

LEAML: Label-Efficient Adaptation to Out-of-Distribution Visual Tasks for Multimodal Large Language Models

Ci-Siang Lin¹ Min-Hung Chen² Yu-Yang Sheng¹ Yu-Chiang Frank Wang^{1,2}
¹Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
²NVIDIA

Abstract

Multimodal Large Language Models (MLLMs) have achieved strong performance on general visual benchmarks but struggle with out-of-distribution (OOD) tasks in specialized domains such as medical imaging, where labeled data is limited and expensive. We introduce LEAML, a label-efficient adaptation framework that leverages both scarce labeled VQA samples and abundant unlabeled images. Our approach generates domain-relevant pseudo question-answer pairs for unlabeled data using a QA generator regularized by caption distillation. Importantly, we selectively update only those neurons most relevant to question-answering, enabling the QA Generator to efficiently acquire domain-specific knowledge during distillation. Experiments on gastrointestinal endoscopy and sports VQA demonstrate that LEAML consistently outperforms standard fine-tuning under minimal supervision, highlighting the effectiveness of our proposed LEAML framework.

Introduction

Large Language Models (LLMs) (Touvron et al. 2023; Achiam et al. 2023; Bai et al. 2023) have demonstrated impressive capabilities across diverse language tasks. By incorporating visual understanding, Multimodal Large Language Models (MLLMs) (Alayrac et al. 2022; Liu et al. 2023a; Wang et al. 2024a) extend these capabilities to visual question answering (VQA), image captioning, and multimodal reasoning tasks. (Wang et al. 2024b) Recent MLLMs have achieved remarkable performance on general visual benchmarks, showing strong generalization within their training distributions. Although scaling laws (Kaplan et al. 2020) show promise for enhancing model performance, they do not eliminate dependence on pre-training data, which limits generalization to novel domains. In real-world deployment, MLLMs will face domain-specific tasks that fall outside their training distribution, such as specialized medical imaging or technical visual content. When confronted with such out-of-distribution (OOD) data, these models often produce erroneous or unreliable outputs, limiting their applicability in specific domains where accuracy is critical.

While existing studies on OOD have primarily focused on detection (Ming et al. 2022; Wang et al. 2023) or domain-specific classification tasks (Liu et al. 2023b; Lei et al. 2023; Zhang et al. 2023a), applying MLLMs to visual question answering in unfamiliar domains introduces

distinct challenges. Unlike detection, which focuses on filtering unfamiliar inputs, VQA requires models to generate accurate and contextually grounded answers for images outside the training distribution. Although many specific domains offer abundant unlabeled images, standard VQA training pipelines depend on paired question-answer annotations, which are costly to obtain due to the need for domain expertise. Furthermore, VQA questions must be carefully constructed based on the visual content. Randomly pairing unrelated questions with images creates confusing training signals and reduces the effectiveness of learning. Due to the scarcity of high-quality labeled data, fully fine-tuning MLLMs often leads to severe overfitting, particularly in VQA where models must produce free-form, semantically appropriate responses rather than select from fixed labels. These challenges underscore the need for adaptation methods that can leverage both limited labeled examples and large pools of unlabeled images in a way that preserves generation capabilities while acquiring domain-specific understanding.

In this paper, we introduce **LEAML**, a two-stage Label-Efficient Adaptation framework for MultiModal LLM designed to effectively leverage both the limited labeled VQA data and the abundant unlabeled images in the target domain. The framework comprises Pseudo QA Generation, which constructs domain-relevant QA pairs from unlabeled images, and OOD VQA Finetuning, which fine-tunes the target MLLM using both the generated pairs and the original labeled data. In Pseudo QA Generation, a QA Generator is trained using the small labeled dataset to capture the domain-specific patterns of question-answer formulations. To mitigate overfitting caused by the scarcity of labeled data, the generator is regularized via caption distillation, where it additionally learns from image captions produced by a large-scale model, providing broader visual-linguistic signals beyond the limited annotations. The generator then synthesizes diverse pseudo QA pairs for the unlabeled domain images, creating a significantly richer training resource. In OOD VQA Finetuning, the target MLLM is fine-tuned with both the original labeled samples and the generated QA pairs. To enhance adaptation efficiency and prevent overfitting, we employ Selective Neuron Distillation, which is motivated by the insight that domain-specific knowledge is often encoded in a subset of neurons; focusing updates on these neurons en-

ables the model to acquire domain-relevant reasoning capabilities while preserving general language generation skills. Experiments conducted on gastrointestinal endoscopy VQA demonstrate that our method significantly improves performance compared to standard fine-tuning approaches.

Our contributions can be summarized as follows:

- We propose a two-stage learning framework LEAML, which leverages both scarce labeled data and abundant unlabeled images to achieve MLLM adaptation for domain-specific VQA .
- We introduce Pseudo QA Generation, which learns a generator to produce pseudo question-answer pairs for unlabeled data, augmenting the training set for finetuning the VQA model on specialized domains.
- We design Selective Neuron Distillation for the QA generator, which performs captioning distillation to acquire domain-related knowledge while selectively updates QA-related neurons, resulting in reliable pseudo QA pairs for finetuning.

Related Works

Out-of-Distribution Data Learning

Out-of-distribution (OOD) data refers to inputs that differ significantly from those seen during training. Early OOD research (Ming et al. 2022; Wang et al. 2023) primarily focused on detection tasks, where the goal is to identify whether an input falls outside the training distribution. These methods often treat OOD detection as anomaly detection. However, these methods aim to identify and reject OOD inputs rather than adapt models to perform well on them. Beyond detection, another line of work explores how vision-language models can be adapted for OOD domains. CLIP and its variants have shown promise in medical imaging applications through domain-specific fine-tuning (Liu et al. 2023b; Lei et al. 2023; Zhang et al. 2023a). However, these approaches are limited to classification settings, where models predict labels from a fixed set rather than generating open-ended responses.

