Fine-Grained Visual Recognition in the Age of Multimodal LLMs

Hari Chandana Kuchibhotla, Abbavaram Gowtham Reddy,
Sai Srinivas Kancheti, Vineeth N Balasubramanian

Indian Institute of Technology Hyderabad, India
{ai20resch11006, cs19resch11002, cs21resch01004, vineethnb}@iith.ac.in

Abstract

Fine-grained Visual Recognition (FGVR) involves differentiating between visually similar categories, and is challenging due to subtle differences between the categories and the need for large, expert-annotated datasets. We observe that recent Multimodal Large Language Models (MLLMs) demonstrate potential in FGVR, but querying such models for every test input is not practical due to high costs and time inefficiencies. To address this, we propose a novel pipeline that fine-tunes a CLIP model for FGVR by leveraging MLLMs. Our approach requires only a small support set of unlabeled images to construct a weakly supervised dataset, with MLLMs as label generators. To mitigate the impact of obtained noisy labels, we construct a candidate set for each image using labels of neighboring images, thereby increasing the likelihood of having the correct label in the candidate set. We then employ a partial label learning algorithm to fine-tune a CLIP model using these candidate sets. Our method sets a new benchmark for efficient fine-grained classification, achieving comparable performance to MLLMs at just $1/100^{th}$ of the inference cost and a fraction of the time taken.

1 Introduction

Fine-grained visual recognition (FGVR) is a task in computer vision that focuses on distinguishing between highly similar categories within a broader class [1]. For instance, traditional image classification aims to differentiate between dogs, cats, and birds, while FGVR aims to distinguish between different bird species, such as *Tennessee Warbler, Yellow-rumped Warbler, Orange-crowned Warbler* and *Sedge Warbler*. FGVR is crucial for applications requiring high specificity, such as medical diagnosis [2, 3] and biodiversity studies [4, 5]. FGVR poses significant challenges due to the subtle differences between categories and the need for large, annotated datasets. Typically, fine-grained classification datasets are annotated by domain experts who meticulously examine each image and assign a corresponding label. Multimodal Large Language Models (MLLMs) are trained on extensive corpora, and excel at zero-shot multimodal tasks. They thus offer a promising avenue for FGVR, especially when domain-specific, curated datasets are unavailable. We observe that directly querying MLLMs as *"Provide a best fine-grained class label for this image."*, results in reasonable performance in FGVR tasks, but querying such models for every test input is costly and time consuming. For instance, for the benchmark datasets used in this study, GPT-4o requires approximately 17.5 hours for querying and incurs a cost of around USD $100 for inference. This underscores the urgent need for an efficient FGVR system that conserves both time and financial resources while maintaining the performance of MLLMs. We propose to label a small support set of unlabeled images by querying an MLLM for each image. The obtained weakly supervised dataset is used to fine-tune a CLIP [6] model, which can then perform inference in a cost-effective manner. Our approach achieves performance comparable to that of MLLMs while incurring only $1/100^{th}$ of the total inference cost.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

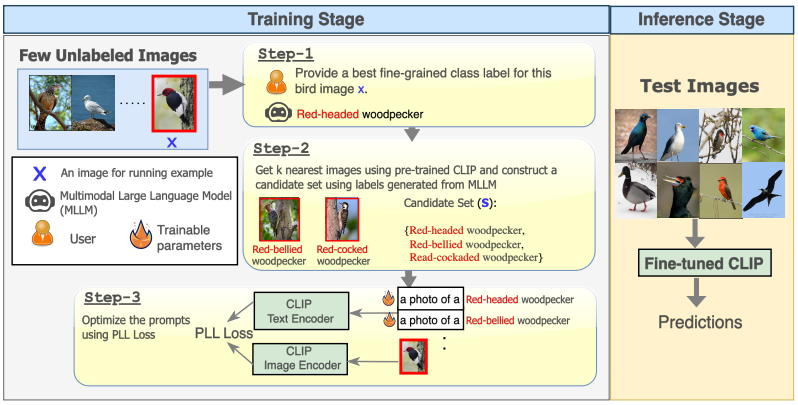

Figure 1: Our proposed training workflow consists of three steps: (Step 1) Querying an MLLM to obtain noisy labels for a small support set of unlabeled images; (Step 2) Forming a candidate set by aggregating MLLM generated labels of top-k nearest images; and (Step-3) Fine-tuning a CLIP model by optimizing learnable prompts using a PLL loss. After fine-tuning the model, inference can be performed on test images with the MLLM label space obtained during training.

