What do vision-language models see (or not) in the context? Investigating multimodal in-context learning

Anonymous authors
Paper under double-blind review

Abstract

In-context learning (ICL) enables Large Language Models (LLMs) to learn tasks from demonstration examples without parameter updates. Although it has been extensively studied in LLMs, its effectiveness in Vision-Language Models (VLMs) remains underexplored. In this work, we present a systematic study of ICL in VLMs, evaluating seven models spanning four architectures on three image captioning benchmarks. We analyze how prompt design, architectural choices, and training strategies influence multimodal ICL. To our knowledge, we are the first to analyze how attention patterns in VLMs vary with an increasing number of in-context demonstrations. Our results reveal that training on image–text interleaved data enhances ICL performance but does not imply effective integration of visual and textual information from demonstration examples. In contrast, instruction tuning improves instruction-following but can reduce reliance on in-context demonstrations, suggesting a trade-off between instruction alignment and in-context adaptation. Attention analyses further show that current VLMs primarily focus on textual cues and fail to leverage visual information, suggesting a limited capacity for multimodal integration. These findings highlight key limitations in the ICL abilities of current VLMs and provide insights for enhancing their ability to learn from multimodal in-context examples.

1 Introduction

Large Language Models (LLMs) have demonstrated notable performance in a wide range of Natural Language Processing (NLP) tasks, showcasing their potential across various domains. As the scale of LLM increases, in-context learning (ICL) emerges as a new ability that enables models learning new tasks from a few demonstration examples Brown et al. (2020); Wei et al. (2022). In this paradigm, the text generation process is conditioned on a set of input-output, i.e., demonstrations, which enhance the prompt with contextual information. Since ICL does not require parameter updates Dong et al. (2024), it has been used as a cost-effective alternative to traditional fine-tuning for many NLP applications.

Although the ICL ability of LLMs has been studied from multiple perspectives Dong et al. (2024), comparatively little attention has been given to understanding this capacity in Vision-Language Models (VLMs) Baldassini et al. (2024); Qin et al. (2024); Yang et al. (2024); Chen et al. (2024). Exploring ICL in VLMs is particularly important because strategies that are effective for LLMs are not necessarily transferable to multimodal settings, as demonstrated by Li et al. (2024). Previous works have mostly focused on investigating demonstration selection and ordering strategies as well as the contribution of each modality to ICL, with an emphasis on tasks such as Visual Question Answering (VQA) and image classification. However, these studies typically evaluate a limited set of models trained on interleaved image-text data (i.e., datasets composed of instances consisting of multiple images and texts interleaved), leaving open questions about the generalization of their findings to other tasks and to VLMs trained on image-text pair datasets, where each instance comprises only an image and an associated text. In addition, it remains unclear how VLMs use their contextual information when performing downstream tasks.

To address this gap, we present a comprehensive study of ICL in VLMs, with a focus on the task of image captioning. We evaluate seven models that cover four distinct architectures across three image captioning

benchmarks. Unlike prior work, our analysis includes both models trained on interleaved image-text data (OpenFlamingo Awadalla et al. (2023), Idefics2 Laurençon et al. (2024b), and LLaVA-Next-Interleave Li et al. (2025)) and models trained on image-text pair datasets (InstructBLIP Dai et al. (2024) and LLaVA v1.5 Liu et al. (2023)). Through a series of controlled experiments, we systematically analyze how different model architectures and training strategies impact ICL performance. Particularly, we conducted experiments varying the instructions, blacking out and removing demonstration images, and we studied their effect on the ICL capacity of different VLMs. We hypothesize that a VLM with robust multimodal ICL capacity can efficiently leverage the textual and visual information from demonstration examples to generate the answer. Then, its performance should be minimally impacted by the changes in instruction. Still, it should be significantly hampered when corrupting demonstrations (in this case, removing or blacking out images). To our knowledge, we are the first to investigate ICL in VLMs through the lens of attention patterns, offering new insights into how VLMs attend to context and revealing limitations in their current ICL capabilities.

Our main findings are as follows:

- Training data structure significantly impacts ICL capacity; in particular, training on image-text interleaved datasets improves models' ICL ability. However, this benefit does not imply effective integration and use of visual and textual information from demonstration examples.
- Through an analysis of attention maps, we find that the evaluated models do not fully exploit in-context visual information; their ICL behavior is primarily driven by textual context, suggesting a limited integration of multimodal cues.
- While instruction tuning improves instruction-following ability, allowing models to comprehend detailed instructions, it can impair ICL by diminishing the model's reliance on in-context demonstration.

These findings highlight crucial limitations in current VLMs that should be addressed to enhance their multimodal ICL ability.

2 Related Work

VLMs. VLMs excel in vision-language tasks due to pre-trained visual encoders and LLMs Yin et al. (2024); Zhang et al. (2024). They comprise three key components: a visual encoder for image features, an LLM for text generation, and a modality projector to align visual and textual data, bridging the modality gap.

