Enhancing Fine-Grained Image Classifications via Cascaded Vision Language Models

Anonymous ACL submission

Abstract

Fine-grained image classification, particularly in zero/few-shot scenarios, presents a significant challenge for vision-language models (VLMs), such as CLIP. These models often struggle with the nuanced task of distinguishing between semantically similar classes due to limitations in their pre-trained recipe, which lacks supervision signals for fine-grained categorization. This paper introduces CascadeVLM, an innovative framework that overcomes the constraints of previous CLIP-based methods by effectively leveraging the granular knowledge encapsulated within large vision-language models (LVLMs). Experiments across various fine-grained image datasets demonstrate that CascadeVLM significantly outperforms existing models, specifically on the Stanford Cars dataset, achieving an impressive 85.6% zero-shot accuracy. Performance gain analysis validates that LVLMs produce more accurate predictions for challenging images that CLIPs are uncertain about, bringing the overall accuracy boost. Our framework sheds light on a holistic integration of VLMs and LVLMs for effective and efficient fine-grained image classification.

1 Introduction

The vision-language model (VLM) landscape is evolving, highlighted by models like CLIP (Radford et al., 2021) and its significant zero/few-shot classification capabilities (Zhou et al., 2022b). Despite these advances, fine-grained image classification remains challenging, particularly in distinguishing closely related sub-classes (Ren et al., 2023a), such as the *watercress* and the *wallflower* in Figure 1, due to the subtle visual differences between classes. To enhance the zero/few-shot fine-grained classification performance of CLIP, recent methods have concentrated on advanced prompt engineering (Zhou et al., 2021, 2022a) and the enhancement of pre-training supervision (Li et al., 2023b; Singh et al., 2023). Additionally,

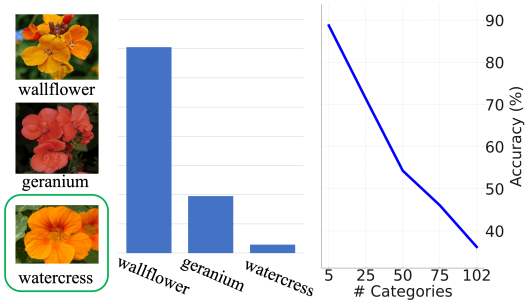

Figure 1: Illustration of model performance: CLIP's misclassification of watercress (left) and the inverse relationship between LVLM accuracy and the number of categories (right).

to enrich CLIP's prompt context, Menon and Vondrick (2022) introduced a technique whereby GPT3 generates detailed class descriptions for enhancing CLIP models. However, this method cannot handle visually similar classes as the generated descriptions tend to be nearly identical, offering limited benefits for fine-grained classification tasks.

In this paper, we instead resort to large vision language models (LVLMs), to better capitalize on their extensive world knowledge. A straightforward way is to directly concatenate all the candidate classes as a prompt for LVLMs. However, the limited capacity for long-context modeling (Zhao et al., 2023) poses challenges, particularly evident when LVLMs grapple with a large candidate image class set. As shown in Figure 1, an advanced open-sourced LVLM, Qwen-VL (Bai et al., 2023), still suffers from a dramatic accuracy decline when the candidate categories increased from 5 to 102. To tackle this, we employ CLIP models for locating potential candidate classes to ease the burden for LVLMs. Due to the contrastive pre-training objective, CLIP is capable of finding a set of possible classes, evidenced by the relatively higher top-K prediction accuracy. For example, in the Flowers102 dataset, CLIP(ViT-B/32) gives a 68.7% Top-1 accuracy, yet the Top-10 accuracy is significantly

boosted to 89.9%. This characteristic validates our motivation to holistically integrate CLIP and LVLM.

Motivated by our previous exploration, in this paper, we introduce CascadeVLM, which integrates the complementary capabilities of CLIP-like models and LVLMs to perform fine-grained image classification. The key idea is to leverage the CLIP-like models as a class filter for LVLMs to fulfill the LVLM potentials. Besides, the results can be further enhanced by leveraging the in-context learning (Dong et al., 2022) of LVLMs to perform few-shot learning. Moreover, the overall inference efficiency can be improved by adopting an entropy threshold as a heuristic mechanism to evaluate the necessity of deploying LVLMs, achieving a dynamic early exiting (Xin et al., 2020; Li et al., 2021b). We conducted zero/few-shot experiments across various fine-grained image datasets, and the evaluation results demonstrate that CascadeVLM surpasses other individual models in performance. For instance, CascadeVLM achieved an 85.6% accuracy rate on the Stanford Cars dataset. Additionally, we utilized two advanced pretrained CLIP-like models as backbones for the iNaturalist and SUN397 datasets, achieving superior results compared to the individual backbone models. Furthermore, our analysis meticulously dissects CascadeVLM's performance improvements, spotlighting the model's sophisticated handling of entropy thresholds to balance computational demands with accuracy. This dual-focused investigation not only clarifies the operational dynamics of CascadeVLM but also its adeptness in integrating with CLIP-like architectures and LVLM for refined final predictions, underlining the model's comprehensive adaptability and strategic efficiency.

In summary, the contributions of this paper could be summarized as follows: (1) We propose the CascadeVLM framework which effectively combines CLIP-like VLMs and LVLMs for zero/few-shot fine-grained image classification. (2) Experiment on six relevant datasets reveals that CascadeVLM achieves performance on par with advanced zero/few-shot methods. Our analysis provides insights into the holistic integration of VLMs and LVLMs

2 Methodology

In this section, we delineate the methodology underpinning our CascadeVLM framework, which is structured into two steps. (1) The first step involves candidate selection facilitated by the CLIP model. (2) The second step encompasses the application of zero-shot or few-shot prediction techniques using large vision-language models (LVLMs). For zero-shot prediction, LVLMs directly engage in classification based on these pre-selected candidates. For few-shot situations, we use the in-context learning strategy to augment the semantic context.

