IMPROVING MULTIMODAL MANY-SHOT LEARNING WITH NEEDLE-IN-A-HAYSTACK TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models have shown significant promise in combining text and image modalities, yet their ability to effectively utilize multimodal many-shot context remains understudied. We study the performance of VLMs on multimodal many-shot classification problems and observe that their performance is inconsistent. On some datasets the classification performance improves linearly for every doubling of the number of shots, while for others, the performance outright degrades as the number of shots increases. Surprisingly, we find that finetuning these models on a set of multimodal many-shot classification tasks does not improve their performance on previously unseen many-shot classification tasks. Instead, we demonstrate that finetuning the models on a simpler but related needle-in-a-haystack problem, where the image to be classified and its label (needle) are hidden in the context (the haystack), significantly improves the performance and scaling with the number of shots. We analyze the attention patterns of these models to demonstrate evidence that needle-in-a-haystack training provides a strong learning signal to search the entire context, which changes the positional bias of the learned attention pattern and thus allows the model to generalize to many-shot settings. Overall, our results show that just 200 gradient steps of needle-in-a-haystack training may greatly improve performance on many-shot learning.

1 INTRODUCTION

Building upon the success of large language models, vision-language models (VLMs) combine the modalities of text and images by encoding images into the embedding space of language models. These VLMs maintain the natural language capabilities of the model, such as the ability to hold human-like conversations, while also gaining new abilities such as visual question answering and image captioning (Antol et al., 2015; Gemma Team, 2025; Yang et al., 2024; Gemini Team, 2025). However, one unresolved question is if the emergent capabilities of language models will again appear in visual domains. For example, an interesting property of language models is their ability to learn from examples, i.e., in-context learning. Initially observed as an emergent capability, providing a language model examples of a task improves its performance on that task, similar to how humans may learn from examples (Wei et al., 2022). This so-called many-shot learning emerged as an alternative to finetuning as it is a pure inference-time approach, and therefore avoids expensive retraining periods (Dong et al., 2024). In this paper, we investigate multimodal many-shot learning for VLMs.

This inference-time adaptation is particularly valuable in settings where finetuning is impractical or unavailable. In specialized domains such as medical imaging and remote sensing, practitioners may have hundreds of labeled examples but lack sufficient data, infrastructure, or permission to finetune a model (Vahdati et al., 2025; Conti, 2025). Similarly, many frontier VLMs are accessible only through APIs that do not support finetuning, leaving in-context examples as the primary mechanism for adaptation. Many applications also involve task definitions or label spaces that change over time (Chalkidis and Søgaard, 2022); for example, content-moderation policies may be updated frequently, making prompt-level adaptation substantially easier than retraining. These settings make it important to understand not only whether VLMs can learn from in-context examples, but how their performance scales as increasingly many examples are provided. For language-only models, prior work has documented an approximate linear-log relationship between the number of shots (i.e. examples) and the performance of the model (Agarwal et al., 2024). In other words, doubling the number of shots

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

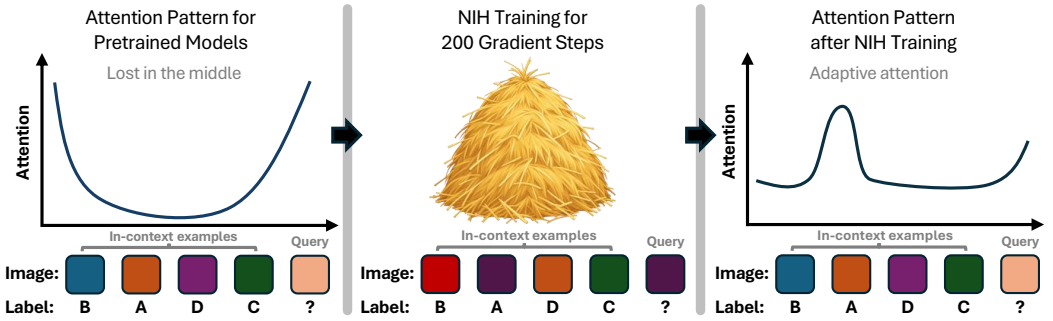


Figure 1: **Needle-in-a-haystack training promotes adaptive attention.** (*left*) Pretrained vision-language models exhibit a lost-in-the-middle pattern, preferentially attending to demonstrations near the beginning and end of the prompt. (*middle*) During needle-in-a-haystack (NIH) training, the model must search the in-context examples for the image matching the query and return its associated label. (*right*) After only 200 gradient steps, the model exhibits less positionally biased attention and focuses on a relevant demonstration regardless of its location, improving its ability to use many-shot context.

leads to a roughly constant increase in performance. Given this motivation, we investigate if this same phenomenon emerges in VLMs.

We use three VLMs as case studies: Gemini 2 (Gemini Team, 2025), Gemma 3 (Gemma Team, 2025), and Qwen 2.5 Omni (Yang et al., 2024). We construct 13 multimodal in-context learning datasets from image classification problems and evaluate pretrained models across varying numbers of shots. We find that performance roughly follows a linear-log relationship with the number of shots, but is neither monotonic nor consistent across datasets: on some datasets, performance decreases as more shots are provided. This is problematic because users expect adding more data should improve performance. To improve the model, an intuitive solution is to finetune directly on many-shot examples. However, empirical results show this produces little improvement in practice. This result suggests that simply providing more many-shot training examples is not sufficient, motivating a different training objective that more efficiently teaches the model the capabilities required for many-shot inference.

Needle-in-a-haystack (NIH) (Ivgi et al., 2023; Liu et al., 2024) is a related but simpler problem in which the query image and its ground-truth label are hidden among the shots. To answer correctly, the model need only search through the context for the matching query image and return its associated label. NIH has been used as an evaluation for testing whether models can retrieve information from long contexts; here, we instead use it as a training objective. This represents a different training paradigm: rather than training directly on the target many-shot classification objective, we train on a simpler proxy objective designed to induce the underlying behavior needed for success. Despite the difference between the training and evaluation objectives, we counterintuitively find that NIH finetuning significantly outperforms both the base model and direct many-shot finetuning on the original many-shot classification task.

To explain this behavior, we hypothesize that the pretrained models struggle with many-shot learning because their attention mechanisms focus on the beginning and end of prompts while ignoring the shots in the middle, i.e., the lost in the middle phenomenon (Liu et al., 2024). When we finetune with many-shot prompts, the model may have a weak gradient signal to learn from the shots, but it may improve its performance by paying more attention to the query image or improving its semantic understanding of the dataset. In contrast, when finetuning on the NIH prompts, there is a strong gradient signal to search for the needle and return its label. Thus, this training method encourages the model to consider every shot regardless of the semantics, thus reducing the positional bias. We investigate this hypothesis by analyzing the attention patterns of the pretrained and finetuned models, and find evidence to support this hypothesis, both in terms of the maximum attention the base models are paying to any one example label, and the entropy of which label the model is focusing on.

More broadly, this finding suggests that NIH objectives should be incorporated into VLM pretraining or post-training to improve long-context utilization beyond the specific tasks studied here. Such an

intervention is particularly attractive because it is highly efficient: we observe substantial gains after only 200 gradient steps of NIH finetuning. We summarize our contributions as follows:

- We systematically study multimodal many-shot in-context learning and find that directly finetuning on many-shot prompts yields little to no improvement in many-shot performance.
- We introduce a new training paradigm based on finetuning on the related needle-in-a-haystack (NIH) task, showing that only 200 gradient steps substantially improve many-shot performance.
- We investigate why NIH training improves many-shot performance, hypothesizing that it reduces positional bias by encouraging the model to search across all demonstrations before focusing on the most relevant shot, and provide attention-based evidence in support of this hypothesis.
- Based on these findings, we advocate incorporating NIH training into multimodal model training mixtures as an efficient mechanism for improving many-shot in-context learning.

