CaMML: Context-Aware Multimodal Learner for Large Models

Yixin Chen[*†]
The Chinese University of Hong Kong
yxchen@cse.cuhk.edu.hk

Shuai Zhang[*]
Amazon Web Services
shuaizs@amazon.com

Boran Han
Amazon Web Services
boranhan@amazon.com

Tong He
Amazon Web Services
htong@amazon.com

Bo Li[†]
The University of Chicago
bol@uchicago.edu

Abstract

In this work, we introduce **C**ontext-**A**ware **M**ulti**M**odal **L**earner (CaMML), for tuning large multimodal models (LMMs). CaMML, a lightweight module, is crafted to seamlessly integrate multimodal contextual samples into large models, thereby empowering the model to derive knowledge from analogous, domain-specific, up-to-date information and make grounded inferences. Importantly, CaMML is highly scalable and can efficiently handle lengthy multimodal context examples owing to its hierarchical design. Based on CaMML, we have developed two multimodal models, CaMML-7B and CaMML-13B, that have shown exceptional performance across an array of benchmark datasets for multimodal tasks. Remarkably, CaMML-13B achieves the state-of-the-art performance on over ten widely recognized multimodal benchmark datasets, surpassing LLaVA-1.5 (13B) with a noticeable margin, without integration of any external resources. Moreover, we have conducted extensive ablative studies to inspect the inner workings of CaMML and performed qualitative analyses to showcase its effectiveness in handling real-world challenging cases. Code and models are available at: `https://github.com/amazon-science/camml`.

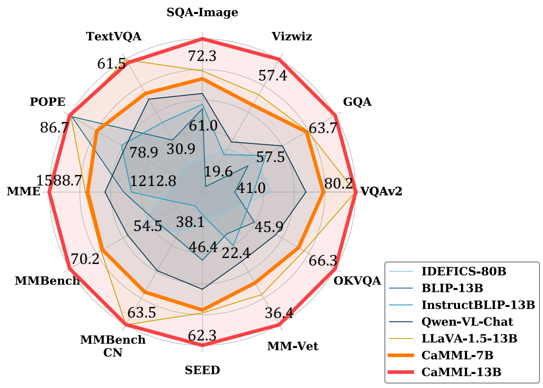

Figure 1: CaMML achieves the state-of-the-art performance on a number of multimodal benchmarks, outperforming LLaVA-1.5 and many other large multimodal models.

1 Introduction

Recently, large multimodal models (LMMs) (Liu et al., 2023b; Zhang et al., 2023a; Wang et al., 2022; Liu et al., 2023a; Zhu et al., 2023; Tsimpoukelli et al., 2021; Alayrac et al., 2022) have demonstrated remarkable performance in a variety of tasks, including but not limited to visual question answering, image captioning, visual grounding, visual-language reasoning, optical character recognition, and visual entailment. Notably, in certain benchmark assessments, these multimodal foundation models have even exceeded human-level performance (Zhang et al., 2023b; Liu et al., 2023b).

Despite the impressive performance, their ability to make inferences is constrained by the knowledge encoded in the model parameters. The inflexible design of these models makes it challenging for them to generalize from contextual examples. For instance, LLaVA-1.5 falls short in processing multiple images and attributes it to the lack of corresponding instruction-tuning training data (Liu et al., 2023a). Nevertheless, learning and making inferences through contextual examples are fundamental elements of our cognitive processes. Human beings frequently tackle intricate problems by relying on past experiences and identifying analogous situations. Taking inspiration from the cognitive process, we hypothesize that empowering large multimodal models with the capability to perceive and derive insights from analogous contextual examples can significantly streamline the inference process and lead to more precise predictions. Nonetheless, the means to replicate the human cognitive processes in LMMs remain unclear.

As such, our goal is to empower multimodal

[*] Co-first Author
[†] Work done at Amazon

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4056–4071
August 11-16, 2024 ©2024 Association for Computational Linguistics

foundational models to harness context-aware learning, thereby enhancing their ability to comprehend and adapt to previously unseen examples. Identifying relevant context examples is relatively straightforward; this process can be facilitated using multimodal embedding models like ImageBind (Girdhar et al., 2023) or CLIP (Radford et al., 2021). This approach simulates the act of recalling similar situations from past experiences. Yet, effectively and efficiently integrating the identified similar context samples into large models poses challenges, particularly given the potential variability in the number of context samples and interleaved modalities, resulting in lengthy and heterogeneous context input.

To this end, we propose a context-aware multimodal learner, dubbed as CaMML, for LMMs. CaMML acts as a crucial intermediary between the contextual examples and a large language model (LLM). Our approach is structured hierarchically, where the initial level establishes connections between the text and image modalities for each example through cross-attention mechanisms. This integration of text and image information enables a deeper understanding of the interleaved context. Following this, another module takes the outputs of the first level and performs cross-attention between the contextual information and a predefined set of learnable, fixed-length tokens. The resulting output from this level is then used as fixed-length input for the LLM, allowing the model to leverage the refined and context-aware information to perform complex multimodal understanding tasks. To summarize, we make the following contributions:

- We propose CaMML, a context-aware multimodal learning approach for finetuning multimodal models. CaMML is lightweight and can be applied to process extremely long multimodal context samples;
- With CaMML, we have developed two multimodal models, CaMML-7B and CaMML-13B. These models have achieved state-of-the-art performance across a diverse range of benchmarks encompassing various multimodal tasks, all without the need for external data integration;
- We conduct comprehensive model analyses and case studies to examine the internal mechanisms of CaMML and showcase how the proposed model can effectively handle real-world challenging cases.