Since MLLM outputs may be noisy, the original images and corresponding labels from an MLLM forms a noisy supervised dataset. To counter this, we propose to build a candidate set for each image using the labels of other similar images. The intuition here is that, for a given image x, even if the label l obtained from an MLLM might be incorrect, by building a size k candidate set $S(x) = \{l, l_1, \ldots, l_{k-1}\}$, using labels of $k-1$ most similar images, the chance of a label that is semantically closer to ground truth is included in S is higher. We then leverage partial label learning paradigm to adapt a VLM to further reduce the ambiguity of the candidate set. The overall idea is shown in Fig. 1. In summary, our contributions include: **(i)** To the best of our knowledge, this is the first work that uses MLLMs to build a cost-efficient vocabulary-free fine-grained visual recognition system, **(ii)** We propose a pipeline that can handle the noisy labels from MLLMs, and **(iii)** We outperform all the existing works and match the performance of MLLMs in a cost-efficient way.

2 Related Work

Fine-Grained Visual Recognition (FGVR). FGVR often requires additional supervision in terms of annotations or domain experts. However such curated data is usually not available for many domains of interest such as e-commerce and medical data. Thus there is a requirement of performing FGVR when no or very little supervised data is available. In this work we tackle the FGVR problem by leveraging MLLMs to annotate a small set of unsupervised images. **Foundation Models for FGVR.** Recent advances in MLLMs have led to models that show strong zero-shot performance on a variety of multimodal tasks [7–9]. Such MLLMs can be directly used for fine-grained classification by treating it as a VQA problem. However, performing inference for every test point is costly and time consuming. Recently, FineR [10] proposed a pipeline model consisting of Visual Question Answering (VQA) systems and Large Language Models (LLMs) to solve the FGVR task using just unsupervised data. However, their architecture is complex and does not utilize the advances of MLLMs, leading to inferior performance. **Prompt Tuning.** Prompt tuning fine-tunes parameters efficiently to enhance the performance of large pre-trained models on specific tasks. Context Optimization (CoOp) [11] was the first to introduce text-based prompt tuning, replacing manually designed prompts like *"a photo of a"* with adaptive soft prompts. On the other hand, Visual Prompt Tuning [12] introduces learnable prompts specifically within the vision branch. Among the various approaches for incorporating visual prompts, we chose a straightforward strategy by implementing the simplest text-based method, CoOp.

3 Methodology

Problem Formalization. Let $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathcal{X}$ be the support set of n unlabeled training images. We assume that the given support set has m-shot samples for each class of the unknown ground-truth label set. In this work, we explore how state-of-the-art Multimodal Large Language Models (MLLMs) can be leveraged for fine-grained visual recognition. Let L denote the MLLM used to generate labels for the training images. For each image x_i, we obtain a label $l_i = L(x_i, p)$, where p is a text prompt that helps the MLLM generate a class label for the image. In

this work, we use the simple prompt, *'Provide a best fine-grained class label for this image'*. The resulting dataset $\mathcal{D} = \{(x_i, l_i)\}_{i=1}^n$ consists of image-label pairs, and the label space is denoted by $\mathcal{Y} = \bigcup_{i=1}^n l_i$. Notably, the label space generated by the MLLM is larger than the true label set, since the labeling is noisy, and increases with the size of the support set.

CLIP Classifier. CLIP [6] consists of an image encoder $\mathcal{I}$ and a text encoder $\mathcal{T}$ trained contrastively on paired image-text data to learn a common multimodal representation space. For the FGVR task, we create a dataset $\mathcal{D}$ with label space $\mathcal{Y}$ as described above. CLIP can now perform zero-shot (ZS) classification of any image x by choosing the class name with the highest cosine similarity from the set of label names $\mathcal{Y}$, i.e, predicted class is $\hat{y}(x) = \arg\max_{t \in \mathcal{Y}} sim(\mathcal{I}(x), \mathcal{T}(t))$. We refer to this simple approach, of generating feasible label names using MLLMs which then form the label set for CLIP, as ZS-CLIP. Note that ZS-CLIP does not make use of the supervision information available in $\mathcal{D}$ and only acts as a simple baseline. In the section below, we present our proposal to utilize labels of nearest-neighbors to learn a better classifier using the weak labeling provided by MLLMs.