2 Related Work

In-context learning (ICL) allows language models to adapt at inference time using input–output demonstrations rather than parameter updates (Brown et al., 2020). Its effectiveness generally increases with scale: larger models are better able to infer unseen input–label mappings, and smaller models rely more on pretrained semantics (Wei et al., 2023). Prior work interprets ICL as Bayesian task inference (Xie et al., 2022) or as an implicit implementation of gradient descent within the forward pass (Akyürek et al., 2023; von Oswald et al., 2023). With long context windows, ICL extends to hundreds or thousands of examples. Agarwal et al. (2024) show that many-shot ICL can outperform few-shot prompting, override pretraining biases, and approach finetuning performance. However, ICL remains sensitive to demonstration ordering and selection (Liu et al., 2021; Lu et al., 2022; Peng et al., 2024), as well as label and prompt biases (Zhao et al., 2021; Jiang et al., 2023). Moreover, demonstrations may specify the input distribution, label space, and output format without teaching the intended mapping (Min et al., 2022).

Multimodal ICL extends this paradigm to combined visual and text inputs. Flamingo established multimodal few-shot learning from interleaved image–text demonstrations, including prompts containing multiple images (Alayrac et al., 2022). Subsequent models such as Otter, MMICL, and Emu2 extended this paradigm through multimodal instruction tuning, improved prompt understanding, and unified multimodal generation (Li et al., 2025; Zhao et al., 2024; Sun et al., 2024). MetaVL transfers ICL ability from language to vision-language models, whereas SINC explicitly trains models to infer mappings from image–label sequences (Monajatipoor et al., 2023; Chen et al., 2023). Extending this setting beyond few-shot prompts, Jiang et al. (2024) evaluate up to nearly 2,000 multimodal demonstrations and find approximately linear-log improvements for closed models such as Gemini 1.5 Pro and GPT-4o, but limited benefits for the evaluated open-weight models. Nevertheless, studies suggest that multimodal ICL is often driven more by demonstration text than images (Baldassini et al., 2024; Chen et al., 2025), although modality-aware example selection improves results (Xu et al., 2025; Zhou et al., 2024). Other works evaluate multimodal needle-in-a-haystack performance and find that the evaluated models perform unsatisfactorily, especially as the context increases in length (Wang et al., 2024; 2025). Prior work has also used text-based NIH objectives for training, finding improvements in long-horizon reasoning, question answering, and retrieval (Xiong et al., 2025; An et al., 2024). Our results complement these findings by showing that NIH training can similarly improve multimodal many-shot learning.

3 Problem Setting

We begin by formally describing the VLM. Let $\mathcal{X}$ be the set of RGB images and $\mathcal{V}$ be a finite set of tokens, referred to as the *vocabulary*. We define the set of all possible token sequences as the Kleene closure of the vocabulary, $\mathcal{T} = \mathcal{V}^*$. A Vision-Language Model (VLM) is characterized by a conditional probability distribution P over the vocabulary. Specifically, given an interleaved sequence of images and tokens (a prompt) $S \in (\mathcal{X} \cup \mathcal{T})^*$, the model predicts the probability of an output

response $r \in \mathcal{T}$ of length N as:

$$P(r|S) = \prod_{j=1}^{N} P(r_j|S, r_{<j}) \tag{1}$$

where $r_j \in \mathcal{V}$ denotes the j-th token of r, and $r_{<j}$ represents the sequence of tokens preceding r_j.

Next, we describe the classification setting. Let $\mathcal{Y} = \{y_i\}_{i=1}^{m} \subset \mathcal{T}$ be a set of natural language labels, and let $f : \mathcal{X} \to \mathcal{Y}$ be the labeling function. Naturally, the goal is to approximate f. Given a query image $x_q \in \mathcal{X}$, the model acts as a classifier by selecting the most-probable label $\hat{y} \in \mathcal{Y}$:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P\left(y \mid S\left(x_q\right)\right), \tag{2}$$

where $S(x_q)$ is a prompt containing x_q. Typically, the VLM requires a textual prefix, $t_{pre} \in \mathcal{T}$, which describes the problem, and a postfix $t_{post} \in \mathcal{T}$, which triggers the model to generate a response. We define the zero-shot (ZS) prompt as the case where the model is given only x_q without any preceding image-label pairs in the input prompt:

$$S_{zs}(x_q) = (t_{pre}, x_q, t_{post}) \tag{3}$$

In the many-shot (MS) setting, the model is also given a sequence $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ of image-label pairs, with $y^{(i)} = f(x^{(i)})$. These so-called *examples* or *shots* are used to improve the model's estimate of $f(x_q)$, similar to how humans learn from examples. The MS prompt is defined as:

$$S_{ms}(x_q) = \left(t_{pre}, \underbrace{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})}_{\text{n shots}}, x_q, t_{post} \right) \tag{4}$$

Lastly, the needle-in-a-haystack (NIH) (Ivgi et al., 2023) setting is similar to the MS setting, except the query image and its label (the needle) are hidden in the shots (the haystack), i.e., $(x^{(k)}, y^{(k)}) := (x_q, f(x_q))$ for some index k. Therefore, the model only needs to search through the shots to find the needle, and return its label. The NIH prompt is defined as:

$$S_{nih}(x_q) = \left(t_{pre}, \underbrace{(x^{(1)}, y^{(1)}), \ldots, \overbrace{(x_q, f(x_q))}^{\text{needle at index } k}, \ldots (x^{(n)}, y^{(n)})}_{\text{n shots}}, x_q, t_{post} \right) \tag{5}$$

4 DATASETS

We collect a diverse set of multimodal classification datasets to create both MS and NIH prompts. To do so, we start with zero-shot multimodal datasets which consist of image-label pairs. These datasets were chosen to span a wide range of classification tasks, spanning topics such as cancer, satellite imagery, textures, facial expressions, rendered text, memes, and traffic signs. They also include variations in class counts from two to as many as 47 with both balanced and unbalanced class distributions. We use the following 13 datasets, many of which are collected from the Massive Image Embedding Benchmark (Xiao et al., 2025): CLEVR, CLEVR Count (Johnson et al., 2017), Describable Textures Dataset (Cimpoi et al., 2014), Eurosat (Helber et al., 2019), Facial Expression Recognition (Goodfellow et al., 2015), German Traffic Sign Recognition Benchmark (Stallkamp et al., 2011), Oxford Pets (Parkhi et al., 2012), Patch Camelyon (Ehteshami Bejnordi et al., 2017), Rendered SST2 (Radford et al., 2021), Remote Sensing Image Scene Classification (RESISC) (Cheng et al., 2017), STL 10 (Coates et al., 2011), Hateful Memes (Kiela et al., 2020), and Ham10k (Tschandl et al., 2018). For more information, see Table 6. In the following sections, Hateful Memes, Eurosat, and German Traffic Sign Recognition Benchmark are held out, while the other 10 datasets are used for training. These heldout datasets were chosen to include a diversity of topic and class counts.

We form ZS, MS, and NIH versions of each dataset using the prompt construction method described in Section 3. The example shots are uniformly randomly selected and placed, meaning that there is no correlation between index and usefulness of a shot. Furthermore, the examples are not class-balanced, and the index k selected for the needle is also uniformly random. Each dataset is split into train and validation splits. For a given split, the example shots are sampled from that same split, meaning that during validation, neither the query image nor the example shots have been seen before.