Recent works explore adapting LLMs to specific domains (Zhang et al. 2023b; Cheng et al. 2024; Bhatia et al. 2024; Cheng, Huang, and Wei 2023). For instance, (Zhang et al. 2023b; Kim et al. 2025) employ retrieval-based approaches to handle out-of-distribution inputs, relying on external sources to provide relevant context for domain adaptation. However, such methods still assume that the LLM has sufficient foundational knowledge to effectively interpret the retrieved information. In the vision-language domain, (Cheng et al. 2024) propose fine-tuning MLLMs to automatically generate question-answer pairs from existing image-caption datasets, subsequently using these synthetic QA pairs for domain-specific VQA training. However, the effectiveness of such methods depends heavily on both the LLM’s domain knowledge and the granularity of source captions. Fine-grained captions enable detailed QA generation, while coarse-grained descriptions yield only generic questions. These factors highlight the need for methods that can produce high-quality domain-specific QA pairs even with limited domain knowledge and varying caption quality.

Semi-Supervised Learning

Semi-supervised learning (SSL) is a strategy to bridge the gap between limited labeled data and abundant unlabeled resources, especially in domains where annotation is expensive or requires domain expertise. Among existing SSL paradigms, pseudo-labeling (Lee et al. 2013; Xie et al. 2020) has been widely adopted. This approach trains models on labeled data, then uses them to pseudo-label unlabeled samples for further training.

Pseudo-labeling has been widely applied to tasks such as image classification (Zeng et al. 2023) and Segmentation (Yang et al. 2023), where labels are either discrete or can be directly generated from the data. In these settings, unlabeled samples can be automatically annotated with relatively reliable supervision. In contrast, applying pseudo-labeling to visual question answering (VQA) is considerably more difficult. VQA requires not only a grounded answer but also a relevant and context-aware question that must be closely aligned with the visual content. Unlike captions or class labels, such question-answer pairs cannot be directly inferred from the image alone. A naïve solution might involve randomly assigning questions to domain images, but this leads to incoherent or misleading training data. As a result, semi-supervised learning remains largely underexplored in VQA, particularly in domain-specific scenarios where labeled data is scarce and generating valid pseudo QA pairs is highly non-trivial.

Neuron-Level Knowledge Attribution in DNNs

Knowledge in neural networks is localized and stored within specific neural components rather than distributed uniformly across parameters. Previous work has shown that the feed-forward network (FFN) layers in Transformers play a key role in storing knowledge (Geva et al. 2020, 2022). These layers have been characterized as performing additive, knowledge-based updates on token representations. (Dai et al. 2021) propose a neuron-level attribution approach to identify knowledge neurons in FFNs responsible for storing factual knowledge. These findings indicate that MLLMs encode different types of knowledge in specialized neural components with distinct activation patterns, enabling targeted manipulation of specific knowledge without affecting the entire network.

To support such targeted interventions, a variety of attribution methods have been proposed to quantify the importance of individual neurons or weights. Some are designed to guide pruning, which reduces model size by removing components deemed less relevant. For example, Wanda (Sun et al. 2023) identifies unimportant weights by combining their magnitude with the norm of the associated input activation. Others (Fang et al. 2024; Pan et al. 2023; Yu and Ananiadou 2023, 2025) aim to inform fine-tuning, where only selected parts of the model are updated to acquire new capabilities or adapt to specific domains. A common strategy in this context is to measure gradient magnitudes during back-propagation, under the assumption that neurons with higher gradients contribute more significantly to the output (Zhang et al. 2024).