Various approaches have been explored for the modality projector, including linear layers and multi-layer perceptrons (MLPs) Koh et al. (2023); Liu et al. (2023); Shukor et al. (2023); Su et al. (2023); Lin et al. (2024); Liu et al. (2024a), which, despite the low training costs, can lead to long sequences of tokens, thereby increasing the inference costs. Pooling strategies help mitigate this issue Cha et al. (2024); Sun et al. (2024); Hu et al. (2024). Advanced methods like Q-Former Li et al. (2023) improve alignment between frozen visual encoders and LLMs Zhu et al. (2024a); Dai et al. (2024); Geigle et al. (2024). Another alternative is to use interleaved cross-attention layers Alayrac et al. (2022); Laurençon et al. (2023); Xue et al. (2024), in which the LLM directly attends to visual features. However, this approach substantially increases the number of trainable parameters, as noted by Laurençon et al. (2024b).

Training these models typically involves pretraining the modality projector on large-scale image-text datasets while keeping the visual encoder and LLM frozen for feature alignment. Subsequently, the LLM can be fine-tuned alongside the modality projector on instruction-following datasets to improve zero-shot generalization. Most works Dai et al. (2024); Liu et al. (2024a; 2023); Zhu et al. (2024a); Hu et al. (2024) train on a mixture of image captioning Lin et al. (2014); Li et al. (2022); Sharma et al. (2018), VQA Goyal et al. (2017); Schwenk et al. (2022); Marino et al. (2019), and instruction-following Liu et al. (2024a) datasets. Some models, such as Flamingo Alayrac et al. (2022), Idefics Laurençon et al. (2023); Laurençon et al. (2024b;a), VILA Lin et al. (2024), MMICL Zhao et al. (2024), MM1 McKinzie et al. (2025), and xGen-MM (BLIP-3) Xue et al. (2024), are trained on interleaved image-text datasets Laurençon et al. (2023); Zhu et al. (2024b) to further enhance multimodal reasoning capabilities.

ICL in VLMs. Although ICL has been widely studied in LLMs, it remains relatively underexplored in VLMs. Recent works have investigated the factors that influence ICL in VLMs, including modality importance, recency bias, demonstration retrieval, and ordering strategies. However, these studies are generally limited to a small set of models trained on interleaved image-text datasets, with a focus primarily on VQA and image classification tasks.

Qin et al. (2024) studied ICL in VLMs trained with interleaved data under different scenarios. They showed that the internal order of the modalities within each demonstration has a greater impact on performance than the arrangement of demonstrations themselves. Also, unlike ICL in LLMs, where increasing the number of demonstrations typically improves performance, they did not observe significant performance gains when providing more demonstrations.

Yang et al. (2024) investigated ICL for image captioning, analyzing different strategies for demonstration retrieval and caption assignment. Their findings suggest that when demonstration images are similar to the query image, VLMs may leverage in-context captions as shortcuts to generate a new one rather than learning the captioning task. They conducted their experiments, however, within a restricted scope, using only MS COCO Lin et al. (2014) and experimenting with only Idefics and OpenFlamingo models, which limited the generalizability of their conclusions.

More related to our work, Chen et al. (2024) and Baldassini et al. (2024) investigated ICL in two Flamingo-based VLMs: Idefics and OpenFlamingo. They showed that textual information plays a more important role than visual information in the demonstrations. Removing images results in only a minor performance decrease, whereas corrupting textual descriptions leads to a significant performance decline, indicating that these VLMs heavily rely on textual cues even when processing multimodal demonstrations. Moreover, Baldassini et al. (2024) found that these models exhibit recency bias, tending to replicate outputs of the most recent demonstrations, even when earlier demonstrations are more semantically relevant.

In this work, we focus on the task of image captioning and present a systematic analysis of ICL in seven VLMs across four distinct architectures and three benchmark datasets. Unlike previous studies, we extend our investigation to include InstructBLIP Dai et al. (2024) and LLaVA v1.5 Liu et al. (2023), originally designed for single image-text pairs. To our knowledge, this is the first comprehensive evaluation of ICL in VLMs that have not been trained on interleaved image-text datasets. Additionally, we are the first to investigate attention patterns across the layers of text decoder blocks in different VLM architectures as the number of demonstrations varies, providing new insights into the limits of their ICL capacity.

3 Methodology

3.1 Experimental Setup

Models. We analyze four distinct families of VLMs: InstructBLIP Dai et al. (2024), LLaVA Liu et al. (2023); Li et al. (2025), OpenFlamingo Awadalla et al. (2023), and Idefics2 Laurençon et al. (2024b). These families were selected to systematically explore how various design choices, such as bridging the modality gap and different training strategies, affect the ICL capabilities of VLMs.

We use model checkpoints with parameter sizes ranging from 4B to 9B for a fair comparison across similar scenarios. Specifically, for InstructBLIP, we evaluate two checkpoints with different LLMs: InstructBLIP FlanT5-XL and InstructBLIP Vicuna 7B. For the other families, we assess LLaVA v1.5 7B, LLaVA-NeXT-Interleave, OpenFlamingo 9B, and two checkpoints of Idefics2, before and after the instruction-tuning phase, namely, Idefics2 (Base) and Idefics2 (IT)[1].