3.1 Learning from weak MLLM labels

We propose to leverage local geometry to mitigate the noise in generated labels l_i. More formally, we make the *manifold assumption*, which suggests that similar images should share similar or identical class labels. This is particularly useful when the label l_i assigned to the image x_i by the MLLM is incorrect. By constructing a candidate label set, we increase the likelihood of including the true label or a semantically closer alternative in the candidate set rather than relying solely on the incorrect label provided by the MLLM. To construct the candidate set in a simple and intuitive manner, we use CLIP's image encoder $\mathcal{I}$ to extract image features of the entire support set X. For each image x_i, we select the top-k most similar images (including x_i itself) and gather their corresponding labels to form the candidate set $S_i = (l_i, l_1, \ldots, l_{k-1})$. In this work, we choose $k = 3$. The resulting dataset is reconstructed as $\mathcal{D} = \{(x_i, S_i)\}_{i=1}^n$, incorporating the candidate sets instead of single labels alone. To fine-tune CLIP in an efficient manner, we adopt prompt-tuning methods that add a small number of learnable tokens to the input token sequence of either modality. Specifically, we follow CoOp [11], which adds prompts to only the text modality. The next step is to define an appropriate training objective to effectively leverage this supervision. Fortunately, existing loss functions designed for partial-label learning (PLL), such as PRODEN [13], can be directly applied to this dataset to fine-tune prompts. We choose PRODEN and CoOp because they are easy to implement and train, allowing us to demonstrate the effectiveness of our pipeline without relying on sophisticated PLL algorithms or prompt-tuning techniques.

4 Experiments and Results

Datasets: We perform experiments on five benchmark fine-grained datasets: CaltechUCSD Bird-200 [14], Stanford Car-196 [15], Stanford Dog-120 [16], Flower-102 [17], Oxford-IIIT Pet-37 [18]. **Baselines:** We compare against three classes of baseline methods. **(i)** We evaluate six MLLMs of varying sizes, including three proprietary models – GPT-4o [8], Gemini Flash, Gemini Pro [9] and three open-source models – BLIP-2 [7], LLaVA-1.5 with 7 billion and 13 billion parameters [19, 20]. To perform classification, we query each MLLM with the prompt *'Provide a best fine-grained class label for this image'* for each test image. **(ii)** We also consider two contemporary baselines which do not require expert annotations but use foundational models to perform FGVR – CaSED [21] and FineR [10]. **(iii)** Lastly, we compare against zero-shot CLIP with corresponding label spaces obtained from querying MLLMs: ZS-CLIP-GPT-4o, ZS-CLIP-GeminiFlash & ZS-CLIP-GeminiPro. We also include ZS-CLIP-WordNet with the label set from WordNet [22]. **Evaluation Metrics:** Following [10] we evaluate using three metrics: **(i)** Semantic Accuracy (sACC), which scores the predicted class by its semantic similarity to the ground-truth class using Sentence-BERT embeddings; **(ii)** Clustering Accuracy (cACC), which measures how well predicted classes cluster the images; and **(iii)** Semantic IOU (sIOU), which measures the intersection-over-union between the predicted and ground-truth class names.

Main Results: In Table 2 we compare the performance of our method against various baselines. The results of the various MLLMs and ZS-CLIP models were run by us, and we show numbers of CaSED and FineR as reported in their papers. We note that the proprietary MLLMs perform the best for all datasets, with the caveat that performing inference for every test sample is costly and time taking. We hence do not directly compare against these baselines, but treat them as upper bounds (in gray). Our method with Gemini Pro labels obtains a $+6.7\%$ & $+3.9\%$ improvement in sACC and cACC over the state-of-the-art fine-grained classification approach FineR [10]. Part of this

Method	Training Time	Inference Time	US$
GPT-4o	-	~ 17.5	~ 100
Gem. Flash	-	$> 24^*$	~ 47
Gem. Pro	-	$> 24^*$	~ 47
ZS-CLIP GPT-4o	-	0.03	~ 1
ZS-CLIP Gem. Flash	-	0.03	~ 1
ZS-CLIP Gem.Pro	-	0.03	~ 1
Ours (GPT-4o)	0.6	0.03	~ 1
Ours (Gem. Flash)	0.6	0.03	~ 1
Ours (Gem. Pro)	0.6	0.03	~ 1

Table 1: Time in hours and cost in US$ incurred by each method. *Only 10000 API calls per day.

