AIM: Let Any Multimodal Large Language Models Embrace Efficient In-Context Learning

Jun Gao[1], **Qian Qiao**[1], **Tianxiang Wu**[1], **Zili Wang**[2], **Ziqiang Cao**[1*], **Wenjie Li**[3]
[1] School of Computer Science and Technology, Soochow University, China
[2] Independent Researcher
[3] Computation Department, The Hong Kong Polytechnic University, Hong Kong
{jgao1106, qqiao, 20224227044}@stu.suda.edu.cn,
ziliwang.do@gmail.com, zqcao@suda.edu.cn, cswjli@comp.polyu.edu.hk

Abstract

In-context learning (ICL) advances Large Language Models (LLMs) exhibiting emergent ability on downstream tasks without updating billions of parameters. However, in the area of multimodal Large Language Models (MLLMs), two problems hinder the application of multimodal ICL: (1) Most primary MLLMs are only trained on **single-image** datasets, making them unable to read extra multimodal demonstrations. (2) With the demonstrations increasing, thousands of visual tokens highly challenge hardware and degrade ICL performance. During preliminary explorations, we discovered that the inner LLM focuses more on the linguistic modality within multimodal demonstrations during generation. Therefore, we propose a general and light-weighted framework **AIM** to tackle the mentioned problems through **A**ggregating **I**mage information of **M**ultimodal demonstrations to the latent space of the corresponding textual labels. After aggregation, AIM substitutes each demonstration with generated **fused virtual tokens** whose length is reduced to the same as its texts. Except for shortening input length, AIM further upgrades MLLMs pre-trained on image-text pairs to support multimodal ICL, as images from demonstrations are disregarded. Furthermore, benefiting from aggregating different demonstrations independently, AIM configures **Demonstration Bank** (**DB**) to avoid repeated aggregation, which significantly boosts model efficiency. We build AIM upon QWen-VL and LLaVA-Next, and AIM is comprehensively evaluated on image caption, VQA, and hateful speech detection. Outstanding results reveal that AIM provides an efficient and effective solution in upgrading MLLMs for multimodal ICL.

Code — https://github.com/jungao1106/AIM.
Extended version — https://arxiv.org/abs/2406.07588

Introduction

In-context learning (ICL) (Brown et al. 2020; Xie et al. 2021; Wang et al. 2023a) enhances the reasoning ability of scaled-up Large Language Models (LLMs) on **training-agnostic** data by prepending a few demonstrations (Wang et al. 2023b; Yang et al. 2023; Wei et al. 2023). Unfortunately, primary multimodal Large Language Models (MLLMs) such as LLaVA (Liu et al. 2024b, 2023), LLaMA-Adapter (Zhang et al. 2023), and BLIP-2 (Li et al. 2023b), only support a single image as the vision input, which are impossible to learn from multimodal demonstrations composed of **[image, instruction text, reference text]** pairs. Additionally, to assist the inner LLM understand visual inputs, most MLLMs utilize Perceiver (Alayrac et al. 2022; Awadalla et al. 2023; Liu et al. 2024b, 2023, 2024a; Li et al. 2023b; Dai et al. 2024) to generate abundant visual tokens according to image features extracted by an existing visual encoder. Therefore, multiple images in multimodal demonstrations inevitably produce thousands of visual tokens, resulting in extreme memory costs as depicted in Figure 1. More importantly, existing work (Alayrac et al. 2022; Awadalla et al. 2023; Bai et al. 2023; Zhao et al. 2023; Li et al. 2023a; Huang et al. 2024) point out that the prompt length surged by demonstration images might be one of the key factors constraining the performance of multimodal ICL.

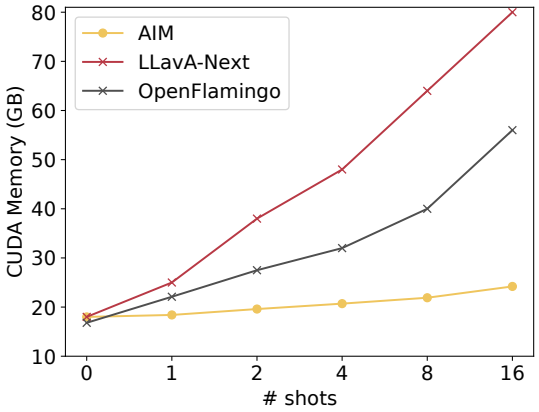

Figure 1: Memory cost comparison between LLaVA-Next, OpenFlamingo, and AIM on Flickr30k. The memory cost of LLaVA-Next and OpenFlamingo occurs a surge, while it almost remains unchanged in AIM.

*Corresponding Author

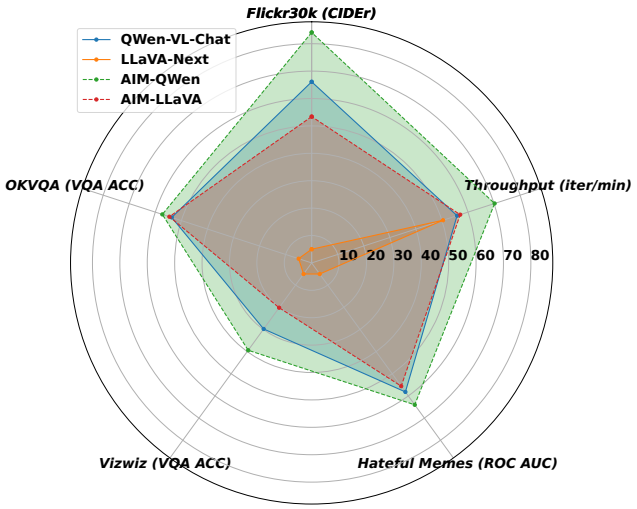

Figure 2: Performance comparison between AIM and its underlying backbone in the 16-shot ICL setting.

