturbation methods. *e.g.*'elephant' to 'anteleph'.
- **Random insert:** randomly pick a few characters from a to z, and insert them into random positions within the class name. *e.g.*'streetcar' to 'sttrKeeutcEayrU'.
- **Shuffle:** shuffle all characters in the class name. *e.g.*'shrew' to 'hsewr'.

3.3 Attribute Description Generation

In the process of LVLM performing FSC, it implicitly leverages its internalized knowledge. Since LVLM's knowledge is beneficial to FSC, we attempt to explicitly utilize it. Considering that LVLM performs well in image captioning, we let LVLM generate image-related descriptions to assist the model inference process in subsequent sections.

Unlike prior studies that merely utilized class names, generated a single global image-text description, or manually picked relevant attributes, we designed an adaptive attribute description generation framework using LVLM to generate high-quality attributes and global descriptions for images in each category. The detailed steps are as follows:

Step 1: Adaptive Attribute Selection. In this step, the type of dataset to be analyzed (*e.g.*, bird species) and the number of attributes desired (k) are specified. The LVLM then suggests k relevant attributes, along with brief explanations of their importance in describing the images.

Step 2: Automatic Prompt Generation. LVLM is required to produce prompts for each attribute as guidelines to generate descriptions in subsequent steps. It provides concise and tailored prompts for all k attributes, ensuring that the generated descriptions remain focused and informative.

Step 3: Attribute Specific Description Generation. For each of the k attributes identified previously, the LVLM is provided with a corresponding attribute prompt from Step 2. In response, the model generates a specific detailed description of the attribute for the image.

Step 4: Global Attribute Description Generation. Finally, the specific attribute descriptions from Step 3 are combined into a single, comprehensive description sentence and fed to the LVLM. The LVLM responds with its overview of the image. This attribute-global description not only captures the essence of the image but also highlights its unique features at multiple perspectives of detail.

Through our adaptive attribute description generation framework, for each support or query image in the meta-task instruction, we can obtain $k+1$ attribute descriptions regarding each image to assist the subsequent model inference process. See the appendix for more details on the attribute descriptions generation process.

3.4 Attribute-Based Candidate Selection

To leverage the generated attribute descriptions, we initially integrated these descriptions with meta-task instructions and then prompted LVLM. However, this method did not yield better results, as it increased the context length and introduced additional complexity. Instead, we designed a simple yet effective semantic-based method for candidate selection (CS) using these descriptions as illustrated in Figure 2. This approach not only reduces the task complexity but also enhances the self-consistency of LVLM.

For each of the $M+1$ samples in the meta-task instruction, there are $k+1$ attribute descriptions. The $j-th$ description for the $i-th$ sample is denoted as T_j^i, where $i \in [1, M], j \in [1, k+1]$. Additionally, T_j^q represents the $j-th$ description for the query sample.

We compute the text similarity $S_j \in \mathbb{R}^{1\times M}$ for description T_j between the query sample and support samples to obtain $k+1$ text similarity matrices. We then aggregated these similarities to obtain an overall text similarity S_{aggr}:

$$S_{aggr} = S_1 + ... + S_{k+1}. \quad (4)$$

Then we utilize S_{aggr} to identify the top $N//2$ classes as candidate categories C_{can}, while the rest are considered unreliable. We then compare C_{can} with the LVLM's initial inference result A_{ini}. If candidate categories contains the initial inference result ($A_{ini} \in C_{can}$), we consider the result to be validated. Otherwise, we reorganize the $N//2+1$ ways ($N//2$ ways from C_{can} and 1 way from A_{ini}) meta-task instruction to prompt LVLM again for a final inference. Final inference eases classification by narrowing candidates. Also, it boosts output A_{fin} reliability via self-consistency.