Datasets & Metrics. We evaluate the models using three image captioning benchmarks: MS COCO Lin et al. (2014), Flickr30K Young et al. (2014) and NoCaps Agrawal et al. (2019) datasets. We conduct our evaluation on the respective validation sets, utilizing the MS COCO training set as the knowledge base

[1] https://huggingface.co/Salesforce/instructblip-flan-t5-xl, https://huggingface.co/Salesforce/instructblip-vicuna-7b, https://huggingface.co/llava-hf/llava-1.5-7b-hf, https://huggingface.co/llava-hf/llava-interleave-qwen-7b-hf, https://huggingface.co/openflamingo/OpenFlamingo-9B-vitl-mpt7b, https://huggingface.co/HuggingFaceM4/idefics2-8b-base, and https://huggingface.co/HuggingFaceM4/idefics2-8b.

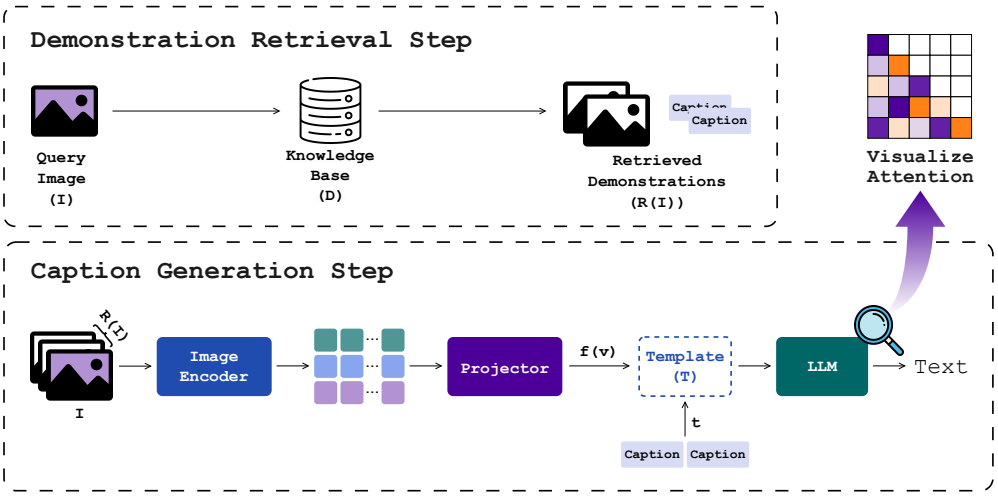

Figure 1: Overview of our evaluation pipeline for assessing the ICL capability of VLMs. We illustrate the demonstration retrieval and caption generation steps.

from which we retrieve similar examples to construct the context. Each demonstration example comprises an image-caption pair, with the caption being randomly sampled from the multiple human annotations available for MS COCO. We employ CIDEr-D Vedantam et al. (2015) as the evaluation metric.

3.2 Evaluation Protocol

Demonstrations Retrieval. Inspired by Yang et al. (2023), we retrieve demonstration examples employing a k-Nearest Neighbor approach based on the similarity distance in the visual feature space. We construct a knowledge base $\mathcal{D} = \{(i_1, t_1), \ldots, (i_n, t_n)\}$, consisting of images i paired with their corresponding texts t different from those in the evaluation sets. In our experiments, we use the MS COCO training set as our knowledge base.

For each query image I, we extract its features $f(I)$ and we retrieve the top-k most similar image-text pairs based on the cosine similarity between visual features, as illustrated in the "Demonstration Retrieval Step" in Figure 1. Formally, the retrieved set $\mathcal{R}(I)$ of image-text pairs is defined as $\mathcal{R}(I) = \{(i,t) \,|\, \text{top-}k_{(i,t)\in\mathcal{D}}\ sim(f_I, f_i)\}$[2], where $sim(\cdot)$ denotes the cosine similarity. We use a ViT Dosovitskiy et al. (2021)[3] to encode the images.

ICL. During the caption generation step (Figure 1), we first extract visual features from the query image and from the images in the retrieved demonstration set $\mathcal{R}(I)$. These features are projected into the LLM's token embedding space, producing visual tokens denoted as $f(v)$. We then construct a multimodal prompt by inserting the visual tokens and their corresponding captions into a predefined template $\mathcal{T}$. This prompt is passed to the LLM to generate the caption.

Although this pipeline is implemented in a relatively straightforward manner for LLaVA, Idefics2, and OpenFlamingo, adapting it to InstructBLIP presents additional challenges. InstructBLIP employs a Q-Former module to extract instruction-aware visual features. To extend it to handle multiple demonstrations, we process each demonstration image independently using the Q-Former, paired with a fixed instruction: `"a short image description"`. The resulting visual query tokens are then embedded in the template $\mathcal{T}$ alongside their corresponding captions.

Templates. To evaluate the models' ability to adapt at inference time, we experiment with two different templates. First, for each model, we construct a straightforward template based on its original training

[2]For simplicity, we denote $f(i)$ as f_i and $f(I)$ as f_I.
[3]`https://huggingface.co/google/vit-large-patch16-224-in21k`

Straightforward Template:

User: **<image$_1$>**Describe this image. [Caption$_1$]
User: **<image$_2$>**Describe this image. [Caption$_2$]
User: **<image3>**Describe this image. [Caption$_3$]
User: **<query image>**Describe this image.

