How Do Multimodal LLMs Really Fare in Classical Vision Few-Shot Challenges? A Deep Dive

Qing Guo[1], **Prashan Wanigasekara**[2], **Jian Zheng**[2], **Zhiyuan Fang**[2]
Xinwei Deng[1], **Chenyang Tao**[2,†]
[1]Virginia Tech [2]Amazon
chenyang.tao@duke.edu

Abstract

Recent advances in multimodal foundational models have demonstrated marvelous in-context learning capabilities for diverse vision-language tasks. However, existing literature have mainly focused on few-shot learning tasks similar to their NLP counterparts. It is unclear whether these foundation models can also address classical vision challenges such as few-shot classification, which in some settings (*e.g.*, 5-way 5-shot) necessitates sophisticated reasoning over several dozens of images – a challenging task for learning systems. In this work, we take a deep dive to probe the potentials and limitations of existing multimodal models on this problem. Our investigation reveals that while these models under careful calibration can outperform dedicated visual models in complex narratable scenes, they can falter with more abstract visual inputs. Moreover, we also investigate the curriculum learning and find out it can mitigate the performance gap via smoothly bridging verbal and nonverbal reasoning for vision language tasks.

1 Introduction

How humans rapidly adapt across different tasks with little supervision has long fascinated the science community [1–3]. In many real-world situations, acquiring large datasets for training is either impractical, infeasible, or cost prohibitive, thus making reliable predictions about unseen cases from sparse exemplars a vital research direction for machine learning [4–7].

As humans, our ability to generalize well without extensive supervision is widely believed to come from prior experiences, knowledge, and the ability to integrate information and *think*. For example, when adults learned about the animal *llama* for the first time, it rarely takes more than a couple of images to register the concept of this species for its resemblance to more common animals such as sheep, camels, or horses[1]. This fast learning happens by (i) activating prior **knowledge** from previously seen tasks (*e.g.*, similar animals); and (ii) extracting useful **information** from the limited exemplars for the new task. To enable robust generalizations, especially in situations of incomplete information, (iii) **reasoning** kicks in: creating higher levels of abstraction to relate disparate pieces of information, formulating/testing hypotheses, and making logical extrapolations.

These intuitions have all been mathematically formalized under various learning theories, such as *Bayesian methods* [8], *weakly supervised learning* [9], *causal machine learning* [10], *information theoretic generalization bounds* [11], *etc*. While different technical notations of knowledge, information, and reasoning can provably improve few-shot generalization, for real-world problems there has been a lack of generic learning frameworks to accommodate knowledge manipulation and complex reasoning at scale. Consequently, domain knowledge driven or regularization based methods have ruled few-shot learning leaderboards in practice [12].

[1]In fact, in some languages *llama* literally translates into *sheep-camel*, *camel-horse*, *etc.*

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

In recent years, the rise of foundational models has initiated a paradigm shift in building general purpose tools for machine learning [13]. Large language models (LLMs) such as `ChaptGPT` have demonstrated human-like problem solving skills when strong reasoning and encyclopedic knowledge are seamlessly integrated, and they are receptive to human instructions to collaboratively complete more complex tasks. This has also ignited significant interests in transcending the domain boundaries to synergize visual and language foundational models [14, 15] to build *Large Multimodal Models* (LLM). Recent studies have shown the effectiveness & efficiency of interfacing different modalities using light-weight adapters [16–19], and the impressive emerging generalization capabilities of these fused models on both traditional and novel vision-language tasks, either with or without few-shot task adaptation.

While existing LMMs perform strongly on certain dimensions (*e.g.*, few-shot task following, captioning, VQA, *etc.*), they are highly reliant on what the base LLMs have been heavily tuned on. This raises an interesting question: to what extent does the broad knowledge and strong logic benefit classical vision challenges? For example, the multi-way few-shot classification also necessitates advanced cognitive function, and it solving this challenge entails several core competence dimensions not adequately covered by popular multimodal benchmarks [20–22]: models need to compose reasoning over a large number of images, possibly with complex visual scenes and subject to heavy confounding (see Figure 1 for example, the same scene can be associated with multiple categories).

In this work, we make the following contributions: (i) ablating different few-shot classification strategies for LMM; (ii) benchmarking on an extensive set of datasets with varying difficulty; (iii) proposing auxiliary tasks that boost performance and interpretability. Our findings show that with proper tuning, LMMs can do exceptionally well on complex narratable visual inputs, and even beat state-of-art dedicated vision models; however, their effectiveness degrades on more abstract visual inputs and struggle on subtle differences that requires very sophisticated, verbose descriptions. These observations reveal current gaps for multimodal models: they have mostly learned shallow object concept mappings during training and heavily rely on verbal reasoning to perform inference. To facilitate future research, we also release synthetic data annotations used in this study.

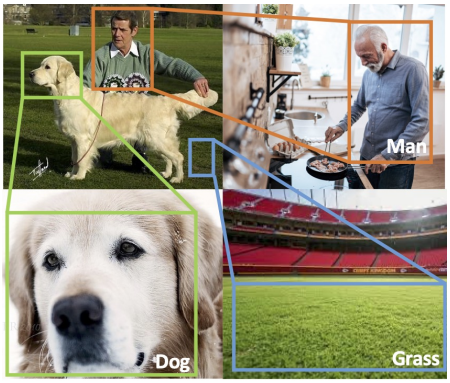

Figure 1: Multiple concepts often co-exist in complex visual scenes, making 1-shot classification an ill-posed problem. Thus, the system must reason from multi-shot examples for accurate concept binding.