4 Experiment

4.1 Implementation Details

Datasets For instruction tuning, we selected 13 datasets from ELEVATER (Li et al. 2022). These datasets span various domains such as remote sensing, scene recognition, stripe recognition, and fine-grained classification. For these datasets, we randomly split the classes into base and novel sets with a 7:3 ratio, using the base sets for fine-tuning. Note that we carefully select datasets to avoid data leakage and make fine-tuning and inference categories have no overlap.

For inference, we evaluate our method on eight established FSL datasets: MiniImageNet (MINI) (Vinyals et al.

Method	MINI	CIFAR	TIERED	CUB
BAVARDAGE (2023)	84.80	87.35	85.20	90.42
EASY 3×ResNet12 (2022)	84.04	87.16	84.29	90.56
PEMnE-BMS* (2022)	85.54	88.44	86.07	94.78
PTMAP-SF-SOT (2022)	85.59	89.94	-	95.80
P>M>F (2022)	95.30	84.30	-	-
CAML (2023)	96.20	70.80	95.40	91.80
Qwen-VL	39.38	37.04	57.14	45.00
Ours	**98.24**	**95.02**	**98.06**	**96.40**

Table 1: Comparison with previous work on three general datasets as well as CUB in the 5-way 1-shot setting.

Method	CUB	Dogs	FGVC	Flowers	Cars
FRN (2021)	83.40	77.53	87.89	81.22	87.63
TDM (2022)	83.25	76.59	87.91	82.31	87.69
MCL-Katz (2023)	85.84	72.07	88.44	76.57	86.12
DeepBDC (2022)	81.85	78.81	85.22	81.07	85.48
LCCRN (2023)	82.80	77.29	88.66	82.86	86.24
SRM (2024)	84.14	77.57	89.14	83.25	88.70
Qwen-VL	45.00	49.34	34.68	47.58	34.82
Ours	**96.40**	**96.68**	**95.64**	**99.58**	**99.72**

Table 2: Comparison with previous work on five fine-grained datasets in the 5-way 1-shot setting.

Meta-Learning	Label Augmentation	Candidate Selection	MINI	CIFAR	TIERED	CUB	Dogs	FGVC	Flowers	Cars	AVG(↑)
✗	✗	✗	39.38	37.04	57.14	45.00	49.34	34.68	47.58	34.82	43.12
✗	✗	✓	72.30	68.22	81.66	59.92	63.52	65.34	70.70	68.78	68.81 (+25.69)
✓	✗	✗	98.20	95.18	97.96	93.60	94.82	93.48	98.32	99.52	96.39 (+53.27)
✓	✗	✓	**98.28**	**95.20**	98.00	94.54	95.40	94.00	98.72	99.56	96.71 (+53.59)
✓	✓	✗	98.24	94.64	**98.06**	96.40	96.40	95.48	99.38	**99.74**	97.29 (+54.17)
✓	✓	✓	98.24	95.02	**98.06**	**96.64**	**96.68**	**95.64**	**99.58**	**99.74**	**97.45 (+54.33)**

Table 3: Contribution of each component on eight datasets under the 5-way 1-shot setting. Meta-Learning refers to instruction tuning in a meta-learning manner. Label Augmentation means adding character perturbation during tuning. Candidate Selection means selecting more reliable candidates and re-evaluating in the inference.

2016), CIFAR-FS (CIFAR) (Bertinetto et al. 2019), Tiered-ImageNet (TIERED) (Triantafillou et al. 2018), CUB (Wah et al. 2011), Stanford Dogs (Dogs) (Khosla et al. 2011), FGVC-Aircraft (FGVC) (Maji et al. 2013), Oxford 102 Flower (Flowers) (Nilsback and Zisserman 2008), and Stanford Cars (Cars) (Krause et al. 2013). Typical FSC methods are tested on the first three datasets as well as CUB, while fine-grained FSC methods are evaluated on the latter five datasets. We also test our methods on ISEKAI (Tai et al. 2024), a fully synthetic dataset generated with Midjourney.