During the early exploration, we surprisedly found that MLLMs attend more to the linguistic modality, namely the texts in demonstrations, than the visual modality for response generation, as shown in Figure 3. Motivated by this finding, we propose the framework AIM, with the aim to make any MLLM embrace efficient multimodal ICL. In contrast to mainstream MLLMs that always treat visual and textual tokens equally for both demonstrations and queries (Bai et al. 2023; Liu et al. 2023, 2024b; Zhao et al. 2023), AIM aggregates the image information of each demonstration into its linguistic modality and then drops their lengthy visual tokens. Thus, AIM approximately reduces image-text demonstrations into text-like demonstrations, while the images and texts in queries are unmodified. Specifically, AIM first applies the inner MLLM to infer each demonstration independently to obtain the hidden states of the image and its text. Then, AIM applies an adapter with 17M trainable parameters to transform the hidden states on top of texts into fused virtual tokens, making them acceptable for LLMs. Therefore, the sizes of demonstrations are reduced to the dimensions of their textual embeddings as hidden states of images are disregarded. Finally, the fused virtual tokens replace the original image-text pair demonstration, serving as a text-like one fused with image information, fed into the inner LLM to guide response generation. In this case, the built-in MLLMs only attend to a single query image as images from demonstrations are removed in the input end. AIM can perform ICL even if the backbones don't develop the ability to understand interleaved multimodal inputs during pre-training. Additionally, as aggregating image information is independent, AIM asynchronously processes different demonstrations within a batch, horizontally concatenating fused virtual tokens from different demonstrations for few-shot settings. Benefiting from this, the aggregated fused virtual tokens can be cached to formulate a **Demonstration Bank** (**DB**) for further reusing, avoiding repeated aggregation for the same demonstration.

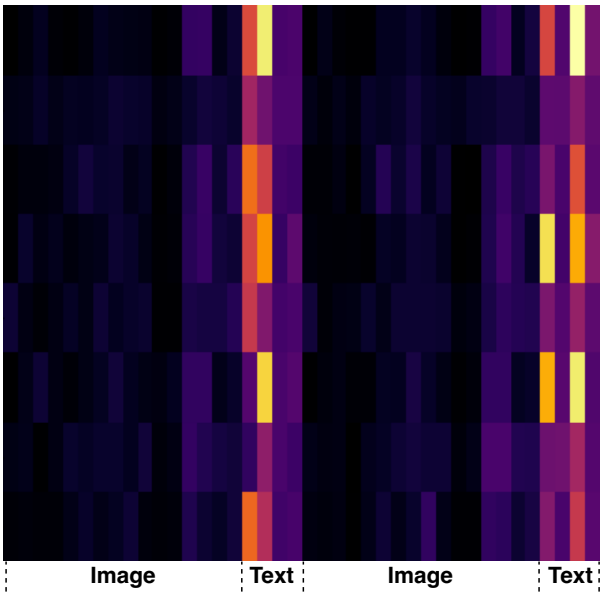

Figure 3: The heat map of attention scores when QWen-VL generates the first token on the hateful memes dataset. The brighter represents that responses to be generated have paid more attention to the current visual/textual tokens. Obviously, the generation relies more on the textual part of a multimodal demonstration.

Considering ICL was proposed in a low-resource setting, previous studies training models on a mixture of downstream task datasets potentially fall into the short-cut answer, resulting in outstanding but not solid results on involved/related tasks. This, following the technical report of OpenFlamingo (Awadalla et al. 2023), we train the adapter on the subset of MMC4 (Zhu et al. 2023), containing 223k images and 1M sequences from websites. We select LLaVA-Next and QWen-VL as the underlying MLLM in AIM to verify the generality, representing MLLMs reading a single image-text pair and images interleaved with texts. Furthermore, we comprehensively evaluate AIM on image caption (Plummer et al. 2015), visual question answering (VQA) (Gurari et al. 2018; Marino et al. 2019), and hateful detection (Kiela et al. 2020), none of the involved datasets occurring in the training data of AIM and mixture downstream pre-training dataset of QWen-VL and LLavA-Next. Outstanding results in Figure 2 reveal that AIM always achieves comparable or better performance than its underlying backbone with less than 10% tokens remaining on average (refer to Table 4). Generally, our main contributions are as follows:

- To the best of our knowledge, we are the first to analyze the attention distribution of multimodal demonstrations during generation, revealing that MLLMs prioritize attention towards the linguistic over the visual modality in multimodal demonstrations.
- Building upon this finding, we propose to transform multimodal demonstrations to text-like representations, enabling any MLLMs qualified for efficient multimodal ICL.

- Our proposed AIM exhibits efficiency in terms of trainable parameters, memory usage, and inference throughput.