2 Background and Problem Setup

Large Multimodal Models (LMM). In this work, we focus on the popular multimodal adapter architecture, where models are comprised of three components: a base LLM to process instructions and generate responses, a visual encoder to embed visual inputs, and an adapter to align visual embeddings to LLM inputs. This modularized design is highly flexible and parameter efficient: one can easily plug in different pre-trained foundational models, and only train the light-weight adapter to achieve impressive performance on a variety of tasks. Specifically, we follow MiniGPT-4's recipe as our starting point [19]: LLaMA-based `Vicuna` model is used as our base LLM, and the vision encoder is the pre-trained Q-former from `BLIP-2` [18] along with its `ViT` backbone [23], and a simple linear adaptation layer is used to connect the two. See Figure 2 for our model architecture. We expect our conclusions to generalize well to other LMMs as MiniGPT-4's architecture and training recipe is simple and representative.

Vision-language alignment and instruction tuning. Similar to their language-only counterparts, LMMs typically need to go through a two-stage training process to become useful. In the first alignment stage, the models are supervised with diverse text-image pairs or interleaved multimodal texts to establish the mapping between corresponding visual and language tokens. Massive amounts of data are used in this phase to teach various visual concepts, and typically, they are noisy and not directly tied to specific tasks. To teach LMM to understand human intents and accomplish regular vision-language tasks, instruction fine-tuning coaches models with high-quality task data in the second stage. Perhaps surprisingly, if the base LLM is already well tuned on diverse instructions,

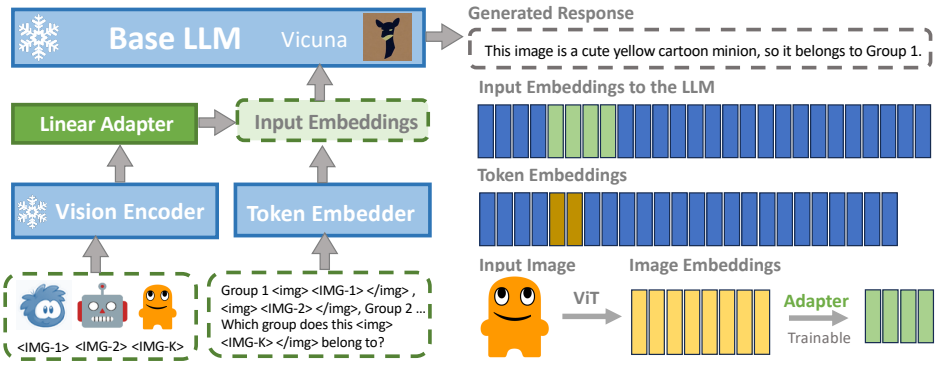

Figure 2: Model architecture of multimodal adapter.

then the second-stage learning can be highly efficient: models tuned with a few hundred examples on simple generic tasks such as image captioning already capable of carrying out a wide range of common visual-language tasks, and it only takes a few minutes on a single GPU. More extensive tuning on a richer set of tasks typically yields further gains on quantifiable dimensions, as evaluated by comprehensive benchmarks [21, 20].

Few-shot classification. In this work, we adopt the classical N-*way* K-*shot* setup for the few-shot classification problem. Specifically, the models are trained and evaluated in an episodic fashion: each episode $\mathfrak{e}$ is composed of a support set $X^{\mathfrak{e}}_{\text{supp}} = \{(x_{nk}, y_{nk}) | n \in [1, N], k \in [1, K]\}$ and a query set $X^{\mathfrak{e}}_{\text{query}} = \{(x_l, y_l) | l \in [1, L]\}$, where n indexes different classes. The goal is to make predictions on the query labels using the K exemplars for each class within the episode. Compared to traditional "static" classification where classes are fixed beforehand at training time, few-shot classification poses challenges on capturing generalizable intra/inter-class patterns dynamically with only a handful of examples. As such, the classes used in our evaluation are disjoint from the classes used in training, *i.e.*, $\mathcal{C}_{\texttt{train}} \bigcap \mathcal{C}_{\texttt{test}} = \emptyset$. If powerful learners fail to come up with a good hypothesis, it easily overfits.

Importance of in-context few-shot classification for large visual systems. In open-world interactive settings, new visual concepts or categories not present during training can emerge (*e.g.*, personal items, pets, *etc.*), and the visual system needs to act on such unseen visual entities. Often in such situations, the system may receive some limited directions, blending both language instruction and a few visual examples. For complex visual scenes, the system also has to deal with additional difficulties such as distracting backgrounds, visual occlusion, confounding elements, *etc.* Also, for large visual systems, it is typically impractical to perform gradient-based task adaptations. Hence, in-context few-shot classification becomes a critical capability, enabling the system to comprehend and reason about these unfamiliar concepts to successfully complete the tasks. To the best of author(s)' knowledge, this critical dimension has not been rigorously investigated in LMM literature.

3 Few-Shot Classification With LMM

Compared to existing few-shot methods, LMMs have several key advantages: (i) it can leverage the enormous parametric knowledge; (ii) powerful reasoning of various kinds (*e.g.*, logical, commonsense, causal, counterfactual, inductive, abductive, *etc.*); (iii) flexibility to receive additional language guidance such as instruction or corrective feedback. It is natural to hypothesize that these models should also work reasonable well on the classic N-way K-shot classification:

- For simpler cases where all input images feature a salient object that has a common name, models can directly invoke their zero-shot classification ability to classify;
- For harder cases where more complex visual scenes are involved, models should activate higher level of cognitive features, first decompose the scene into more elementary components, then apply reasoning to summarize the intra-class commonalities & inter-class differences to derive decision rules.