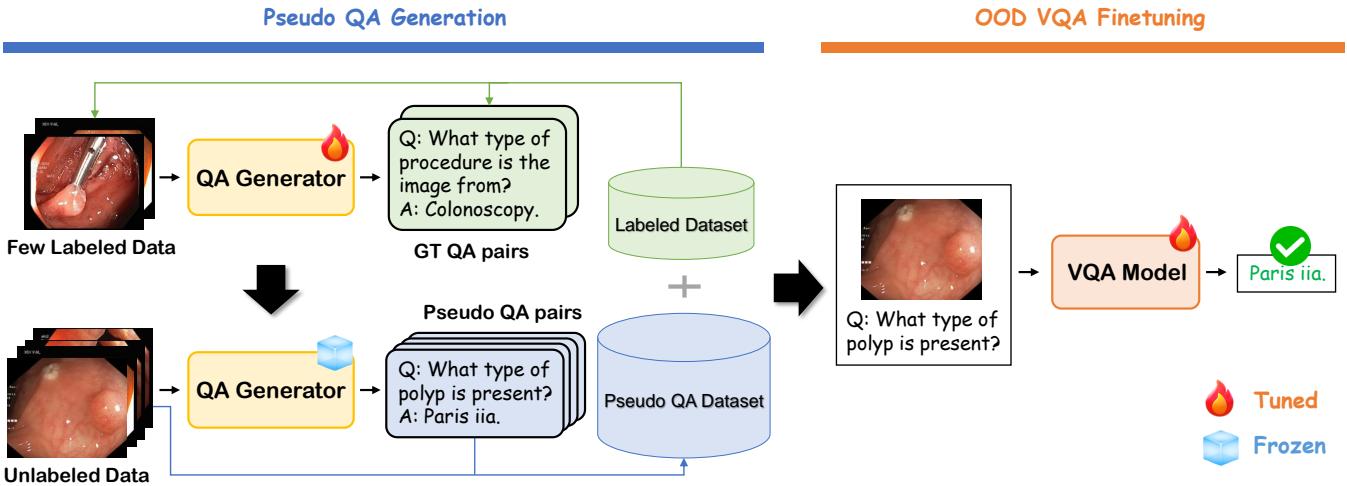


Figure 1: Overview of the proposed two-stage LEAML framework for OOD VQA adaptation. In Pseudo QA Generation, the QA Generator is trained using a small set of labeled question-answer pairs and then used to generate pseudo QA pairs for a large collection of unlabeled images. In OOD VQA Finetuning, the VQA model is fine-tuned with both the original labeled data and the produced pseudo QA pairs of unlabeled data, enabling label-efficient adaptation to out-of-distribution visual-question answering. We will detail the learning of our QA Generator in Figure 2.

Method

Problem Formulation and Framework Overview

Problem Formulation. We first define the problem settings and notations used in this paper. For the out-of-distribution visual question-answering problem, we consider domain-specific tasks (e.g., gastrointestinal endoscopy) which are still based on RGB images or videos but are rarely or not covered in the pretraining data of general-purpose multimodal large language models. Since such domain-specific tasks generally require great expense to obtain manual annotations from domain experts, we further consider a realistic yet challenging setting, where only few data are annotated in the training dataset D . That is, the training dataset D contains a set of labeled data D_l of few samples and also a set of abundant unlabeled data D_u , which is similar to traditional semi-supervised learning. Each data instance in the labeled dataset D_l contains an image or video V , an associated question Q , and the corresponding answer A , while each sample in the unlabeled dataset D_u contains only the visual part V .

Framework Overview. Given the above training data and a pretrained general MLLM, our goal is to adapt this pretrained MLLM to address out-of-distribution visual question-answering with only few labeled instances. To achieve this goal, we propose a two-stage learning framework LEAML as shown in Figure 1. Our LEAML framework includes two stages: Pseudo QA Generation and OOD VQA Finetuning. The former stage aims to generate proper pseudo question-answer pairs for the unlabeled dataset D_u , while the latter stage leverages both the labeled dataset D_l and the produced pseudo QA pairs for the unlabeled data D_u to finetune a pretrained MLLM for addressing out-of-

distribution visual question-answering.

Specifically, since the unlabeled data only contain the visual information but lack the corresponding textual annotations for training a VQA model, our LEAML framework first employ a QA Generator G which is directly supervised by the labeled data to produce pseudo QA pairs for unlabeled data. However, in this manner, the QA Generator may overfit on those few labeled samples and fail to generalize well to produce reliable QA pairs for unlabeled data. To overcome this challenge, we further propose Selective Neuron Distillation to leverage also the unlabeled data during the training of the QA Generator. As shown in Figure 2, in addition to the QA generation supervised by labeled data, we perform captioning distillation using unlabeled data to distill related knowledge from a large MLLM. More importantly, during training, we update only QA-related neurons for both the QA generation and captioning distillation. In this way, such captioning distillation is designed to focus solely on gaining QA-related knowledge, and hence the QA Generator G is able to produce reliable pseudo QA pairs for unlabeled data during inference, benefiting the following finetuning for the VQA model. We now detail our learning framework in the following subsections.

Pseudo QA Generation and OOD VQA Finetuning

Recently, multimodal large language models (MLLMs) have achieved impressive generalization on a wide range of vision-language tasks. However, their performance often degrades when applied to specialized domains such as medical imaging, where domain-specific visual cues and terminology differ significantly from pretraining data. A significant barrier in adapting multimodal large language models (MLLMs) to these specialized domains is the extreme

