Enhancing Fine-Grained Image Classifications via Cascaded Vision Language Models

Anonymous ACL submission

Abstract

Zero-shot fine-grained image classification poses significant challenges for vision language models (VLMs), primarily due to the subtle distinctions among closely related classes. This paper introduces CascadeVLM, a cascading framework that seamlessly integrates CLIP with large vision language models (LVLMs), harnessing the strengths of both models in addressing fine-grained image classification. Our methodology involves two primary steps. Initially, CLIP is employed to identify potential class candidates based on prediction confidence. Then, LVLMs are adopted for zero/few-shot prediction, focusing on these candidate classes. Empirical evaluations on four fine-grained image classification benchmarks demonstrate CascadeVLM's superior performance compared to individual models. For example, on the StanfordCars dataset, CascadeVLM achieves an impressive 85.6% zero-shot accuracy. Further efficiency analysis uncovers a trade-off between inference speed and prediction accuracy, and error analysis indicates that failed samples primarily stem from LVLMs' prediction errors, even when provided with the correct candidate class options.

1 Introduction

In the dynamic landscape of vision-language models (VLMs), models such as CLIP (Radford et al., 2021) have demonstrated impressive capabilities in broad image classification tasks (Zhou et al., 2022). However, their efficacy diminishes in fine-grained image classification, where the need to distinguish between highly similar subclasses poses a formidable challenge (Ren et al., 2023). The left of Figure 1 illustrates the perplexing classification decisions made by the CLIP model when presented with flower images exhibiting subtle visual nuances. A potential solution lies in turning to large vision-language models (LVLMs) such as GPT-4V (OpenAI, 2023), endowed with the ability to harness vast world knowledge within their extensive language model backbone (Petroni et al., 2019; Dai et al., 2022) for the task. 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 is often the case in fine-grained tasks such as classifying a flower from 100 candidate categories. The right part of Figure 1, illustrates this struggle, as the performant open-sourced LVLM, Qwen-VL (Bai et al., 2023), experiences a dramatic accuracy decline when the candidate categories increase from 5 to 102. These inherent challenges within VLMs and LVLMs prompt a critical research question: Can we effectively harness the strengths of both paradigms to address these limitations?

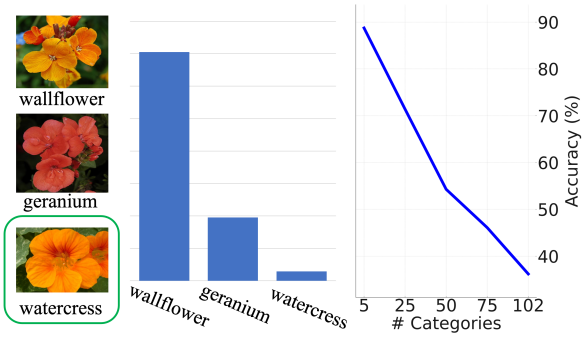

Figure 1: Illustration of model performance: CLIP's misclassification of watercress (left) and the inverse relationship between LVLM accuracy and the number of categories (right).

In this paper, we propose a cascade framework that integrates the complementary capabilities of CLIP and LVLMs to perform fine-grained image classification. The key idea is to leverage the CLIP as a class filter for LVLMs to fulfill the LVLM potentials, elaborated in the following two steps. *Step 1*: The CLIP model performs zero-shot classification on the input image. Based on the output class distribution, we narrow the candidate image

labels to a manageable subset based on the model's prediction confidence. *Step 2*: The LVLMs are responsible for the final prediction within this narrowed label set. It can be performed in a zero-shot manner by asking the LVLMs to classify the image into the class based on the potential classes. Besides, the results can be further enhanced by leveraging the in-context learning (Dong et al., 2022) of LVLMs. We randomly select one image from the training set and construct a demonstration to better inform LVLMs of each class's visual characteristics. The overall inference efficiency can be further improved by adopting a heuristic mechanism to evaluate the necessity of deploying LVLMs. Inspired by the idea of dynamic early exiting (Xin et al., 2020; Schwartz et al., 2020; Li et al., 2021b) which allocate adaptive computation for samples with different complexity, we use entropy threshold functions as a heuristic mechanism to evaluate the necessity of deploying LVLMs. Samples with highly confident CLIP predictions can skip step 2 and the CLIP results are adopted as the final prediction. This approach reduces the computational cost by invoking LVLMs only in scenarios where the CLIP predictions are confusing.