Architecture and Training Details We utilized the interactive Qwen-VL-Chat model as our LVLM. Its large language model was initialized with the pre-trained weights of Qwen-7B (Bai et al. 2023), the visual encoder adopted the ViT architecture and was initialized with the pre-trained weights of Openclip's ViT-bigG (Cherti et al. 2023), and the visual-language adapter consisted of a single-layer cross-attention module with random initialization.

To improve training efficiency and reduce training costs, we chose the quantized version Qwen-VL-Chat-Int4 (Qwen-VL for simplicity), froze the LLM and visual encoder, and used Q-LoRA (Dettmers et al. 2024) to fine-tune the model's adapter. Specifically, the learning process utilized a cosine learning rate scheduler with a base learning rate of 1×10^{-5} and a warm-up ratio of 0.01. Optimization was performed using the Adam optimizer, with a weight decay of 0.1 and a β_2 parameter set to 0.95, which ensured stability in convergence. The maximum sequence length of the model was set to 2048 tokens to effectively handle long sequences. Additionally, we directly used the frozen SBERT (all-MiniLM-L6-v2) (Reimers and Gurevych 2019) as the text encoder used in the semantic aided inference step to measure the similarity between sentences, which had been trained on a 1B sentence pair data set and could effectively capture the semantic information of sentence vectors.

Evaluation Protocol Due to LVLMs' tendency to generate lengthy content and complex class names, we employ three metrics for flexible and comprehensive evaluation:

- Acc: Applying regular expression filters to remove non-alphanumeric symbols from both the candidate class names C_{can} and LVLM outputs A_{LVLM}, then performing strict matching.
- Acc_{occur}: After filtering, if the gold appears anywhere in A_{LVLM}, it is considered correct.
- Acc_{clip}: Leveraging CLIP-L14 to map A_{LVLM} to C_{can} and then matching the result against gold.

Since models that are not fine-tuned often produce unreliable outputs, we use Acc_{clip} as the default metric. In contrast, fine-tuned models provide more stable results, for which we use Acc as the metric. The results for all three evaluation protocols will be detailed in the appendix.

4.2 Comparison With the State-of-the-Art

To evaluate the effectiveness of our approach, we conduct experiments on eight datasets in a 5-way 1-shot setting. Table 1 compares our results with state-of-the-art (SOTA) methods for general FSC, while Table 2 contrasts our approach with methods specialized for fine-grained FSC.

As shown in Table 1, our method exceeds existing SOTA methods with gains of 2.02%, 5.08%, 2.66%, and 0.60% on the MINI, CIFAR, TIERED, and CUB datasets, respectively. Moreover, our method achieves an average accuracy of 96.93% across these four datasets, surpassing the highest average accuracy of 90.44% achieved by PTMAP-SF-SOT by 6.49%. It also outperforms the vision transformer-based methods P>M>F and CAML, which have average accuracies of 89.80% and 88.55%, respectively. As shown

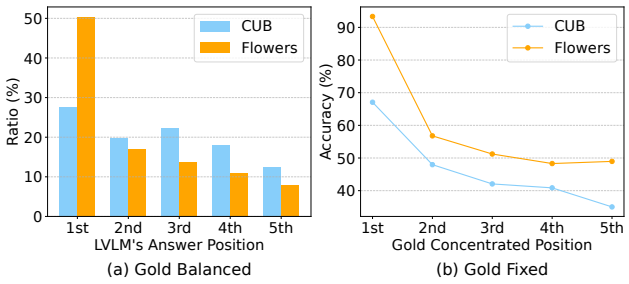

Figure 3: Position bias on CUB and Flowers under the 5-way setting. (a) The distribution of answers output by LVLM when gold answers are evenly distributed. (b) The accuracy of LVLM when gold answers are fixed in a specific position.

Method	MINI	CIFAR	TIERED	CUB
Baseline	20.81	20.62	20.17	20.70
Meta-Learning	90.31	89.79	84.98	56.16

Table 4: Evaluation results of Flamingo.