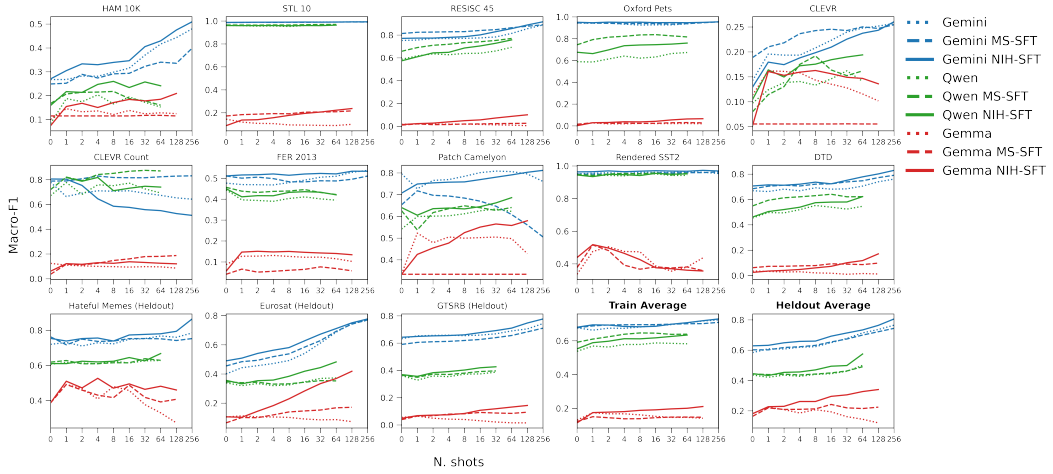


Figure 2: **Evaluation Results.** Evaluation of model families and checkpoint types across all validation splits using MS prompts. Performance is measured in Macro-F1 (y-axis) against the number of shots on a log scale (x-axis). The upper two rows display results for in-distribution validation datasets, while the bottom row illustrates performance on heldout datasets alongside the mean results.

5 Training Procedure

To investigate if the observed behavior is general to VLMs or specific to a model, we use three instruction tuned VLMs: Gemini 2, Gemma 3 (4B), and Qwen 2.5 Omni (7B). Note that implementation details and model sizes differ between these models (See Appendix C), so comparisons *between* models should be interpreted with caution. Instead, our experiments highlight how the same behavior emerges for all three models. In our experiments, we use supervised fine tuning (SFT) to train the models. Specifically, we use a completion-only loss objective, and thus the model learns to maximize the probability of the correct label. The models are trained on a uniform mixture over the 10 training datasets with a batch size of 32. Crucially, we impose a minor intervention of only 200 gradient steps.

As described in the introduction, we evaluate three checkpoints per model: 1) Base, 2) MS-SFT, and 3) NIH-SFT. Notably, NIH-SFT is evaluated on MS prompts despite never seeing them during training. For clarity, we refer to architectures such as Qwen as *model families*, and to their individual pretrained or finetuned variants, such as MS-SFT and NIH-SFT, as *checkpoints*.

6 Results

After training, we evaluate each model family and checkpoint on the validation splits of all datasets. Note that all evaluations are on MS prompts, meaning that *NIH checkpoints are being evaluated on a different task than they were trained on*. See Figure 2.

The pretrained checkpoints display linear-log scaling on some datasets, but fail on others. We observe that the base checkpoint performance does improve as the number of shots increases, roughly following an n-shot scaling law. However, this is not consistent across datasets. For example, Qwen's performance on the facial expression recognition dataset decreases relative to its ZS performance. On the other hand, Qwen and Gemini show scaling behavior on the RESISC 45 dataset (Appendix B).

The MS-SFT checkpoints improve on the validation splits of the training datasets, but do not learn generalizable many-shot transfer. We find that MS-SFT outperforms the base checkpoints on the validation splits of the training datasets. On the other hand, their performance is roughly equal to the base checkpoints on the heldout datasets. This implies that MS-SFT has not improved its MS learning capabilities in general, but rather has overfit to the semantics of these specific datasets. We will continue the discussion of the implications in the following section.

The NIH-SFT checkpoints outperform both the base checkpoints and the MS-SFT checkpoints on the heldout datasets. From the in-distribution validation results, NIH-SFT does not seem better compared to MS-SFT. However, the difference on the heldout datasets is striking. Here, we see that NIH-SFT is outperforming both the pretrained checkpoint and MS-SFT, especially as the number of shots increases, for all three model families (Gemini, Gemma, Qwen). This result is surprising; NIH-SFT was not trained on MS prompts, and yet its MS performance exceeds the MS-SFT models. In the following section, we propose a hypothesis to explain this result, and use the attention scores to analyze the differences between the model checkpoints.

7 HYPOTHESIS

To explain this phenomenon, we first define the potential solution space available to a model during either MS or NIH supervised finetuning. We categorize these strategies into three distinct approaches:

(1) **Focus on the query image:** Improve the semantic understanding of the query image, i.e., specialize the model to the provided datasets without using the examples.

(2) **Focus on the examples:** Improve the ability to learn from examples. For example, learn to understand the labels' trend, or focus on a single representative example.

(3) **Find the needle:** Search for the needle, then return the ground truth label.

Note that these strategies are not necessarily mutually exclusive; a model could improve by simultaneously learning how to make better use of examples while also improving its semantic understanding of the query image. However, it is unlikely the learning signals will have the same magnitude, meaning that a training approach will tend to focus on one strategy more than the others.

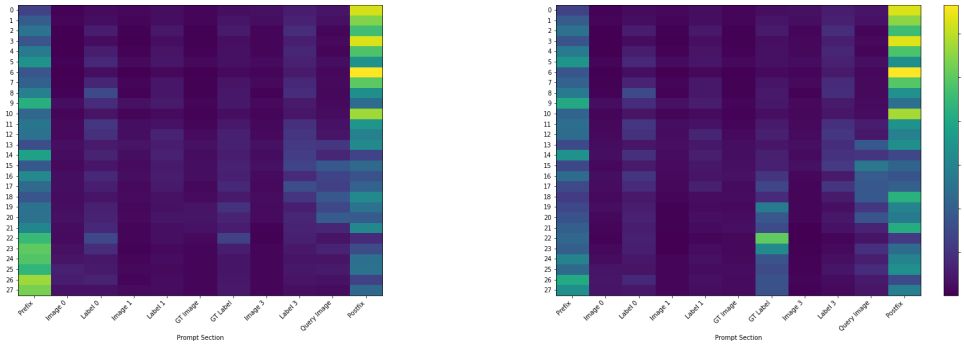
In the case of MS-SFT, the model is limited to options (1) and (2). However, option (2) only provides a weak learning signal; for any given prompt, the relevance of specific examples is inconsistent, and it is even possible that no examples are relevant. Consequently, we expect MS-SFT to primarily learn solution (1), focusing on the semantic understanding of the images in the dataset.

In contrast, NIH-SFT possesses a significantly more direct path via solution (3). Because a needle is guaranteed to exist in every NIH prompt by definition, the training data provides a constant, high-strength learning signal toward this heuristic. By simply matching the query image to the needle's image to extract the label, the model can achieve high performance without deeper semantic integration. Thus, we expect NIH-SFT predominantly adopts solution (3).

While it is clear why strategy (3) would perform well on NIH prompts, it is not obvious why learning strategy (3) would improve the model's performance on MS prompts. Consider the implication of a model that learns strategy (3). In order to find the needle, the model must search for it. Since the needle is distributed uniformly, the model must search all of the example shots without a positional bias. Once the needle is found, the model is likely to allocate a large amount of attention to the needle, since that is all it needs to get the correct output. In the MS setting, this positionally unbiased search is useful as it allows the model to consider all of the shots and then allocate a large amount of attention to a "primary" shot, potentially selected by relevance to the query. In other words, learning strategy (3) acts as a useful regularizer or bias which encourages the model to correct for the lost in the middle phenomenon, and consider all of the example shots rather than just the first and/or last.