2 Related Work

Large Multimodal Models The success of LLMs has sparked a burgeoning interest in scaling up multimodal models. One prevalent strategy involves the incorporation of vision encoders, such as ViT (Dosovitskiy et al., 2021) and CLIP, into existing LLMs (*e.g.*, LLaMA (Touvron et al., 2023), Vicuna (Zheng et al., 2023)). For example, LLaMA-adapter (Zhang et al., 2023a) inserts learnable adaption prompts into LLaMA and make it to follow multimodal instructions for multimodal reasoning. BLIP2 (Li et al., 2023b) bridges different modalities via Q-Former and employs a two-stage training method to bootstrap both representation learning and LLM (*e.g.*, OPT (Zhang et al., 2022), Flan-T5 (Chung et al., 2022)) generative learning capabilties. LLaVA (Liu et al., 2023b,a) projects encoded image features to text token space using linear layers and shows state-of-the-art performance on a variety of multimodal tasks. Another noteworthy series of strategy is to unify multimodal input data and pretrain the model from scratch, exemplified by OFA (Wang et al., 2022), Perceiver (Jaegle et al., 2021, 2022), Uni-Perceiver (Zhu et al., 2022), Unival (Shukor et al., 2023), and Unified-IO (Lu et al., 2023a). These models map images, text, and other modalities into the same IO space and use a causal language model objective for training.

Multimodal Few-shot Learning Extensive research has been conducted on learning from a limited number of multimodal examples. Among them, Flamingo (Alayrac et al., 2022) introduces gated xatten-dense layers to establish cross-modal interactions between visual input and text input for few-shot learning. Frozen (Tsimpoukelli et al., 2021) trains a vision encoder to produce a sequence of image embeddings and input it to frozen language models for multimodal tasks. An alternative approach for tackling multimodal few-shot learning involves leveraging retrieval augmented generations (RAG) (Lewis et al., 2020; Borgeaud et al., 2022; Ram et al., 2023; Khandelwal et al., 2020). This technique enables the models to access external knowledge repositories, databases, or structured data sources, allowing them to access the up-to-date as well as domain-specific expertise when crafting responses. Recently, there has been a surging interest for harnessing RAG within the realm of multimodal models (Chen et al., 2022; Yang et al., 2023; Yasunaga et al., 2023; Liu et al., 2023c). Amalgamating RAG with multimodal models helps provide contextually-rich responses and opens up exciting possibilities across domains such as image captioning (Yang et al., 2023), image generation (Yasunaga et al., 2023), etc. Among them, Chen et al. (2022) propose MuRAG, a model of size 527M, for multimodal-QA. MuRAG concatenates the visual embeddings (extracted by ViT) and text word embeddings of retrieved multimodal image-text pairs and requires the language model to handle a very lengthy input sequences. Re-ViLM (Yang et al., 2023) is tailored for image captioning and it follows the Flamingo architecture and only uses the retrieved text for augmentation. RA-CM3 (Yasunaga et al., 2023) leverages retrieval to augment the CM3 (Aghajanyan et al., 2022) pipeline which can do both text-to-image and image-to-text generations.

In contrast to previous approaches, our primary objective is to develop a lightweight module capable of processing multimodal context information efficiently and effectively, with superior generalization capabilities. This novel approach enables LLM to proficiently reason from multimodal in-context examples, leading to more accurate and precise inferences. As discussed in (Liu et al., 2023a), LLaVA-1.5 faces limitations when confronted with scenarios involving multiple images and lengthy contexts. CaMML can effectively address these challenges, a fact further corroborated by its superior performances across multiple benchmarks.

3 Context-Aware Multimodal Learner

3.1 Architecture of CaMML

The structure of CaMML is depicted in Figure 2. We provide a detailed explanation of each component below.

3.1.1 Datastore and Context Retriever

The datastore is created from either the training set or external resources. We adopt an embedding encoder, ImageBind (Girdhar et al., 2023), to extract dense vector representations for every multimodal sample contained in the datastore. Following this, we build an index using Faiss (Johnson et al., 2019a), a highly efficient similarity search library, to enable rapid and efficient search operations. During both the training and inference phases, our approach involves identifying the top N most closely related multimodal samples, denoted as $C_1, C_2, \ldots, C_N$, from the datastore using Faiss, based on a given multimodal query q[1]. It is worth noting that the retriever remains frozen throughout all stages of the process.

3.1.2 Multimodal CaMML Perceiver

To seamlessly integrate interleaved multimodal samples, we introduce a novel module denoted as CaMML Perceiver, which revolves around two key design principles: (1) The module should accommodate a dynamic number of context samples without imposing a significant computational burden. It is worth noting that directly concatenating the image tokens and text tokens will result in $L = \sum_{i=1}^{N}(T_i^{IMG} + T_i^{TXT})$ tokens, where T_i^{IMG} is the number of image tokens and T_i^{TXT} is the number of text tokens. The formulation results in a linear increase of L with respect to N, posing scalability concerns; (2) It is crucial that the text and image in each individual sample remain tightly coupled.

To achieve this, we embrace a hierarchical design, facilitating the transformation of the N context samples into M tokens, where $M < L$. Formally, the multimodal CaMML Perceiver comprises three key modules: a *Vision Perceiver* (VP) denoted as $f_{\theta^{VP}}$, a *Language Perceiver* (LP) denoted as $f_{\theta^{LP}}$, and a *Context Perceiver* (CP) denoted as $f_{\theta^{CP}}$. Each of these modules follows a similar architectural pattern with Perceiver Layers as the core component, as depicted in Figure 2.

Vision Perceiver The Vision Perceiver takes the image feature extracted with ViT, $\mathbf{V}^{(C_i)} \in R^{T_i^{IMG} \times d}$, as input, and it undergoes a sequence of transformations. It starts with a self-attention layer to capture inherent visual relationships. Subsequently, cross-attention layers are employed to interact with the text token embeddings $\mathbf{U}^{(C_i)}$, allowing the model to enrich its understanding by integrating information from the text domain. Finally, the processed data passes through a feed-forward layer, resulting in an output denoted as $f_{\theta^{VP}}(\mathbf{V}^{(C_i)}) \in R^{T_i^{IMG} \times d}, i = 1, ..., N$. Subsequently, we concatenate the outputs generated by the Vision Perceiver for each of the N context samples, yielding a feature matrix of shape $R^{(\sum_{i=1}^{N} T_i^{IMG}) \times d}$ $\mathrm{E}^{IMG} = \{f_{\theta^{VP}}(\mathbf{V}^{(C_1)}, \mathbf{U}^{(C_1)}), ..., f_{\theta^{VP}}(\mathbf{V}^{(C_N)}, \mathbf{U}^{(C_N)})\}$, where $\{\cdot\}$ denotes the concatenation operation.