Despite the above reasons for optimism, we also foresee a few risks that may prevent LMMs from being performant in few-shot image classification:

- **Hallucination:** LMMs also suffer from hallucinations, and such errors can propagate through the decision chain and lead to wrong conclusions;

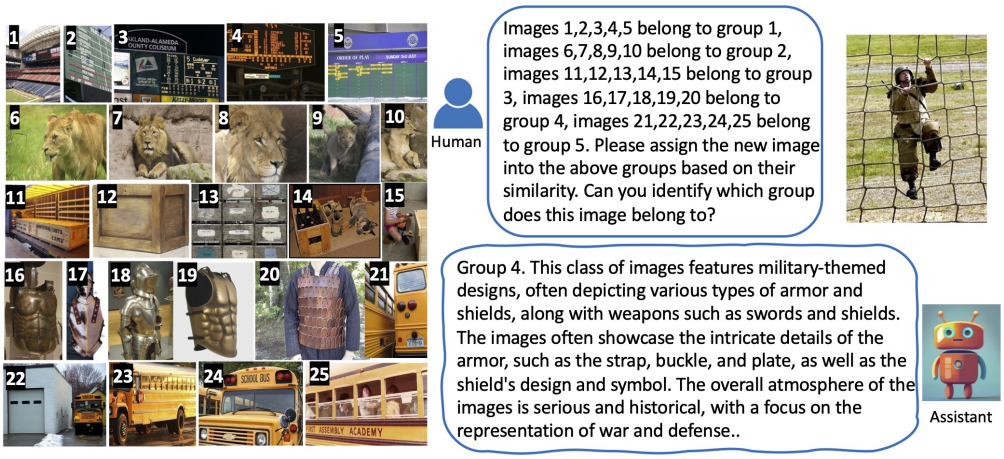

Figure 3: An example of reasoning augmented few-shot classification using LMM.

- **Unstructured visual tokens:** It is unclear how LLMs perceive and process visual inputs that are presented as cluttered semantic embeddings, the standard reasoning process learned from structured texts may break down in such scenarios;
- **Mismatched correspondence:** LMMs are generally tuned to optimize text-image correspondence during training, while few-shot vision classification mainly assesses the understanding of image-image correspondence;
- **Visual information overflow:** At training time, LMMs typically see no more than a couple of images in the input; while in the standard 5-way 5-shot setup for example, the models need to reason over $5 \times 5 + 1 = 26$ images, which is cognitively overwhelming.

With these doubts in mind, we ran a quick sanity test on the 5-way 5-shot `miniImageNet` classification task using the `OpenFlamingo` model. Not surprisingly, the emerging few-shot task following capability failed this classical challenge: we only observed a meager accuracy of 27.5%, slightly better than random guessing.

Having established that emergence alone is inadequate for few-shot classification, we are interested in exploring how additional supervised training can help. Below, we describe a series of carefully designed experiments to study how to activate the models' potentials on this task. Training details and prompt templates used in these experiments can be found in the Appendix A.

Caption-guidance (two-step). Given most LMMs are pretrained on image captioning, we propose a two-step captioning guided classification strategy as our baseline: in the first step, we generate a descriptive caption for the query image; in the second step, we prompt the model with the support image embeddings and the generated query caption to predict the query image label.

Reasoning-guidance (one-step). Our second strategy is motivated by chain-of-thought (CoT) reasoning and the consideration to make the model's decision more interpretable: in addition to the label prediction, models also need to generate the reasoning for the decision (Figure 3). We adopt an inductive setup to synthesize reasoning examples for training: sample 5 images from the query class, caption them using LMM, then use a rewriter LLM to summarize captions' commonalities as the supporting argument. To improve reasoning quality, we both inject manually authored seed examples to guide rewriter and apply *ad-hoc* rule-filtering. After experimenting with multiple candidates, we pick `Vicuna-`$7B$ v1.5 as our rewriter.

Non-verbal reasoning (one-step). *"It's easier to show than to tell."* While well-tuned LLMs are proficient in verbal reasoning, many vision-language tasks also involve non-verbal reasoning steps which are hard to put into words (*e.g.*, abstract shapes, analogies and transformations that are ineffable). Applying verbal reasoning as in the previous two strategies will trigger hallucinations and harm accuracy. This motivates us to skip all intermediate textual steps (captions or reasoning) and directly predict with image embeddings.

Selective focusing (auxiliary task). One key observation we made is that vision-adapted LLMs suffer cognitive overloading when processing dense image inputs, and they easily confuse the contents

Table 1: MiniImageNet 5-way 5-shot test classification accuracy (%).

Baselines	Acc	Baselines	Acc	Baselines	Acc	MM-LLM	Acc
ProtoNet [24]	79.4	FRN [25]	82.8	PAL [26]	84.4	CAP	79.5
FEAT [27]	82.0	BML [28]	83.6	COSOC [29]	85.2	NVR	85.3
DeepEMD [30]	82.4	Meta-NVG [31]	83.8	CNL [32]	83.4	REASON+SF	89.3
MELR [33]	83.4	MetaQDA [34]	84.3	FewTURE [35]	86.4	NVR+SF	**92.8**

`CAP`: caption-guided; `NVR`: non-verbal reasoning; `REASON`: reasoning-guided; `SF`: selective focusing.