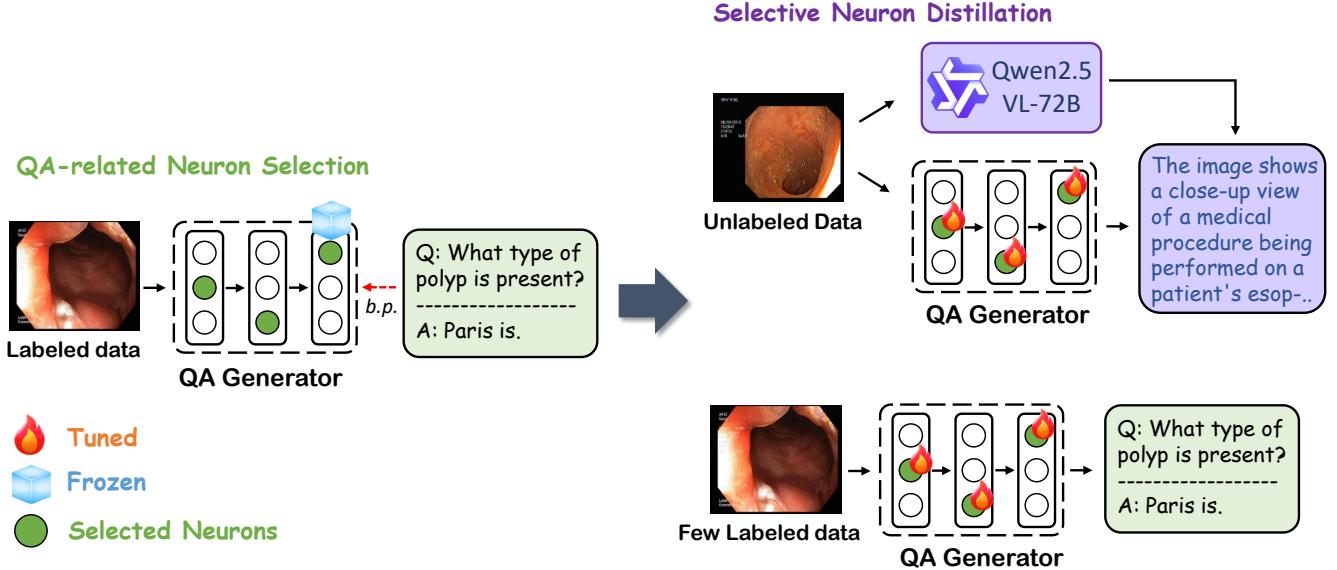


Figure 2: Illustration of our Selective Neuron Distillation for the QA Generator. The QA-relevant parameters are first selected based on gradient scores from labeled QA data. During training, only these selected parameters are updated using auxiliary caption supervision from unlabeled images, allowing QA-related knowledge distillation for the QA Generator.

scarcity of annotated question-answer (QA) pairs. Obtaining such labeled data from domain experts is often expensive, time-consuming, and infeasible at scale, leading to severe limitations on direct fine-tuning. On the other hand, large volumes of unlabeled visual data typically exist, representing a potential yet largely untapped source of supervision. While traditional semi-supervised learning have demonstrated that unlabeled data can be highly beneficial, such success is largely limited to classification tasks. For unlabeled data in visual question-answering, both the question and answer are not accessible, and hence traditional semi-supervised learning cannot be easily applied in VQA to generate the pseudo labels without knowing or providing a corresponding question.

To overcome the limitation of requiring fully annotated data, we introduce a pseudo QA generation stage that transforms unlabeled visual data into useful supervision for the VQA model. The core idea is to train a QA Generator using limited labeled examples D_l , and then apply it to unlabeled inputs D_u to produce additional VQA samples. Specifically, for each labeled instance (V, Q, A) in D_l , we formulate the corresponding target text sequence Y by concatenating of the question Q and answer A with special tokens:

$$Y = [<\text{q}>Q<\text{q}><\text{a}>A<\text{a}>]. \quad (1)$$

Then, the QA generator G , initialized from a pretrained general MLLM, is trained to autoregressively generate Y conditioned on the visual input V . The training objective \mathcal{L}_{QA} is defined as a standard autoregressive negative log-likelihood:

$$\mathcal{L}_{QA} = \sum_{t=1}^{|Y|} \log p(y_t | V, y_{<t}), V \in D_l \quad (2)$$

where y_i denotes the i th text token of the sequence Y and p is the likelihood of text tokens estimated by the QA Generator G . Once the training is complete, the QA generator G would be applied to produce pseudo QA pairs $[<\text{q}>\hat{Q}<\text{q}><\text{a}>\hat{A}<\text{a}>] = G(V)$ for the visual input V from unlabeled data, resulting in the dataset $\hat{D}_u = \{(V, \hat{Q}, \hat{A})\}$. This dataset is then used to augment the original labeled data D_l for finetuning the final VQA model with similar autoregressive objective \mathcal{L}_{VQA} :

$$\mathcal{L}_{VQA} = \sum_{t=1}^{|A|} \log p(a_t | V, Q, a_{<t}), V \in D_l, \hat{D}_u \quad (3)$$

Note that the VQA model is again initialized from a pre-trained general MLLM. While the above pseudo QA generation enables the model to leverage unlabeled visual data for OOD VQA adaptation, the quality of generated QA pairs is tightly bound to the small number of labeled examples. With limited supervision, the model may produce noisy outputs on unlabeled data, especially under the out-of-distribution settings.

Selective Neuron Distillation for QA Generator

To further mitigate the aforementioned problem of Pseudo QA Generation, our LEAML framework introduces a distillation mechanism that directly incorporates abundant unlabeled data into the QA Generator training process. In addition to the QA generation part supervised by labeled data D_l , we further consider captioning distillation on unlabeled D_u to gain domain-related knowledge from a large MLLM. Specifically, for each unlabeled visual input V , we employ

a large pre-trained general MLLM to automatically generate a descriptive caption C . While the large MLLM may still be poor when directly applied to perform OOD VQA tasks, these state-of-the-art large models are pre-trained on vast and diverse corpora such as medical literature, Wikipedia articles, and other authoritative web resources. As a result, the captions generated by such models are often infused with rich semantic information, domain-specific terminology, and contextual knowledge, which are beneficial to the learning of the QA Generator G . Formally, the QA Generator G is jointly optimized with the QA objective \mathcal{L}_{QA} in Equation 2 using labeled data D_l and the captioning distillation objective \mathcal{L}_C using unlabeled data D_u :

$$\begin{aligned}\mathcal{L}_C &= -\sum_{t=1}^{|C|} \log p(c_i | V, c_{<i}), V \in D_u \\ \mathcal{L}_G &= \mathcal{L}_{QA} + \mathcal{L}_C\end{aligned}\quad (4)$$

While caption distillation enables the QA Generator G to absorb rich semantic knowledge from unlabeled data, it remains essential that the model’s capacity is focused on generating question-answer pairs rather than captioning. Under this motivation, we introduce a neuron selection strategy to ensure such captioning distillation is solely for enhancing QA-related knowledge. Specifically, we identify parameters in the QA Generator that contribute most to the question-answer generation and restrict parameter updates during training to this subset. For each parameter θ in the QA Generator G , its importance score s is quantified by the average magnitude of the gradient of the QA loss \mathcal{L}_{QA} :

$$s = \left| \frac{1}{|D_l|} \sum_{(V,Q,A) \in D_l} \frac{\partial \mathcal{L}_{QA}}{\partial \theta} \right|. \quad (5)$$