[1] Specifically, we compute the similarity between query text and context image, and top-k images along with their corresponding texts are recalled from the datastore.

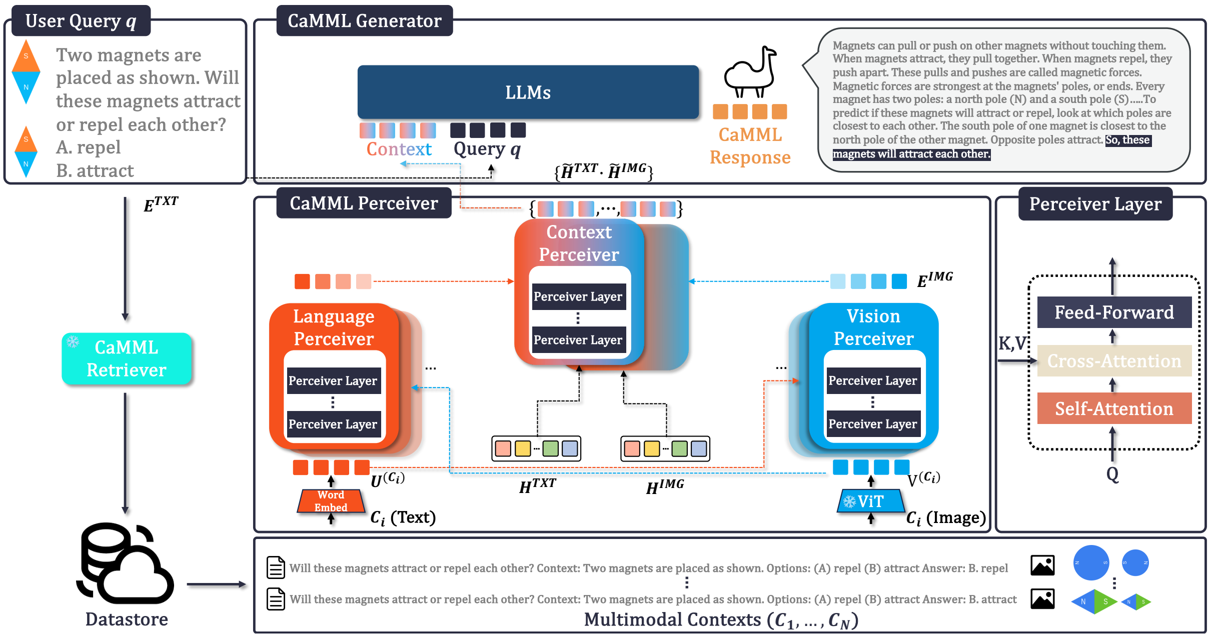

Figure 2: CaMML framework, which consists of retriever, perceiver and generator. Once receiving user query q, CaMML retriever identifies relevant multimodal contexts C from datastore, then CaMML Perceiver seamlessly integrates various modalities, effectively encodeing long-context information and injecting it into the CaMML generator. This allows for the prediction of responses that are conditioned on both the context and the query q.

Language Perceiver The Language Perceiver, on the other hand, takes as input the text token embeddings $\mathbf{U}^{(C_i)} \in R^{T_i^{TXT} \times d}$, which encapsulate the linguistic content of the sample. The processing pipeline mirrors that of the Vision Perceiver, starting with a self-attention mechanism to capture textual relationships, followed by interactions with the visual embeddings $\mathbf{V}^{(C_i)}$ via a cross-attention layer to create a holistic understanding of the multimodal context. The final step involves a feed-forward layer, resulting in an output denoted as $f_{\theta LP}(\mathbf{U}^{C_i}) \in R^{T_i^{TXT} \times d}$. Similarly, we concatenate the outputs of all Language Perceivers to obtain a feature matrix of shape $R^{(\sum_{i=1}^{N} T_i^{TXT}) \times d}$: $\mathrm{E}^{TXT} = \{f_{\theta LP}(\mathbf{U}^{(C_1)}, \mathbf{V}^{(C_1)}), ..., f_{\theta LP}(\mathbf{U}^{(C_N)}, \mathbf{V}^{(C_N)})\}$.

Context Perceiver The Context Perceiver takes the lengthy outputs from either the Vision Perceiver or the Language Perceiver as input and produces a condensed set of contextual representations. This is achieved by using a fixed number of learnable embeddings of size $R^{\frac{M}{2} \times d}$ as the input of the Context Perceiver, where M is typically significantly smaller than L. These embeddings undergo processing through a self-attention layer, followed by a cross-attention layer with the feature matrix from either the Vision Perceiver or the Language Perceiver, and a feedforward layer. Specifically, for the Vision Perceiver, the learnable embeddings are denoted as $\mathbf{H}^{IMG} \in R^{\frac{M}{2} \times d}$, and the output after applying the Context Perceiver is denoted as: $\tilde{\mathbf{H}}^{IMG} = f_{\theta CP}(\mathbf{H}^{IMG}, \mathbf{E}^{IMG}) \in R^{\frac{M}{2} \times d}$. Similarly, for the Language Perceiver, the learnable embeddings are denoted as $\mathbf{H}^{TXT} \in R^{\frac{M}{2} \times d}$, and the output after applying the Context Perceiver is denoted as: $\tilde{\mathbf{H}}^{TXT} = f_{\theta CP}(\mathbf{H}^{TXT}, \mathbf{E}^{TXT}) \in R^{\frac{M}{2} \times d}$.

Following this, $\tilde{\mathbf{H}}^{IMG}$ and $\tilde{\mathbf{H}}^{TXT}$ are concatenated, resulting in a contextual feature matrix $\tilde{\mathbf{H}} \in R^{M \times d}$. This matrix is prepended to the beginning of the sequence, along with the text token embeddings of q and the image embeddings of q (transformed by the LLaVA multimodal projector). This integrated data serves as the input for the Large Language Model (LLM) to facilitate further processing.