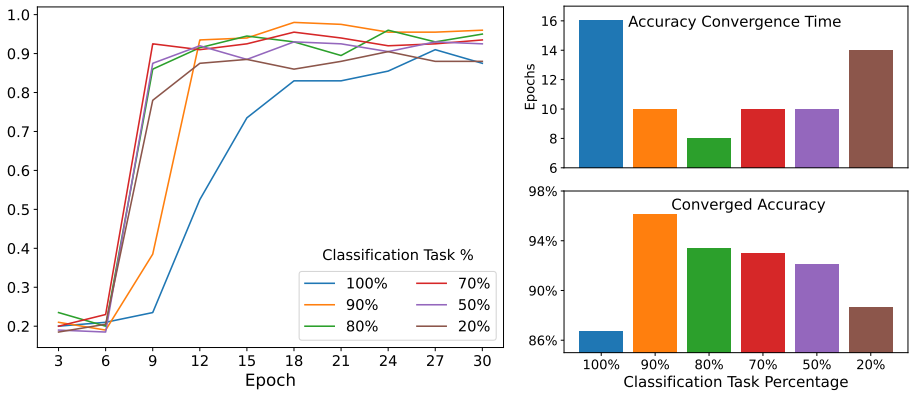

Figure 4: Ablation on the effect of adding different portion of selective focusing tasks. A small fraction of description teaches the model to better understand subtle difference between the images

from different images. This inspired us to augment the training with an auxiliary task we call selective focusing, where the models are instructed to generate detailed description for one randomly selected image from its inputs. Given the absence of such granular level annotations, we use synthetic dense captions generated by the LMM as training data.

Curriculum training. Our initial results reveal substantial performance gaps on datasets with unnarratable visual inputs or visually similar categories. We conjecture this is due to the sharp transition from verbal to non-verbal reasoning traps the model in bad local optima. To remedy, we adopt curriculum training for these more challenging settings, where we warm-start the model on narratable classifying diverse narratable scenes (*i.e.*, `ImageNet`).

4 Experiments

We used the following popular few-shot classification benchmarks in our experiments: *miniImageNet* [6], *tieredImageNet* [36], *CIFAR-FS* [37], and *Meta-Dataset* [38]. Limited by space, we present only the key results in the main text, and defer details of training & evaluation setups to the Appendix B. Our code will be available from `https://github.com/qingguo666/MMFSL`.

We start by ablating different learning strategies on the `miniImageNet` with 5-way 5-shot classification. Table 1 summarizes the main results along with baseline numbers from prior arts [2]. Despite being a general purpose tool, LMM performs strongly on the few-shot classification task: without exploiting any special architecture, feature engineering, or inductive bias, they are able to match SOTA performance of dedicated vision FSL models. Our close analyses on the less performant caption-guided strategy revealed errors are mostly due to hallucinated objects in the caption or the target object not visually salient.

Table 2: 5-way 5-shot accuracies (%)

	tiredIN	miniIN	CIFAR-FS
ProtoNet [24]	84.01	79.46	-
FEAT [27]	84.79	82.05	-
MetaQDA [34]	89.56	84.28	88.79
FewTURE [35]	89.96	86.38	88.90
MM-LLM	**91.83**	**92.8**	**92.5**
cross-domain	[Source]	97.17[†]	89.17

[†] miniIN might overlap with tiredIN.

[2] We exclude works using visual encoder trained from full ImageNet data due to information leakage concerns.

Table 3: 5-way 5-shot accuracies (%) and 95% confidence interval on `Meta Dataset`

Method	Flower	Airplane	Fungi	Birds	Texture	Omniglot
Train on ImageNet only						
k-NN [38]	83.10±0.68	46.81±0.89	36.16±1.02	50.13±1.00	66.36±0.75	37.07±1.15
MatchingNet [38]	80.13±0.71	48.97±0.93	33.97±1.00	62.21±0.95	64.15±0.85	52.27±1.28
ProtoNet [40]	86.96±0.73	58.04±0.96	40.73±1.15	74.07±0.92	68.76±0.77	68.50±1.27
fo-MAML [38]	81.74±0.83	56.24±1.11	32.10±1.10	63.61±1.06	68.04±0.81	55.55±1.54
RelationNet [38]	68.76±0.83	40.73±0.83	30.55±1.04	49.51±1.05	52.97±0.69	45.35±1.36
BOHB [41]	87.34±0.59	54.12±0.90	41.38±1.12	70.69±0.90	68.34±0.76	67.57±1.21
TSA [42]	94.05±0.45	80.13±1.01	51.38±1.17	83.39±0.80	**79.61±0.68**	82.58±1.11
Train on MetaDataset						
MatchingNet [38]	81.90±0.72	69.17±0.96	33.70±1.04	56.40±1.00	61.80±0.74	78.25±1.01
ProtoNet [38]	86.85±0.71	71.14±0.86	40.26±1.13	67.01±1.02	65.18±0.84	79.56±1.12
RelationNet [38]	76.08±0.76	69.71±0.83	32.56±1.08	54.14±0.99	56.56±0.73	86.57±0.79
fo-Proto-MAML [38]	88.72±0.67	75.23±0.76	41.99±1.17	69.88±1.02	68.25±0.81	82.69±0.97
CNAPs [43]	88.90±0.50	83.70±0.60	50.20±1.10	73.60±0.90	59.50±0.70	91.70±0.50
SUR-pnf [44]	90.00±0.60	79.70±0.80	49.80±1.10	75.90±0.90	72.50±0.70	90.00±0.60
URT [45]	88.20±0.60	85.80±0.60	63.50±1.00	76.30±0.80	71.80±0.70	94.40±0.40
FLUTE [46]	91.60±0.60	87.20±0.50	58.10±1.10	79.20±0.80	68.80±0.80	93.20±0.50
URL [47]	92.11±0.48	88.59±0.46	68.75±0.95	80.54±0.69	76.17±0.67	94.51±0.41
TSA [42]	92.18±0.52	89.33±0.44	67.40±0.99	81.42±0.74	76.74±0.72	**94.96±0.38**
LMM (starting from `tiredImageNet` trained checkpoint)						
Direct transfer	69.6±1.11	27.6±2.31	30.5±1.71	64.8±1.34	59.3±2.16	34.2±1.42
Curriculum on each dataset	96.7±0.62	89.9±0.85	81.2±0.61	96.7 ±1.09	77.0±1.58	94.3±0.77
Curriculum on all datasets	**97.3±0.30**	**91.6±0.50**	**83.2±1.80**	**96.9±0.58**	78.2±0.09	91.9±1.40