Related Work

Multimodal Large Language Models

Recently, the development of LLMs significantly advanced the iterations of MLLMs, and the inner LLMs play crucial roles. Researchers first trained the visual encoder to align to the frozen language models (Tsimpoukelli et al. 2021), performing vision-language tasks. Predominate MLLMs can be abstracted to Perceiver & LLM architecture, where the Perceiver is usually composed of a Vision Transformer (ViT) (Dosovitskiy et al. 2020) to extract image features and an adapter, concatenating visual and language tokens in the input end of the built-in LLM. Specifically, the perceiver in QWen-VL (Bai et al. 2023) and the Q-Former in BLIP-2 (Li et al. 2023b) apply learnable queries to extract visual information based on cross-attention, while the Connector in LLaVA (Liu et al. 2023, 2024b,a) directly projects the visual features extracted from the pre-trained ViT-L/14 in CLIP (Radford et al. 2021). Considering the further alignment within LLMs, Flamingo (Alayrac et al. 2022) introduced the XATTN layer to align visual tokens originating from the Resampler and textual embeddings within the LLM.

However, the perceiver in MLLM will introduce hundreds or even thousands of visual tokens to the inner LLM in ICL, resulting in over-length multimodal prompts and thereby bringing enormous memory costs.

In-Context Learning

In the field of NLP, LLMs including ChatGPT (Luo, Xie, and Ananiadou 2023), GPT-4 (OpenAI 2023), and LLaMA (Meta 2023), exhibit general spectacular emergent abilities on downstream tasks that provide a novel paradigm for generative models known as "Pre-training, Prompting, and Prediction". Within this paradigm, ICL assumes a pivotal role, bolstering the generalization capability of LLMs (Wang et al. 2023b; Yang et al. 2023; Wei et al. 2022; Gao et al. 2024a), without necessitating billions of parameters gradient updating.

The success of ICL in NLP boosts studies focusing on transforming it into the multimodal setting (Tsimpoukelli et al. 2021; Zhao et al. 2023; Alayrac et al. 2022; Yang et al. 2024). Additionally, researchers extensively explore the influence of diverse demonstration configurations in captioning (Yang et al. 2024). As far as we know, recent studies focusing on multimodal ICL (Alayrac et al. 2022; Awadalla et al. 2023; Liu et al. 2023; Zhao et al. 2023; Li et al. 2023a) overlook deployment challenges to some extent. QWen-VL (Bai et al. 2023) and MMICL (Zhao et al. 2023) treat visual and textual tokens equally during training, brought serious length challenges, and restricted model performance due to modeling enormous vision tokens. Flamingo (Alayrac et al. 2022) treated image information as informative noises adding to textual embeddings through extra introduced adapters within selectional inner LLM layers, resulting in additional module latency. However, visual tokens of different images still share the same input window, bringing extra memory costs. LLaVA-Next (Liu et al. 2024a) specialized in single-image inference, and it achieved outstanding performance on popular multimodal benchmarks and textual-only ICL, while its excellent performance failed to extrapolate to practical multimodal ICL settings, exhibiting poor ability of instruction following. Additionally, LLaVA-Next connected pre-trained ViT and LLM via an MLP that resulted in thousands of visual tokens for high-resolution pictures, causing more serious length disasters than perceivers based on cross-attention.

Methods	Inner LLM	# Trainable Para.
Flamingo	Chinchilla (7B)	7B
QWen-VL	QWen (7B)	7B
LLaVA-Next	Vicuna (7B)	7B
AIM	QWen/Vicuna (7B)	17M

Table 1: Quality comparison of recent MLLMs and AIM in ICL mode, LLM size, and trainable parameters.

Efficient In-Context Demonstration

Considering the huge inference costs brought by ICL, researchers recently focused more on formulating efficient in-context demonstrations (Wingate, Shoeybi, and Sorensen 2022). Similar to prefix tunning (Li and Liang 2021), Gist Tokens (Mu, Li, and Goodman 2023) were proposed to replace various task instructions. AutoCompressor (Chevalier et al. 2023) first randomized segmented the texts with thousands of words into model-accepted range and then recursively generated soft prompts for each segment. Similarly, ICAE (Ge et al. 2023) employed a LoRA-adopted Llama-7b (Touvron et al. 2023) to compress the processed demonstrations to compact virtual tokens. Gao, Cao, and Li (2024a,b) propose to compress over-limit prompts into virtual tokens via a frozen LLM and a linear layer. Correspondingly, researchers also endeavored to shorten prompts by extracting informative tokens from the original ones (Li 2023; Jiang et al. 2023), namely token pruning (Kim et al. 2022) or token merging (Bolya et al. 2022; Qiao et al. 2024). LLMLingua (Jiang et al. 2023) and Selective Context (Li 2023) shared similarities but diverged on whether to eliminate tokens with high or low PPL.

Unfortunately, these outstanding studies only focus on the textual modality, which did not suffer from the modal gap in MLLMs. To our best known, AIM is the first to explore the construction of efficient multimodal demonstrations.