Detailed Template:

User: I am an intelligent image captioning bot. Here are the features extracted for similar images along with their captions, following the format: [visual query tokens] [caption].
<image$_1$> [Caption$_1$], **image$_2$>** [Caption$_2$], <**image$_3$>** [Caption$_3$]. **<query image>** A short caption I can generate to describe this image is:

Figure 2: Investigated templates.

instructions, into which we directly embed the demonstration examples $\mathcal{R}(I)$, i.e., image-caption pairs. Additionally, building upon the Socratic Models, we further explore detailed prompts based on Socratic templates Zeng et al. (2023); Ramos et al. (2023) that specify the task and the format for presenting demonstration examples (Figure 2). Following Baldassini et al. (2024), in our experiments, we sort the demonstrations in increasing order of similarity to the query image, as the models tend to assign greater importance to the last demonstrations.

4 Results and Discussions

Effect of Prompt Structure on ICL. To investigate the influence of prompt structure on ICL, we evaluate models on the image captioning task using prompts designed with two levels of detail. The first is built using the straightforward template, where the demonstration image-caption pairs are simply concatenated with an instruction. The second employs the detailed template, clearly specifying the format of the examples and including the phrase "`I am an intelligent image captioning bot`" (Figure 2). Figure 3 shows the results.

Overall, we observed that Idefics2 (Base) and OpenFlamingo perform better when using the straightforward template compared to the detailed one. However, as we increase the number of demonstrations (shots), Idefics2 (Base) begins to perform similarly across both templates. This trend does not hold for OpenFlamingo, which consistently achieves higher scores with the straightforward template regardless of the number of shots. Interestingly, after the instruction-tuning step, Idefics2 exhibits similar performance for both templates. Also, a closer look reveals that Idefics2 (IT) demonstrates significant gains in the zero-shot scenario compared to Idefics2 (Base). However, it converges to a point below that of its base checkpoint, indicating a deterioration of ICL.

InstructBLIP models and LLaVA models[4], both instruction-tuned models, perform similarly with straightforward and detailed templates. These results align with previous works Liu et al. (2024a); Wu et al. (2024), indicating that instruction tuning enhances a model's ability to follow instructions, as evidenced by similar

[4]Due to computational constraints, we evaluate LLaVA models up to 16 demonstrations.

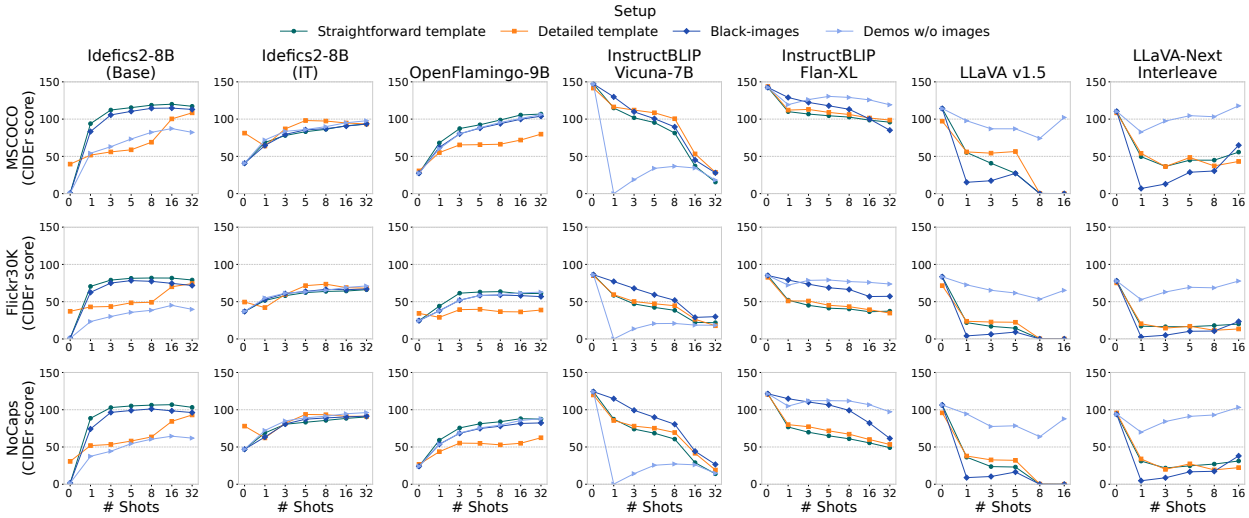

Figure 3: **Evaluating ICL capacity of VLMs across different scenarios**. We evaluate the models using both straightforward and detailed templates. Additionally, we explore scenarios where demonstration captions are provided. However, the demonstration images are blacked out, as well as cases where only the captions are available as context. "Idefics2 8B (IT)" refers to the instruction-tuned checkpoint of the Idefics2 architecture.

performance across different prompts after fine-tuning. On the other hand, our results suggest that it can also significantly hamper the model's ICL ability, as seen in Idefics2 models, where CIDEr-D scores drop by 20 points after instruction tuning.

Interleaved vs. Paired Image-Text Training. Additionally, we investigate how the ICL capacity of VLMs trained on image-text interleaved datasets (Idefics2, OpenFlamingo, and LLaVA-Next-Interleave) differs from those trained on datasets composed of image-text pairs (InstructBLIP and LLaVA v1.5). As shown in Figure 3, these two training paradigms yield opposite trends as the number of demonstrations increases. Models trained on image-text interleaved datasets perform consistently better with more demonstrations, indicating strong ICL capabilities. In contrast, the performance of models trained on image-text paired datasets significantly declines as the number of shots increases, showing they have a limited ICL capacity. Notably, LLaVA v1.5's performance drops to zero when given eight or more demonstrations, whereas LLaVA-Next-Interleave shows stable performance across the number of demonstrations.