While we saw the generated reasoning does generalize reasonably well for unseen inputs, contrary to our original expectation, reasoning-guided learning is less accurate compared to their no reasoning counterpart. The fact that text generation free strategies worked better indicates: (1) harmful hallucination happens in the text decoding step; (2) the visual embeddings received by the model does capture richer signals that can be conveyed in text. We note visual information overflow seems to be the main blocker, as adding selective-focusing enables LMM to beat SOTA results by a large margin.

Figure 4 further ablates how different balancing ratio between classification and selective-focusing impact final accuracy and convergence speed, a small fraction ($10 \sim 20\%$) of focusing can be most beneficial. Based on the learnings from `miniImageNet`, we apply the best setting to `tieredImageNet` and `CIFAR-FS`. Table 2 shows that the results are consistent with our observations from the `miniImageNet` experiments, and the knowledge appears to transfer well across domains.

Next we move to more challenging settings where the models must discriminate between: (i) categories that are visually similar (*i.e.*, `Flower`, `Airplane`, `Fungi`, `Birds`); and (ii) categories that are hard to describe (*e.g.*, `Omniglot`, `Texture`). In Table 3, direct cross-domain transfer from `tiredImageNet` trained model does not do well, which is consistent with our observation that descriptions for images from these datasets are often too broad to make a discriminative call with high confidence (see examples in the Appendix C). However, if we warm start from the *tiredImageNet* checkpoint, they become highly competitive to the SOTA results only after a few epochs' training. We hypothesize this is because after being proficient in leveraging verbal reasoning to solve the problems, it becomes easier to grasp non-verbal reasoning skills on top of that.

As a final remark, recent studies have warned potential catastrophic forgetting after the fine-tuning of LMMs [39]. We therefore interacted with the few-shot optimized models and noticed models can still perform other vision-language tasks, but less compelling than before.

5 Conclusion

We have demonstrated LMM provides a simple, effective, interpretable approach to address the challenge of few-shot image classification. Our results showed model experienced hardships of visual overloading and non-verbal reasoning, which can be mitigated via smoothing the learning curves through introducing auxiliary tasks and adopting curriculum learning.

Limitations and future work. Limited by time and resources, this work narrowly focused on improving one task performance without quantifying the potential regressions in other dimensions. In future work, we will investigate how target improving essential functionalities such as few-shot and compositional reasoning in the alignment-stage to holistically enhance end-task performance.

References

[1] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[2] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[3] I. Biederman, "Recognition-by-components: a theory of human image understanding.," *Psychological review*, vol. 94, no. 2, p. 115, 1987.

[4] S. Thrun, "Lifelong learning algorithms," in *Learning to learn*, pp. 181–209, Springer, 1998.

[5] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *CVPR workshop*, IEEE, 2004.

[6] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *NeurIPS*, vol. 29, 2016.

[7] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, pp. 1126–1135, PMLR, 2017.

[8] N. Ding, X. Chen, T. Levinboim, S. Goodman, and R. Soricut, "Bridging the gap between practice and pac-bayes theory in few-shot meta-learning," *NeurIPS*, 2021.

[9] J. Robinson, S. Jegelka, and S. Sra, "Strength from weakness: Fast learning using weak supervision," in *ICML*, pp. 8127–8136, PMLR, 2020.

[10] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[11] Q. Chen, C. Shui, and M. Marchand, "Generalization bounds for meta-learning: An information-theoretic analysis," in *NeurIPS*, vol. 34, 2021.

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

[13] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "Gpts are gpts: An early look at the labor market impact potential of large language models," *arXiv preprint arXiv:2303.10130*, 2023.

[14] K. Lu, A. Grover, P. Abbeel, and I. Mordatch, "Pretrained transformers as universal computation engines," *arXiv preprint arXiv:2103.05247*, vol. 1, 2021.

[15] OpenAI, "Gpt-4v(ision) system card."

[16] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *NeurIPS*, 2021.

[17] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, *et al.*, "Language is not all you need: Aligning perception with language models," *arXiv preprint arXiv:2302.14045*, 2023.

[18] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[19] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[20] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, *et al.*, "Mmbench: Is your multi-modal model an all-around player?," *arXiv preprint arXiv:2307.06281*, 2023.

[21] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models," *arXiv preprint arXiv:2306.09265*, 2023.

[22] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

[23] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *CVPR*, 2023.

[24] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[25] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *CVPR*, 2021.

[26] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, and W. Abd-Almageed, "Partner-assisted learning for few-shot image classification," in *CVPR*, 2021.

[27] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *CVPR*, 2020.

[28] Z. Zhou, X. Qiu, J. Xie, J. Wu, and C. Zhang, "Binocular mutual learning for improving few-shot classification," in *CVPR*, 2021.