Then, we select and update only parameters with top- K scores in each neuron (i.e., each row in a linear weight matrix) while keep others frozen:

$$\theta \leftarrow \begin{cases} \theta - \eta \cdot \frac{\partial \mathcal{L}_G}{\partial \theta} & \text{if } s \text{ is among top-}K \\ \theta, & \text{otherwise} \end{cases} \quad (6)$$

By restricting updates to only the most QA-relevant neurons or parameters, we ensure that the auxiliary knowledge gained from caption distillation is efficiently integrated to support question-answer generation, rather than generic captioning or unrelated model capacities. With the above learning, the QA Generator is not limited to patterns seen in the small labeled set, but is instead exposed to the full diversity and richness of the target domain through the captions of unlabeled images. This enables the QA Generator to produce question-answer pairs that are more reflective of domain-specific semantics present in the unlabeled corpus. Hence, the resulting pseudo dataset \hat{D}_u would be more accurate and reliable with our proposed Selective Neuron Distillation, benefiting the subsequent learning of the VQA model in Equation 3.

Experiments

Datasets

Kvasir-VQA Kvasir-VQA (Gautam et al. 2024) is a large-scale visual question answering dataset curated for research in the gastrointestinal (GI) medical imaging domain. The dataset consists of 6,500 endoscopic images sourced from the HyperKvasir and Kvasir-Instrument datasets, covering a wide spectrum of clinically relevant GI findings, anatomical sites, and medical instruments. The images in Kvasir-VQA reflect real-world clinical scenarios, including both normal findings and a variety of pathological conditions such as polyps, ulcers, and esophagitis, as well as procedure-related scenes featuring different instruments and interventions. Each image is paired with one or more expert-annotated question-answer (QA) pairs. These QA annotations are diverse, encompassing multiple question types including yes/no, multiple-choice, and counting, thus enabling comprehensive evaluation of both recognition and reasoning abilities for VQA.

To standardize the evaluation process, we further convert all question-answer pairs in Kvasir-VQA into a multiple-choice format. For each question, we define the set of candidate options as all possible answers that appear for that question type throughout the dataset, so that the number of answer choices is not fixed for different questions. For experimental setup, we partition the dataset into training and testing splits, containing 18,499 and 18,075 QA pairs, respectively. To simulate the limited annotation scenario when learning our LEAML framework, only 1% QA pairs in the training split are used as labeled data while the remaining training images are treated as unlabeled data.

SPORTU SPORTU (Xia et al. 2024) is a recently released benchmark designed to evaluate the sports understanding and reasoning abilities of multimodal large language models (MLLMs). The dataset comprises 1,701 slow-motion sports video clips, spanning seven popular sports: *American football, badminton, baseball, basketball, ice hockey, soccer, and volleyball*. The questions in SPORTU are diverse and challenging, covering rule comprehension, tactical analysis, prediction of outcomes, and recognition of actions and fouls. All questions are categorized into three levels of difficulty: **easy**, which focuses on basic recognition tasks such as identifying the sport or counting players; **medium**, which requires knowledge of player roles and basic tactics; and **hard**, which involves deep reasoning about rules, foul detection, and scenario-based understanding. This tiered design enables comprehensive evaluation of MLLMs, from simple perception (in-distribution) to advanced, domain-specific sports reasoning (out-of-distribution). We partition the dataset into training and testing splits, containing 5,525 and 5,478 QA pairs, respectively. Similarly, we only consider 1% QA pairs in the training split when learning our LEAML framework, while the remaining training images are treated as unlabeled data.

Implementation Details

Our entire implementation is based on the PyTorch framework. For simplicity, we use NVILA-Lite-2B (Liu et al.

Table 1: Quantitative results of different learning strategies on the Kvasir-VQA dataset. All scores are reported in percentage (%) of VQA accuracy. Only 1% data are labeled during training and we use NVILA-Lite-2B as the MLLM backbone.

Method	Image Category				Average
	Colitis	Esophagitis	Instrument	Polyps	
Zero-Shot	35.4	21.1	49.0	47.8	38.3
LoRA	95.6	57.7	48.9	47.3	62.4
Full-Tuning	89.0	60.8	51.2	51.3	63.1
LEAML (Ours)	95.8	83.9	61.6	65.5	76.7
Zero-Shot (Qwen2.5-VL-72B)	59.7	48.6	53.1	55.9	54.3
Fully-Supervised	97.4	97.4	85.8	82.1	90.7

Table 2: Quantitative results of different learning strategies on the SPORTU dataset. All scores are reported in percentage (%) of VQA accuracy. “Easy”, “Medium”, and “Hard” denote different level of difficulty for sport understanding and reasoning questions. Only 1% data are labeled during training and we use NVILA-Lite-2B as the MLLM backbone.