Methodology

We propose AIM as illustrated in Figure 4, which is a training- and inference-efficient framework that aggregates image information of multimodal demonstrations into their latent space of texts. Considering different details of popular MLLMs, we present an empirical comparison in Table 1. Specifically, The Flamingo is distinguished from others by its Gated XATTN layer inserted in the LLM blocks to fuse

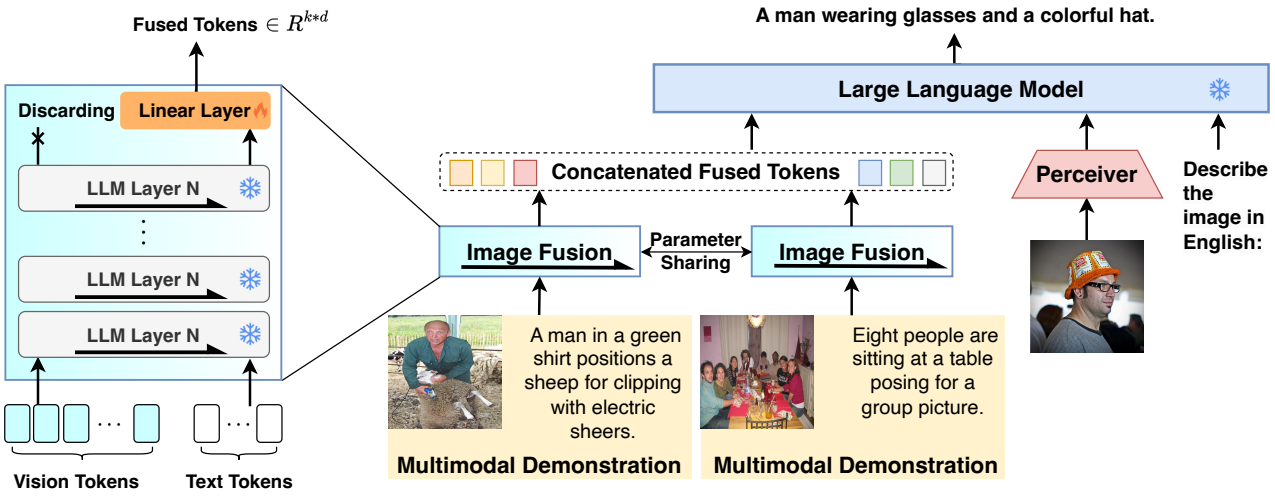

Figure 4: The architecture of AIM. Fused tokens from different demonstrations are concatenated and fed into the inner LLM, discarding original visual tokens.

image information into embeddings, **sacrificing inference efficiency to memory usage**. LLaVA-Next directly projects visual features extracted from pre-trained ViT to thousands of visual tokens, **resulting in higher memory cost increment** than other methods based on Q-Former. Additionally, LLaVA-Next can read only a single image and thus it doesn't support multimodal ICL.

AIM discards substantial visual tokens in multimodal demonstrations after aggregating demonstrated image information into its text, resembling a multimodal ICL prompt approximately containing a single query image. Therefore, AIM operates seamlessly with any MLLMs, regardless of whether they initially understand multimodal demonstrations well. AIM employs the 7B version of QWen-VL and LLaVA-Next as the built-in backbone, representing the two coarse-grained types of MLLMs divided by input form, to verify the effectiveness of fused virtual tokens.

Preliminary

A multimodal ICL prompt encompasses several demonstrations consisting of image-text pairs $(< X_1, Y_1 >, < X_2, Y_2 >, ..., < X_n, Y_n >)$ and a query denoted as $< X_{query}, ins. >$, where X_i and X_{query} representing the i-th demonstration image and the query image. Additionally, we use manually designed instructions $ins.$ to wrap the bare label for each demonstration. Thus, the demonstrated texts in $i-$th demonstration are composed of instruction $ins.$ and its reference label Y_i, e.g., **[IMG]** Describe the image in a sentence in English: **[Caption]**.

Image Information Aggregation

Taking into account our discoveries, AIM configures efficient demonstrations by disrupting the original parity between visual and linguistic modalities. AIM signals the linguistic space to gather image information via the forward propagation of the inner LLM, as illustrated in the left part of Figure 4. Practically, we split the original concatenated n-shot demonstrations into n separate image-text pairs, decorating labels with manually designed instructions. Then, we feed images to the Perceiver, the Visual Prompt Generator (VPG), to obtain the visual tokens $(X_1^v, X_2^v, ..., X_n^v)$ normally. Consequently, the first N LLM layers infers X_i^v attached with the i-th textual embeddings Y_i, obtaining last hidden states corresponding to Y_i, while dropping the others and $\oplus$ means token-level concatenation:

$$\_, H_i^Y = f_\theta(X_i^v \oplus Y_i), \tag{1}$$

where θ is the parameters of the first N LLM layers. Due to the inner transformer layers, H_i^Y is compelled to attend to the preceding visual tokens, making the latter textual tokens able to aggregate visual information. However, H_i^Y is still in the output space that the LLM can't understand directly although it has fused with image information. We insert a learnable projection layer serving as the adapter to convert H_i^Y into the LLM-acceptable fused tokens, similar to the perceiver converting visual features from visual encoder to visual tokens:

$$\hat{Y_i} = W_p \cdot H_i^Y, \tag{2}$$

where W_p is the parameters of the projection layer. Notably, aggregating image information of each image is independent of other demonstrations. Thus, AIM supports obtaining all $\hat{Y_i}$ in a batch asynchronously or repeating this process for each demonstration synchronously to trade off memory costs with time.