3.2 Model Training

In the training process, we freeze the retriever and vision encoder, and train CaMML by minimizing the following causal language modeling loss. $\ell = -\sum_{i=1}^{|y|} \log p_\theta(y_i|\hat{y}_{1:i-1}, q, C_1, ..., C_N)$, where $\theta \leftarrow (\theta^{LLM}, \theta^{VP}, \theta^{LP}, \theta^{CP}, \mathbf{H})$ is the model trainable parameter (θ^{LLM} is the parameter of LLM); y_i is the ground-truth target, and $\hat{y}_{1:i-1}$ is the $i-1$ preceding tokens of output y_i; Specifically, we use Vicuna-7B and Vicuna-13B as our backbone language model, resulting in two models

Method	AVG.	IMG	TXT
Human Average	88.40	87.50	89.60
UnifiedQA (Lu et al., 2022a)	74.11	66.53	66.42
GPT-3 CoT (Lu et al., 2022a)	75.17	67.43	74.68
GPT-4 CoT (Lu et al., 2023b)	83.99	71.49	82.65
LLaMA-Adapter (Zhang et al., 2023a)	85.19	80.32	83.72
$\text{MMCoT}_{\text{Base}}$ (Zhang et al., 2023b)	84.91	82.90	87.88
$\text{MMCoT}_{\text{Large}}$ (Zhang et al., 2023b)	91.68	88.80	95.26
LLaVA-7B (Liu et al., 2023b)	89.28	87.32	90.96
LLaVA-13B (Liu et al., 2023b)	90.90	88.00	89.49
CaMML-7B	**91.32**	89.24	93.21
CaMML-13B	**92.03**	89.94	93.84

Table 1: Comparison with state-of-the-art methods on ScienceQA benchmark: CaMML finetuned on `train` split and evaluated on `test` split. "AVG." represents the average accuracy of all ScienceQA questions. "IMG" refers to the questions that include image contexts, while "TXT" refers to the questions without any images.

CaMML-7B and CaMML-13B.

4 Experiment

Table 1, Table 2, and Table 3 provide a comprehensive summary of the performance achieved by CaMML-7B and CaMML-13B across various multimodal benchmarks. In comparison to previous state-of-the-art approaches, CaMML exhibits noticeable improvement and establishes a new state-of-the-art. We elaborate on the experimental settings in the following sections and offer a comprehensive description (*e.g.*, hyperparamter configuration) of the settings in the supplementary material.

4.1 Multimodal Reasoning on ScienceQA

ScienceQA, as introduced in (Lu et al., 2022a), serves as a multimodal benchmark designed for the task of science question answering. It comprises 21,000 multiple-choice questions, encompassing various domains including biology, physics, chemistry, Earth science, and more. We use the ScienceQA train split to build the datastore search index and for the model training. In specific, Vicuna-v1.3 is used as the LLM backbone, and CLIP-ViT-L-14 (Radford et al., 2021) is adopted as the visual feature encoder. By default, we retrieve three samples from the datastore as the contexts. We conduct evaluation of the ScienceQA on test split and adopt the same datastore (*i.e*, train split) search index for contextual samples retrieval.

We compare CaMML with various baselines including UnifiedQA (Lu et al., 2022b), GPT-4 CoT (Lu et al., 2023b), LLaMA-Adapter (Zhang et al., 2023a), $\text{MMCoT}_{\text{Base}}$ (Zhang et al., 2023b), $\text{MMCoT}_{\text{Large}}$ (Zhang et al., 2023b), LLaVA (Liu et al., 2023b) 7B and 13B. As shown in Table 1, CaMML surpass baselines by a noticeable margin. It is worth noting that our model attains the best performance on average, surpassing the previous state-of-the-art model $\text{MMCoT}_{\text{Large}}$[2]. We also notice that, on the IMG questions, CaMML-13B outperforms LLaVA-13B and $\text{MMCoT}_{\text{Large}}$ by a wide margin, which underscores CaMML's strong capability in handling challenging questions that incorporate scientific images. Furthermore, we conduct detailed ablative studies with the ScienceQA dataset using CaMML. See Section 5.1 for more details.

4.2 Multimodal Instruction Tuning

CaMML can be enhanced by tuning to follow multimodal instructions to handle a more diverse set of multimodal tasks. To substantiate this, we perform instruction tuning following (Liu et al., 2023a) on LLaVA-665K, an instruction-tuning data mixture of LLaVA-1.5. Following (Liu et al., 2023a), we use Vicuna-v1.5 as the LLM backbone and CLIP-ViT-L-14-336px as the visual encoder. The datastore is built using the exact training set, LLaVA-665K, without external injected information for a fair comparison. LLaVA-665K comprises 665K multimodal samples, and is made up of LLaVA-Instruct-158K, ShareGPT-40K (sha, 2023), VQAv2 (Antol et al., 2015), GQA (Hudson and Manning, 2019), OKVQA (Marino et al., 2019), OCRVQA (Mishra et al., 2019), A-OKVQA (Schwenk et al., 2022), TextCaps (Sidorov et al., 2020), RefCOCO (Yu et al., 2016) and VG (Krishna et al., 2017). We adopt a training strategy in which each individual sample might have a distinct N, spanning a range from 1 to 3 and denote it as "mixed-shots training". During inference, we retrieve three relevant supporting multimodal samples from the datastore index. To evaluate the model effectiveness, we traverse 11 comprehensive benchmarks, including VQAv2, GQA, TextVQA (Singh et al., 2019), MME (Fu et al., 2023), POPE (Li et al., 2023c), MM-Vet (Yu et al., 2023), MMBench (Liu et al., 2023d), MMBench-CN (Liu et al., 2023d), SEED-Bench (Li et al., 2023a), and Vizwiz (Gurari et al., 2018). Results on ScienceQA image (denoted as SQA^{I}) is also reported following (Liu et al., 2023a).