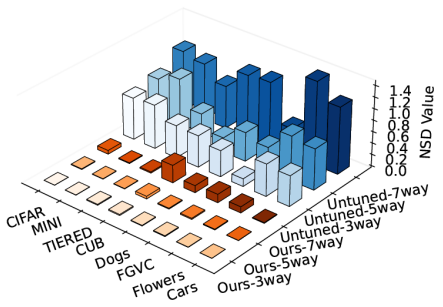

Figure 4: Comparison of answer position distributions: Our Method vs. Untuned LVLM. NSD values indicate the normalized standard deviation between the model's actual output positions and uniformly sampled gold positions.

	Qwen-VL	Meta-Learning	Meta-Learning w/LA
ISEKAI	33.38	87.92	95.18

Table 5: Evaluation results on the ISEKAI dataset.

in Table 2, in the fine-grained domain, our method improves upon SOTA methods with gains of 10.56%, 17.87%, 6.50%, 16.33% and 11.02% on the CUB, Dogs, FGVC, Flowers and Cars datasets, respectively. Our method reaches an average accuracy of 97.60% across these five fine-grained datasets, significantly surpassing the highest average accuracy of 84.56% achieved by the SRM method by 13.04%.

It is evident that Qwen-VL performs poorly on both tasks. We will provide a detailed analysis of this phenomenon in the Analysis Studies section. Nevertheless, our approach can improve LVLM to achieve SOTA performance across diverse downstream datasets without requiring additional adjustments on the base set. This advantage indicates the superiority of applying LVLM in FSC tasks.

4.3 Analysis Studies

What Is Each Component's Contribution? To validate the effectiveness of our proposed methods, we conducted ablation studies on meta-learning, label augmentation (LA), and candidate selection (CS), as detailed in Table 3. Across eight datasets, we observed the following average improvements: 1) Meta-learning fine-tuning alone led to a substantial 53.27% improvement in model performance; 2) Adding LA, CS, and both LA and CS together resulted in performance gains of 54.17%, 53.59%, and 54.33%, respectively; 3) Incorporating CS provided significant benefits to training-free methods, achieving 25.69% accuracy gains. These results demonstrate that each component contributes to enhancing the model's few-shot classification capabilities.

Since the meta-learning contributes the most to the improvements, we implemented it on another VLM flamingo-4B (Awadalla et al. 2023) and reported the result in Table 4. We attained an average gain of 59.74% across four datasets. This result indicates that meta-learning fine-tuning can benefit different VLM models.

Why Does LVLM Initially Perform Badly? We observed severe position bias in Qwen-VL across all eight datasets. Specific illustrations for CUB and Flowers can be seen in Figure 3. We compare Gold Balanced and Gold Fixed settings for detailed investigation. In Gold Balanced, gold answers are uniformly spread among 5-way candidates (e.g. 1000 times per position in 5000 tests). But LVLM's outputs cluster in early candidate positions(e.g. over half of LVLM's responses favored the first position on the Flowers dataset). In Gold Fixed, gold answers are in the same position among all candidates. For example, Gold Concentrated 3rd Position means all gold answers are in the third. When gold answers are in the first, model accuracy is higher; as gold answers move later, accuracy drops, showing the model struggles with farther candidates. This phenomenon suggests the reason for LVLMs' poor FSC performance: their inability to effectively utilize information from support samples.

Does Our Approach Mitigate Position Bias? We extend our experiments from the 5-way setting to the 3-way and 7-way settings, where the gold distribution is balanced. We calculate the normalized standard deviation (NSD) of the model's actual answer positions compared to the balanced gold distribution. A higher NSD indicates a worse position bias problem where model and gold answers have greater position differences.