We evaluate the proposed CascadeVLM framework on four fine-grained image classification datasets, achieving superior results over individual models. Notably, in the StanfordCars dataset, CascadeVLM achieved an 85.6% accuracy rate, significantly surpassing the baselines of 76.2% for CLIP (ViT-L/14). In few-shot scenarios, this performance enhancement is consistently replicated across datasets leveraging GPT-4V as the LVLM. Our approach yields a 94.5% accuracy in the Flower102 dataset and 88.5% in the StanfordCars dataset, using CLIP (ViT-L/14) and GPT-4V with a 1-shot demonstration for each class. Further analysis uncovers an inherent accuracy-computation trade-off by varying the threshold. Additionally, an in-depth error analysis exposes the bottlenecks of CascadeVLM, primarily stemming from inaccuracies in candidate options provided by CLIP and misclassifications by LVLM, even when the correct label is present in the candidate set.

Our study makes a two-fold contribution: (1) We present a cascade framework for fine-grained image classification, effectively leveraging the strengths of VLMs and LVLMs. (2) The proposed CascadeVLM framework achieves superior results across diverse benchmarks, and our analysis provides insights for future 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. This phase focuses on narrowing down the potential candidate categories for a given input image, leveraging the robust classification capabilities of CLIP. (2) The second step encompasses the application of zero-shot or few-shot prediction techniques using large vision-language models (LVLMs). In this stage, candidates initially filtered by CLIP undergo further analysis. For zero-shot prediction, LVLMs directly engage in classification based on these preselected candidates. In scenarios requiring few-shot learning, additional images corresponding to each filtered candidate are procured to augment the semantic context, thereby enhancing the learning process and predictive accuracy.

2.1 CLIP-based Candidate Selection

As a pivotal component of our CascadeVLM framework, the CLIP model serves a crucial role in identifying probable class candidates. CLIP's operational mechanism 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.

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, thus facilitating the assessment of an image's compatibility with various textual descriptors or category labels. 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 thus reflect the relative confidence of the CLIP model in associating the given image with each category within the context of the entire set C. For a comprehensive exploration of CLIP's underlying mechanism, we refer readers to the original CLIP paper (Radford et al., 2021)