[29] X. Luo, L. Wei, L. Wen, J. Yang, L. Xie, Z. Xu, and Q. Tian, "Rectifying the shortcut learning of background for few-shot learning," *NeurIPS*, 2021.

[30] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *CVPR*, 2020.

[31] C. Zhang, H. Ding, G. Lin, R. Li, C. Wang, and C. Shen, "Meta navigator: Search for a good adaptation policy for few-shot learning," in *CVPR*, 2021.

[32] J. Zhao, Y. Yang, X. Lin, J. Yang, and L. He, "Looking wider for better adaptive representation in few-shot learning," in *AAAI*, 2021.

[33] N. Fei, Z. Lu, T. Xiang, and S. Huang, "Melr: Meta-learning via modeling episode-level relationships for few-shot learning," in *ICLR*, 2020.

[34] X. Zhang, D. Meng, H. Gouk, and T. M. Hospedales, "Shallow bayesian meta learning for real-world few-shot recognition," in *CVPR*, 2021.

[35] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," *NeurIPS*, 2022.

[36] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *ICLR*, 2018.

[37] L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *ICLR*, 2019.

[38] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, *et al.*, "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *ICLR*, 2020.

[39] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y. J. Lee, and Y. Ma, "Investigating the catastrophic forgetting in multimodal large language models," *arXiv preprint arXiv:2309.10313*, 2023.

[40] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: spatially-aware few-shot transfer," *NeurIPS*, 2020.

[41] T. Saikia, T. Brox, and C. Schmid, "Optimized generic feature learning for few-shot classification across domains," *arXiv preprint arXiv:2001.07926*, 2020.

[42] W.-H. Li, X. Liu, and H. Bilen, "Cross-domain few-shot learning with task-specific adapters," in *CVPR*, 2022.

[43] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner, "Fast and flexible multi-task classification using conditional neural adaptive processes," *NeurIPS*, 2019.

[44] N. Dvornik, C. Schmid, and J. Mairal, "Selecting relevant features from a multi-domain representation for few-shot classification," in *ECCV*, Springer, 2020.

[45] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," in *ICLR*, 2021.

[46] E. Triantafillou, H. Larochelle, R. Zemel, and V. Dumoulin, "Learning a universal template for few-shot dataset generalization," in *ICML*, PMLR, 2021.

[47] W.-H. Li, X. Liu, and H. Bilen, "Universal representation learning from multiple domains for few-shot classification," in *CVPR*, 2021.

[48] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, pp. 8748–8763, PMLR, 2021.

[50] C. Tao, L. Chen, S. Dai, J. Chen, K. Bai, D. Wang, J. Feng, W. Lu, G. Bobashev, and L. Carin, "On fenchel mini-max learning," in *NeurIPS*, 2019.

[51] J. Chen, Z. Gan, X. Li, Q. Guo, L. Chen, S. Gao, T. Chung, Y. Xu, B. Zeng, W. Lu, *et al.*, "Simpler, faster, stronger: Breaking the log-K curse on contrastive learners with FlatNCE," in *NeurIPS 2021 Workshop: Self-Supervised Learning - Theory and Practice*, 2021.

[52] Q. Guo, J. Chen, D. Wang, Y. Yang, X. Deng, F. Li, L. Carin, and C. Tao, "Tight mutual information estimation with contrastive Fenchel-Legendre optimization," in *NeurIPS*, 2022.

[53] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *NeurIPS*, 2022.

[54] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, 2023.

[55] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023.

[56] Z. Xu, Y. Shen, and L. Huang, "Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning," in *ACL*, 2023.

[57] B. Zhao, B. Wu, and T. Huang, "Svit: Scaling up visual instruction tuning," *arXiv preprint arXiv:2307.04087*, 2023.

[58] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[59] W. Dai, J. Li, D. Li, A. Meng Huat Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[60] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[61] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *CVPR*, 2023.

[62] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

[63] A. Zhang, H. Fei, Y. Yao, W. Ji, L. Li, Z. Liu, and T.-S. Chua, "Transfer visual prompt generator across llms," *arXiv preprint arXiv:2305.01278*, 2023.

[64] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *CVPR*, pp. 18030–18040, 2022.

[65] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[66] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *NeurIPS*, vol. 35, pp. 2507–2521, 2022.

[67] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[68] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[69] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to learn*, pp. 3–17, Springer, 1998.

[70] H. Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.

[71] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[72] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[73] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid learning or feature reuse? towards understanding the effectiveness of maml," *arXiv preprint arXiv:1909.09157*, 2019.

[74] A. Rajeswaran, C. Finn, S. Kakade, and S. Levine, "Meta-learning with implicit gradients," 2019.

[75] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[76] T. Munkhdalai and H. Yu, "Meta networks," in *ICML*, pp. 2554–2563, PMLR, 2017.

[77] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*, pp. 10657–10665, 2019.

[78] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *NeurIPS*, 2019.

[79] M. Boudiaf, Z. I. Masud, J. Rony, J. Dolz, P. Piantanida, and I. B. Ayed, "Transductive information maximization for few-shot learning," in *NeurIPS*, 2020.

[80] T. Teshima, I. Sato, and M. Sugiyama, "Few-shot domain adaptation by causal mechanism transfer," in *ICML*, pp. 9458–9469, PMLR, 2020.

[81] Z. Xiu, J. Chen, R. Henao, B. Goldstein, L. Carin, and C. Tao, "Supercharging imbalanced data learning with energy-based contrastive representation transfer," in *NeurIPS*, 2021.