Lastly, consider the generalization and MS implications of each strategy. If a model learns strategy (1), its MS performance will improve on the training datasets, as it has specialized its knowledge to these types of images. However, it will not generalize to new datasets. For example, the model may learn to better recognize tumors in the Patch Camelyon dataset, but this will not improve the model's performance on classifying satellite imagery. Note that the MS-SFT models exhibit a similar lack of generalization in Figure 2. If a model learns strategy (2), it will generalize to new datasets (as long as the pattern matching is not overspecialized to certain semantics). This is exactly the behavior we are targeting: many-shot learning. If a model learns strategy (3), its implicit search over the shots is potentially useful for unseen, MS datasets so long as training does not degrade the model's pretrained knowledge. Under this chain of reasoning, we arrive at the following hypothesis:

Hypothesis 1. *The base model's MS performance is limited by positional bias in its attention pattern; MS-SFT does not improve on heldout MS datasets because it learns strategy (1); NIH-SFT improves on heldout MS datasets because it learns strategy (3), reducing the prevalence of positional bias.*



(a) Base (Eurosat) (b) NIH-SFT (Eurosat)

Figure 3: **Positionally-Aggregated Attention Plots.** These plots visualize attention allocation from the perspective of the final prompt token across all model layers (y-axis). Attention values are summed per prompt segment (x-axis) to simplify analysis. Both plots show the same 4-shot NIH prompt from Eurosat, demonstrating that Gemma NIH-SFT models correctly identify the needle with high attention scores on the ground truth (GT) label in the final layers unlike the base model.

We analyze attention patterns and learning curves from the two models, Gemma and Qwen, to provide evidence for this hypothesis. Because raw attention weights across thousands of tokens are difficult to interpret, we introduce an analytical lens that aggregates attention by position into interpretable segments before presenting the results. See Appendix A for a background on attention.

8 Attention Analysis

To analyze the behavior of the models, we first forward pass the model generating an $H \times H$ attention matrix for each layer, where H is the number of tokens. In the case of multiple heads, we compute the mean attention matrix over heads. We then focus on the last row of the attention matrices (vectors of size H), which correspond to the attention scores of the last token in the prompt. This token is responsible for predicting the class label. Next, we group the tokens by prompt segments and sum their attention scores, e.g., all tokens corresponding to t_{pre} are summed. Lastly, we stack the results for all layers and plot the heatmap, shown in Figure 3. This processing technique allows us to interpret what the model is paying attention to for a given prompt. For example, Figure 3 suggests that Qwen NIH-SFT correctly identifies the needle in a NIH prompt because the last 10 layers have high attention scores for the ground truth label, something that we do not observe in the base model.

To extend this analysis from qualitative to quantitative, we apply the same process to 100 prompts for each model family and checkpoint, and then compute aggregated results. Note that attention analysis is a useful heuristic for interpreting model behavior rather than a mechanistic explanation. We qualitatively observed that the difference between checkpoints largely appears only in the final layers of the model. This aligns with prior work, which has observed that LLMs use the first layers for low-level semantics and later layers for higher levels of abstraction (Tenney et al., 2019). Hence, the quantitative results are averaged over the last 11 layers of Qwen and the last 3 quadratic attention layers of Gemma[1]. For each prompt, we also identify the *primary example*, the example shot that is most heavily paid attention to by the model (not including the needle, if present). This allows us to interpret how heavily the model focuses on a single crucial example rather than paying attention to all examples at once. Furthermore, we compute the entropy of the index of the primary example. This entropy tells us how positionally biased the model is. For example, if the primary example is always at index 1, then the model will have low entropy. On the other hand, if the model exhibits a flexible attention pattern, i.e., the primary example shows no positional bias, then entropy will be high. We focus on the heldout Eurosat dataset in Table 1. Results on other datasets are found in Appendix D.

[1]Gemma interleaves quadratic and sliding window attention layers. Since the sliding window attention layers only attend to the latest 3 shots, we ignore these layers in our analysis.

Table 1: **8-Shot EuroSAT Attention Analysis.** Results represent mean ± std attention scores aggregated over 100 evaluation prompts. Attention scores are expressed as percentages, while entropy is measured in bits. Note that attention scores may not sum to 100 as the table reports only the mean attention allocated to the primary example, excluding the others. For 8 shots, the maximum entropy is 3 bits. From left to right, the prompt sections are the prefix, primary example image, primary example label, ground-truth image, ground-truth label, query image, and postfix.

Model	Prompt	SFT	Pre	Pri. Ex. I	Pri. Ex. L	GT I	GT L	Q. I	Post	Ent.
Qwen	NIH	Base	33.3 ±3.1	2.9 ±0.8	5.3 ±1.3	2.3 ±1.2	5.6 ±2.1	8.5 ±1.6	19.4 ±1.1	–
		MS-SFT	33.2 ±3.2	2.8 ±0.8	5.0 ±1.3	2.3 ±1.2	5.9 ±2.1	9.2 ±1.6	19.6 ±1.0	–
		NIH-SFT	20.4 ±3.2	2.1 ±0.7	5.0 ±2.1	2.3 ±1.3	15.2 ±2.3	7.9 ±1.4	31.0 ±1.8	–
	MS	Base	37.6 ±2.6	2.9 ±0.7	5.4 ±1.4	–	–	10.1 ±1.9	19.0 ±1.2	1.3
		MS-SFT	37.8 ±2.7	2.9 ±0.8	5.2 ±1.4	–	–	10.9 ±2.0	19.0 ±1.1	1.5
		NIH-SFT	34.0 ±5.3	2.9 ±0.9	6.4 ±3.0	–	–	11.7 ±2.2	21.9 ±2.4	2.3
Gemma	NIH	Base	18.4 ±2.3	9.9 ±1.5	8.5 ±2.6	6.1 ±2.0	5.9 ±4.1	4.6 ±1.2	3.2 ±0.4	–
		MS-SFT	13.5 ±3.7	8.2 ±1.4	4.9 ±3.0	8.5 ±2.2	20.2 ±4.5	8.9 ±1.5	2.2 ±0.3	–
		NIH-SFT	2.2 ±1.0	4.1 ±0.7	3.8 ±3.5	7.4 ±2.0	50.0 ±6.2	7.1 ±2.1	5.5 ±0.8	–
	MS	Base	25.6 ±2.8	12.1 ±1.2	7.5 ±2.3	–	–	4.3 ±1.0	2.3 ±0.3	0.8
		MS-SFT	23.6 ±5.6	9.8 ±1.7	7.8 ±4.3	–	–	11.7 ±2.3	1.9 ±0.3	2.9
		NIH-SFT	6.6 ±3.8	6.3 ±1.5	22.8 ±13.5	–	–	19.8 ±8.9	6.6 ±1.8	3.0

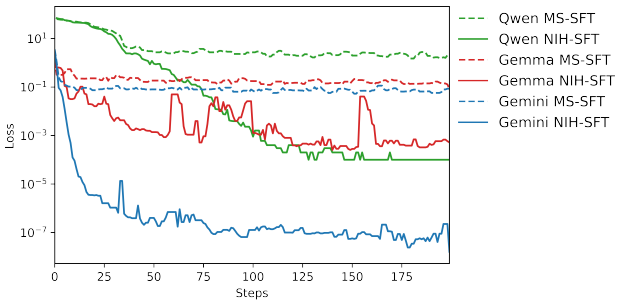


Figure 4: **Learning Curves.** We plot the learning curves for Qwen, Gemma, and Gemini. The NIH-SFT checkpoint converges to orders of magnitude lower loss than the MS-SFT checkpoint. Note the loss curves are smoothed with a median filter with a kernel size of 5.