This finding suggests that the training data structure plays a critical role in shaping a model's ICL capacity, with image-text interleaved datasets contributing to enhancing such capacity. Furthermore, we observe that the two variants of InstructBLIP behave differently. While InstructBLIP Flan-XL appears to plateau at eight shots, InstructBLIP Vicuna-7B continues to decline. We hypothesize that this difference arises from the presence of few-shot templates in the Flan-T5 training set, which enhances the LLM's ICL capabilities in the textual domain. Then, part of the ICL ability learned by the VLM's text decoder can also be leveraged in the multimodal setup.

Do VLMs "See" In-Context Images? To investigate the contribution of the visual modality to model performance, we evaluate two ablation scenarios. First, given a query image, we retrieve similar demonstrations but replace the retrieved images with black ones while preserving their captions. These modified demonstrations are then inserted into the straightforward template (Figure 2). In the second scenario, rather than blacking out the demonstration images, we simply remove them from the context, including only their associated captions into the straightforward template. The results of these experiments are shown in Figure 3.

Comparing the performance of models using demonstrations with original images (straightforward template) against those with blacked-out images (black-images), we find that most models perform similarly across different shot counts. Particularly, both InstructBLIP variants exhibit improved performance on Flickr30K and NoCaps when images are blacked out, although still with a downward trend with respect to shot. In contrast, when we remove demonstration images from the context (demos w/o images), models behave differently. Idefics2 (Base) and InstructBLIP Vicuna-7B exhibit a sharp performance drop, especially at low shot counts. For Idefics2 (Base), this performance degradation is likely because when we pass only similar captions as context, we disrupt the image-text interleaved structure on which it was originally trained, resulting in a prompt out of the training distribution and confusing the model. This issue seems to be mitigated in Idefics2 (IT), indicating that instruction tuning also enhances robustness to such structural changes. OpenFlamingo, in turn, does not exhibit a significant difference in performance when using or not using demonstration images. Conversely, LLaVA models and InstructBLIP Flan-XL perform better when only captions are included, suggesting that the ICL capacity of these models relies mostly on text while visual content may distract their LLMs during text decoding. Overall, these findings suggest that the evaluated models do not "see" images in the context; instead, their ICL ability is predominantly based on the textual modality.

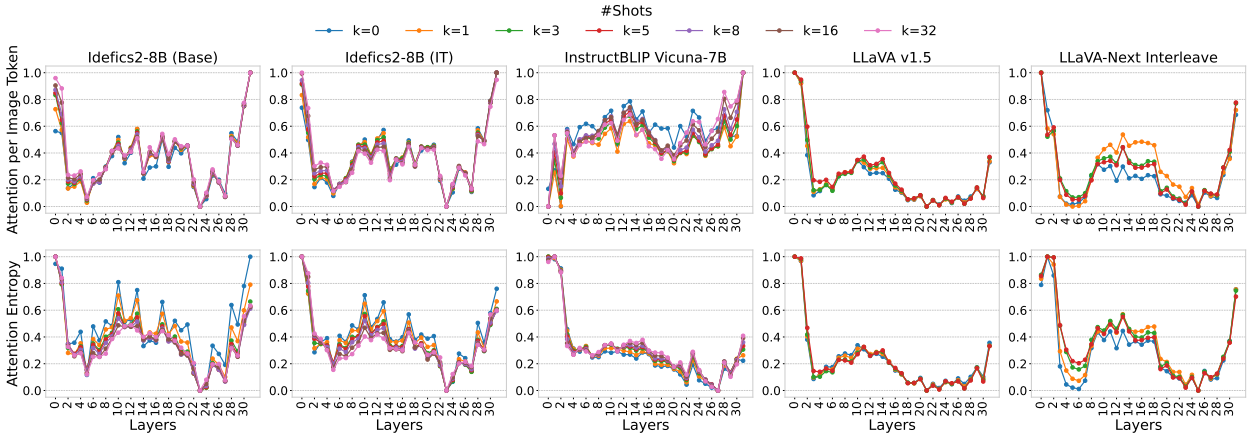

Figure 4: **Layer-wise attention analysis**. The upper row presents the variation of mean attention weight assigned to a visual token across the models' LLM layers. The lower row shows the attention entropy across all tokens at each LLM layer, reflecting the diffuseness of attention distribution. For comparability, the charts plot min-max normalized values.

Analysis of Attention Patterns. To further analyze what models "see" and how different training strategies impact ICL capacity, we select Idefics2 models, InstructBLIP Vincuna-7B, LLaVA v1.5, and LLaVA-Next-Interleave for a close analysis. The choice of these models took into account the similarity in terms of architecture of these models, as both of them use decoder-only LLMs and pass the visual information as input tokens to the LLM. We investigate how the attention weights assigned to visual tokens and the entropy of attention across tokens vary throughout the LLMs layers and whether the patterns are consistent as the number of in-context demonstrations increases. In this experiment, we use the straightforward template.