[82] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[83] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[84] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *ACL*, 2021.

Appendix

Table of Contents

A Prompt Templates

We present the prompt templates used in our experiments below. Note placeholders such as `<ImageHere>`, `<ImageDescriptionHere>` , `<Answer>`, `<ReasonHere>` will be replaced with actual contents, and highlighted parts (*e.g.*, Group <Answer>.) will be used for the loss computation.

Caption generation:

> ###Human: <Img><ImageHere></Img> Describe this image in detail. ###Assistant:

Caption-guided classification.

> ###Human: You will be given a few different images. Image 1: <Img><ImageHere></Img>; Image 2: <Img><ImageHere></Img>; Image 3: <Img><ImageHere></Img>; ... Image 24: <Img><ImageHere></Img>; Image 25: <Img><ImageHere></Img>. Images 1,2,3,4,5 belong to group 1, images 6,7,8,9,10 belong to group 2, images 11,12,13,14,15 belong to group 3, images 16,17,18,19,20 belong to group 4, images 21,22,23,24,25 belong to group 5. Please assign the following images into the above groups based on their similarity. Here is a description of a new image: <ImageDescriptionHere> Can you identify which group does this image belong to? ###Assistant: Group <Answer>.

Non-verbal classification.

> ###Human: You will be given a few different images. Image 1: <Img><ImageHere></Img>; Image 2: <Img><ImageHere></Img>; Image 3: <Img><ImageHere></Img>; ... Image 24: <Img><ImageHere></Img>; Image 25: <Img><ImageHere></Img>. Images 1,2,3,4,5 belong to group 1, images 6,7,8,9,10 belong to group 2, images 11,12,13,14,15 belong to group 3, images 16,17,18,19,20 belong to group 4, images 21,22,23,24,25 belong to group 5. Please assign the following images into the above groups based on their similarity. Here is a new image: <Img><ImageHere></Img>. Can you identify which group does this image belong to? ###Assistant: Group <Answer>.

Reasoning-guided classification.

###Human: You will be given a few different images. Image 1: <Img><ImageHere></Img>; Image 2: <Img><ImageHere></Img>; Image 3: <Img><ImageHere></Img>; ... Image 24: <Img><ImageHere></Img>; Image 25: <Img><ImageHere></Img>. Images 1,2,3,4,5 belong to group 1, images 6,7,8,9,10 belong to group 2, images 11,12,13,14,15 belong to group 3, images 16,17,18,19,20 belong to group 4, images 21,22,23,24,25 belong to group 5. Please assign the following images into the above groups based on their similarity. Here is a new image: <Img><ImageHere></Img>. Can you identify which group does this image belong to? Group <Answer>. <ReasonHere>

Reasoning generation.

Remark. We have used `Vicuna` V1.5 as our rewriter model, which is using different prompt template compared to the `Vicuna` V0 base model used by MiniGPT4. Specifically, speaker tags are

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: We have a lot of images coming from 50 different classes, where images from the same class share some distinctive common features (each image belongs to only one class). Give a list of detailed description of images from one of the class, you need to summarize the common features for these descriptions in one or two sentence. This summary will be used to instruct data annotators to classify new images to this class. A good example of the summary will be something like "this is the domes/buildings class, the images consistently mention architectural structures, particularly domes, and often provide details of intricate designs, decorations, and historical context."

Here are the image descriptions for you to summarize:
1 <ImageDescriptionHere>
2 <ImageDescriptionHere>
...
5 <ImageDescriptionHere> ASSISTANT:

Selective focusing.

###Human: You will be given a few different images. Image 1: <Img><ImageHere></Img>; Image 2: <Img><ImageHere></Img>; Image 3: <Img><ImageHere></Img>; ... Image 24: <Img><ImageHere></Img>; Image 25: <Img><ImageHere></Img>. Describe image <ID> in detail. ###Assistant: <ImageDescriptionHere>

B Experiment Details

Platform. We use `PyTorch` in our experiments. All training jobs are done on one Nvidia A100-80G GPU, and some of the evaluations are offloaded to Nvidia V100-32G GPU.

Base model. We have used MiniGPT4 for all our fine-tuning experiments. Specifically, we use the original official release built on `Vicuna V0`, and pick the 7B model for quick experimentation. Since MiniGPT4 only saw single pair of text-image data during training, we need to find a model trained on interleaved image-text to baseline few-shot classification accuracy. Since canidate models like `FROZEN`, `Flamingo`, `KOSMOS` are not publicly available, we tested the 9B version of `OpenFlamingo` model [48] built on top of `MPT`.

Training. In our experiments, we freeze all parameters except for the linear adapter. Models are trained for 30 epochs unless otherwise specified, with 3k iterations in each epoch. To avoid OOM error, we restrict the batch size to 1 and use gradient accumulation to update weights every 16 iterations. Cosine learning rate scheduler is used, with 200 warmup steps from 10^{-6} to 3×10^{-5}, then gradually anneal to 10^{-5} towards the end. A weight decay factor of 0.05 is also applied. Inputs

are formatted using the templates from Appendix A, and only the `Assistant` response tokens are used in the loss computation.

Evaluation. Given the limited computation resource we can use, different evaluation strategies have been applied to each experiment. For the main `ImagetNet` and `Cifar` results (Table 1 & 2), we used 600 episodes to compute the accuracy. For the ablation on the task mixing ratio and learning dynamics, we evaluated every three epochs and used 200 episodes for each checkpoint. Note evaluating with 200 episodes has higher variance, so the numbers in Figure 4 is slightly different from the numbers reported in previous tables (which used 600 episodes). But our observations are reliable as we saw consistent trends across different experiments. Finally, for the `Meta Dataset` experiments, we run a total of 1,800 episodes for each checkpoint. The additional episodes evaluated helped us to derive a reliable estimate of the 95% confidence interval. For the curriculum with all `Meta Dataset`, we trained 50 epochs.