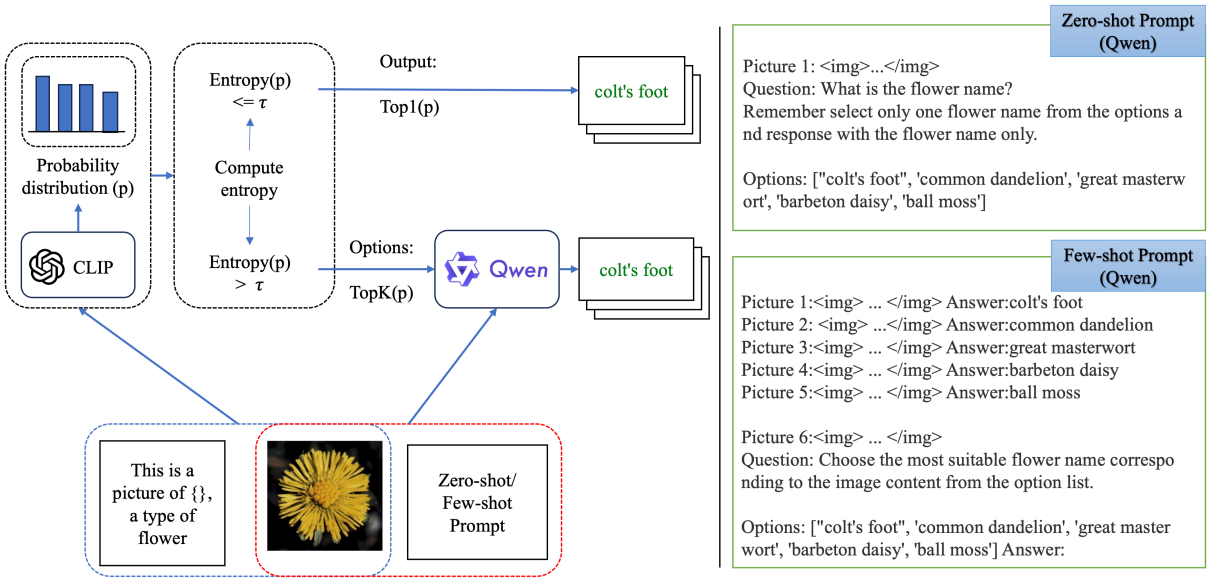

Figure 2: CascadeVLM commences with CLIP for initial image analysis and probabilistic categorization, integrating an entropy threshold, τ, to balance efficiency and accuracy, culminating in LVLM's adaptive classification.

Based on the probability computation, denoted as $P(c_i \mid x)$, specified in Equation (1), we extract the top-k categories from C ensuring 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:

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

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.

In conclusion, the integration of CLIP in our CascadeVLM framework efficiently streamlines the initial selection of class candidates, setting a solid foundation for the subsequent detailed classification process. This step not only highlights the synergy between advanced vision-language technologies but also prepares the ground for the next phase of our methodology, where LVLMs leverage this refined input for precise classification.

2.2 LVLMs Prediction on Reduced Candidate Set

In this subsection, we examine the utilization of Large Vision-Language Models (LVLMs) for the final classification within our CascadeVLM framework. Capitalizing on a subset of candidates preselected 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-rich or 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^*:

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

This function capitalizes on the model's inherent understanding and the context provided by the reduced candidate set.

Thus, 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, our method effectively addresses 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 leverage Large Vision-Language Models (LVLMs) in data-rich scenarios. This approach 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 meticulously 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 manually 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, effectively creating a contextual framework for the LVLMs. This assemblage of prompts and images forms the contextual basis, which we succinctly denote 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 as follows:

<IMG: $x_{c_1'}$>
Question: What is the class of the image? Answer: c_1'
<IMG: $x_{c_1'}$>
Question: What is the class of the image? Answer: c_1'
⋮
<IMG: $x_{c_1'}$>
Question: What is the class of the image? Answer: c_1'

Table 1: Few-shot Prompt

Step 2 - Prediction with Contextual Information: In this step, the comprehensive context set E is seamlessly integrated with the input image x and fed into the Large Vision-Language Model (LVLM). This integration enables the LVLM 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.

Thus, we tailor our methodology to few-shot scenarios in data-rich environments. Our approach is designed to leverage the abundance of data, providing a substantial scope for enhancing the accuracy of LVLM's predictions.

2.3 Speed-up via Adaptive Entropy

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.

In summary, we first employ CLIP for initial candidate class selection, followed by LVLMs for precise zero-shot or few-shot classification, effectively addressing the challenge of fine-grained image categorization. An adaptive entropy-based approach

Dataset	# of Class	# of Test
Flowers102	102	818
StanfordCars	196	8041
FGVC Aircraft	100	3333
Birds525	525	2625

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

further optimizes the process, enhancing inference speed and computational efficiency by judiciously determining when to bypass LVLM processing.

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 our CascadeVLM framework's 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, CLIP-VIT-B/16, or CLIP-VIT-L/14—alongside either Qwen-VL-Chat (Bai et al., 2023) or GPT-4V as the LVLM. Qwen-VL-Chat was selected for its capabilities in detailed visual tasks, while GPT-4V (OpenAI, 2023) was chosen for its proficiency in integrating text and image data. This strategic pairing of models aims to explore their collective effectiveness in fine-grained image classification, offering insights into their collaborative strengths within the CascadeVLM context.