8.1 Evidence

The base model's attention pattern is positionally biased which harms its ability to make use of the shots. In Table 1, we observe low entropy in the Base models on MS prompts. This low entropy means that the primary example label tends to lie at the same index. However, the shots are uniformly randomly placed, meaning that the most useful shot should also be uniformly distributed. Thus, we may interpret this low entropy as a sign that the base models' attention patterns are positionally biased. In particular, we find that Gemma pays the most attention to the first shot 87% of the time on Eurosat, 75% of the time on Ham10k, and 55% of the time on Patch Camelyon. Qwen pays most attention to the first shot 95% of the time on Ham 10k, and pays attention to the last shot 50% of the time on Patch Camelyon and 76% of the time on Eurosat. Thus, we may conclude that the base models exhibit strong positional biases, either in favor of the first or last shot, depending on the dataset.

For all model families, MS-SFT and NIH-SFT exhibit radically different training regimes. All model families use a form of cross entropy in their loss, although their accounting may be different, e.g., if they sum or mean the entropy over labels. In any case, we see the same behavior for all model families. The NIH-SFT converges to a loss orders of magnitude lower than the MS-SFT. This is because even a well-trained MS-SFT will not be perfectly accurate, as the MS prompts may be ambiguous. In contrast, all NIH-SFT models do indeed learn to find the needle and return its label, leading to an extremely low loss. Effectively, these models are almost perfectly accurate on NIH prompts, suggesting that NIH-SFT indeed leverages strategy (3). See Figure 4.

NIH-SFT successfully learns to find the needle. See Table 1 as an example. Compared to the base Qwen model, we see a nearly 3x increase in attention paid to the ground truth label for the NIH-SFT model. For Gemma, the difference is closer to 8x, with 50% of the attention paid to the ground truth label. This is further evidence that NIH-SFT is indeed solving NIH prompts using strategy (3).

Attention analysis suggests MS-SFT learns strategy (1). For Qwen, the MS-SFT has only barely improved entropy, while the attention paid to the primary example label has slightly decreased. This suggests the model's attention pattern is still positionally biased, and thus it has not learned to correctly use the shots. This strongly suggests Qwen MS-SFT has learned strategy (1). For Gemma, the MS-SFT has greatly increased entropy, but its attention paid to the primary example label has barely increased. This suggests its attention is more evenly spread out among the shots, but it is not focusing on a single shot in particular. Thus, we may interpret this model as paying more attention to the set of labels rather than focusing on a single, relevant label, or it could imply Gemma has learned strategy (2). However, a large change in Gemma MS-SFT is how much it pays attention to the query image, increasing more than 2x relative to the base model. In combination with its poor heldout dataset performance, this suggests Gemma MS-SFT has primarily learned strategy (1).

Attention analysis suggests NIH-SFT learns strategy (3). In contrast to MS-SFT, NIH-SFT improves both attention entropy and attention to the primary example label across both model families and datasets, with only one exception. Qwen NIH-SFT exhibits slightly lower entropy on GTSRB (Table 5), indicating that the reduction in positional bias does not hold universally across datasets. Nevertheless, the trend is consistent across all other datasets and models. Combined with the markedly different training regimes and the near-perfect NIH performance, the attention analysis suggests NIH-SFT learns strategy (3).

Takeaway: These results suggest that NIH-SFT checkpoints 1) exhibit more adaptive attention, with the primary attended label distributed more uniformly across context positions, and 2) allocate greater attention to the primary example label. Overall, these findings are consistent with NIH-SFT learning strategy (3) and transferring the resulting positionally unbiased search behavior to MS prompts.

9 Broader Impact

Our results have implications for how VLMs are trained to use long multimodal contexts. NIH-SFT outperforms both the base model and MS-SFT on MS prompts, and our analysis suggests that this improvement comes from learning to search uniformly across demonstrations, mitigating the lost-in-the-middle phenomenon. Importantly, NIH training does not teach the downstream MS classification tasks. Instead, it improves a general capability: retrieving and using relevant information regardless of where it appears in the context. This suggests that **NIH objectives could be incorporated into VLM training to improve long-context utilization beyond the tasks studied here**.

This intervention is also attractive because it is highly efficient. We observe **substantial gains after only 200 gradient steps of NIH-SFT**. It is possible that MS-SFT would improve with substantially more training or more diverse datasets, but that possibility further highlights the advantage of NIH training. Because training capacity must be divided across many competing tasks and capabilities, objectives that target general failure modes and produce large gains with relatively little training are especially valuable. Our findings therefore suggest that improving how models use their context may be more efficient than directly training on every downstream many-shot behavior of interest.

10 Conclusion

In this work, we investigated the many-shot in-context learning capabilities of vision-language models, including Gemini 2, Gemma 3, and Qwen 2.5 Omni. Many-shot in-context learning is crucial when finetuning is limited by data, compute, access, regulation, or rapidly changing task definitions. Our results show that while pretrained VLMs can scale performance with the number of shots, this behavior is often inconsistent and hampered by positional biases in their attention mechanisms. Furthermore, we find that training on MS prompts does not necessarily improve held-out performance, while training on NIH prompts does. We provide evidence that this results from different strategies learned during training, where an NIH-SFT checkpoint learns an approximately unbiased search over example shots. Overall, our results suggest that a small amount of NIH training during VLM post-training may greatly improve many-shot in-context learning at minimal cost.

AI Use Statement

We used generative modeling to lightly edit the paper for clarity and conciseness, and to write code.

Reproducibility Statement

The dataset generation method is described in Sections 3 and 4. The training code is standard completion objectives using cross entropy. We will release the scripts upon publication.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie C. Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal M. P. Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. In *NeurIPS*, 2024.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *ACL*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *ICLR*, 2023.

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

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188, 2024.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433. IEEE Computer Society, 2015.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? *NeurIPS*, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 2020.

Ilias Chalkidis and Anders Søgaard. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, 2022.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can multimodal large language models truly perform multimodal in-context learning? *IEEE WACV*, 2025.

Yi-Syuan Chen, Yun-Zhu Song, Cheng Yu Yeo, Bei Liu, Jianlong Fu, and Hong-Han Shuai. Sinc: Self-supervised in-context learning for vision-language tasks. *ICCV*, 2023.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE*, 105(10):1865–1883, 2017.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.

Silvia Conti. Advancing earth observation with a multi-modal remote sensing foundation model: Remote sensing. *Nature Reviews Electrical Engineering*, 2(9):590–590, 2025.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *EMNLP*, pages 1107–1128. Association for Computational Linguistics, 2024.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 2017.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Gemma Team. Gemma 3 technical report, 2025.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron C. Courville, Mehdi Mirza, Benjamin Hamner, William Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Tudor Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Chuang Zhang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 2015.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient long-text understanding with short-text models. *Trans. Assoc. Comput. Linguistics*, 11:284–299, 2023.

Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H. Chen, and Andrew Y. Ng. Many-shot in-context learning in multimodal foundation models. *ICML*, 2024.

Zhongtao Jiang, Yuanzhe Zhang, Cao Liu, Jun Zhao, and Kang Liu. Generative calibration for in-context learning. *EMNLP*, 2023.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 1988–1997. IEEE Computer Society, 2017.

Dhiraj D. Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. A study of BFLOAT16 for deep learning training, 2019.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *ACL*, 2021.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *ACL*, 2022.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *EMNLP*, 2022.

Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin F. Yang, and Kai-Wei Chang. Metavl: Transferring in-context learning ability from language models to vision-language models. *ACL*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting demonstration selection strategies in in-context learning. *ICL*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, 2018.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IJCNN*, 2011.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *CVPR*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *ACL (1)*, pages 4593–4601. Association for Computational Linguistics, 2019.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018.