As illustrated in Figure 4, the untuned LVLM manifests a severe position bias even in the 3-way setting, and this bias exacerbates as the way count rises. In contrast, our model maintains balanced output distributions across 3-way, 5-way, and 7-way settings, without difficulty in accessing answers at the end of the candidate list. Notably, although our instruction-following dataset is organized for 5-way 1-shot tasks, the fine-tuned model performs well in the 7-way setting, where the candidate list is longer than that in pre-

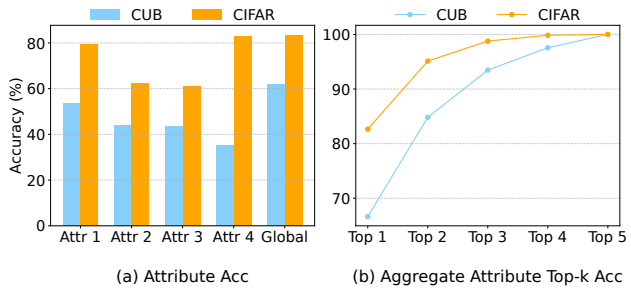

Figure 5: (a) Accuracy of each individual attribute description and a global description. (b) Top-k accuracy of aggregated attribute description from top1 to top5.

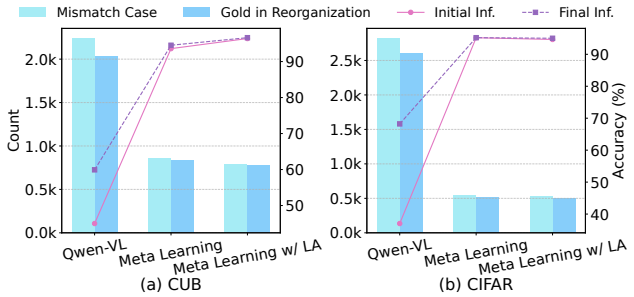

Figure 6: Candidate Selection process. Mismatch Case means the number of times the LVLM output is not among the top-2 attribute-based candidates. Gold in Reorganization is the count of gold in reorganized candidates for mismatch cases. Initial and Final Inference refer to the LVLM results before and after applying CS, respectively.

training. These results demonstrate that our method effectively mitigates the position bias problem.

How and Why Does Our Approach Mitigate Position Bias? In meta-learning fine-tuning, the ground truth positions in candidates are uniformly distributed, enabling models to handle each position more fairly. This process significantly reduces NSD (e.g. from 0.9724 to 0.0051 on the MINI dataset), highlighting it as a key component in mitigating position bias. Furthermore, label augmentation (LA) encourages the model to focus more on the support information, especially for unseen categories (e.g. ISEKAI in Table 5). It assists in enhancing the effectiveness of meta-learning fine-tuning. Since LVLMs face challenges in processing long sequences and capturing distant references, having more candidates can exacerbate the position bias problem, as illustrated in Figure 4. Regarding this problem, candidate selection (CS) simplifies the task by filtering out unreliable candidate classes. With fewer candidates in the final inference, there is a lower risk of encountering position bias.

How Does the Model Perform on Completely Novel Classes? Most of the datasets used for testing are composed of natural images that are widely accessible on the internet, making it challenging to ensure that the novel sets are entirely unseen by the pre-trained models. To investigate this issue, we also assess our approach on the ISEKAI dataset, which consists of virtual images generated by Midjourney. The concepts in ISEKAI are fantastical, sourcing from legends, myths, or fictional media, resulting in minimal prior exposure for LVLMs. Additionally, it introduces real-world but distinct categories, such as "octopus vacuum" to increase classification difficulty. As demonstrated in Table 5, our method significantly enhances the performance of LVLMs on this highly challenging task. Moreover, our method further achieved a 7.26% performance gain owing to LA, underscoring the value of using support information.

Why Does Aided Semantic Strategy Work? For each attribute, we calculate the similarity between support and query samples, taking the maximum as the classification outcome. The results for CUB and CIFAR are presented in Figure 5 (a). Evidently, a single attribute fails to offer adequate intra-class similarity and inter-class discriminability for effective LVLM classification. The performance on CUB is poorer than on CIFAR, likely because the finer granularity of CUB demands more distinctive attribute descriptions.