Figure 4 shows that overall models exhibit similar trends for the attention entropy. It is maximum in the lower layers, then there is a sharp fall followed by an inverted U-shaped curve in the middle layers, and finally it rises again. Moreover, considering normalized values, entropy trends remain relatively stable across different shot counts. However, Idefics2 (Base) presents higher entropy in the final layers under the zero- and one-shot condition, consistent with its weaker performance in those settings. The results show that, for all evaluated models, attention is more diffuse among tokens in early and also in late layers for Idefics2 models.

In terms of attention per image token, the variation of normalized attention scores across layers shows a similar pattern for all shots, consistently observed across all evaluated models. Moreover, models assign maximum attention to the visual tokens in the early layers with a sharp fall followed by an inverted U-shaped

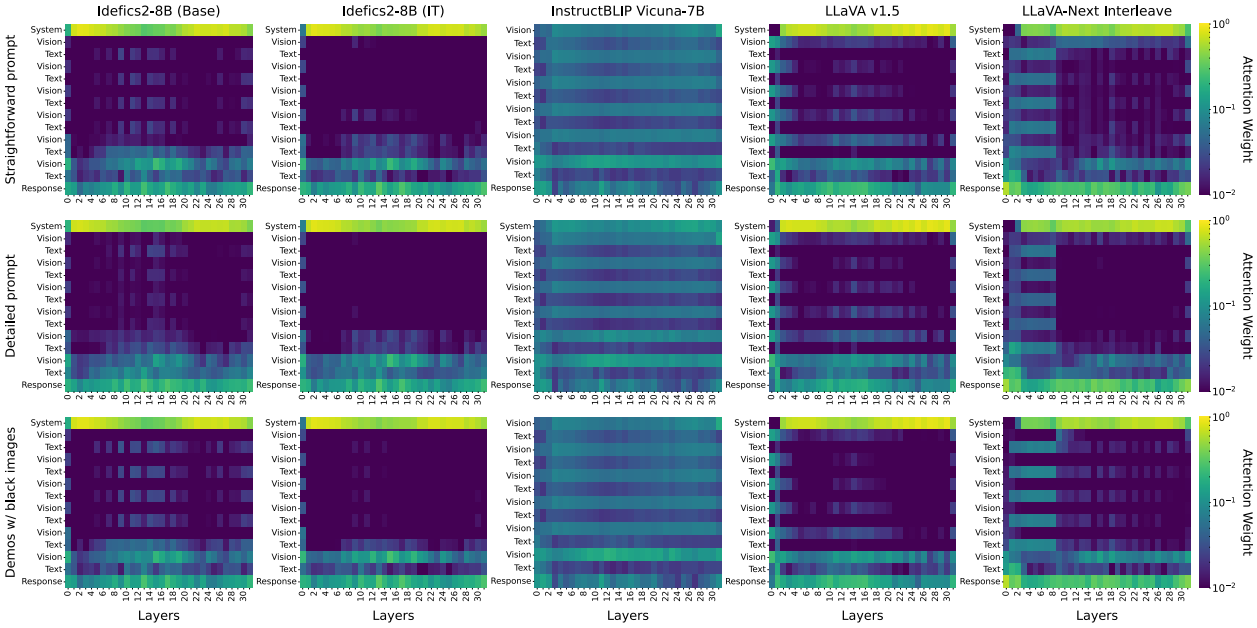

Figure 5: **Attention maps with scores aggregated by token type** (log-scale). Maps are plotted for InstructBLIP Vicuna-7B and Idefics2 models, comparing the 5-shot setting across prompts built on the straightforward template (first row), detailed template (second row), and demonstrations with blacked-out images (third row). Columns correspond to the respective models.

curve in the middle layers, and finally it turns to an increase in the later layers. Conversely, InstructBLIP Vicuna-7B assigns minimal attention to visual tokens in early layers, which increases in deeper layers. This difference can be explained by the fact that InstructBLIP Vicuna-7B was the only model that kept the text decoder frozen during the whole training. Then, the LLM possibly treats the visual tokens in the same way as textual tokens. On the other hand, when the LLM is unfrozen during VLM's training, it may learn to assign higher attention to visual tokens in early layers in order to extract relevant visual information. This finding is consistent with the previous conclusion Zhang et al. (2025), and we demonstrate that it also generalizes to Idefics2 models, in addition to LLaVA, and holds across different shot numbers.

Next, we further analyze how the models attend to individual images and text segments in the context. We plot the total attention weight assigned to each subsequence, i.e., textual or visual tokens, and previously generated tokens (response), across all layers. We conduct this analysis in the 5-shot setting, using straightforward and detailed templates, as well as demonstrations with blacked-out images.

As shown in Figure 5, InstructBLIP Vicuna-7B's behavior contrasts with Idefics2 and LLaVA models. InstructBLIP Vicuna-7B distributes attention more uniformly across the tokens in the early layers, also evidenced by the high entropy, and concentrates attention on visual tokens in the middle and final layers. In contrast, Idefics2 and LLaVA models assign the highest attention to the first textual segment, system prefix[5], across all layers. Moreover, these models tend to concentrate attention near the end of the token sequence, particularly on the query image, task instruction, and previously generated tokens. This observation aligns with the conclusion of Liu et al. (2024b) that LLMs give more importance to information at the beginning and end of context, and instruction fine-tuned models tend to assign a high attention score to the system prefix. It is also consistent with the improved performance reported when demonstrations are sorted by increasing similarity to the query image Baldassini et al. (2024).