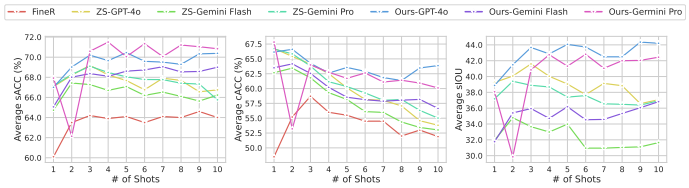

Figure 2: We illustrate the performance of our method compared to other baselines when varying the number of shots m per class, for $k = 3$.

success can be attributed to the advancements in the latest MLLMs, as their zero-shot baselines are already significantly stronger than prior methods. We also show a significant improvement of 3% over the corresponding ZS baselines using our simple proposed pipeline. Our proposed method is cost-effective, fast, and achieves comparable performance to the MLLM upper bounds, showing a drop of just 2.3% in average sACC. Table 1 presents the training time, inference time, and associated costs for running the method, summarizing its effectiveness in terms of time and cost savings relative to other baselines while considering performance gains.

Method	Bird-200			Car-196			Dog-120			Flower-102			Pet-37			Average		
	sACC	cACC	sIOU	sACC	cACC	sIOU	sACC	cACC	sIOU	sACC	cACC	sIOU	sACC	cACC	sIOU	sACC	cACC	sIOU
†GPT-4o	85.2	68.8	68.4	61.5	37.4	16.8	80.4	71.1	61.4	51.6	50.5	19.6	83.5	68.2	69.8	72.4	59.2	47.2
†Gemini Flash	74.8	49.3	49.9	60.2	51.2	16.0	76.3	62.3	53.8	57.9	31.1	16.6	77.8	61.1	49.6	69.4	51.0	37.2
†Gemini Pro	82.7	66.1	64.9	62.8	35.4	18.1	81.2	65.8	57.3	54.3	45.3	20.1	85.7	71.3	71.7	73.3	56.8	46.5
‡BLIP-2	56.8	30.9	-	57.9	43.1	-	58.6	39.0	-	**59.1**	61.9	-	60.5	61.3	-	58.6	47.2	
‡LLaVA-1.5-7B	45.6	5.6	2.3	48.3	11.9	5.3	45.0	3.8	2.8	44.3	8.6	4.9	42.8	6.6	0.1	45.2	7.3	3.0
‡LLaVA-1.5-13B	42.8	10.0	2.5	12.1	4.4	0.1	33.0	13.0	0.9	35.2	20.7	35	37.9	1.0	12.9	31.5	10.1	1.1
CaSED	50.1	25.6	-	41.4	26.9	-	55.9	38.0	-	52.3	67.2	-	63.6	60.9	-	52.6	43.7	-
FineR	69.5	51.1	-	**63.5**	49.2	-	64.9	48.1	-	51.3	63.8	-	72.4	72.9	-	64.3	57.0	-
ZS-CLIP-WordNet	57.7	39.3	-	33.3	18.3	-	70.6	53.9	-	49.8	42.1	-	61.9	55.4	-	54.7	41.8	-
ZS-CLIP-GPT-4o	72.5	48.8	46.6	59.5	42.9	14.4	69.1	51.0	43.8	53.0	62.1	18.2	78.7	68.2	59.9	66.6	54.6	36.6
ZS-CLIP-GeminiFlash	71.8	47.4	44.4	58.8	44.9	13.3	68.8	50.9	40.2	53.5	53.1	13.4	75.3	70.9	44.2	65.6	53.4	31.1
ZS-CLIP-GeminiPro	74.6	51.7	50.5	61.7	41.6	16.3	72.6	58.9	40.7	49.1	57.7	13.7	78.6	71.7	60.8	67.3	56.3	36.4
Ours (GPT-4o)	**78.7**	56.6	**58.1**	60.5	49.6	15.7	75.5	63.9	**53.7**	52.3	**69.0**	**21.8**	**84.7**	**78.3**	**72.5**	70.3	**63.5**(+5.4)	**44.4**(+6.5)
Ours (Gemini Flash)	74.6	50.0	48.2	58.8	**51.1**	14.9	72.0	57.7	47.4	58.4	57.3	18.1	79.1	75.1	51.6	68.6	58.2	36.1
Ours (Gemini Pro)	76.9	**56.9**	56.1	61.6	42.8	**17.0**	**75.9**	**65.2**	46.2	56.8	64.4	21.0	84.0	75.4	70.0	**71.0**(+5.8)	60.9	42.0