Sanaz Vahdati, Elham Mahmoudi, Ali Ganjizadeh, Chiehju Chao, and Bradley J Erickson. Decoding large language models for radiology: strategies for fine-tuning and prompt engineering. *Radiology advances*, 2(4):umaf024, 2025.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. *ICML*, 2023.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *ACL*, 2025.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. Needle in a multimodal haystack, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, 2023.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth C. Enevoldsen, and Niklas Muennighoff. MIEB: massive image embedding benchmark. *ICCV*, 2025.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *ICLR*, 2022.

Zheyang Xiong, Vasilis Papageorgiou, Kangwook Lee, and Dimitris Papailiopoulos. From artificial needles to real haystacks: Improving retrieval capabilities in llms by finetuning on synthetic data. In *International Conference on Learning Representations*, volume 2025, pages 7276–7291, 2025.

Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. *ACL*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. *ICLR*, 2024.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *ICML*, 2021.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860, 2023.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. *ACL*, 2024.

A Background on Attention

Attention is a mechanism for processing sets, where elements are referred to as tokens. Let $X \in \mathbb{R}^{n \times k}$ be the input data, where n is the number of tokens and k is the data size. Attention first computes a query, key, and value for each token via a linear projection,

$$Q = XW^Q \in \mathbb{R}^{n \times d_k}, \qquad K = XW^K \in \mathbb{R}^{n \times d_k}, \qquad V = XW^V \in \mathbb{R}^{n \times d_v},$$

where W^Q, W^K, W^V are learned matrices, d_k is the dimensionality of the keys, and d_v is the dimensionality of the output. Then the output is computed as

$$\mathbb{R}^{n \times d_v} \ni Y = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.$$

The softmax term, referred to as the attention matrix, is a $n \times n$ matrix in which the (i, j)-th entry represents the weight assigned by token i to token j. The softmax is computed over rows and the output is a weighted sum of the values. This allows information from any token to propagate to any other token. For language modeling, positional information is encoded into the input X via positional embeddings, and tens of these layers are interspersed with other processing layers such as multi-layer perceptrons and residual connections. Additionally, the attention calculation itself is slightly modified for the sake of efficiency, e.g., grouped-query attention Ainslie et al. (2023).

For the purposes of this paper, we are concerned with the attention matrix. Since this matrix controls how strongly information flows from one token to another, it provides some mechanism for interpreting how the model functions. We therefore compute statistics using the attention matrices (over different layers) to understand how certain training schemes have changed the model's behavior.

B CLEVR Count

The CLEVR Count dataset exhibits different behaviors than the other datasets. We observe that the base models' performance degrades as the number of shots increases. This degradation occurs for the NIH-SFT checkpoints too. On the other hand, MS-SFT is stable and even slightly improves as the number of shots increases. The results on this dataset are different enough they may even change the trends of the training set average results. See Figure 5.

We believe this behavior is due to the specific task in this dataset; Counting requires a logical and methodical approach. Thus, including shots may act as a distractor. In particular, if the NIH-SFT model is using strategy (3), then its increased attention to the nearest neighbor may disrupt this counting process. In contrast, if MS-SFT has learned strategy (1), it may be largely ignoring the shots (i.e., not performing MS learning), and hence its behavior is mostly unaffected. In general, this poses interesting questions about which tasks are amenable to in-context learning. Possibly, other tasks requiring logical or systematic approaches are also not amenable to MS learning. This is an interesting direction that merits further investigation. We leave it to future work.

C Model-Specific Implementation Details

In this section, we describe implementation details. Since the implementations are different between model families, any comparison across model families is not perfect. However, we use the same implementation between model checkpoints of a given model family, meaning that it *is* fair to compare between the base model, NIH-SFT, and MS-SFT for a given family.

C.1 General Details

The training mixture is an equal split among the training datasets. NIH prompts are guaranteed to contain the query image as one of the examples; MS may contain the query image in a negligible number of samples. Models are queried without restricting their outputs for the acceptable labels. We find that all three model families are capable of faithfully responding with one of the desired labels with descriptive prompts, meaning that restricting their outputs or manually choosing the argmax label was not necessary. The class is therefore taken as the nearest string match among the viable

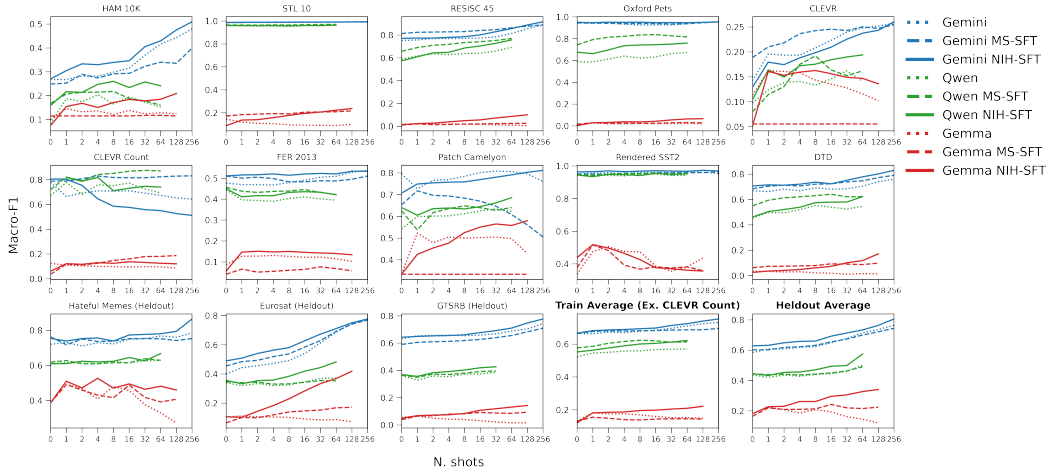


Figure 5: **Inference Results.** This figure shows the same inference results as Figure 2, except the training dataset average results exclude CLEVR Count.

class labels using the difflib algorithm. The MS and NIH prompts have approximately the same prompt length, meaning that an equal number of gradient steps implies an equal amount of compute. NIH prompts contain additional text in their prompt mentioning that the needle exists, though this text is not included when evaluating NIH-SFT models in a MS setting. Due to expense, we only train one seed per model.

C.2 Gemini 2

We finetune Gemini with a batch size of 32 and a learning rate of 5e-7 on a completion-only supervised objective. Unlike the other two models, Gemini is trained on variable numbers of shots for each prompt, with the number of shots ranging from 1 to 100.

C.3 Gemma 3

We use the 4B version of Gemma 3. We train using 4 TPUv7s with a global batch size of 1 and 32 gradient accumulation steps. We use an Adafactor optimizer (Shazeer and Stern, 2018) with a learning rate of 1e-3 and FSDP (Zhao et al., 2023) sharding. During training, all images are resized to 256x256 due to implementation constraints, and we only train on 16 shots.

C.4 Qwen 2.5 Omni

We use the 7B version of Qwen 2.5 Omni. We train using 16 A100s with a per-device batch size of 1, using 2 gradient accumulation steps for a total batch size of 32. We train using bf16 (Kalamkar et al., 2019) with gradient checkpointing and FSDP (Zhao et al., 2023). We finetune with a learning rate of 5e-7 and an AdamW optimizer. During training, we only train on 16 shots. Images are resized to 256x256.

D Attention Analysis (Continued)

In this section, we remark on the attention results for the other datasets. We observe across all datasets that the NIH-SFT model indeed finds the needle because the attention paid to the ground truth label is significantly higher than either the base model or MS-SFT for both model families. Thus, we will focus on our analysis on the MS prompts. We will also remark on which model empirically performs better, as there may be correlations worth noting.