Datasets In our evaluation of the CascadeVLM framework, we utilize a collection of datasets sourced from Kaggle, 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), and Birds525 (Berg et al., 2014), collectively encompassing a wide range of categories, from botanical and ornithological species to intricate mechanical designs. Each dataset presents its own set of challenges, with a varying number of classes and test images, ranging from 100 to 525 classes. This variety ensures a comprehensive assessment across different domains, testing the framework's capability to handle fine-grained classifications effectively.

Baselines In our experimental analysis, baseline performances are established using a range of CLIP models and Qwen-VL to benchmark against the capabilities of our CascadeVLM framework. The comprehensive performance metrics of these models are detailed in Table 3. In the case of GPT-4V, constrained by API call limitations and budgetary considerations, we conduct evaluations on a strategically chosen subset of 200 random samples from each dataset to maintain a balanced class representation. The results of this targeted assessment are compiled in Table 4.

3.2 Zero-shot Learning Results

Table 3 delineates the zero-shot prediction results, showcasing the superior performance of our CascadeVLM framework across various benchmarks. Notably, in the StanfordCars dataset, CascadeVLM achieved a remarkable accuracy of 85.57%, underscoring its effectiveness in integrating CLIP and LVLMs for fine-grained image classification. A detailed examination of the results reveals that while the baseline performance of LVLMs alone is modest, the accuracy is significantly enhanced merely by sorting the classes based on their probability in descending order. Furthermore, the implementation of top-k selection within our framework further amplifies this improvement, thereby validating the efficacy of our cascade approach in optimizing fine-grained classification tasks.

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.

Given the constraints of OpenAI's API rate limits and our budget considerations, our experiments with GPT-4V were limited to a subset of 200 samples per dataset. These experiments, encompassing full class categorization and a top-k (k = 5)

Model	Flower102	StanfordCars	FGVC Aricraft	Birds525	Avg.
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) Qwen Cascade (full classes)	72.7	74.3	22.7	20.3	47.5
CLIP(ViT-B/32) Qwen Cascade (top k)	**74.2**	**79.2**	**27.1**	**56.7**	**59.3**
CLIP(ViT-B/16)	73.0	64.4	24.5	52.5	53.6
CLIP(ViT-B/16) Qwen Cascade (full classes)	70.5	74.1	26.2	20.2	47.8
CLIP(ViT-B/16) Qwen Cascade (top k)	**73.3**	**79.1**	**30.7**	**56.6**	**60.0**
CLIP(ViT-L/14)	**81.3**	76.2	30.9	62.2	62.7
CLIP(ViT-L/14) Qwen Cascade (full classes)	75.8	78.9	30.1	21.0	51.5
CLIP(ViT-L/14) Qwen Cascade (top k)	78.2	**85.6**	**37.0**	**63.0**	**66.0**

Table 3: Zero-shot prediction 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.

Model	Flower102	StanfordCars	FGVC Aricraft	Birds525	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 4: 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

approach, were instrumental in validating the CascadeVLM approach. Here, the initial category filtering by CLIP, followed by GPT-4V's targeted application, markedly improved classification accuracy. Consistent with our zero-shot findings in Section 3.2, we observed that even a basic reordering of categories by their probabilities enhanced GPT-4V's performance, with the application of a top-k selection further amplifying this effect.

Furthermore, integrating few-shot learning into this cascade framework yielded 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.

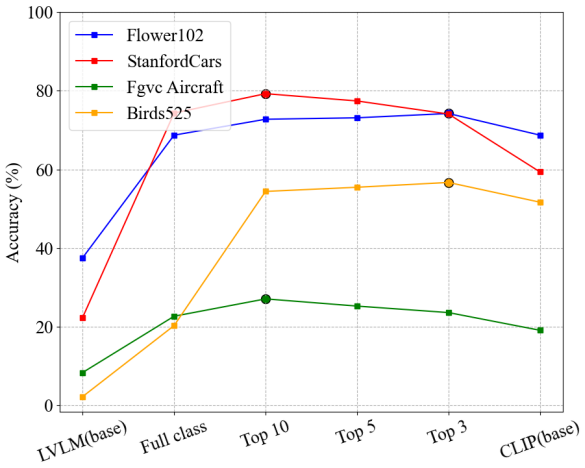

Figure 3: Performance changes with varied k with CLIP-ViT-B/32.