Method	Easy	Medium	Hard	Average
Zero-Shot	50.5	34.0	37.5	40.7
LoRA	82.5	59.2	22.3	54.7
Full-Tuning	82.7	59.0	21.3	54.3
LEAML (Ours)	82.5	60.4	46.3	63.1
Fully-Supervised	98.1	75.4	66.1	79.9

2025) as our MLLM backbone for both our QA Generator and the VQA model. As for the large MLLM used for captioning distillation, we consider the state-of-the-art open source model, Qwen2.5-VL-72B-Instruct (Bai et al. 2025). All the model weights are initialized with official pretrained checkpoints. During training, we use the AdamW optimizer with an initial learning rate of 0.00001 and a cosine annealing schedule for learning rate decay. The batch size is set as 16. As for neuron selection, we set the number of parameter $K = 1000$ on the Kvasir-VQA dataset and $K = 3000$ for SPORTU, respectively. The training is performed on 16 NVIDIA A100 GPUs with 80GB memory each. During inference, we use deterministic (greedy) decoding for all answer generation for the VQA model, i.e., at each generation step, the model always selects the token with the highest probability. This ensures that the outputs are fully reproducible and comparable across runs. As for pseudo QA Generation, we choose to use nucleus sampling on our QA Generator to produce several different question-answer pairs for each single visual input.

Quantitative and Qualitative Experiments

While multimodal large language models (MLLMs) have demonstrated remarkable abilities on general visual understanding benchmarks, their robustness and adaptability often fall short when transferred to specialized domains such as medical image question answering. This gap arises primarily from two factors: the limited availability of expert-annotated data, and the significant distribution shift between general pretraining data and domain-specific imagery. In the context of medical imaging, acquiring comprehensive labeled datasets is both costly and time-consuming, which mo-

tivates the need for approaches that can effectively leverage abundant unlabeled data. To evaluate our method under such kind of scenarios, we design experiments on the Kvasir-VQA dataset, a large-scale benchmark of expert-annotated QA pairs paired with real-world endoscopic images.

Table 1 presents a detailed comparison of different learning strategies on the Kvasir-VQA datasets, where only 1% of the training data is labeled and the remaining ones are treated as unlabeled. From this Table, we see that zero-shot inference using NVILA-Lite-2B achieves only 38.3% average accuracy, indicating the difficulty of the task without any fine-tuning. Fine-tuning with LoRA (Hu et al. 2022) or full parameter updates on just 1% labeled data yields moderate improvements (62.4% and 63.1% accuracy, respectively), but these methods still struggle on less-represented or out-of-distribution categories. By augmenting the training with pseudo QA pairs generated from the unlabeled images, our approach significantly boosts average accuracy to 76.7%, especially on challenging categories such as Esophagitis. As a reference, the fully supervised model (using all labeled data) achieves 90.7%.

We further evaluate our method on the SPORTU dataset to assess its generalization ability in the sports domain, which poses unique challenges compared to the medical setting. Unlike static medical images, SPORTU consists of dynamic sports video clips that require not only visual recognition but also temporal reasoning, action understanding, and comprehension of complex game rules. These factors make sports VQA a particularly demanding task for multimodal large language models. As shown in Table 2, our LEAML framework achieves consistent improvements over all baselines. In particular, our method raises the average ac-

Table 3: Ablation study of our Selective Neuron Distillation for the QA Generator on the Kvasir-VQA dataset.

Method	Image Category				Average
	Colitis	Esophagitis	Instrument	Polyps	
Baseline	93.0	88.2	55.7	56.2	73.3
Baseline+Distill.	95.1	80.3	59.7	60.4	73.9
Baseline+Distill.+QA Neurons	95.8	83.9	61.6	65.5	76.7

Table 4: Ablation study of our Selective Neuron Distillation for the QA Generator on the SPORTU dataset.

Method	Easy	Medium	Hard	Average
Baseline	75.5	60.3	44.4	60.1
Baseline+Distill.	82.9	61.2	42.3	62.2
Baseline+Distill.+QA Neurons	82.5	60.4	46.3	63.1

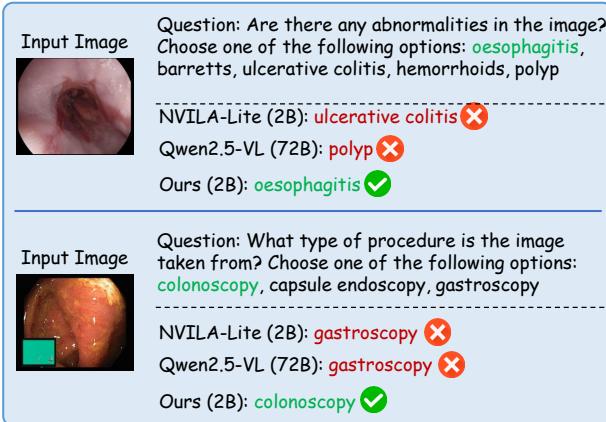


Figure 3: Qualitative results on the Kvasir-VQA dataset.

curacy to 63.1%, outperforming both LoRA and full-tuning by a clear margin. Notably, the accuracy on “Hard” questions—which require advanced reasoning—improves significantly from around 22.3% to 46.3%. In addition to quantitative results, we also provide qualitative comparisons as shown in Figure 9 and 4. We see that, our LEAML framework is able to produce accurate answers compared to state-of-the-art MLLMs on the challenging medical and sport domains. Through the above experiments, we verify that our LEAML framework, which directly incorporates unlabeled data through pseudo QA generation, can substantially enhance VQA performance on out-of-distribution domain with limited annotations.