We compare CaMML with large multimodal baselines such as BLIP-2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023), Shikra (Chen et al., 2023),

[2]CaMML-13B attains the highest AVG scores on the ScienceQA leaderboard when GPT4-review is not used

Method	LLM	Data	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED	MM-Vet
IDEFICS-9B (Laurençon et al., 2023)	LLaMA-7B	353M+1M	50.9	38.4	35.5	-	25.9	-	-	48.2	25.2	-	-
InstructBLIP (Dai et al., 2023)	Vicuna-7B	129M+1.2M	-	49.2	34.5	60.5	50.1	-	-	36.0	23.7	53.4	26.2
Qwen-VL (Bai et al., 2023)	Qwen-7B	1.4B+50M	78.8	59.3	35.2	67.1	**63.8**	-	-	38.2	7.4	56.3	-
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	1.4B+50M	78.2	57.5	38.9	68.2	61.5	-	1487.5	60.6	56.7	58.2	-
LLaVA-1.5 (Liu et al., 2023a)	Vicuna-7B	558K+665K	78.5	62.0	50.0	66.8	58.2	85.9	**1510.7**	64.3	58.3	58.6	30.5
CaMML-7B	Vicuna-7B	558K†+665K	**79.4**	**62.7**	**51.2**	**67.9**	58.0	**86.4**	1506.9	**66.9**	**60.6**	**60.4**	**32.2**
IDEFICS-80B (Laurençon et al., 2023)	LLaMA-65B	353M+1M	60.0	45.2	36.0	-	30.9	-	-	54.5	38.1	-	-
BLIP-2 (Li et al., 2023b)	Vicuna-13B	129M	41.0	41.0	19.6	61.0	42.5	85.3	1293.8	-	-	46.4	22.4
InstructBLIP (Dai et al., 2023)	Vicuna-13B	129M+1.2M	-	49.5	33.4	63.1	50.7	78.9	1212.8	-	-	-	25.6
Shikra (Chen et al., 2023)	Vicuna-13B	600K+5.5M	77.4	-	-	-	-	-	-	58.8	-	-	-
LLaVA-1.5 (Liu et al., 2023a)	Vicuna-13B	558K+665K	80.0	63.3	53.6	71.6	61.3	85.9	1531.3	67.7	**63.6**	61.6	35.4
CaMML-13B	Vicuna-13B	558K†+665K	**80.2**	**63.7**	**57.4**	**72.3**	59.9	**86.7**	**1588.7**	**70.2**	**63.6**	**62.3**	**36.4**

Table 2: Comparison with state-of-the-art large multimodal models on 11 benchmarks (VQA^{v2} (Antol et al., 2015), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018), SQA^{I} (Lu et al., 2022a): ScienceQA-IMG, VQA^{T} (Singh et al., 2019): TextVQA, POPE (Li et al., 2023c), MME (Fu et al., 2023), MMB (Liu et al., 2023d): MMBench, MMB^{CN} (Liu et al., 2023d): MMbench-Chinese, SEED (Li et al., 2023a): SEED-Bench, MM-Vet (Yu et al., 2023)). CaMML achieves the best performance on 10/11 tasks. † denotes BLIP558K-pretrained projector is initialized for instruction tuning.

Method (shots)	COCO Caption CIDEr	Flickr30k CIDEr	OKVQA Acc	VQAv2 Acc	Vizwiz Acc
Retrieval-augmented models:					
RA-CM3 (Yasunaga et al., 2023) (2)	89.1	-	-	-	-
ReViLM (Yang et al., 2023) (0)	60.8	52.1	-	-	-
ReViLM (Yang et al., 2023) (2)	77.2	-	-	-	-
ReViLM (Yang et al., 2023) (4)	90.5	-	-	-	-
ReViLM (Yang et al., 2023) (8)	90.2	-	-	-	-
Zero/Few-shot models:					
Uni-Perceiver (0) (Zhu et al., 2022)	109.8	41.2	-	-	-
Flamingo-9B (4)	93.1	72.6	49.3	56.3	34.9
Flamingo-9B (32)	106.3	72.8	51.0	60.4	44.0
Flamingo-80B (4)	103.2	75.1	57.4	63.1	39.6
Flamingo-80B (32)	113.8	75.4	57.8	67.6	49.8
KOSMOS-1 (4) (Huang et al., 2023)	101.7	75.3	-	51.8	35.3
MMICL (4) (Zhao et al., 2023)	-	72.0	-	70.6	50.3
CaMML-7B (3)	111.4	82.7	64.7	79.4	51.2
CaMML-13B (3)	**116.8**	**84.5**	**66.3**	**80.2**	**57.4**

Table 3: Comparison with state-of-the-art zero/few-shot models including retrieval-augmented counterparts on Captioning and VQA tasks: CaMML are the ones that trained in instruction tuning (Sec 4.2) and are evaluated with 3-shots. CaMML-13B achieves the best performance on 5/5 tasks, even outperforming Flamingo-80B model (Alayrac et al., 2022) under 32 shots.

Method	AVG.	IMG	TXT
CaMML-7B Baseline	**91.3**	**89.2**	**93.2**
- *w/o Perceiver*	89.7	85.8	93.2
- *w/o Vision Perceiver*	89.8	87.4	92.1
- *w/o Language Perceiver*	90.0	87.7	92.0
- *w/o Shared weights*	91.3	89.1	93.2

Table 4: Ablation Experiments on CaMML perceiver components and shared-weights Context Perceiver: CaMML-7B models are evaluated on ScienceQA-`test`.

IDEFICS-9B (Laurençon et al., 2023), IDEFICS-80B (Laurençon et al., 2023), Qwen-VL (Bai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), and LLaVA-1.5 (Liu et al., 2023a). As shown in Table 2, CaMML-13B demonstrates the best performance across all 11 benchmarks, surpassing LLaVA-1.5, which uses the same training data. Notably, when employing Vicuna-13B as its backbone, CaMML-13B outperforms BLIP-2, InstructBLIP-13B, Shikra, and LLaVA-1.5-13B. Additionally, CaMML-7B delivers impressive results, outperforming all other baseline models except for LLaVA-1.5-13B on ten of the benchmark datasets. On POPE, CaMML-7B achieves performance on par with that of CaMML-13B. These encouraging observations showcase CaMML's efficacy in harnessing multimodal contextual information and following multimodal instructions. In contrast to Liu et al. (2023a)'s assertion that LLaVA-1.5 cannot handle multiple images due to the absence of instruction-tuning data and context length limitations, our design demonstrates feasibility of training a model to effectively process multiple images merely with the same training set and efficiently manage long context length.