4 Analysis

In this section, we undertake a series of investigative experiments to elucidate various facets of our CascadeVLM framework. Initially, we delve into the influence of the top-k variable on our model's performance in Section 4.1. Subsequently, we examine the implications of the entropy threshold in Section 4.2, focusing on its role as a balancing factor between computational efficiency and accuracy. Lastly, we will conduct a thorough error analysis 4.3 and present case studies 4.4 to further contextualize our findings and insights into the framework's operational dynamics.

4.1 Influence of candidate classes number k

The validation performance of our cascade framework exhibits a dependency on the number of candidate classes, k, considered during the classification process. One of a setting of our experiments, using

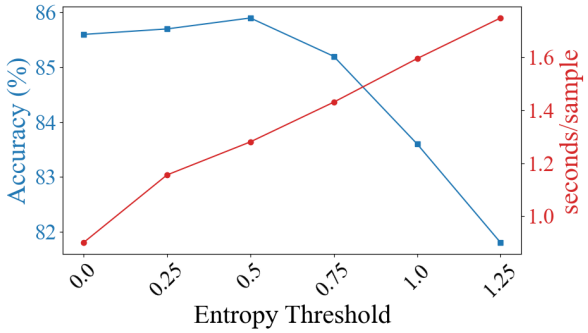

Figure 4: Performance variation in the StanfordCars dataset with varying entropy thresholds using CLIP-ViT-L/14 for cascading, set at top-k=10. An increase in entropy threshold results in decreased inference speed and reduced accuracy.

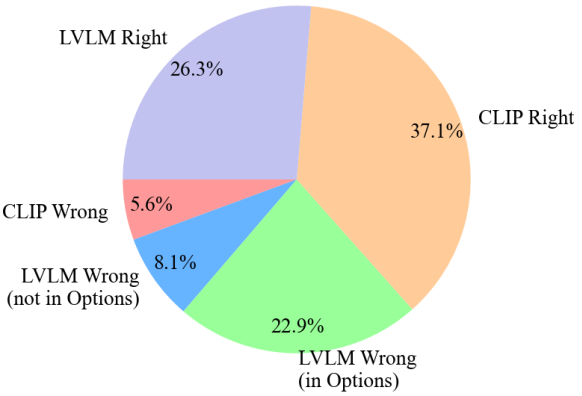

Figure 5: Error analysis of the Birds525 dataset with an entropy threshold of 1.25 and top-k=10. The analysis reveals that despite CLIP including correct options, LVLM frequently misclassifies.

CLIP ViT-B/32 and Qwen as components in the framework, represented graphically in Figures 3, demonstrate that as k is varied, the classification accuracy shifts. Interestingly, the optimal k value appears to be dataset-specific, suggesting that the intrinsic properties of each dataset may favor a different range of candidate classes. For instance, while the Flower102 and Birds525 dataset shows a gradual improvement as k decreases, indicating that a narrower focus enhances accuracy, the StanfordCars and Fgvc Aircraft dataset peaks at $k = 10$ before seeing a decline, implying a sweet spot in the balance between too few and too many options, and a further reduction in candidate classes does not confer additional benefits. This nuanced behavior underscores the importance of tailoring the cascade framework's parameters to the specific dataset at hand to achieve optimal performance.

4.2 Efficiency of threshold

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.

4.3 Error Analysis

An in-depth error analysis was conducted on the Birds525 dataset using the cascade framework, which incorporates CLIP (ViT-L/14) for initial classification and Qwen as the LVLM for refined categorization with $k = 10$, as shown in Figure 5. 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*).

4.4 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, validating the efficacy of our framework in rectifying initial misclassifications.