2.1 CLIP-based Candidate Selection

As a pivotal component of our CascadeVLM framework, CLIP's operational mechanism (Radford et al., 2021) allows it to effectively discern potential correct classes, making it an ideal choice for the initial phase of candidate filtering from an extensive array of class labels.

Specifically, in our approach, the function $f_{\text{CLIP}}(x, c_i)$ denotes the score outputted by the CLIP model for a specific category c_i when given an image x. CLIP's core functionality lies in its ability to align image and text representations within a unified embedding space, and gives scores for each pairs. Upon acquiring raw scores from CLIP for each category in the label set C, we employ a softmax function to transform these scores into a probability distribution, as delineated by:

$$P(c_i \mid x) = \frac{\exp(f_{\text{CLIP}}(x, c_i))}{\sum_{c_j \in C} \exp(f_{\text{CLIP}}(x, c_j))}, \quad (1)$$

The resulting probabilities could reflect the relative confidence of the CLIP model in associating the given image with each category within the context of the entire set C.

Based on the probability computation, denoted as $P(c_i \mid x)$, specified in Equation 1, we extract the top-k categories from C to ensure that they are sorted in descending order of probability. This selection and sorting process, crucial for the framework's efficacy, is denoted as a function $s_{(topk)}$. Not only does this step condense the pool of candidate classes, but it also addresses the sensitivity of LVLMs to the sequence in which these categories are presented. Our empirical results 3.2 affirm that simple sorting based on probability significantly bolsters the predictive precision of LVLMs. The generalized representation of this procedure is as follows:

$$C^* = \{c'_1, c'_2, \ldots, c'_k\} = s_{\text{topk}}(P(c_i \mid x), C), \quad (2)$$

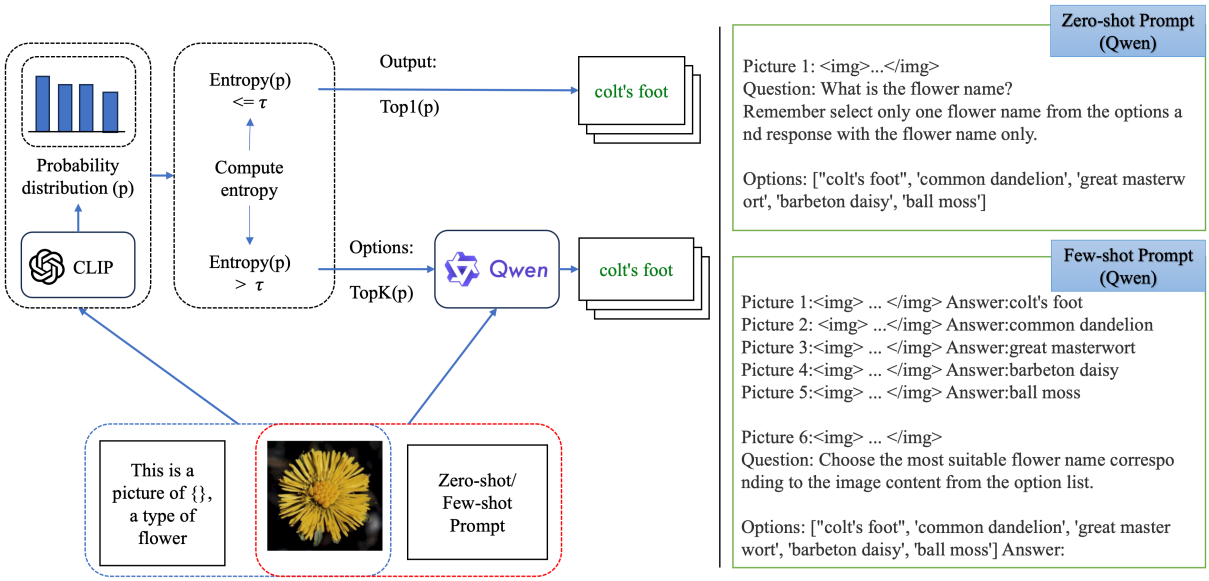

Figure 2: CascadeVLM commences with CLIP for initial image analysis and probabilistic categorization, integrating an entropy threshold, τ, to balance efficiency and accuracy, culminating in LVLM's adaptive classification.

where C^* encapsulates the optimally sorted candidates, with c'_1, c'_2 through to c'_k representing the elements in descending order of their computed probabilities.

2.2 LVLMs Prediction with Candidate Set

In this section, we seek to leverage Large Vision-Language Models (LVLMs) for the final classification within our CascadeVLM framework. Capitalizing on a subset of candidates pre-selected by CLIP, LVLMs overcome the challenge of extensive context and improve prediction accuracy through adaptable zero-shot and few-shot learning strategies, tailored to the data-sparse environments.

Zero-Shot Prediction Zero-shot learning (Socher et al., 2013) enables models to predict unseen classes without specific training examples, leveraging pre-existing knowledge from broader contexts or related tasks. This method is particularly beneficial in data-scarce scenarios, where it effectively infers new categories despite limited training data.

In the context of our CascadeVLM framework, zero-shot prediction is executed after identifying the top-k candidate classes using CLIP. The LVLM then selects one candidate $c*$, as the final prediction. Here, we generalize the process of LVLM prediction as function $f_{(LVLM)}$, given the input image x and the top-k candidate set C^*:

$$c^* = f_{\text{LVLM}}(x, C^*) \tag{3}$$

The zero-shot prediction phase in our CascadeVLM framework highlights LVLMs' proficiency in utilizing pre-trained knowledge for unseen data while adeptly managing contextual complexities. Focusing on a select set of candidates, the proposed methods coude effectively address the intricacies of fine-grained classification, ensuring precise and dependable outcomes even without class-specific examples.