Dataset	Training	# Instances	Eval. Set	Metric
Flickr30k	✗	1000	Test	CIDEr
OKVQA		5046	Val	VQA acc.
Vizwiz		4319	Val	VQA acc.
Hateful Memes		815	Test	ROC AUC

Table 2: Details of involved evaluating datasets.

Response Generation in multimodal ICL

AIM applies the entire frozen inner LLM to respond to current queries, with the guidance of our proposed fused tokens. For n-shot ICL, AIM obtains $(\hat{Y_1}, \hat{Y_2}, ..., \hat{Y_n})$ independent from each other, concatenating them to configure an efficient demonstration sequence $D = \hat{Y_1} \oplus \hat{Y_2} \oplus ... \oplus \hat{Y_n}$. Finally, the demonstration sequence D together with query image X_{query} and the instruction $ins.$ are fed into the inner LLM, performing auto-regressive generation:

$$y_t = argmaxP(y|D; X_{query}; ins.; y_{<t}). \quad (3)$$

Training

The trainable parameters in AIM are merely 17M originating from the projection layer W_p. We supervised-tune the projection layer under the language modeling objective of its built-in LLM. We collect 56k instances from the web multi-image dataset MMC4. Each instance includes several images $[X_1, X_2, ..., X_k]$, and each image corresponds to a most similar text $[Y_1, Y_2, ..., Y_k]$ assigned by existing ViT-L/14 in CLIP, constructing an interleaved image-text training instance. During data preprocessing, we ensure each training instance has non-overlapping remaining texts and concatenate them, denoted as $Y^R = Y_{k+1} \oplus Y_{k+2} \oplus ....$

AIM first independently aggregates X_i to its corresponding text Y_i, obtaining $\hat{Y_i}$. Then, the language modeling loss can be formulated as:

$$loss = -\frac{1}{|Y^R|} \sum_{t=0}^{|Y|} logP(Y_t^R | \hat{Y_1}, ..., \hat{Y_k}; Y_{<t}^R). \quad (4)$$

Notably, the carefully designed training approach separates the information aggregation of different images, breaking the inner relevance among images crawled from the same web page, increasing learning difficulty, and scaling up model robustness. Specifically, this gets rid of the influence of other image-text pairs, allowing each image to focus on the given text only and better match the patterns of ICL intuitively. More importantly, assuming k image-text pairs of length l, breaking their inner relevance will reduce the memory complexity from $\mathcal{O}(k^2l^2)$ to $\mathcal{O}(kl^2)$. It also brings AIM crucial merit that each aggregated result can be cached independently, formulating a **Demonstration Bank** (**DB**) for further reusing without aggregating image information into its latent space of demonstrated texts every time.

Experiment

Setting

We briefly illustrate the involved dataset of AIM in Table 2. The test data was carefully filtered to exclude datasets that were encountered during the training of backbone MLLMs according to their technical reports, aiming to obtain more reliable ICL results.

During training, we set the maximum number of pictures to 5 per step for efficiency and filtered images to see if they were similar to all texts below 0.24 following the usage of Multimodal C4 (MMC4) in OpenFlamingo. We fix the learning rate to 3e-5 and use Adam as the optimizer, and the effective batch size is 16 (4 GPUs data parallelism and 4 steps gradient accumulation). The number of epochs is set to 10 and we get a checkpoint per 3400 training steps. Additionally, we conduct all experiments on a single DGX node with 8*Nvidia H800 GPUs. LLaVA-Next supports processing any resolution image by splitting it into sub-images, bringing several times visual tokens. We ignore this character and require LLaVA to pad each image to 336*336 resolution since AIM introduces mass pictures as demonstrations[1].

We borrow some crafted prompts from previous studies. For captioning, we format demonstrations as "[image] Describe the image in English in one sentence: [caption]"; For VQA we format demonstrations as "[image]\n[question] Answer in a word: [answer]"; For Hateful Memes, we prompt the model with "[image] is an image with [text] written on it. Is it hateful? Answer: [answer]". Notably, following the previous studies (Alayrac et al. 2022; Awadalla et al. 2023), we explicitly provide OCR text as inputs of AIM and baselines, and we don't extra prompt AIM can respond "unanswerable" in Vizwiz, reducing inappropriate induction.

Baselines

Considering our aim to enable any MLLMs to embrace efficient ICL, the underlying backbones within AIM are convinced baselines to compare their efficiency and performance, namely, QWen-VL and LLaVA-Next. We also cite the results of the Flamingo (Alayrac et al. 2022) and OpenFlamingo (Awadalla et al. 2023) from their published studies for reference.

Result

We filter the benchmarks occurring in the training of our selected backbones to simulate the practical in-context learning situation. Interestingly, when provided QWen-VL with demonstrations including textual information merely (*w/o visual* in Table 3), it even outperforms the large shots situation provided both visual and textual. Additionally, QWen-VL produces significant performance degradation in all 4 benchmarks when provided with over 8 demonstrations. This further highlights that treating visual and textual tokens equally limits MLLMs from exhibiting outstanding ICL performance, despite QWen-VL having developed multi-image understanding ability during training. When concentrating on LLaVA-Next, especially in the close-ended evaluation, perplexities concerning golden labels become **NotaNumber (NaN)** in 8- and 16-shot settings, occurring overflow while calculating PPL. In other vision language tasks, LLaVA-Next fails to generate when provided over one demonstration and occurs <1 metric in evaluation since it didn't learn to understand interleaved image-text prompts during pre-training.