C Examples Where Verbal Reasoning Struggle

Figure S1 gave a few examples of image descriptions generated for `Omniglot` characters. These abstract hieroglyph characters are difficult to be described verbally, which explains the bad transfer performance for the models trained on the more narratable `ImageNet`.

Original Minigpt4: The image shows a small black and white picture of a person in profile with their eyes closed. The person is standing in front of a white background with their head tilted to the left and their eyes closed. The image appears to be a simple line drawing of the person.

TieredImagenet trained: This is a simple black and white drawing of a cross. It has a simple design and is easy to recognize. The cross is made up of four intersecting lines, with two vertical lines on the top and two horizontal lines on the bottom. The cross is not very detailed, but it is still easy to recognize.

Ominiglot trained: The image shows an illustration of a horizontal line with a square in the middle. The line has a horizontal extension at one end and a vertical extension at the other end.

Original Minigpt4: This image appears tend. The triangle has one long side and two short sides, with the point facing downwards. It could be interpreted as a symbol for stability or balance, as it has a solid, structured form. The triangle could also be seen as a representation of the human body, with the long side representing the spine and the short sides representing the limbs. However, without more context, it is difficult to determine the exact meaning or significance of this image.o be a black and white depiction of a triangular shape with a small, pointed

TieredImagenet trained: The image shows a letter 'a' written in the Arabic script.

Ominiglot trained: The image shows a triangle made out of two lines. The lines are connected at their tips and have one point on top. The point is not visible in the image. The triangle could be used to represent a tent, a pyramid, or a point in space.

Figure S1: Image descriptions for `Omniglot` characters generated by different models.

D Related Works

To give readers a more comprehensive view of the fields, we briefly review related literature here.

Multimodal LLM. The unification of architectures and the growing availability of pre-trained foundational models has catalyzed surging interests in general purpose large multimodal foundation models. Substantial progress has been made in the past few years. Pioneering works such as CLIP [49] leverage rich multimodal data for self-supervised representation learning, which lays the foundation for various multimodal applications [23]. Note CLIP learning essentially optimizes the mutual information between text and visual representations, and its learning efficiency can be further enhanced via leveraging more advanced contrastive learning algorithms [50–52]. While pioneering investigations on augmenting LLM with visual perception are very compute intensive [16, 53, 17, 18], they demonstrated the potentials of leveraging visual augmented LLM as a general purpose tooling for diverse visual language tasks (*e.g.*, VQA, captioning, chat, creative writing, *etc.*). GPT4 further showed how visual understanding can multiply powerful LLM to massively boost productivity [15, 54]. Following that, the adoption of adapter architectures, the availability of

powerful open LLMs and visual instruction tuning recipes [55–57] have made multimodal LLM more accessible for academic studies and practical deployments [19, 58–65]. Relevant to our work are the works of [66, 67], where chain-of-thought reasoning are applied for question answering in multimodal settings. Note [16, 53, 17] also presented few-shot classification examples with multimodal LLM, but those are simple proof-of-concept demonstrations; as demonstrated in our work, such emergent capabilities break down in the face of complex scenes and dense visual inputs.

Few-shot learning. Quick and robust learning from limited examples has been a long standing challenging for CV [5, 68]. It can be framed under *meta-learning* which broadly refers to techniques that facilitate the learning of a new task in sample-scarce scenarios using prior knowledge from previously seen tasks, such as *learning to learn* [69], *domain adaptation* [70], *transfer learning* [71], *causal machine learning* [10], *zero/few-shot learning* [72], *weakly supervised learning* [9], *etc.* Here we focus on methods that are developed in the deep learning era. Works such as `MAML` [7, 73–75], `MetaNet` [76], and `FewTURE` [35] builds a meta-learned a base model during episodic training then adapt to a task-specific model at inference time using gradient updates. Methods like `MatchingNet` [6], `ProtoNet` [24], and `FEAT` [27] applied different notations of affinity metrics to align query samples with support sets in the neural embedding spaces, while `MetaOptNet` [77] advocates the use of hyperplanes to separate classes. Motivated by the success of transformers, attention based few shot learners have also been proposed [78, 40]. Other interesting directions include using Info-Max regularization [79] or causal representation learning [80, 81] to address the sample scarcity issue.

Adaptation. Inference time adaptation is important for improving low-shot learning performance [12]. In the context of few-shot image classification, task adaptation can be considered as a procedure to reduce cross-class similarities and highlight discriminative features for the given task. This is often achieved by computing the class-prototypes [24], or applying (several) gradient updates to edit the input weights [35] or model parameters [7]. Applying closed-form or iterative solvers can further boost the efficiency and efficacy of the gradient-based adaptations [37, 77]. A major drawback with the gradient-based adaptation is that error back-propagation is costly especially for large networks.

For LLMs, task adaptation can be achieved more flexibly via prompting [82, 72, 83]. By consuming proper prompts (such as task examples, instructions, or combination of both), language models are better conditioned for individual tasks, thus yielding substantial performance gains for zero/few-shot applications [84]. This is more appealing operational wise as a single generalist model can simultaneously serve many different tasks. Prior studies have shown vision-adapted LLMs also exhibit emergent few-shot learning capability on diverse vision-language tasks [16, 53, 17].