Ham 10k (Table 2). Consider the checkpoints' behavior on the MS prompts. We observe that, as in Eurosat, the NIH-SFT models have higher attention scores for the primary example label, i.e., it

is focusing on one example more than either the MS-SFT or base models are. Note that NIH-SFT outperforms MS-SFT on this dataset as shown in Figure 2.

Patch Camelyon (Table 3). Similar to Ham 10k and Eurosat, the NIH-SFT shows higher primary example label attention scores and higher entropy than the MS-SFT and base checkpoints. However, for Qwen, the results on this dataset are less pronounced. This may influence the empirical results, where the Qwen family shows less difference between checkpoints than Gemma on this dataset.

Hateful Memes (Table 4). On this dataset, Qwen shows almost no difference between checkpoints in terms of entropy or attention scores. Gemma shows improved entropy and primary example label attention, as in the other datasets, but it is significantly less pronounced. Interestingly, the empirical results show that all checkpoints are performing similarly for a given model family. Potentially, the lack of variation in attention patterns between checkpoints explains this behavior.

German Traffic Sign Recognition Benchmark (Table 5). For Gemma, the results on this dataset are largely the same as Eurosat. NIH-SFT has higher entropy and attention with respect to the primary example label. However, Qwen is different. The NIH-SFT has lower entropy, but still higher attention with respect to the primary example label. This suggests the Qwen NIH-SFT has learned to focus on a key example, but has actually increased the positional bias on this dataset. Thus, we may conclude that NIH-SFT training does not universally reduce positional bias, although it did on the previous datasets.

Table 2: **8-Shot Ham 10k Attention Analysis.** Mean ± std across 100 evaluation prompts. Scores in %, entropy in bits.

Model	Prompt	SFT	Pre	Pri. Ex. I	Pri. Ex. L	GT I	GT L	Q. I	Post	Ent.
Qwen	NIH	Base	29.9 ±1.8	3.3 ±0.8	4.7 ±1.2	2.4 ±1.0	3.9 ±2.4	5.1 ±0.9	29.2 ±0.9	–
		MS-SFT	29.8 ±2.1	3.3 ±0.9	3.9 ±0.7	2.6 ±1.0	3.7 ±1.5	6.1 ±1.1	28.8 ±0.9	–
		NIH-SFT	20.8 ±2.3	2.4 ±0.7	4.1 ±1.7	2.4 ±0.9	11.9 ±3.1	5.5 ±1.1	36.5 ±1.3	–
	MS	Base	30.2 ±1.8	3.2 ±0.3	4.8 ±0.7	–	–	5.2 ±1.1	31.0 ±0.8	0.3
		MS-SFT	31.1 ±2.4	3.4 ±0.4	3.7 ±0.6	–	–	7.1 ±1.5	29.0 ±0.9	1.6
		NIH-SFT	25.8 ±3.8	2.8 ±0.5	5.0 ±1.6	–	–	8.8 ±2.1	32.6 ±1.4	2.4
Gemma	NIH	Base	17.4 ±2.2	12.2 ±3.0	8.5 ±2.3	6.7 ±4.1	5.0 ±3.6	2.8 ±0.5	6.7 ±1.0	–
		MS-SFT	12.4 ±4.6	6.8 ±1.4	5.1 ±1.7	8.7 ±2.2	16.6 ±3.9	10.7 ±1.9	5.5 ±1.2	–
		NIH-SFT	1.4 ±0.4	4.3 ±1.0	3.0 ±1.8	9.2 ±1.9	53.2 ±7.0	5.1 ±1.5	9.5 ±2.1	–
	MS	Base	26.4 ±2.0	18.5 ±1.7	5.6 ±1.3	–	–	2.5 ±0.4	6.9 ±0.7	0.9
		MS-SFT	16.3 ±5.9	8.6 ±1.7	8.4 ±3.1	–	–	13.9 ±2.5	5.3 ±1.2	2.7
		NIH-SFT	3.3 ±2.1	5.9 ±1.6	19.2 ±14.1	–	–	20.8 ±7.9	19.8 ±6.9	2.7

Table 3: **8-Shot Patch Camelyon Attention Analysis.** Mean ± std across 100 evaluation prompts. Attention scores in %, entropy in bits.

Model	Prompt	SFT	Pre	Pri. Ex. I	Pri. Ex. L	GT I	GT L	Q. I	Post	Ent.
Qwen	NIH	Base	23.2 ±6.0	3.0 ±0.7	4.2 ±2.6	1.9 ±1.0	3.3 ±2.9	4.4 ±1.0	43.5 ±7.1	–
		MS-SFT	19.1 ±4.4	3.5 ±0.8	5.7 ±2.4	2.4 ±1.0	3.7 ±2.8	5.2 ±1.0	39.5 ±6.2	–
		NIH-SFT	19.5 ±6.4	2.7 ±0.6	4.0 ±2.9	2.0 ±1.1	6.6 ±5.5	4.4 ±0.9	47.1 ±8.2	–
	MS	Base	21.1 ±5.3	3.2 ±0.6	4.9 ±2.9	–	–	5.2 ±1.5	43.0 ±7.6	2.1
		MS-SFT	16.9 ±2.3	3.8 ±0.6	6.7 ±1.8	–	–	6.4 ±1.1	37.5 ±4.3	2.2
		NIH-SFT	16.1 ±4.0	3.5 ±0.6	6.9 ±2.6	–	–	6.9 ±1.4	39.7 ±5.8	2.6
Gemma	NIH	Base	13.2 ±2.2	9.0 ±1.5	10.4 ±4.1	5.3 ±2.3	6.3 ±4.8	3.3 ±0.8	13.3 ±3.4	–
		MS-SFT	11.7 ±4.1	7.3 ±1.9	8.1 ±3.3	7.7 ±2.0	16.5 ±5.8	8.6 ±1.4	6.4 ±1.7	–
		NIH-SFT	2.5 ±1.1	6.0 ±2.4	3.4 ±2.1	8.7 ±2.4	44.2 ±8.9	8.5 ±2.4	6.4 ±1.9	–
	MS	Base	17.1 ±1.9	11.3 ±1.1	9.9 ±3.0	–	–	3.2 ±0.5	13.3 ±3.7	1.9
		MS-SFT	12.4 ±2.6	8.3 ±2.1	12.5 ±3.4	–	–	11.5 ±2.4	6.3 ±1.4	2.5
		NIH-SFT	3.5 ±1.3	7.3 ±2.6	20.4 ±11.9	–	–	21.8 ±8.2	10.7 ±3.3	2.7

Table 4: **8-Shot Hateful Memes Attention Analysis.** Mean ± std across 100 evaluation prompts. Attention scores in %, entropy in bits.