Few-Shot Prediction In the Few-Shot Prediction phase of our CascadeVLM framework, we capitalizes on LVLMs' in-context learning (Brown et al., 2020) ability, where additional relevant samples significantly enhance performance, allowing LVLMs to deepen their understanding and improve predictive accuracy.

In the integration of few-shot learning within our cascade framework, we undertake a two-step process for candidate categories set C^*:

Step 1: Context Generation: In this initial phase, for each category c'_i in C^*, we randomly select an example image $x_{c'_i}$ from the training dataset, and design a prompt to contextualize the input image x for the LVLMs. Here, each candidate class c'_i and its corresponding example image $x_{c'_i}$ are integrated with the prompt template, creating a contextual framework for the LVLMs. We denote this assemblage of prompts and images form the contextual basis as E in the subsequent step of our methodology. For instance, within the context of the GPT4-V scenario, the contextual basis denoted

as E is formulated in Table 1.

```
<IMG: x_{c'_1}>
Question: What is the class of the
image? Answer: c'_1
<IMG: x_{c'_2}>
Question: What is the class of the
image? Answer: c'_2
⋮
```

Table 1: Few-shot prompt used in our experiments.

Step 2 - Prediction with Contextual Information: In this step, the context E is integrated with the input image x and fed into the LVLMs. This integration enables the LVLMs to utilize the rich contextual information embedded in E to enhance and refine its predictive process for the image x. Consequently, the final classification outcome, denoted as c^*, emerges from this enriched inferential framework. The process can be mathematically represented as:

$$c^* = f_{\text{LVLM}}(x, C^*, E) \tag{4}$$

where f_{LVLM} represents the LVLM prediction based on provided image x, the top-k candidate set C^* and the context set E.

2.3 Speed-up via Adaptive Entropy Threshold

In our CascadeVLM framework, we introduce an adaptive entropy-based approach aimed at enhancing inference speed, reducing the computational load on LVLMs, and accelerating overall throughput. The entropy $H(x)$ of the probability distribution, a measure of uncertainty or predictability within the distribution, is calculated as follows:

$$H(x) = -\sum_{c_i \in C} P(c_i \mid x) \log P(c_i \mid x) \tag{5}$$

This computation serves as a critical decision point in our methodology. If the calculated entropy $H(x)$ falls below a predefined threshold, it signifies a high confidence level in the top-1 category as determined by CLIP. In such cases, we expedite the process by directly outputting this top-1 category, thereby bypassing the need for further LVLM processing. Conversely, if $H(x)$ exceeds the threshold, indicating a lower level of confidence and greater uncertainty, we proceed to the subsequent steps involving LVLMs for refined classification. This adaptive mechanism effectively balances speed and accuracy, streamlining the framework while ensuring reliable classification outcomes.

Dataset	# of Class	# of Test
Flowers102	102	818
StanfordCars	196	8,041
FGVC Aircraft	100	3,333
BirdSnap	500	2,444
SUN397	397	19,850
iNat18 (iNaturalist 2018)	8,142	24,426

Table 2: Statistics of the evaluated fine-grained image classification benchmarks.

3 Experiments

In this section, we rigorously evaluate the performance of our CascadeVLM framework across diverse benchmarks. Initially, we detail the experimental setup in Section 3.1, followed by an in-depth analysis of the framework's efficacy in zero-shot learning scenarios in Section 3.2, and subsequently in few-shot learning contexts in Section 3.3.

3.1 Experimental Settings

Models For experimental evaluation, we employed various CLIP models in combination with specific Large Vision-Language Models (LVLMs). The experiments utilized one of the CLIP variants CLIP ViT-B/32, ViT-B/16, or ViT-L/14 alongside either Qwen-VL-Chat (Bai et al., 2023) or GPT-4V (OpenAI, 2023) as the LVLM. Additionally, we explore the framework's adaptability by integrating it with two robust CLIP-like models, MAWS-CLIP (Singh et al., 2023) and CLIP (ViT-G/14) pretrained with OpenCLIP (Ilharco et al., 2021) using the Laion2B (Schuhmann et al., 2022) dataset, as backbones cascading with GPT-4V.

Datasets We utilize a collection of datasets each offering unique characteristics and significance for fine-grained image classification, as summarized in Table 2. These datasets include Flowers102 (Nilsback and Zisserman, 2008), StanfordCars (Krause et al., 2013), FGVC Aircraft (Maji et al., 2013), BirdSnap (Berg et al., 2014), SUN39 (Xiao et al., 2010), and iNaturalist(2018) (Horn et al., 2018), collectively encompassing a wide range of categories.

Model	Flower102	StanfordCars	FGVC Aricraft	BirdSnap	Avg.
Supervised	99.8 (2021)	96.3 (2021)	95.4 (2022)	90.1 (2020)	-
Qwen (full classes)	37.5	22.4	8.4	2.3	17.6
CLIP ViT-B/32	68.7	59.6	19.1	51.7	49.8
CLIP ViT-B/32 Cascade (Qwen, full classes)	72.7	74.3	22.7	20.3	47.5
CLIP ViT-B/32 Cascade (Qwen, top k)	**74.2**	**79.2**	**27.1**	**56.7**	**59.3**
CoOp ViT-B/16 (Zhou et al., 2021)	68.7	64.5	18.5	-	-
CoCoOp ViT-B/16 (Zhou et al., 2022a)	71.9	65.3	22.9	-	-
POMP ViT-B/16 (Ren et al., 2023b)	72.4	66.8	25.6	-	-
CLIP ViT-B/16	73.0	64.4	24.5	52.5	53.6
CLIP ViT-B/16 Cascade (Qwen, full classes)	70.5	74.1	26.2	20.2	47.8
CLIP ViT-B/16 Cascade (Qwen, top k)	**73.3**	**79.1**	**30.7**	**56.6**	**60.0**
FLIP ViT-L/14 (Li et al., 2023b)	75.0	**90.7**	29.1	**63.0**	64.5
CLIP ViT-L/14	**81.3**	76.2	30.9	62.2	62.7
CLIP ViT-L/14 Cascade (Qwen, full classes)	75.8	78.9	30.1	21.0	51.5
CLIP ViT-L/14 Cascade (Qwen, top k)	78.2	85.6	**37.0**	**63.0**	**66.0**

Table 3: Zero-shot results comparison with different CLIP models as the backbone. The k is selected based on the validation set. Our CascadeVLM achieves the best overall performance on four benchmarks. The best scores are shown in **bold**.