Moreover, LLaVA v1.5 assigns higher attention to visual tokens in early layers and less significant attention in the middle layers, while ignoring the textual information of demonstrations. Conversely, LLaVA-Next-Interleave assigns insignificant attention to demonstration images and higher attention to textual tokens

[5]All models except InstructBLIP Vicuna-7B use a system prompt (e.g., `"USER:"`) before user instruction.

of demonstrations in early and middle layers, focusing on the query image and system prefix in the late layers. Idefics2 (Base), in turn, assigns attention to the query and the last few images in early layers, and it distributes attention among demonstration captions in the middle and final layers. After instruction tuning, Idefics2 (IT) seems to ignore the information in the middle of the context after the first layer. This is consistent with the weaker ICL ability of Idefics2 (IT) compared to Idefics2 (Base), suggesting that instruction tuning may reduce the use of demonstration content. We observe no substantial differences in attention distribution between the straightforward and detailed templates, nor between original and blacked-out demonstrations.

These findings further support our hypothesis that the evaluated VLMs have limited capacity to leverage multimodal in-context information. For Idefics2 and LLaVA-Next-Interleave models, ICL appears to rely predominantly on textual information. Moreover, instruction tuning in Idefics2 may reduce reliance on demonstration captions, potentially impacting ICL performance. In contrast, InstructBLIP Vicuna-7B concentrates attention on images while ignoring the captions. These insights underscore the importance of achieving balanced attention across modalities to fully exploit ICL potential in multimodal settings.

5 Conclusions

In this paper, we conduct a comprehensive study of ICL in VLMs, evaluating seven models spanning four distinct architectures on several image captioning benchmarks. We investigate how prompt design, model architecture, and training data structure influence ICL performance. In contrast to prior work, we go beyond models trained solely on interleaved image-text data; we also analyze VLMs trained on datasets composed of image-text pairs. We find that instruction tuning can enhance instruction-following ability, allowing models to comprehend detailed instructions, but it can hamper ICL capacity by diminishing the model's reliance on in-context demonstration. In contrast, training on interleaved image-text datasets enhances ICL ability. However, the benefits do not necessarily extend to multimodal settings. Our attention map analysis reveals that these models do not fully leverage visual inputs; their ICL behavior is largely driven by textual information, indicating limited capacity for integrating multimodal information.

Our findings uncover critical limitations in current VLMs. Future research should investigate whether our findings generalize to larger-scale models and explore to what extent VLMs inherit and utilize the ICL capabilities of their underlying LLMs. Designing more effective modality projectors may be the key to better transferring these abilities. Finally, investigating training strategies that combine instruction tuning with interleaved image-text supervision to support both instruction following and contextual learning is a promising direction.

Limitations

Although our analysis focuses on VLMs with up to 9B parameters due to computational constraints, studying larger models would be important to determine whether our conclusions hold at a greater scale. Furthermore, to better understand the role of instruction-tuning and training of interleaved image-text datasets, it would be interesting to evaluate more models before and after instruction-tuning. Finally, our analysis is limited to VLMs trained in English-language texts. However, evaluating the ICL capacity of multilingual models is essential. It would be necessary to study whether ICL can improve VLMs' performance on low-resource languages.

Ethics Statement

This study systematically analyzes the ICL capabilities of publicly available VLMs. Our analysis is based solely on publicly available image captioning datasets, and we fully comply with the terms of use and licensing agreements associated with each model and dataset. We do not conduct any fine-tuning or modifications in the models that could introduce unintended risks. However, we recognize that our work reflects the existing limitations and potential risks of the evaluated models, including, but not limited to, gender, racial, and cultural biases, as well as the potential for generating misinformation or disinformation.

LLM Usage Statement

We used GPT-5 solely for language assistance during the preparation of this paper, including improving grammar and clarity and polishing the author's original content. The model was not used to generate scientific ideas, to design the method, to conduct experiments, to analyze results, or to draw conclusions. All technical content, claims, and interpretations were developed and verified by the authors, who take full responsibility for the final manuscript.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. NoCaps: Novel object captioning at scale. In *IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 23716–23736, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1539–1550, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, pp. 1877–1901, 2020.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13817–13827, 2024.

Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. In *Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Benoit Courty, Victor Schmidt, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, SabAmine, inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Amine Saboni, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, alencon, Michał Stęchły, Christian Bauer, Lucas-Otavio, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, 2024. URL `https://doi.org/10.5281/zenodo.11171501`.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 2024.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. In *Workshop on Advances in Language and Vision Research*, pp. 7–25, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics*, pp. 3096–3120, 2024.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pp. 17283–17300, 2023.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Advances in Neural Information Processing Systems*, pp. 71683–71702, 2023.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Leo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024a. URL `https://openreview.net/forum?id=iSL0FHZStr`.

Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems*, 2024b.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *International Conference on Learning Representations*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26710–26720, 2024.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, June 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2024a.