Case 2 depicts a situation where, despite CLIP's inclusion of the correct answer—striped owl—in

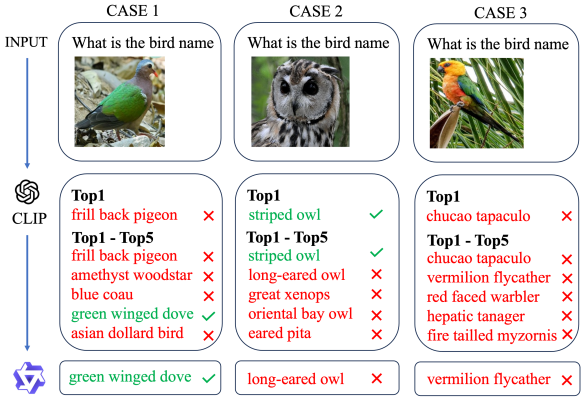

Figure 6: Three case studies demonstrating the cascade process from CLIP predictions to LVLM refinement for bird species classification.

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

Our work closely relates to recent studies building vision language models and fine-grained image classification.

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 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., 2019; 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; Yao et al., 2022; Li et al., 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., 2023), 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.

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.

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.

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.

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. 2019. 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.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

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)*.

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. 2023. 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*.

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.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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. 2023. 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.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651.

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.

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.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: fine-grained interactive language-image pre-training. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

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. 2022. 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: ...
```

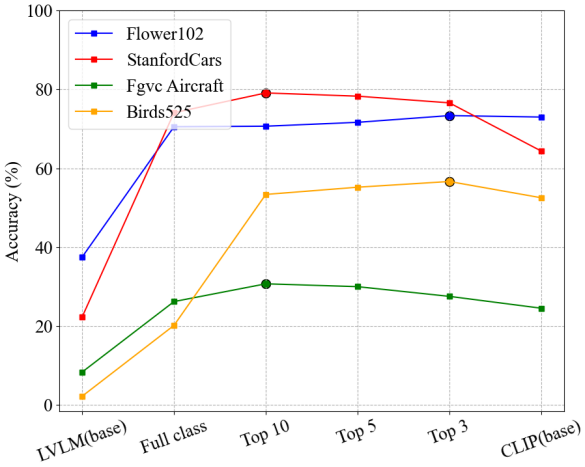

Figure 7: Performance changes with varied k with CLIP-ViT-B/16.

```
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 Influence of candidate classes number k

In complementing the analysis presented in Section 4.1, Figure 7 elucidates the impact of varying the number of candidate classes, k, within the CLIP-ViT-B/16 configuration. In this scenario, the StanfordCars and FGVC Aircraft datasets exhibit peak performance at a top-10 setting, whereas the Flower102 and Birds525 datasets demonstrate optimal results at a top-3 setting. Conversely, Figure 8 reveals distinctive trends, particularly in the StanfordCars dataset, which maintains its maximal accuracy at a top-10 setting, and the Birds525

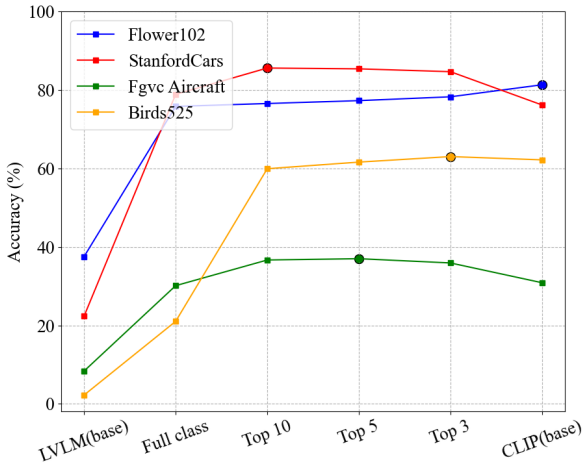

Figure 8: Performance changes with varied k with CLIP-ViT-L/14.

dataset, which continues to show peak performance at top-3. Notably, for the FGVC Aircraft dataset, an optimal shift to top-5 is observed, possibly attributed to the baseline performance enhancement of CLIP with the ViT-L/14 model. Furthermore, in the Flower102 dataset, the intrinsic fine-grained image classification ability of the CLIP-ViT-L/14 supersedes that of the Qwen LVLM in this specific dataset, leading to superior accuracy when deployed as a standalone model.