Model	iNat18	SUN397
MAWS-CLIP	20	71.0
MAWS-CLIP Cascade	**26.8**	**75.4**
OpenCLIP (ViT-G/14)	6.6	74.4
OpenCLIP (ViT-G/14) Cascade	**11.8**	**77.6**

Table 4: Zero-shot prediction result comparison with MAWS and OpenCLIP ViT-G/14 as backbone, cascading GPT-4V.

Baselines In our experimental evaluation, we benchmark the performance of various CLIP variants both individually and within our cascade framework. This comparative analysis allows us to demonstrate the enhancement in classification accuracy achieved by integrating these models into the CascadeVLM framework. Moreover, we extend our comparison to include diverse methodologies, providing a comprehensive evaluation of our proposed approach.

3.2 Zero-shot Learning Results

Table 3 showcases the zero-shot prediction capabilities of CascadeVLM, demonstrating its superior performance across several benchmarks. Remarkably, without necessitating any training, CascadeVLM outperforms established methods such as CoOp, CoCoOp, and POMP. Moreover, it is comparable to and often outperforms the more strongly supervised method FLIP, which leverages fine-grained supervision for training. Given the flexibility of CascadeVLM, there is potential for further accuracy enhancements by employing FLIP as the underlying architecture. Consequently, we propose an experimental framework that integrates two advanced pre trained CLIP-like models as the backbone to explore this potential.

Table 4 presents the results of experiments that applied two advanced CLIP-like models, MAWS CLIP and OpenCLIP (ViT-G/14), cascading to GPT-4V to challenging datasets: iNaturalist and SUN397, with random samples of 500 from each due to budget and rate limit constraints. The imbalanced nature of iNaturalist typically poses challenges for CLIP models. However, the findings reveal significant performance enhancements when utilizing a cascade framework. Specifically, MAWS CLIP achieved improvements of 6.8% on iNaturalist and 3.6% on SUN397. Similarly, OpenCLIP's performance increased by 5.2% on iNaturalist and 3.2% on SUN397. These improvements highlight the cascade framework's adaptability and effectiveness in enhancing the capabilities of CLIP-like models.

3.3 Few-shot Learning Results

In our initial explorations, we assessed Qwen-VL's capacity for few-shot learning within fine-grained image classification domains. However, it became apparent that Qwen-VL struggled to optimally utilize in-context demonstrations and instructions in this setting. Consequently, we turned our focus to GPT-4V, anticipating its better alignment with our framework's requirements.

Our experiments with GPT-4V were limited to

Model	Flower102	StanfordCars	FGVC Aricraft	BirdSnap	Avg.
CLIP(ViT-L/14)	82.0	75.0	30.0	60.5	61.9
GPT4-V (full classes)	67.5	74.0	61.5	46.0	62.3
CLIP(ViT-L/14) + GPT4-V (k=full classes)	82.0	82.5	**64.5**	55.5	71.1
CLIP(ViT-L/14) + GPT4-V (k=5)	86.5	85.5	56.0	62.0	72.5
CLIP(ViT-L/14) + GPT4-V (k=5) + 1-shot	**94.5**	**88.5**	63.0	**72.5**	**79.7**

Table 5: Few-shot learning results with GPT-4V as the LVLM. GPT-4V can better utilize the in-context demonstrations to achieve superior results for fine-grained classification. The result of CasecadeVLM is superior overall datasets.

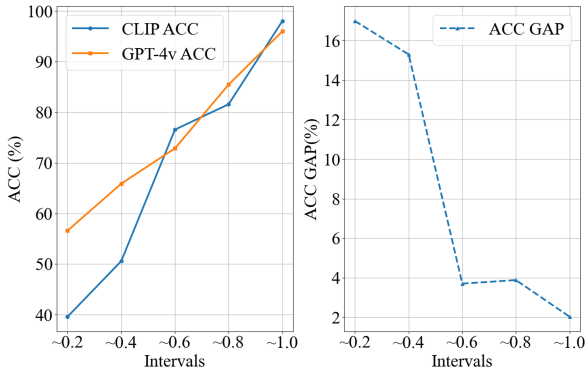

Figure 3: Comparative Analysis of ACC performance between CLIP and GPT-4V across different intervals of classification certainty. The left graph shows the ACC of both models across varying levels of margin. The right graph presents the ACC gap between the two models.

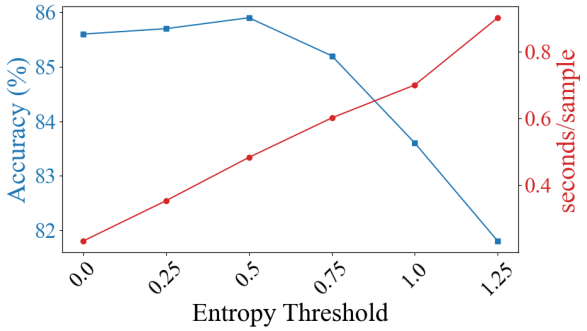

Figure 4: Performance variation in the StanfordCars dataset with varying entropy thresholds using CLIP-ViT-L/14 for cascading, set at top-k=10. An increase in entropy threshold results in decreased inference speed and reduced accuracy.

a random subset of 200 samples per dataset due to budget and rate limit constraints. Here we utilize top-k (k=5) for few-shot experiments and yield even more pronounced improvements in predictive accuracy. For instance, with few-shot learning applied, the Flower102 dataset achieved an impressive 94.5% accuracy, while the StanfordCars dataset attained 88.5%. These results not only reaffirm the effectiveness of our cascade framework but also highlight its adaptability and efficiency in leveraging few-shot learning for fine-grained classification tasks.