4.3 Multimodal Few-shot Learning

CaMML essentially operates as a few-shot learning method. Few-shot models leverage the synergy of multiple intertwined images and texts to guarantee precise inference. Among these, retrieval-augmented models are widely recognized for their ability to retrieve highly similar samples, thereby improving few-shot prediction. In Table 3, we conduct a comparison between CaMML, retrieval-augmented models (RA-CM3 (Yasunaga et al., 2023) and ReViLM (Yang et al., 2023)), and other popular few-shot LLMs such as Flamingo (Alayrac et al., 2022), KOSMOS-1 (Huang et al., 2023), MMICL (Zhao et al., 2023) on tasks including visual captioning (COCO caption, Flickr30k), and visual question answering (OKVQA, VQAv2, Vizwiz). For CaMML, we employ the same model described in Section 4.2, with LLaVA-665K serv-

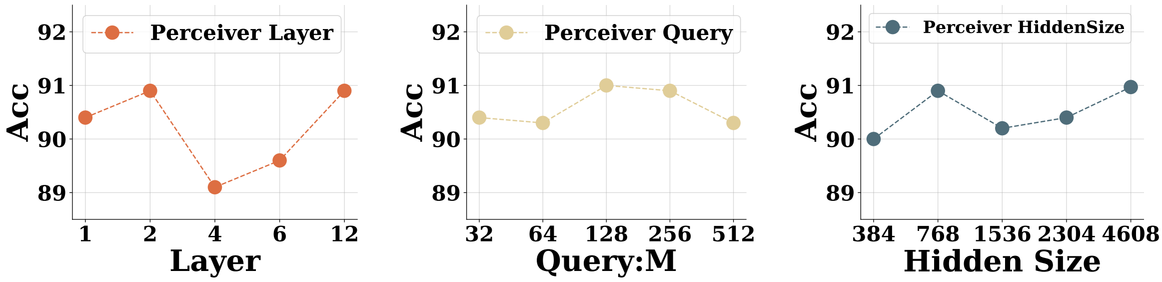

Figure 3: Ablation Experiments on CaMML perceiver hyper-parameters: layers, query number M and hidden sizes. CaMML-7B with different settings are evaluated on ScienceQA `test`.

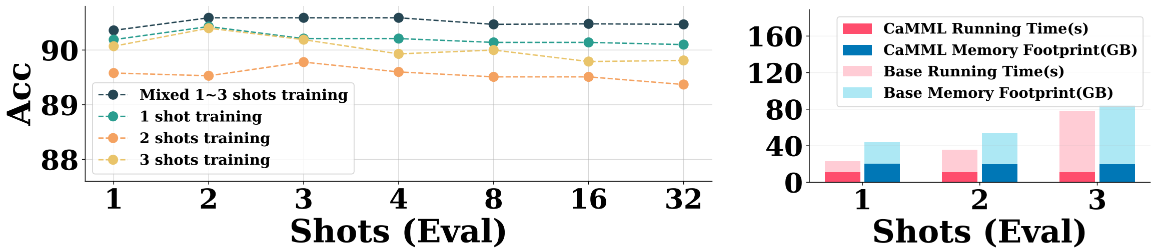

Figure 4: Ablation Experiments on CaMML context number N. Left: different CaMML models trained on N shots are evaluated under 1$\sim$32 shots. Right: comparison between CaMML and CaMML without perceiver in terms of inference running time and memory footprint, the statistic is averaged on 100 samples from CaMML-7B, which are tested on NVIDIA A100-80G GPU using ScienceQA dataset.

ing as the datastore.

Several observations merit attention: (1) It is worth mentioning that while previous visual language models augmented with retrieval capabilities have been predominantly optimized for visual captioning tasks, CaMML exhibits broad applicability across a wide range of tasks, including visual question answering and visual captioning. (2) Notably, CaMML-13B outshines the significantly larger Flamingo-80B, highlighting CaMML's superior efficiency and architectural effectiveness despite its smaller size. (3) Remarkably, CaMML, even with a mere three-shot setup, surpasses models that necessitate 32 shots for equivalent tasks, underscoring its exceptional data efficiency and learning prowess.

5 Model Analyses

5.1 Quantitative Analyses

Here, we conducted an analysis of CaMML, using the CaMML-7B model for expeditious experiments, on the ScienceQA dataset. Our goal was to discern the significance of each module within the CaMML architecture, as well as the influence of critical hyperparameters such as the number of layers, the selection of M, and the quantity of retrieved context samples. The default settings for CaMML-7B baseline are as follows: the number of layers is set to 2 for all perceivers, $M = 128$, $N = 3$, and hidden size is set to 768. We also conduct experiments in Sec 5.1.3 about CaMML computation (inference speed & memory footprint) compared with baseline approach, where all tokens are directly fed into LLMs.

5.1.1 Contribution of Each Components

We conducted ablation studies on the Perceiver, Vision Perceiver, Language Perceiver and Context Perceiver (shared weights or not) within CaMML to evaluate the individual contributions of these components. The outcomes are documented in Table 4. Our results highlight the critical role played by each component. The removal of Perceivers, in particular, leads to a marked deterioration in performance, indicating its significant influence on the model's overall effectiveness. Also, we have observed nearly identical results for both alternatives of the Context Perceiver, whether the weights are shared or not. Therefore, we prefer selecting the shared-weights option to save on computation.