Ablation Studies

As shown in Table 3, our ablation study on Kvasir-VQA evaluates the effects of caption distillation and selective neuron updates. Caption distillation alone leads to only a minor increase in average accuracy (an improvement of just 0.6% over the baseline). In contrast, when selective neuron updates are applied, the accuracy increases substantially, yielding a total gain of 3.4% over the baseline. A similar trend is observed on the SPORTU dataset (Table 4). These results highlight that targeted neuron selection is crucial for effec-

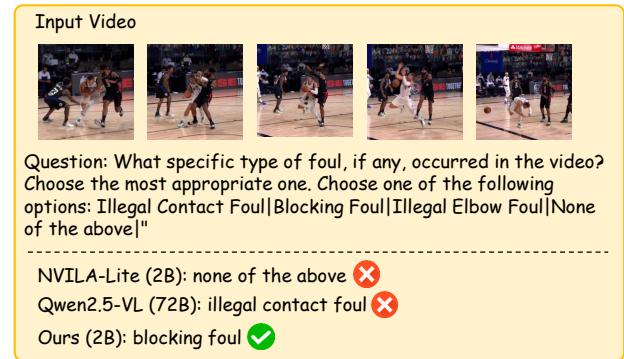


Figure 4: Qualitative results on the SPORTU dataset.

tive distillation for the QA Generator, and that substantial performance gains can only be achieved when both components are used together.

Conclusion

In this paper, we introduce LEAML, a label-efficient adaptation framework designed to transfer MLLMs to out-of-distribution (OOD) domains that extend far beyond their original pretraining distribution. Our approach first utilizes Pseudo QA Generation to generate domain-relevant pseudo question-answer pairs from large pools of unlabeled visual data, effectively expanding the training set in scenarios where annotated examples are extremely limited. In addition, we employ Selective Neuron Distillation for the QA Generator, a selective neuron updating strategy that identifies and updates only knowledge-relevant neurons while acquiring domain knowledge from state-of-the-art large MLLMs. We conduct extensive experiments on OOD benchmarks, including gastrointestinal endoscopy and sports VQA, demonstrating the applicability of our framework. Detailed ablation studies and both quantitative and qualitative analyses consistently show that our proposed LEAML framework achieves substantial improvements over conventional fine-tuning methods, validating its effectiveness and robustness for adapting MLLMs to challenging domain-specific visual tasks with limited annotations.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Bhatia, G.; Nagoudi, E. M. B.; Cavusoglu, H.; and Abdul-Mageed, M. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv preprint arXiv:2402.10986*.
- Cheng, D.; Huang, S.; and Wei, F. 2023. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.
- Cheng, D.; Huang, S.; Zhu, Z.; Zhang, X.; Zhao, W. X.; Luan, Z.; Dai, B.; and Zhang, Z. 2024. On Domain-Adaptive Post-Training for Multimodal Large Language Models. *arXiv preprint arXiv:2411.19930*.
- Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.
- Fang, J.; Bi, Z.; Wang, R.; Jiang, H.; Gao, Y.; Wang, K.; Zhang, A.; Shi, J.; Wang, X.; and Chua, T.-S. 2024. Towards neuron attributions in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37: 122867–122890.
- Gautam, S.; Storås, A. M.; Midoglu, C.; Hicks, S. A.; Tham-bawita, V.; Halvorsen, P.; and Riegler, M. A. 2024. Kvadir-vqa: A text-image pair gi tract dataset. In *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, 3–12.
- Geva, M.; Caciularu, A.; Wang, K. R.; and Goldberg, Y. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.
- Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Kim, K.; Park, G.; Lee, Y.; Yeo, W.; and Hwang, S. J. 2025. VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3295–3305.
- Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.
- Lei, Y.; Li, Z.; Shen, Y.; Zhang, J.; and Shan, H. 2023. Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 403–412. Springer.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, J.; Zhang, Y.; Chen, J.-N.; Xiao, J.; Lu, Y.; A Landman, B.; Yuan, Y.; Yuille, A.; Tang, Y.; and Zhou, Z. 2023b. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 21152–21164.
- Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. 2025. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4122–4134.
- Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35: 35087–35102.
- Pan, H.; Cao, Y.; Wang, X.; Yang, X.; and Wang, M. 2023. Finding and editing multi-modal neurons in pre-trained transformers. *arXiv preprint arXiv:2311.07470*.
- Sun, M.; Liu, Z.; Bair, A.; and Kolter, J. Z. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Wang, H.; Li, Y.; Yao, H.; and Li, X. 2023. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1802–1812.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, Y.; Chen, W.; Han, X.; Lin, X.; Zhao, H.; Liu, Y.; Zhai, B.; Yuan, J.; You, Q.; and Yang, H. 2024b. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.
- Xia, H.; Yang, Z.; Zou, J.; Tracy, R.; Wang, Y.; Lu, C.; Lai, C.; He, Y.; Shao, X.; Xie, Z.; et al. 2024. Sportu: A comprehensive sports understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2410.08474*.

- Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698.
- Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7236–7246.
- Yu, Z.; and Ananiadou, S. 2023. Neuron-level knowledge attribution in large language models. *arXiv preprint arXiv:2312.12141*.
- Yu, Z.; and Ananiadou, S. 2025. Locate-then-Merge: Neuron-Level Parameter Fusion for Mitigating Catastrophic Forgetting in Multimodal LLMs. *arXiv preprint arXiv:2505.16703*.
- Zeng, Q.; Xie, Y.; Lu, Z.; and Xia, Y. 2023. Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15671–15680.
- Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. 2023a. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.
- Zhang, Y.; Wang, Y.; Cheng, F.; Kurohashi, S.; et al. 2023b. Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on Chinese legal domain. *arXiv preprint arXiv:2310.03328*.
- Zhang, Z.; Zhang, Q.; Gao, Z.; Zhang, R.; Shutova, E.; Zhou, S.; and Zhang, S. 2024. Gradient-based parameter selection for efficient fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28566–28577.