4 Analysis

In this section, we explore various aspects of CascadeVLM, highlighting the underlying reasons for its enhanced performance, the trade-off of the entropy threshold, and more. More investigations can be found in the Appendix B.

4.1 Performance Gain Analysis

This analysis seeks to demonstrate why cascading CLIP with an LVLM model leads to improved accuracy in classification tasks. Using the Flowers102 dataset as a case study, we assess the performance of CLIP and the enhancement brought by LVLM. The margin (Settles, 2009), i.e., the difference between the top1 and top2 probability scores from CLIP, serves as an indicator of the model's certainty about its prediction, where smaller margins suggest greater ambiguity in the image classification.

We divide the range of margins into five intervals, from 0 to 1, to analyze the effects systematically. The data reveals that GPT-4V, representing LVLM, significantly outperforms CLIP with margins less than 0.4, where CLIP experiences confusion. This is evident in the consistently high ACC for GPT-4V in these instances. When the margin exceeds 0.6, the ACC for CLIP improves, indicating that the model is more confident and accurate in its predictions, thus reducing the gap in performance between CLIP and LVLM. The accompanying Figure 3 illustrates this trend, with the ACC gap decreasing sharply as the margin increases. This pattern suggests that while LVLM provides a significant advantage in cases of high ambiguity, the benefit tapers off as CLIP's confidence in its classifications rises.

Q: What is the bird name in picture?
Choose one from the options and explain your selection briefly.
Options:
['titmouse', 'parus major', 'go away bird', 'european goldfinch', 'black-capped chickadee']

A: The bird in the picture is the 'Parus major', commonly known as the Great Tit. I can determine this because the Great Tit has distinctive coloring with a black head and neck, white cheeks, a yellow underpart with a black stripe down the center, and greenish back and wings, which match the bird in the picture.
The other options do not match the coloring and markings:
•A 'titmouse' generally refers to birds in the family Paridae in North America, which have a different appearance, typically with a plain head without the black and white pattern.
•The 'go away bird' is a type of turaco, which is completely different in color and has a crest on its head.
•The 'European goldfinch' has a red face and a very different pattern of colors, including a bright yellow wing bar.
•The 'black-capped chickadee' has a solid black cap and lacks the yellow coloring on the underparts.
Therefore, the correct identification for the bird in the picture is 'Parus major'.

Figure 5: CLIP's prediction is incorrect, while LVLM corrects the answer and makes the reasoning.

4.2 Sensitivity of Option Orders

As delineated in Tables 3 and 5, we find, surprisingly, that the arrangement of options provided by CLIP plays a pivotal role in the efficacy of Language-Vision Language Models (LVLMs). While it may seem a minor detail to supply LVLMs with the entire class set from CLIP, this procedure is significantly impactful. For example, as shown in Table 3, presenting all classes in a random order to Qwen results in an average accuracy of only 17.6%. Conversely, when the classes are organized according to the probabilities assigned by CLIP, there is a notable enhancement in performance. Specifically, in the case where CLIP(ViT-L/14) cascades with Qwen, offering a fully ordered class set, there is a substantial accuracy increase of 33.9% across various datasets. This improvement is consistently observed with GPT-4V as well, where organizing the classes based on CLIP's guidance boosts accuracy by 8.8% across the datasets.

4.3 Inference Efficiency Analysis

This subsection critically evaluates the efficacy of implementing an entropy threshold within the CascadeVLM framework. Functioning as a heuristic determinant, this threshold crucially dictates the juncture at which processing shifts from CLIP's initial evaluation to the computationally demanding LVLM analysis. This strategic integration plays a pivotal role in augmenting inference speed, adeptly balancing expeditious processing with the need for in-depth LVLM processing. Our experiments, conducted in a 1GPU (V100) environment, are illustrated in Figure 4. Results indicate a direct correlation between an increase in entropy threshold and heightened inference speed, albeit at the cost of reduced accuracy.

4.4 Towards Explainable Predictions

Understanding the rationale behind the predictions in image classification provides valuable insights for a robust decision (Chen et al., 2019a). Employing the cascade framework, we design an explanation prompt that inquires why the LVLM prioritizes one option over others. For instance, as depicted in Figure 5, CLIP initially predicts an incorrect answer. The LVLM not only corrects this but also explains its prediction, allowing us to dissect the LVLM's reasoning ability. This enables us to differentiate between the terms *titmouse* which is CLIP's prediction, and *parus major*, which is the LVLM's corrected choice. This case highlights the great potential of our CascadeVLM beyond producing accurate classification results.

4.5 Case Study

Our case study analysis presents an examination of three distinct scenarios encountered during experimentation with CLIP ViT-L/14 and Qwen as the LVLM in a $k = 5$ setting.

Case 1 illustrates a scenario where CLIP's top-1 prediction is incorrect; however, the ground truth is present within its top-5 predictions. Leveraging the LVLM's discernment, the accurate class—green-winged dove—is selected.

Case 2 depicts a situation where, despite CLIP's inclusion of the correct answer—striped owl—in its top-5 predictions, the LVLM fails to identify it correctly. This instance highlights potential areas for refinement within the LVLM's decision-making process.

Case 3 demonstrates a complete misalignment where both CLIP and LVLM fail to recognize the correct class within the top-5 predictions, leading to a compounded error in the final outcome.

These cases underscore the nuanced complexities of fine-grained image classification and reaffirm the necessity for integrated approaches like CascadeVLM to capitalize on the strengths of both CLIP and LVLMs. They also provide valuable insights into the decision-making dynamics of the models, offering pathways for future enhancements.