Model	Prompt	SFT	Pre	Pri. Ex. I	Pri. Ex. L	GT I	GT L	Q. I	Post	Ent.
Qwen	NIH	Base	25.6 ±2.0	2.6 ±0.5	4.4 ±1.1	2.2 ±0.8	3.9 ±1.7	5.9 ±1.0	34.6 ±1.3	–
		MS-SFT	25.4 ±1.9	2.7 ±0.5	4.3 ±1.1	2.4 ±0.8	4.1 ±1.7	6.6 ±1.1	33.6 ±1.2	–
		NIH-SFT	19.7 ±2.8	1.8 ±0.3	3.4 ±1.0	2.4 ±0.8	8.5 ±2.3	7.0 ±0.9	41.5 ±1.8	–
	MS	Base	24.2 ±1.6	2.6 ±0.4	4.8 ±0.9	–	–	5.9 ±0.9	38.0 ±1.2	2.5
		MS-SFT	24.8 ±1.8	2.8 ±0.5	4.8 ±1.0	–	–	7.0 ±1.0	35.8 ±1.0	2.5
		NIH-SFT	23.4 ±2.1	2.5 ±0.4	4.6 ±1.2	–	–	8.4 ±1.1	39.2 ±1.5	2.6
Gemma	NIH	Base	37.7 ±2.0	15.5 ±4.0	5.7 ±1.5	4.9 ±4.0	1.8 ±1.9	2.3 ±0.5	5.2 ±0.6	–
		MS-SFT	15.3 ±4.1	7.6 ±0.8	5.1 ±1.4	8.4 ±2.6	17.7 ±4.4	7.3 ±1.1	6.8 ±1.0	–
		NIH-SFT	2.4 ±1.1	4.3 ±0.7	1.8 ±0.9	9.1 ±1.7	50.6 ±4.1	6.7 ±1.5	10.9 ±1.6	–
	MS	Base	31.7 ±3.1	13.6 ±2.0	7.3 ±2.0	–	–	3.0 ±0.7	6.3 ±1.0	1.5
		MS-SFT	21.6 ±2.8	9.5 ±1.1	7.4 ±1.7	–	–	9.8 ±1.5	6.8 ±1.4	2.0
		NIH-SFT	5.5 ±2.5	4.9 ±0.6	9.1 ±6.2	–	–	23.2 ±5.1	28.3 ±5.7	2.5

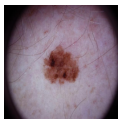
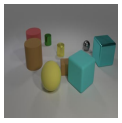
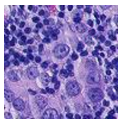
Table 5: **8-Shot GTSRB Attention Analysis.** Mean ± std across 100 evaluation prompts. Attention scores in %, entropy in bits.

Model	Prompt	SFT	Pre	Pri. Ex. I	Pri. Ex. L	GT I	GT L	Q. I	Post	Ent.
Qwen	NIH	Base	32.0 ±4.6	2.0 ±0.9	3.2 ±1.8	2.7 ±2.2	5.1 ±3.9	8.0 ±4.5	34.7 ±17.3	–
		MS-SFT	31.6 ±4.3	2.2 ±0.9	4.6 ±1.6	3.0 ±1.9	6.9 ±2.8	10.5 ±2.8	25.8 ±4.7	–
		NIH-SFT	22.0 ±3.2	1.7 ±0.7	5.1 ±1.9	2.7 ±1.6	12.4 ±3.6	8.0 ±2.0	34.7 ±2.9	–
	MS	Base	37.0 ±4.7	2.3 ±1.1	5.2 ±2.2	–	–	12.6 ±4.4	24.9 ±4.7	2.3
		MS-SFT	35.3 ±4.1	2.3 ±1.0	5.7 ±1.9	–	–	11.0 ±3.7	26.6 ±2.0	1.9
		NIH-SFT	33.0 ±5.1	2.1 ±1.0	6.5 ±2.6	–	–	9.6 ±2.2	30.7 ±1.5	1.9
Gemma	NIH	Base	33.7 ±5.6	7.5 ±1.0	5.0 ±2.0	5.5 ±1.6	11.0 ±4.2	4.2 ±1.4	2.7 ±0.4	–
		MS-SFT	23.0 ±4.7	6.8 ±0.8	4.7 ±2.9	7.7 ±2.2	22.0 ±4.5	9.3 ±2.3	2.1 ±0.4	–
		NIH-SFT	6.1 ±1.3	3.8 ±0.6	4.6 ±3.9	9.0 ±2.5	49.7 ±6.0	6.1 ±1.7	4.7 ±0.8	–
	MS	Base	49.7 ±7.3	8.1 ±1.1	5.5 ±3.1	–	–	4.2 ±1.9	2.3 ±0.4	2.6
		MS-SFT	36.6 ±9.3	7.9 ±1.3	9.4 ±5.5	–	–	13.0 ±4.0	1.9 ±0.4	2.8
		NIH-SFT	13.0 ±10.1	7.7 ±3.8	25.0 ±16.0	–	–	16.3 ±8.6	8.0 ±3.4	2.9

E EXAMPLE PROMPT

You are an expert classifier. Your task is to label the following image as one of the following classes: "a photo of a abyssinian, a type of pet.", "a photo of a american bulldog, a type of pet.", "a photo of a american pit bull terrier, a type of pet.", "a photo of a basset hound, a type of pet.", "a photo of a beagle, a type of pet.", "a photo of a bengal, a type of pet.", "a photo of a birman, a type of pet.", "a photo of a bombay, a type of pet.", "a photo of a boxer, a type of pet.", "a photo of a british shorthair, a type of pet.", "a photo of a chihuahua, a type of pet.", "a photo of a egyptian mau, a type of pet.", "a photo of a english cocker spaniel, a type of pet.", "a photo of a english setter, a type of pet.", "a photo of a german shorthaired, a type of pet.", "a photo of a great pyrenees, a type of pet.", "a photo of a havanese, a type of pet.", "a photo of a japanese chin, a type of pet.", "a photo of a keeshond, a type of pet.", "a photo of a leonberger, a type of pet.", "a photo of a maine coon, a type of pet.", "a photo of a miniature pinscher, a type of pet.", "a photo of a newfoundland, a type of pet.", "a photo of a persian, a type of pet.", "a photo of a pomeranian, a type of pet.", "a photo of a pug, a type of pet.", "a photo of a ragdoll, a type of pet.", "a photo of a russian blue, a type of pet.", "a photo of a saint bernard, a type of pet.", "a photo of a samoyed, a type of pet.", "a photo of a scottish terrier, a type of pet.", "a photo of a shiba inu, a type of pet.", "a photo of a siamese, a type of pet.", "a photo of a sphynx, a type of pet.", "a photo of a staffordshire bull terrier, a type of pet.", "a photo of a wheaten terrier, a type of pet.", "a photo of a yorkshire terrier, a type of pet.". You will now be given a sequence of images and a label following each image that specifies its classification. Using these examples, please label the test image.
<image>
Label: a photo of a japanese chin, a type of pet.
<image>
Label: a photo of a shiba inu, a type of pet.
<image>
Label: a photo of a egyptian mau, a type of pet.
<image>
Label: a photo of a wheaten terrier, a type of pet.
<image>
Label:

Table 6: **Dataset Overview.** Summary of dataset characteristics, including representative samples, number of classes, if the samples are class-balanced or not, their respective roles in training or evaluation, train set size, and validation set size.

Dataset	Example	N. Class	Bal.	Heldout	N. Train	N. Val.
Ham10k (Tschandl et al., 2018)		7	✗	✗	9,577	2,492
STL-10 (Coates et al., 2011)		10	✓	✗	5,000	8,000
RESISC45 (Cheng et al., 2017)		45	✓	✗	18,900	6,300
Oxford Pets (Parkhi et al., 2012)		37	✓	✗	3,680	3,669
CLEVR (Johnson et al., 2017)		6	✓	✗	6,300	15,000
CLEVR Count (Johnson et al., 2017)		8	✓	✗	6,300	15,000
Facial Expression Recognition (FER) (Goodfellow et al., 2015)		7	✗	✗	28,709	7,178
Patch Camelyon (Ehteshami Bejnordi et al., 2017)		2	✓	✗	262,144	32,768
Rendered SST2 (Radford et al., 2021)		2	✓	✗	6,920	1,821
Describable Textures Dataset (DTD) (Cimpoi et al., 2014)		47	✓	✗	3,760	1,880
Hateful Memes (Kiela et al., 2020)		2	✗	✓	8,500	540
Eurosat (Helber et al., 2019)		10	✗	✓	16,200	5,400
German Traffic Sign Recognition (GTSRB) (Stallkamp et al., 2011)		43	✗	✓	26,640	12,630