5.1.2 Impact of Hyperparameters

In Figure 3, we report the performance by varying the number of layers, hidden sizes, and query number M. Our observations reveal several key insights: (1) Increasing the number of layers in CaMML can have a detrimental effect on model performance. While we observe similar performance with both 2 layers and 12 layers, we opt for 2 layers due to its smaller model size; (2) Increasing the value of M does not consistently lead to performance improvements; (3) Larger hidden sizes yield more favorable results in our analysis.

Furthermore, we investigated the impact of the number of context samples, N, used in the training stage as well as the inference stage. We conduct experiments where we varied the value of N during training and inference, and the corresponding performance is reported in Figure 4. We observe that: (1) It is easy for CaMML to accommodate a large number of shots; (2) Increasing the value of N during inference does not consistently result in improved performance. Similar trends have been reported in previous works such as (Alayrac et al., 2022) and (Yang et al., 2023). One plausible explanation for this phenomenon could be that a longer context might introduce complexity and potentially convolute the inference process; (3) The "mixed-shots" training strategy (Sec 4.2) has shown the potential to yield constant superior performance when compared to using a fixed value for N.

5.1.3 Inference Cost

As outlined in the model design, our context model adeptly manages large samples by condensing a significant number of raw tokens into a more streamlined representation. This efficient process leads to a faster forward pass in subsequent LLMs. We compared CaMML with a baseline approach, where all context tokens are directly input into LLMs without CaMML perceivers, and we present the inference speed and memory footprint in Figure 4 (right). We have noticed that CaMML only incurs negligible additional cost when increasing the number of context sample; As the number of context samples grows, the efficiency and memory benefits of CaMML become more significant compared with the baseline method.

5.2 Qualitative Analysis

5.2.1 The Importance of Context-Awareness

CaMML effectively processes a wide range of contextual inputs. In Figure 5, we compare the response from CaMML with context sample support and LLaVA-1.5, emphasizing the importance of context-awareness. The comparison illustrates that the inclusion of relevant context examples is crucial for ensuring the accuracy of answers. CaMML can not only infer using highly relevant context samples (The Great Wall example), but also draw insights from analogous domain-specific samples (The Metamorphosis example) and conduct multimodal analogy-based learning (Antonietti, 2012).

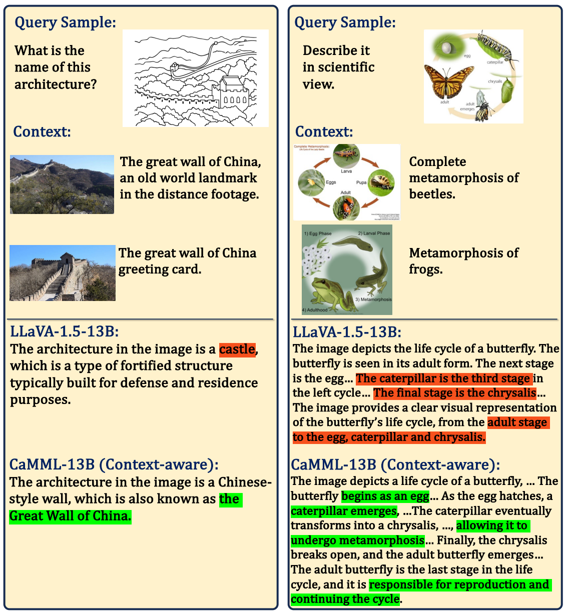

Figure 5: Visualization of context-aware CaMML *vs* no-context LLaVA-1.5. Left: sketch drawing of the Great Wall. Right: depiction of metamorphosis of a butterfly.

5.2.2 Handling Image Sequences

We also demonstrate that CaMML possesses the ability to handle image sequences, even in the absence of explicit training for this specific task. In Figure 6, instead of retrieval, we consider the last image of the image sequence as the q and the preceding images as context samples. A comparison with LLaVA-1.5 and GPT-4V in Figure 6 on the QA task on image sequences reveals that while LLaVA-1.5 and GPT-4V struggle to fully comprehend the image sequences and learn from the priors, resulting in incorrect answers, CaMML can effectively understand the image sequences and give accurate responses. More importantly, CaMML can provide accurate inferences given access to the most up-to-date information regardless of LLM's internal knowledge, yet both LLaVA-1.5 and GPT-4V lack the capability to incorporate real-time updates of world knowledge.

5.2.3 Tackling Multimodal Hallucination

Previous research indicates that RAG is effective in reducing hallucinations in content generation (Shuster et al., 2021). Here, we aim to investigate if CaMML can handle hallucinations in a multimodal setting. Figure 7 illustrates a multi-

Figure 6: Visualization of CaMML *vs* GPT-4V & LLaVA-1.5. CaMML demonstrates a strong understanding of contexual sequences. We directly input contexual sequences (Left: consecutive Homer Simpson video frames and Right: Lionel Messi winning Ballon d'or in October 2023.) to CaMML w/o additional contexts.

Figure 7: Visualization of CaMML *vs* GPT-4V & LLaVA-1.5. CaMML effectively mitigates hallucinations and accurately identifies objects. On the left, there is an illusion of muffin and chihuahua. On the right, there is an illusion of fried chicken and labradoodle.

modal QA task involving the recognition of similar objects: Muffins and Chihuahuas; Labradoodles and Fried Chicken. We treat the grid of images as a single image. It is evident that LLaVA-1.5 and GPT-4V face challenges in distinguishing between these objects, leading to hallucinated answers. In contrast, CaMML accurately answers the questions with the groundness of relevant context samples.

6 Conclusion

We present a new methodology called CaMML, a context-aware multimodal learner designed for the fine-tuning of large multimodal models. With CaMML, we build two multimodal models, CaMML-7B and CaMML-13B. CaMML empowers them to draw insights from analogous, domain-specific, and up-to-date context samples to make grounded inferences. Moreover, it employs a lightweight multimodal perceivers to seamlessly integrate these context samples, enabling an efficient processing of lengthy context tokens. Our proposed CaMML-13B model achieves state-of-the-art results across more than ten prominent multimodal benchmarks, surpassing previous methods by a substantial margin. These achievements underscore CaMML's effectiveness in various multimodal applications.