5 Related Work

Vision Language Models Building vision language models (VLMs) for understanding the multimodal world has been an active research area. Pilot studies leverage pre-training concepts from

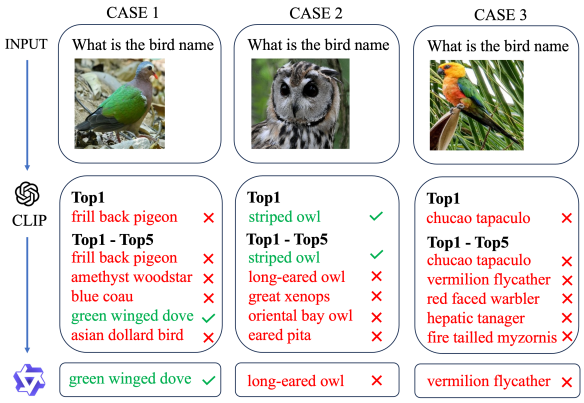

Figure 6: Three case studies demonstrating the cascade process from CLIP predictions to LVLM refinement for bird species classification.

NLP (Devlin et al., 2019), learning shared representations across modalities from mixed visual and language inputs (Li et al., 2019; Tan and Bansal, 2019; Su et al., 2020; Chen et al., 2019b; Li et al., 2020). Among these, Radford et al. (2021) introduced CLIP, a contrastive language-image pre-training framework that employs language as supervision, demonstrating potential for multi-modal tasks and inspiring subsequent variants for improvement (Jia et al., 2021; Li et al., 2022b, 2023b, 2021a, 2022a). The evolution of large language models like ChatGPT (OpenAI, 2022) has motivated the development of large vision language models (LVLMs), combining powerful vision encoders like CLIP with large language models such as LLaMa (Touvron et al., 2023) and Vicuna (Chiang et al., 2023). Achieved through large-scale modality alignment training on image-text pairs (Alayrac et al., 2022; Awadalla et al., 2023) and supervised fine-tuning on multi-modal instruction tuning datasets (Liu et al., 2023; Li et al., 2023a), resulting LVLMs like GPT-4V (OpenAI, 2023) and Qwen-VL (Bai et al., 2023) exhibit promising perceptual and cognitive abilities (Yang et al., 2023) for engaging user queries. This paper identifies limitations in CLIP and LVLMs for fine-grained image recognition and proposes the CascadeVLM framework to effectively enhance prediction accuracy by harnessing the advantages of both models.

Fine-grained Image Classification Fine-grained image recognition, involving categorization into subordinate classes within a broader category, such as cars (Krause et al., 2013) and aircraft models (Maji et al., 2013), demands fine-grained feature learning. Previous work explores diverse strategies, including local-global interaction modules with attention mechanisms (Fu et al., 2017; Zheng et al., 2017), end-to-end feature encoding with specialized training objectives (Dubey et al., 2018; Chang et al., 2020), and the incorporation of external knowledge bases or auxiliary datasets (Chen et al., 2018; Xu et al., 2018). These approaches offer potential enhancements similar to our CLIP model, which we identify as a future exploration for improved performance.

CLIP Enhancements for Fine-grained Image Classification Recent studies have enhanced the CLIP primarily via prompt engineering and pre-training techniques. In prompt engineering, CoOp (Zhou et al., 2021) introduces an innovative method by learning context words as continuous vectors. Extending this idea, CoCoOp (Zhou et al., 2022a) incorporates a lightweight neural network to generate input-specific tokens for images, further improving model performance. POMP (Li et al., 2023b) proposes pre-training a general soft prompt on the ImageNet-21K dataset for universal visual tasks. Besides, Menon and Vondrick (2022) instead employs GPT-4 to generate better descriptive prompts for classification, enriching the prompting context for CLIP models. In the realm of pre-training, MAWS (Singh et al., 2023) combines Masked Autoencoder (MAE) pre-training with weakly supervised learning, significantly enhancing the learning efficacy. Similarly, FLIP (Li et al., 2023b) increases prediction accuracy by masking substantial portions of image patches, facilitating the processing of more image-text pairs within the same timeframe and boosting performance across various tasks. Different from previous studies, our CascadeVLM explores the integration of LVLMs for leveraging their world knowledge to effectively handle similar classes.

6 Conclusion

In this paper, we propose CascadeVLM, harnessing the advantages of CLIP and LVLMs for fine-grained image classification. By utilizing CLIP for selecting the potential candidate class, LVLM can make more accurate predictions for image classes with subtle differences. Experimental results on four benchmarks demonstrate the effectiveness of our proposed framework. Further extension to the few-shot setups showcases the great potential of the cascading framework to leverage the in-context learning ability of LVLMs.

Limitations

The efficacy of our CascadeVLM framework hinges critically on the symbiotic interplay between the CLIP model and LVLMs. A key limitation emerges when CLIP's top-K accuracy is insufficient, failing to encompass correct options in LVLM's narrowed candidate set, thereby limiting the scope for enhanced accuracy. Moreover, if CLIP outperforms the LVLM in fine-grained classification, incorporating an LVLM with relatively inferior capabilities may inadvertently diminish overall accuracy. These dynamics underscore the imperative for meticulous selection and alignment of models, ensuring each component's strengths are effectively leveraged within the cascade architecture.

The CascadeVLM framework mainly utilizes the LVLM's extensive familiarity and common knowledge with the dataset. If lack of such knowledge, the accuracy would not be good enough. But we believe the integration of Agent or RAG approches (Lewis et al., 2020) may offer a solution to this limitation. Besides, augmenting the context for LVLM with additional information may also be a fact to be explored. Besides the few-shot learning we applied in experiments, we also tested sending the CLIP's prediction scores to LVLM. Unfortunately, this led to a reduction in accuracy, with a nearly 4% increase in erroneous predictions, attributed to LVLM's over-reliance on CLIP's scores. This highlights the need for a balanced approach to information feeding within the cascade framework.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, abs/2204.14198.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv preprint*, abs/2308.01390.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv preprint*, abs/2308.12966.

Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. 2022. Sr-gnn: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Transactions on Image Processing*, 31:6017–6031.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. 2020. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29:4683–4695.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019a. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo. 2018. Knowledge-embedded representation learning for fine-grained image recognition. In *International Joint Conference on Artificial Intelligence*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019b. Uniter: Learning universal image-text representations. *ArXiv*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey for in-context learning.

Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. 2018. Maximum-entropy fine-grained classification. *ArXiv*, abs/1809.05934.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware minimization for efficiently improving generalization. *CoRR*, abs/2010.01412.

Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484.

Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. 2021. Escaping the big data paradigm with compact transformers. *CoRR*, abs/2104.05704.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *4th IEEE Workshop on 3D Representation and Recognition, at ICCV 2013 (3dRR-13)*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022a. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.

Lei Li, Yankai Lin, Deli Chen, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021b. Cascadebert: Accelerating inference of pre-trained language models via calibrated complete models cascade. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 475–486.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023a. M^3IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv preprint*, abs/2306.04387.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. of ECCV*.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023b. Scaling language-image pre-training via masking.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4v(ision) system card.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Shuhuai Ren, Lei Li, Xuancheng Ren, Guangxiang Zhao, and Xu Sun. 2023a. Delving into the openness of CLIP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9587–9606, Toronto, Canada. Association for Computational Linguistics.

Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, and Xu Sun. 2023b. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition.

Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Burr Settles. 2009. Active learning literature survey.

Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, and Ishan Misra. 2023. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *ICCV*.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. Deebert: Dynamic early exiting for accelerating bert inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251.

Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. 2018. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1100–1113.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5219–5227.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *CoRR*, abs/2109.01134.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Appendix

A Prompt Tuning of Qwen

A.1 Zero-shot Prompt Tunning of Qwen

In our experiments, we experimented with various prompt designs to optimize the performance of Qwen in selecting the top-k categories. Two representative prompt styles were identified, each with distinct characteristics and performance implications.

The first prompt style, while intuitive, occasionally led to non-compliant responses where Qwen would select a flower name not listed in the given options, or use an alias instead of the specified name. This approach yielded suboptimal results.

Subsequently, we adapted our prompts to align more closely with the training data of Qwen, where the use of the keyword "options" was prevalent. This adaptation significantly improved compliance and accuracy in the model's responses. Thus for the overall experiment, we use 'PROMPT2'. And for GPT-4V, we applied a similar prompt style but followed the API requirement.

PROMPT 1:

```
Picture 1: <img>....jpg</img>
Please examine the flower image
    ↪ and identify the most
    ↪ suitable flower name
    ↪ corresponding to the image
    ↪ content from the list of
    ↪ flower names below.
    ↪ Remember select only one
    ↪ flower name from the list,
    ↪ and response with the
    ↪ flower name ONLY. Available
    ↪  flower names: [...]
```

PROMPT 2:

```
Picture 1: <img>...jpg</img>
Question: What is the flower name
    ↪ ? Remember select only one
    ↪ flower name from the
    ↪ options and response with
    ↪ the flower name only.
    ↪ Options: [...]
```

A.2 Few-Shot Prompt Tunning of Qwen

In the domain of few-shot learning, we conducted experiments with Qwen-VL and observed challenges in its ability to effectively utilize in-context demonstrations and follow instructions. Our experimentation involved different prompt structures in the context of the CLIP-ViT B/32 model with a top-$k = 10$ setting on the Flower102 dataset.

The initial two prompts led to moderate success, achieving an accuracy of approximately 50%. However, the implementation of the final prompt design demonstrated a notable improvement, yielding an accuracy close to 68%. This highlights the impact of prompt design on the model's ability to leverage few-shot learning effectively.

To corroborate the versatility of our Cascade-VLM framework, we conducted few-shot learning experiments with GPT-4V. These trials demonstrated the framework's adaptability across different LVLMs, reinforcing its effectiveness in diverse data-rich scenarios.

PROMPT 1:

```
<img>...jpg</img> Question: What
    ↪ is the flower name? Options
    ↪ : [...] Answer: ...
<img>...jpg</img> Question: What
    ↪ is the flower name? Options
    ↪ : [...] Answer: ...
<img>...jpg</img> Question: What
    ↪ is the flower name? Options
    ↪ : [...] Answer: ...
...
<img>...jpg</img> Question: What
    ↪ is the flower name? Answer:
    ↪  ...
```

PROMPT 2:

```
Picture 1: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Options: [...]
    ↪  Answer: ...
Picture 2: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Options: [...]
    ↪  Answer: ...
Picture 3: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Options: [...]
    ↪  Answer: ...
...
Picture 4: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Options: [...]
    ↪  Answer:
```

PROMPT 3:

```
Picture 1: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Answer: ...
Picture 2: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Answer: ...
Picture 3: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Answer: ...
...
Picture 4: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Options: [...]
    ↪  Answer:
```

PROMPT 4:

```
Picture 1: <img>...jpg</img>
    ↪ Answer: ...
Picture 2: <img>...jpg</img>
    ↪ Answer: ...
Picture 3: <img>...jpg</img>
    ↪ Answer: ...
...
Picture 4: <img>...jpg</img>
    ↪ Question: What is the
    ↪ flower name? Options: [...]
    ↪  Answer:
```