Table 2: ZS-Zero Shot, ‡-Open-source models used for inference. †-SOTA models used for inference, in our case they act as oracle models. The best numbers are highlighted in bold with a mint background. The second-best numbers are underlined with a yellow background. **Green** numbers show the improvement compared to previous published SOTA FineR method. Our results shown here are for $k = 3$ and $m = 3$.

Ablation Studies: To evaluate the manifold hypothesis, we compare our nearest-neighbor candidate sets against randomly sampling labels from the label space to generate a candidate set, which we denote as Random CS. The results presented in Table 3 indicate that nearest neighbor candidate sets perform better over all metrics. We also study the effect of querying the MLLM to directly generate a candidate set of most relevant class labels as opposed to a single class label. There are two main disadvantages with this approach – the label space becomes prohibitively large, and we lose similarity information. As shown in Table 4, our proposal for obtaining candidate sets shows superior performance while avoiding Out-of-Memory (OOM) issues that arise due to the larger label space of the former method. In figure 2, we study the effect of varying the number of unlabeled samples per class in the training set. Our method shows good performance over a wide range of shots, and outperforms FineR at all shots.

Method	Average		
	sACC	cACC	sIOU
Random CS + GPT-4o	69.4	58.7	42.4
Random CS + Gemini Flash	66.9	55.5	33.6
Random CS + Gemini Pro	69.3	58.0	40.1
Our CS + GPT-4o	70.3	**63.5**	**44.4**
Our CS + Gemini Flash	68.6	58.2	36.1
Our CS + Gemini Pro	**71.0**	60.9	42.0

Table 3: Our proposed NN candidate set performs better than Random sampling.

Method	Datasets				
	Birds-200	Cars-196	Dog-120	Flower-102	Pet-37
GPT-4o CS	76.4	OOM	74.02	51.43	83.84
Ours(GPT-4o)	**78.7**	**63.5**	**75.5**	**52.3**	**84.7**

Table 4: Our proposed NN candidate set performs better while being more efficient when compared to directly querying the MLLM to generate k=3 candidates.

5 Conclusion

In this work we describe a pipeline to leverage MLLM knowledge to perform efficient fine-grained recognition for domains where expert annotations are unavailable. We propose to obtain labels for a small unlabeled support set from state-of-the-art MLLMs, which are used to fine-tune a CLIP model that can perform efficient inference. We achieve performance close to that of directly querying MLLMs for all test images, at a fraction of compute time and cost.

References

[1] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44:8927–8948, 2021.

[2] Muhammad Ridzuan, Ameera Bawazir, Ivo Gollini Navarrete, Ibrahim Almakky, and Mohammad Yaqub. Self-supervision and multi-task learning: Challenges in fine-grained covid-19 multi-class classification from chest x-rays. In *Annual Conference on Medical Image Understanding and Analysis*, pages 234–250. Springer, 2022.

[3] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 19339–19352, 2020.

[4] Yu Liu, Yaqi Cai, Qi Jia, Binglin Qiu, Weimin Wang, and Nan Pu. Novel class discovery for ultra-fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17679–17688, 2024.

[5] Guofeng Yang, Yong He, Yong Yang, and Beibei Xu. Fine-grained image classification for crop disease based on attention mechanism. *Frontiers in Plant Science*, 11:600854, 2020.

[6] 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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[9] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[10] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348, 2021. URL https://api.semanticscholar.org/CorpusID:237386023.

[12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser Nam Lim. Visual prompt tuning. *ArXiv*, abs/2203.12119, 2022. URL https://api.semanticscholar.org/CorpusID:247618727.

[13] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:211171790.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech ucsd bird dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[15] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[18] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[21] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. *Advances in Neural Information Processing Systems*, 36:30662–30680, 2023.

[22] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38:39–41, 1995.