Input Image



Question: What type of polyp is present? Choose one of the following options:
paris is|none|paris iia|paris ip|
NVILA-Lite (2B): paris ip ✗
Qwen2.5-VL (72B): paris iia ✗
Ours (2B): none ✓

Input Image



Question: What is the size of the polyp? Choose one of the following options:
5-10mm|>20mm|<5mm|none|11-20mm|>20|
NVILA-Lite (2B): <5mm ✗
Qwen2.5-VL (72B): none ✗
Ours (2B): >20mm ✓

Figure 5: Qualitative results on the Kvasir-VQA dataset.

Input Video



Question: What kind of foul does this video show? Choose one of the following options: elbow foul|offside foul|push foul|kick foul|
NVILA-Lite (2B): elbow foul ✗
Qwen2.5-VL (72B): kick foul ✗
Ours (2B): push foul ✓

Figure 6: Qualitative results on the SPORTU dataset.

Additional Visualization

Comparison with Baseline in VQA

We provide additional qualitative results comparing our LEAML method with baseline approaches on both medical and sports VQA tasks. These examples demonstrate our method’s superior performance in generating accurate, domain-specific answers. The results show that our approach better understands specialized terminology and visual patterns, leading to more precise responses compared to standard MLLMs.

Generated Pseudo-QA Examples

We show examples of pseudo question-answer pairs generated by our QA Generator. These synthetic QA pairs effectively expand the training dataset and provide domain-relevant supervision for the subsequent fine-tuning stage. The generated questions are contextually appropriate and cover diverse aspects of the visual content, while the answers demonstrate proper domain knowledge and terminology usage.

Captions from Large MLLM

We present captions generated by the large MLLM (Qwen2.5-VL-72B) used in our distillation process. These captions successfully capture important visual details and domain-specific information, which helps improve the QA Generator’s learning through knowledge distillation.

Input Image



Question: Are there any abnormalities in the image? Check all that are present. Choose one of the following options: polyp|ulcerative colitis|barretts|oesophagitis|hemorrhoids|
Answer: ulcerative colitis

Input Image



Question: What type of procedure is the image taken from? Choose one of the following options: gastroscopy|colonoscopy|capsule endoscopy|
Answer: gastroscopy

Figure 7: Qualitative results of generated pseudo-QA on the Kvasir-VQA dataset.

Input Video



Question: What specific type of foul, if any, occurred in the video? Choose the most appropriate one. Choose one of the following options:
Blocking Foul|Tripping Foul|Pushing Foul|None of the above|
Answer: Blocking Foul

Figure 8: Qualitative results of generated pseudo-QA on the SPORTU dataset.

Input Video



The video captures an intense moment during a professional basketball game between two teams, identifiable by their distinct uniforms. The team in white jerseys is on offense, with player **number 3**, named Paul, dribbling the ball near the three-point line. He is closely **guarded by a player** from the opposing team, who is wearing a black jersey with the number 24. As Paul attempts to drive towards the basket, he is met with strong defensive pressure. The defender from the black team extends his arm, **attempting to block** or steal the ball. In the background, other players from both teams are positioned strategically, ready to react to the play’s outcome. The court is well-lit, and the audience can be seen in the stands, adding to the atmosphere of the game. The action is fast-paced, with all players fully engaged in the competitive moment.

Figure 9: Qualitative results of generated caption on the SPORTU dataset.

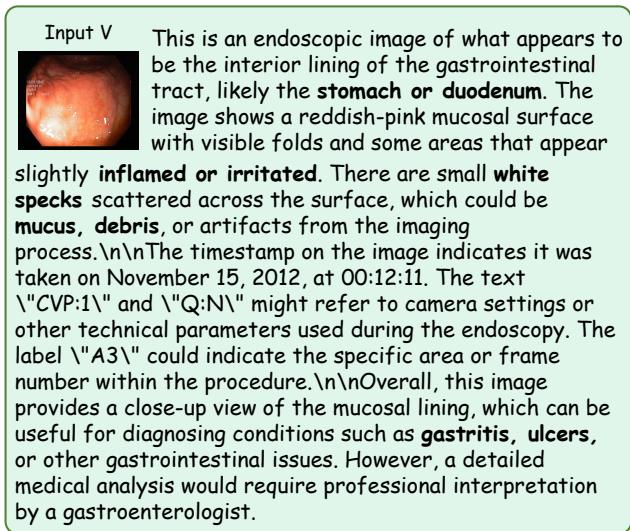


Figure 10: Qualitative results of generated caption on the Kvasir-VQA dataset.

Input Image



Question: What type of polyp is present? Choose one of the following options:
paris is|none|paris iia|paris ip|

NVILA-Lite (2B): paris ip ✗
Qwen2.5-VL (72B): paris iia ✗
Ours (2B): none ✓

Input Image



Question: What is the size of the polyp? Choose one of the following options:
5-10mm|>20mm|<5mm|none|11-20mm|>20|

NVILA-Lite (2B): <5mm ✗
Qwen2.5-VL (72B): none ✗
Ours (2B): >20mm ✓

Figure 11: Qualitative results on the Kvasir-VQA dataset.

Input Video



Question: What kind of foul does this video show? Choose one of the following options: elbow foul|offside foul|push foul|kick foul|

NVILA-Lite (2B): elbow foul ✗
Qwen2.5-VL (72B): kick foul ✗
Ours (2B): push foul ✓

Figure 12: Qualitative results on the SPORTU dataset.