The input of AIM resembles prompts containing a single query image, effectively bridging the modality gap, as image information has integrated into its linguistic space. In this

[1]Set image_aspect_ratio to pad.

Method	Flickr30k (CIDEr ↑)				OKVQA (VQA-ACC ↑)				Vizwiz (VQA-ACC ↑)				Hateful Memes (ROC-AUC ↑)			
	#-shots				*#-shots*				*#-shots*				*#-shots*			
	0	4	8	16	0	4	8	16	0	4	8	16	0	4	8	16
Flamingo†⋈	61.5	72.6	-	-	44.7	49.3	-	-	28.8	34.9	-	-	57.0	62.7	-	-
Open Flamingo†⋈	39.2	52.2	58.7	60.6	38.3	42.0	44.1	45.1	34.6	41.0	45.0	46.2	67.1	70.1	71.2	73.2
-Random†⋈	59.5	65.8	62.9	62.8	37.8	40.1	41.1	42.7	27.5	34.1	38.5	42.5	51.6	54.0	54.7	53.9
QWen-VL ⋆	73.4	77.1	75.1	63.0	46.3	52.6	53.2	53.3	27.6	30.6	30.1	28.7	56.5	56.2	57.1	59.5
QWen-VL		80.4	74.6	66.1		56.4	55.1	53.7		31.3	31.7	29.8		59.0	59.3	58.2
-w/o visual		74.7	74.3	71.7		54.7	55.3	55.1		28.7	30.5	29.8		58.2	58.2	57.1
AIM	74.3	78.1	78.8	82.3	55.3	55.8	56.1	56.0	28.4	35.1	35.6	36.1	58.3	57.8	58.9	57.1
-16		**81.2**	**83.4**	**84.1**		**57.9**	**58.2**	**57.3**		**37.6**	**38.3**	**39.4**		**59.6**	**62.9**	**64.0**
-24		73.3	75.8	78.1		55.2	54.3	53.5		34.3	34.6	34.6		56.9	57.4	59.1
LLaVA-Next ⋆	32.7	23.5	<1	<1	33.5	29.8	28.6	<1	17.8	14.5	15.9	<1	54.0	52.8	52.4	NaN
LLaVA-Next		<1	<1	<1		<1	<1	<1		11.2	<1	<1		55.6	NaN	NaN
-w/o visual		48.1	47.6	44.3		52.9	51.5	52.8		21.7	20. 5	19.9		55.9	55.6	55.0
AIM	46.2	**58.2**	**57.5**	**53.4**	49.7	**55.3**	**55.0**	**54.6**	19.8	**24.1**	**21.5**	**20.2**	55.9	**58.6**	**57.8**	**55.6**
-16		44.3	30.8	26.5		51.7	50.6	48.7		21.8	17.4	14.8		56.3	53.4	53.0
-24		46.0	39.2	38.0		54.3	**55.0**	51.4		21.6	20.0	19.0		55.3	54.7	53.3

Table 3: Main results of AIM. *w/o visual* stands for providing textual label only. -16/24 represents the number of LLM layers applied to aggregate image information. † stands for the results from previous works and ⋈ indicates extra providing 2 textual label in 0-shot. ⋆ represents further tune backbones with LoRA. The 0-shot results of AIM and its backbone are the same and merged. <1 indicates LLaVA-Next fails to respond to interleaved inputs.

case, MLLMs are only required to attend to the query image while fused tokens still guide generation, thus bringing more concise responses. Additionally, the valuable merit artfully unlocks the ICL ability of MLLM trained on the single image-text pair. It is verified by the successful deployment on LLaVA-Next that the fused tokens combined with image and text information are harmless for the inner Vicuna. For vision language tasks involved in this spectrum, QWen-based AIM outperforms backbones provided with concrete visual features that achieve +18 CIDEr gains in 16-shot in Flickr30k. In the Vizwiz dataset, over 33% answers are answerable in the statistic. AIM exhibits relatively lower performance compared with other multimodal ICL methods since we don't prompt AIM to output 'unanswerable', avoiding not solid short-cut answers.

Notably, both OpenFlamingo and AIM employ MMC4 as the multi-image training set, but AIM, even applying LLaVA-Next as the backbone, still achieves comparable or even outstanding performance via aggregating image information when provided with random demonstrations (refer to the *Random* row). Results of AIM using RICES (Retrieval-based In-Context Example Selection) applied in Flamingo (Alayrac et al. 2022) and OpenFlamingo (Awadalla et al. 2023) are accessible in our extended version (Gao et al. 2024b).

Analysis

Training Data Abalation

To avoid the model learning to generate short-cut answers because of in-domain or task-relevant training data, we train the linear layer on MMC4, which is a popular image-text-interleaved pre-training dataset used in OpenFlamingo. However, training AIM on MMC4 still potential to help MLLMs better understand image-text-interleaved inputs, achieving performance gains.