As shown in Figure 5 (b), the accuracy of aggregated attributes surpasses that of any single attribute at the top-1 level and shows significant improvement at the top-2 level. At the top-3 level, the accuracy consistently exceeds 90%. According to the experimental results, we utilize aggregated attributes to select $N//2$ candidates for the final reference. It is less likely to omit the gold answer and can effectively simplify the classification task in the final reference.

We also illustrate the CS process in Figure 6. We compare LVLM's output with the top-2 attribute-based candidates. If the LVLM's output is not among these top-2 candidates, it is considered a mismatch, and we reorganize the candidates. We count how often the gold label appears in these reorganized candidates, this phenomenon reflects the potential classification performance. The results show that the gold label appears in up to 90% of the new candidates. We also report the accuracy of the initial inference and the accuracy after applying CS. The experiments demonstrate the effectiveness of our candidate selection method, which helps retain the correct answer and simplifies the task for better performance, especially benefiting the unfine-tuned Qwen-VL.

5 Conclusion

In this paper, we explored the challenges and opportunities of applying LVLM to FSC. We found untuned LVLMs could not effectively learn from support samples as well as suffered from the position bias problem. To enhance the few-shot learning ability of LVLMs, we organized a meta-learning-based few-shot classification instruction-following dataset. We designed label augmentation to force the model to focus more on support information in the instruction tuning. For the inference phase, we developed an attribute-based candidate selection strategy to simplify task complexity. Our method achieves SOTA performance on both general and fine-grained FSC benchmarks. In the future, we will explore more applications of LVLM in FSC.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62372155 and Grant 62302149, in part by the Fundamental Research Funds for the Central Universities under Grant B240201077, in part by the Aeronautical Science Fund under Grant 2022Z071108001, in part by the Qinglan Project of Jiangsu Province, and in part by Changzhou Science and Technology Bureau under Project 20231313.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bendou, Y.; Hu, Y.; Lafargue, R.; Lioi, G.; Pasdeloup, B.; Pateux, S.; and Gripon, V. 2022. Easy—ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components. *Journal of Imaging*, 8(7): 179.

Bertinetto, L.; Henriques, J.; Torr, P.; and Vedaldi, A. 2019. Meta-learning with differentiable closed-form solvers. In *ICLR*.

Cai, Q.; Pan, Y.; Yao, T.; Yan, C.; and Mei, T. 2018. Memory matching networks for one-shot image recognition. In *CVPR*, 4080–4088.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Chen, W.; Si, C.; Zhang, Z.; Wang, L.; Wang, Z.; and Tan, T. 2023b. Semantic prompt for few-shot image recognition. In *CVPR*, 23581–23591.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2818–2829.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B. A.; Fung, P.; and Hoi, S. C. H. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *ArXiv*, abs/2305.06500.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2024. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 36: 10088–10115.

Elsken, T.; Staffler, B.; Metzen, J. H.; and Hutter, F. 2020. Meta-learning of neural architectures for few-shot learning. In *CVPR*, 12365–12375.

Fifty, C.; Duan, D.; Junkins, R. G.; Amid, E.; Leskovec, J.; Ré, C.; and Thrun, S. 2023. Context-aware meta-learning. *arXiv preprint arXiv:2310.10971*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. PMLR.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *CVPR*, 4367–4375.

Hospedales, T.; Antoniou, A.; Micaelli, P.; and Storkey, A. 2021. Meta-learning in neural networks: A survey. *TPAMI*, 44(9): 5149–5169.

Hsieh, C.-Y.; Chuang, Y.-S.; Li, C.-L.; Wang, Z.; Le, L.; Kumar, A.; Glass, J.; Ratner, A.; Lee, C.-Y.; Krishna, R.; and Pfister, T. 2024. Found in the middle: Calibrating Positional Attention Bias Improves Long Context Utilization. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *ACL*, 14982–14995. Bangkok, Thailand: Association for Computational Linguistics.

Hu, S. X.; Li, D.; Stühmer, J.; Kim, M.; and Hospedales, T. M. 2022. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *CVPR*, 9068–9077.