B Analysis

B.1 Influence of candidate classes number k

In the analysis of the influence of the number of candidate classes, k, on classification performance, two distinct configurations of the CLIP model, namely CLIP-ViT-B/16 and CLIP-ViT-L/14, as well as the integration of CLIP ViT-B/32 with Qwen in a cascade framework, have been explored. The investigation reveals dataset-specific optimal settings for k. Specifically, for the StanfordCars and FGVC Aircraft datasets, peak performance is observed at a top-10 setting across different configurations, with an interesting shift to top-5 for the FGVC Aircraft dataset when using the ViT-L/14 model, highlighting an enhancement in baseline performance. In contrast, the Flower102 and BirdSnap datasets exhibit optimal results at a top-3 setting, with the Flower102 dataset showing a superior accuracy with the CLIP-ViT-L/14 model, attributed to its intrinsic fine-grained image classification capability. This suggests that the CLIP-ViT-L/14 model's performance surpasses that of the Qwen LVLM in specific cases. Furthermore, the validation performance across different datasets demonstrates a dependency on the chosen value of k, indicating a nuanced behavior where the intrinsic properties of each dataset may favor a different range of candidate classes. This behavior underscores the importance of tailoring the cascade framework's parameters to the specific dataset at hand to achieve optimal performance, as evidenced by the gradual improvement in accuracy with a narrower focus in candidate classes for certain datasets, and a discernible peak before a decline in others, suggesting a balance between too few and too many options is crucial for maximizing classification accuracy.

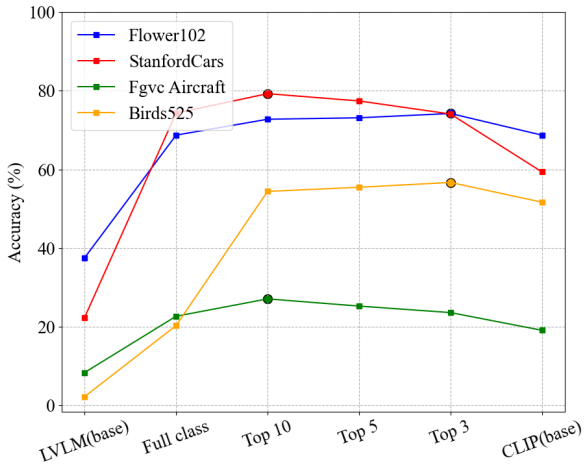

Figure 7: Performance changes with varied k with CLIP-ViT-B/32.

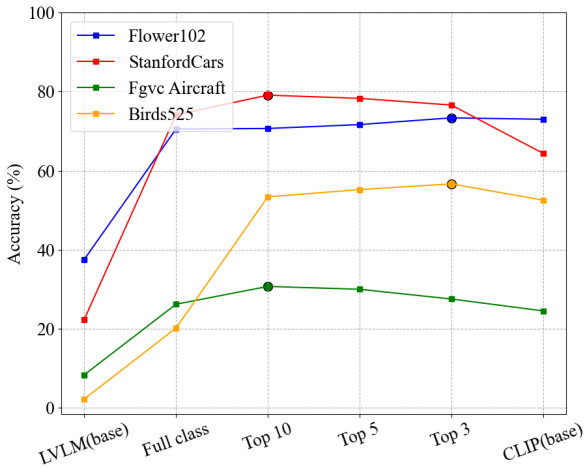

Figure 8: Performance changes with varied k with CLIP-ViT-B/16.

B.2 Error Analysis

An error analysis was conducted on the BirdSnap dataset using the cascade framework, which incorporates CLIP (ViT-L/14) for initial classification

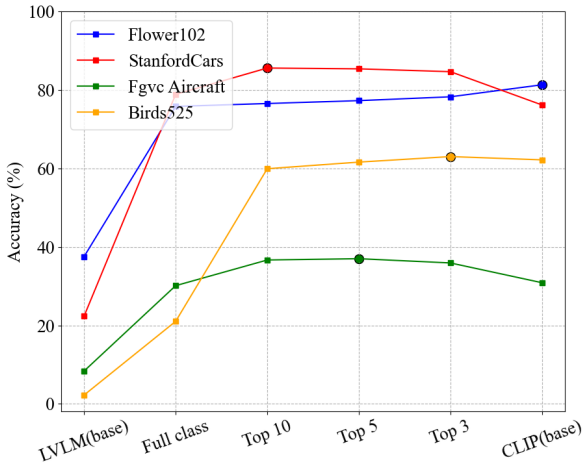

Figure 9: Performance changes with varied k with CLIP-ViT-L/14.

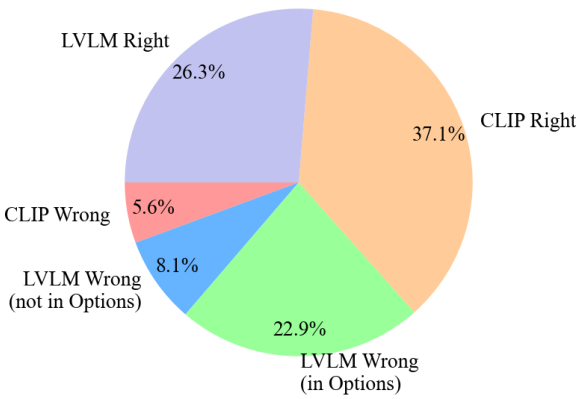

Figure 10: Error analysis of the BirdSnap dataset with an entropy threshold of 1.25 and top-k=10. The analysis reveals that despite CLIP including correct options, LVLM frequently misclassifies.

and Qwen as the LVLM for refined categorization with $k = 10$, as shown in Figure 10. When entropy is lower than the threshold, prediction is only processed by CLIP, in this case, 148 misclassifications were noted (*CLIP WRONG*). Otherwise, after the CLIP narrows down the options of classes, the LVLM Qwen would do the final classification. In this case, LVLM resulted in 812 misclassifications (*LVLM Wrong*), which further breaks down into two categories: 212 instances where the correct option was not present in the top-10 candidates given by CLIP(*LVLM Wrong not in Options*), and 600 instances where the correct option was present, but the LVLM failed to identify it (*LVLM Wrong in Options*).