Due to AIM modifies the input form of MLLMs, we adopt LoRA to tune the built-in MLLMs, the QWen-VL and the LLaVA-Next, with comparable trainable parameters (17M) as AIM in Table 3 by setting the LoRA rank to 32 to simulate these gains(**referring to ⋆ in Table 3**). Further tuning on MMC4 improves LLaVA-Next in reading ICL prompts to some extent, but LLaVA-Next still underperforms providing pure text demonstrations (*w/o visual*). Due to QWen-VL having developed the multi-image understanding ability during pre-training, further training on MMC4 is not necessary, exhibiting even poorer performances since the web corpora is quite different from captioning, VQA, or image understanding.

LLM Layer Count for Aggregation

Considering the first layers directly interact with pre-trained embeddings, we perform ablation experiments on the first half (*16*), and the first three-quarters (*-24*) to explore the number of LLM layers required to aggregate image information in Table 3. It is interesting that QWen-VL prefers the first 16 layers, while LLaVA-Next is inclined to use full layers as the aggregator. Therefore, the label words claim (Wang et al. 2023a) that shallow layers (first half) focus on information aggregation is not completely applicable for LLaVA-Next in multimodal settings.

From a posterior view, LLaVA-based AIM obtains stable performance gains with the aggregating layers become deeper. We attribute this conflict to LLaVA being pre-trained on single images, requiring deeper LLM layers to fuse image information into corresponding label space thoroughly, thus reducing the understanding difficulty for the built-in LLM. Additionally, The Connecter in the LLaVA-Next project visual features from ViT to 576 visual tokens for a 336*336 image, while QWen-VL has 256 learnable queries. There-

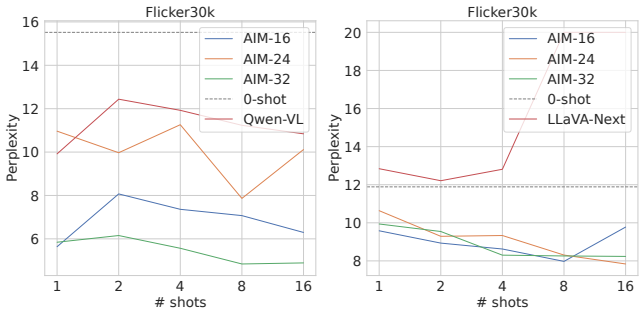

Figure 5: The perplexity variation tendency corresponds to the number of demonstrations. 0-shot server as the baseline.

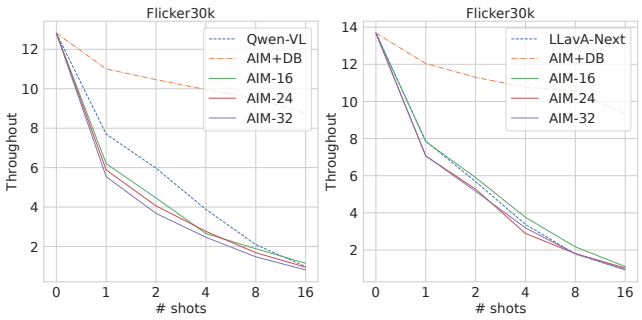

Figure 6: The throughput (iter/s) variation tendency is evaluated on a single H800, with the number of demonstrations increasing from 0 to 16.

fore, LLaVA-based AIM requires deeper LLM layers to perform visual information gathering.

Perplexity Tendency of ICL

We briefly illustrate the perplexity variation tendency concerning golen labels of Flickr30k in Figure 5 with the number of demonstrations changing from 0 to 16. Notably, the perplexity blast occurring in both QWen-VL and LLaVA-Next in the large shot setting indicates that the provided demonstration sequences significantly confused the underlying backbones, resulting in bad responses. While in the scope of AIM, the perplexity presents a decreasing tendency in general with some noise brought by randomly sampled demonstrations. Additionally, the most perplexity values are inferior to the 0-shot ones, indicating the provided demonstrations have a positive effect on helping MLLM generate current golden label responses.

Inference Throughput of AIM

AIM utilizes the inner LLM of existing MLLMs to complete image information aggregation operation. This character makes AIM not need to load the other "aggregator", alleviating the memory costs. However, image information aggregation requires inevitable but minimal time costs due to the parallel computation of forward propagation. What's more, AIM drops all the visual tokens after aggregating them into the dense label space, which compensates for aggregation time costs to some extent by reducing the input length during generation. We evaluate the throughput (iter/s) of AIM on Flick30k in Figure 6. In the few shot settings (less than 8), naive MLLMs are more efficient than AIM, but AIM has a lower inference latency increment. AIM becomes more efficient than the underlying backbone when provided with over 16 demonstrations. Additionally, with the introduction of Demonstration Bank (DB) (AIM+DB), AIM skips the aggregation and performs generation directly, which significantly boosts model efficiency.

Method	# Visual Tokens	Avg. Textual Tokens	Avg. Ratio
AIM-QWen	256	22.7	8%
AIM-LLaVA	576	26.5	4%

Table 4: Quantity statistics of visual and textual tokens in multimodal demonstrations.