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. *Transactions of the Association for Computational Linguistics*, 11:157–173, 2024b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer ision and Pattern Recognition*, pp. 3195–3204, 2019.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: Methods, analysis and insights from multimodal LLM pre-training. In *European Conference on Computer Vision*, pp. 304–323, 2025.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. What factors affect multi-modal in-context learning? an in-depth exploration. In *Advances in Neural Information Processing Systems*, 2024.

Rita Ramos, Bruno Martins, and Desmond Elliott. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics*, pp. 1635–1651, 2023.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018.

Mustafa Shukor, Corentin Dancette, and Matthieu Cord. eP-ALM: Efficient Perceptual Augmentation of Language Models. In *IEEE/CVF International Conference on Computer Vision*, pp. 22056–22069, 2023.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. In Devamanyu Hazarika, Xiangru Robert Tang, and Di Jin (eds.), *Workshop on Taming Large Language Models: Controllability in the Era of Interactive Assistants!*, pp. 11–23, 2023.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision andPattern Recognition*, pp. 4566–4575, 2015.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2341–2369, 2024.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S. Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xGen-MM (BLIP-3): A Family of Open Large Multimodal Models. *arXiv preprint arXiV:2408.08872*, 2024.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 2024.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning. In *Findings of the Association for Computational Linguistics*, pp. 11844–11857, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, pp. 1–18, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations*, 2023.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics*, pp. 12401–12430, 2024.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, 2025.

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. In *International Conference on Learning Representations*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *International Conference on Learning Representations*, 2024a.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *Advances in Neural Information Processing Systems*, 2024b.

A Appendix

A.1 Details on Experimental Setup

To facilitate the reproducibility of our work, we report in Table A1 the models we analyzed, along with details on their number of parameters and training dataset size, as well as the energy consumption and carbon emissions from our experiments estimated with `codecarbon` Courty et al. (2024). Table A2 shows the datasets used in our experiments, including statistics on their size. Additionally, we outline the main text decoding hyperparameters used in our experiments in Table A3. Note that we use the hyperparameters reported for each model for the image captioning task. However, LLaVA v1.5 does not report hyperparameters for this task, then we use the ones from InstructBLIP, which led to the best results in our preliminary experiments. We conducted our experiments in a heterogeneous infrastructure; however, the majority were performed on a single NVIDIA A100 80GB GPU.

A.2 Formalization

In Section 4, we analyze the attention patterns in different VLMs. Here, we formalize how we aggregate the attention scores in our experiments.

Table A1: VLMs investigated in this work. For each model, we report the number of parameters, the size of the training dataset, and the estimated energy consumption and carbon emissions from our experiments.

Model	#Params (B)	Training Set Size (M)	Energy (kWh)	Emission (CO_2eq in kg)
Llava v1.5-7B	7.1	0.15	504.4	1.2
InstructBLIP Vicuna-7B	7.9	15.1	143.1	14.1
InstructBLIP Flan-XL	4.0	15.1	43.4	0.1
Idefics2-8B	8.4	351.2	52.8	5.2
OpenFlamingo-9B	8.1	2,101.0	1,285.0	126.4

Table A2: Datasets used in our experiments. For each dataset, we report the number of samples in each split and the specific task it is used for. Note that we do not use Flickr or NoCaps training sets, as we rely on the MS COCO training set as the knowledge base for these datasets. "Val." stands for the validation dataset.

Dataset	Size
MS COCO	Train: 118.2K/ Val: 5.0K
Flickr30K	Val: 1.0K
NoCaps	Val: 4.5K

Table A3: Text decoding hyperparameters.

Hyperparameters	InstructBLIP / LLaVA	Idefics2	OpenFlamingo
# Beams	5	–	3
Max. New Tokens	30	20	20
Min. Length	10	–	–
Repetition Penalty	1.5	–	–
Length Penalty	1.0	–	–

A.2.1 Attention per Token Type

Formally, given $a_{i,j}^{(h,d)}$ the attention score from i-th to j-th token at h-th attention head for the d-th sample, a sequence of textual and visual tokens $T = (t_1, v_1, \ldots, t_k, v_k)$ and a sequence of response tokens R, we define the attention score per token type T_k as follows:

$$\bar{a}_{i,j} = \frac{1}{|\mathcal{D}|} \frac{1}{H} \sum_d \sum_h a_{i,j}^{(h,d)}, \tag{1}$$

$$a_{T_k} = \frac{1}{|R|} \sum_{i \in R, j \in T_k} \bar{a}_{i,j}, \tag{2}$$

where R denotes the set of response tokens, and $|\mathcal{D}|$ and H represent the dataset size and the number of heads, respectively.

A.2.2 Attention per Visual Token

Given the average attention matrix across all heads and samples, denoted as $\bar{a}$ (Eq. 1), we define the average attention per visual token as follows:

$$a_V = \frac{1}{|V|} \sum_{0 \leq i < N, j \in V} \bar{a}_{i,j}, \tag{3}$$

where V is the set of visual tokens and N is the number of tokens in the sequence.

A.2.3 Attention Entropy

Given the average attention matrix across all heads and samples, denoted as $\bar{a}$ (Eq. 1), we define the attention entropy as follows:

$$\mathcal{H} = -\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \bar{a}_{i,j} log(\bar{a}_{i,j} + \epsilon), \tag{4}$$

where N is the number of tokens in the sequence, and $\epsilon = 10^{-8}$.