Hu, Y.; Pateux, S.; and Gripon, V. 2022. Squeezing backbone feature distributions to the max for efficient few-shot learning. *Algorithms*, 15(5): 147.

Hu, Y.; Pateux, S.; and Gripon, V. 2023. Adaptive dimension reduction and variational inference for transductive few-shot classification. In *AIST*, 5899–5917. PMLR.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR*, volume 2.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV*, 554–561.

Lee, S.; Moon, W.; and Heo, J.-P. 2022. Task discrepancy maximization for fine-grained few-shot classification. In *CVPR*, 5331–5340.

Li, C.; Liu, H.; Li, L.; Zhang, P.; Aneja, J.; Yang, J.; Jin, P.; Hu, H.; Liu, Z.; Lee, Y. J.; et al. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *NeurIPS*, 35: 9287–9301.

Li, X.; Li, Z.; Xie, J.; Yang, X.; Xue, J.-H.; and Ma, Z. 2024. Self-reconstruction network for fine-grained few-shot classification. *Pattern Recognition*, 153: 110485.

Li, X.; Song, Q.; Wu, J.; Zhu, R.; Ma, Z.; and Xue, J.-H. 2023. Locally-enriched cross-reconstruction for few-shot fine-grained image classification. *TCSVT*, 33(12): 7530–7540.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *NeurIPS*, 36.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *ICVGIP*, 722–729. IEEE.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35: 27730–27744.

Peng, Z.; Li, Z.; Zhang, J.; Li, Y.; Qi, G.-J.; and Tang, J. 2019. Few-shot image recognition with knowledge transfer. In *ICCV*, 441–449.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Ravi, S.; and Larochelle, H. 2016. Optimization as a model for few-shot learning. In *ICLR*.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 49–58.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *ICML*, 1842–1850. PMLR.

Saunders, L.; and Wong, M. A. 2020. Learning theories: Understanding how people learn. *Instruction in Libraries and Information Centers*.

Shalam, D.; and Korman, S. 2022. The self-optimal-transport feature transform. *arXiv preprint arXiv:2204.03065*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *NeurIPS*, 30.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*, 1199–1208.

Tai, Y.; Fan, W.; Zhang, Z.; and Liu, Z. 2024. Link-Context Learning for Multimodal LLMs. In *CVPR*, 27176–27185.

Triantafillou, E.; Larochelle, H.; Snell, J.; Tenenbaum, J.; Swersky, K. J.; Ren, M.; Zemel, R.; and Ravi, S. 2018. Meta-learning for semi-supervised few-shot classification. In *ICLR*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *NeurIPS*, 29.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*.

Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Ashok, A.; Dhanasekaran, A. S.; Naik, A.; Stap, D.; et al. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM*, 53(3): 1–34.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-shot classification with feature map reconstruction networks. In *CVPR*, 8012–8021.

Xie, J.; Long, F.; Lv, J.; Wang, Q.; and Li, P. 2022. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *CVPR*, 7972–7981.

Xing, C.; Rostamzadeh, N.; Oreshkin, B.; and O Pinheiro, P. O. 2019. Adaptive cross-modal few-shot learning. *NeurIPS*, 32.

Xu, J.; and Le, H. 2022. Generating representative samples for few-shot classification. In *CVPR*, 9003–9013.

Yan, K.; Zhang, C.; Hou, J.; Wang, P.; Bouraoui, Z.; Jameel, S.; and Schockaert, S. 2022. Inferring prototypes for multi-label few-shot image classification with word vector guided attention. In *AAAI*, volume 36, 2991–2999.

Yue, K.; Chen, B.-C.; Geiping, J.; Li, H.; Goldstein, T.; and Lim, S.-N. 2024. Object Recognition as Next Token Prediction. In *CVPR*, 16645–16656.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2023a. Large language models are not robust multiple choice selectors. In *ICLR*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36: 46595–46623.