Memory Cost of AIM

The normal attention mechanism is known as $O(W^2)$ space complexity concerning a sequence with W words. Therefore, the length challenge brought by in-context demonstrations stimulates memory explosion straightforwardly. AIM drops the visual tokens after image information aggregation and the remaining ratios of fused tokens $\mathcal{R}$ can be calculated according to the number of visual and textual tokens, denoted as $|V|$ and $|T|$:

$$\mathcal{R} = \frac{|T|}{|V| + |T|}. \tag{5}$$

We demonstrate the quantity statistics over four datasets in Table 4, indicating that LLaVA-based AIM merely retains about 4% origin tokens in each multimodal demonstration. Although LLaVA-Next integrates FlashAttention, dropping visual tokens still saves noticeable memory costs as illustrated in Figure 1. Notably, even if vision-language tasks have extremely long textual labels in assumption, AIM is capable of performing efficient ICL as normal with $\mathcal{R}$ close to 1 since visual tokens have been dropped and the textual tokens are required anyway.

Conclusion

In this paper, our initial exploration delves into the attention distribution within the multimodal ICL, revealing that the MLLM exhibits a greater emphasis on the linguistic modality. Built upon this discovery, we propose a light multimodal framework AIM aiming to let any MLLMs embrace efficient ICL, which aggregates the image information of demonstrations into their dense latent space of demonstrated texts. Generally, AIM transforms the multimodal ICL demonstration sequence into a form resembling a single query image accompanied by textual tokens. Thereby, AIM successfully coordinates with any MLLMs regardless of their initial support for multimodal ICL. Except for the outstanding performance of AIM compared with MMLMs specifically trained on multimodal ICL, AIM is both training and inferencing efficient due to its frozen backbone and dropping hundreds of visual tokens.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to all the authors for their valuable contributions to this research. This work was supported by the National Natural Science Foundation of China (NSFC 62106165) and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, China.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

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

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *arXiv preprint arXiv:2308.12966*.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chevalier, A.; Wettig, A.; Ajith, A.; and Chen, D. 2023. Adapting Language Models to Compress Contexts. *arXiv preprint arXiv:2305.14788*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, J.; Cao, Z.; and Li, W. 2024a. SelfCP: Compressing over-limit prompt via the frozen large language model itself. *Information Processing & Management*, 61(6): 103873.

Gao, J.; Cao, Z.; and Li, W. 2024b. Unifying demonstration selection and compression for in-context learning. *arXiv preprint arXiv:2405.17062*.

Gao, J.; Li, Y.; Cao, Z.; and Li, W. 2024a. Interleaved-Modal Chain-of-Thought. *arXiv preprint arXiv:2411.19488*.

Gao, J.; Qiao, Q.; Cao, Z.; Wang, Z.; and Li, W. 2024b. AIM: Let Any Multi-modal Large Language Models Embrace Efficient In-Context Learning. *arXiv preprint arXiv:2406.07588*.

Ge, T.; Hu, J.; Wang, X.; Chen, S.-Q.; and Wei, F. 2023. In-context Autoencoder for Context Compression in a Large Language Model. *arXiv preprint arXiv:2307.06945*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; et al. 2024. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36.

Jiang, H.; Wu, Q.; Lin, C.-Y.; Yang, Y.; and Qiu, L. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624.

Kim, S.; Shen, S.; Thorsley, D.; Gholami, A.; Kwon, W.; Hassoun, J.; and Keutzer, K. 2022. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 784–794.

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

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

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Li, Y. 2023. Unlocking Context Constraints of LLMs: Enhancing Context Efficiency of LLMs with Self-Information-Based Content Filtering. *arXiv preprint arXiv:2304.12102*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Luo, Z.; Xie, Q.; and Ananiadou, S. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.

Meta, A. 2023. Introducing LLaMA: A foundational, 65-billion-parameter large language model. *Meta AI. https://ai. facebook. com/blog/large-language-model-llama-meta-ai*.

Mu, J.; Li, X. L.; and Goodman, N. 2023. Learning to compress prompts with gist tokens. *arXiv preprint arXiv:2304.08467*.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Qiao, Q.; Xie, Y.; Gao, J.; Wu, T.; Huang, S.; Fan, J.; Cao, Z.; Wang, Z.; and Zhang, Y. 2024. DNTextSpotter: Arbitrary-shaped scene text spotting via improved denoising training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10134–10143.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212.

Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; and Sun, X. 2023a. Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning. *arXiv preprint arXiv:2305.14160*.

Wang, Z.; Xie, Q.; Ding, Z.; Feng, Y.; and Xia, R. 2023b. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Xie, P.; Xu, J.; Chen, Y.; Zhang, M.; et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Wingate, D.; Shoeybi, M.; and Sorensen, T. 2022. Prompt compression and contrastive conditioning for controllability and toxicity reduction in language models. *arXiv preprint arXiv:2210.03162*.

Xie, S. M.; Raghunathan, A.; Liang, P.; and Ma, T. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Yang, X.; Li, Y.; Zhang, X.; Chen, H.; and Cheng, W. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Yang, X.; Wu, Y.; Yang, M.; Chen, H.; and Geng, X. 2024. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36.

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

Zhao, H.; Cai, Z.; Si, S.; Ma, X.; An, K.; Chen, L.; Liu, Z.; Wang, S.; Han, W.; and Chang, B. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Zhu, W.; Hessel, J.; Awadalla, A.; Gadre, S. Y.; Dodge, J.; Fang, A.; Yu, Y.; Schmidt, L.; Wang, W. Y.; and Choi, Y. 2023. Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved With Text. *arXiv preprint arXiv:2304.06939*.