7 Limitations

While CaMML effectively integrates multimodal retrieval into large models, the presence of irrelevant examples may impede its performance. In addition, CaMML maintains a datastore for experimental analysis, yet it's unable to encompass all data domains, such as medical images. Consequently, when the data distribution in the datastore diverges significantly from user or test queries, CaMML struggles to leverage its retrieval-augmented capability effectively. Lastly, despite its proficiency in inference and memory utilization, CaMML demands substantial training on large language models (similar to LLaVA) and relies on access to a sizable datastore to demonstrate its effectiveness.

8 Potential Risks

CaMML entails certain potential risks, including: Environmental unfriendliness and escalated computational costs, particularly noticeable with larger models. Despite being entirely encapsulated within its pipeline, with users unable to access the datastore directly (limited to embedding and data entry index), there remains a possibility for users to utilize the CaMML I/O API to instruct LLM to generate texts that may raise concerns or cause privacy leakage.

References

2023. Sharegpt. https://sharegpt.com.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. 2022. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Alessandro Antonietti. 2012. *Analogy-Based Learning*, pages 235–237. Springer US, Boston, MA.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *CVPR*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2022. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. 2021. Perceiver: General perception with iterative attention. In *ICML*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019a. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019b. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. *The International Conference on Learning Representations*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *EMNLP*.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Thirty-seventh Conference on Neural Information Processing Systems*.

Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. 2023c. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhnag, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2023a. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023b. Chameleon: Plug-and-play compositional reasoning with large language models.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

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

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*.

Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *AAAI*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*.

Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. 2023. Unified model for image, video, audio and language tasks.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioningwith reading comprehension.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *NeurIPS*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*. PMLR.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *ECCV*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *CVPR*.

A Extended Experiments

A.1 Experimental Setup

Summary of Datasets and Benchmarks. ScienceQA (Lu et al., 2022a) train split is used for ScienceQA finetuning and ablation study. In our instruction-tuning training, we adopt

LLaVA-665K dataset, which contains LLaVA-158K, ShareGPT-40k (sha, 2023), VQAv2 (Antol et al., 2015), GQA (Hudson and Manning, 2019), OKVQA (Marino et al., 2019), OCRVQA (Mishra et al., 2019), A-OKVQA (Schwenk et al., 2022), TextCaps (Sidorov et al., 2020), RefCOCO (Yu et al., 2016) and VG (Krishna et al., 2017). In our evaluation, VQAv2, GQA, TextVQA (Singh et al., 2019), MME (Fu et al., 2023), POPE (Li et al., 2023c), MM-Vet (Yu et al., 2023), ScienceQA, MMBench (Liu et al., 2023d), MMBench-CN (Liu et al., 2023d), SEED-Bench (Li et al., 2023a) and Vizwiz (Gurari et al., 2018) are considered as benchmarks. We also conduct evaluation on COCO Caption (Chen et al., 2015), Flickr30k (Plummer et al., 2015), OKVQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), and RefCOCO/+/g (Yu et al., 2016). In our qualitative visualization, we adopt another 2M datastore as the source of context examples, which comprised from external resources incorporates 2,348K multimodal samples, ranging from captioning with BLIP-LAION's 558K entries and Local Narrative (Pont-Tuset et al., 2020), to knowledge-based QA with KVQA (Sanket Shah and Talukdar, 2019), narrative-driven QA from VCR (Zellers et al., 2019) and Visual7W (Zhu et al., 2016), visual grounding via RefCOCOPlus (Yu et al., 2016) and RefCOCOg (Yu et al., 2016), OCR from TextOCR (Singh et al., 2021), along with the in-domain LLaVA-665K set.

Summary of Evaluation Metrics. We illustrate the evaluation metrics on Table 2:

- Accuracy: VQAv2, GQA, Vizwiz, SQA-Image, TextVQA, MMBench, MMBench-CN, SEED.
- GPT4-Assisted Evaluation Score: MM-Vet.
- F1 Score (POPE Paper Sec 5.1 Metrics): POPE.
- Figures in MME represent sum of the scores (which measures QA accuracy and accuracy+, illustrated in MME Paper (Fu et al., 2023) Sec 2.2) of all MME perception subtasks, including existence, count, position, color, poster, celebrity, scene, landmark, artwork, and OCR. The full score of each subtask is 200, and that of all perception is 2000.

Summary of Pretrained Checkpoints. We utilize Vicuna-7B/13B (Zheng et al., 2023) as our foundation Large Language Model (LLM), ViT-L-14 architecture is used as the vision encoder. In detail, Vicuna-v1.3[3] and CLIP-ViT-L-14[4] (Radford et al., 2021) initialization is for ScienceQA finetuning and Ablation studies, Vicuna-v1.5[5], CLIP-ViT-L-14-336px[6], and LLaVA-1.5 multimodal projector[7] is initialized for instruction finetuning. For the source of contexts, the ImageBind[8] (Girdhar et al., 2023)-Huge model is adopted to embed texts and images, computing their similarities for indexing top-k samples.

Implementation Details. [9] In our ScienceQA finetuning, we train for 12 epochs, with batch size 4 per GPU and learning rate 2e-5. We illustrate the hyperparameter configurations here for ScienceQA finetuning in main paper Table 1:

- **CaMML-7B:** {number of query M=128, number of perceiver layer 2, perceiver layer hidden size 768, number of shots N=1},
- **CaMML-13B:** {number of query M=256, number of perceiver layer 2, perceiver layer hidden size 4608, number of shots N=3}.

In our instruction finetuning, we train for 1 epoch on 8GPUs, with batch size 8 per GPU and learning rate 2e-5, the LLM is activated as well as CaMML perceivers, while the vision encoder and CaMML retriever is frozen. We illustrate the hyperparameter configurations here for instruction finetuning in main paper Table 2&3:

- **CaMML-7/13B:** {number of query M=128, number of perceiver layer 2, perceiver layer hidden size 768, mixed-training shots 1$\sim$3 and inferenced shots N=3}.

[3]https://huggingface.co/lmsys/vicuna-7b-v1.3, https://huggingface.co/lmsys/vicuna-13b-v1.3
[4]https://huggingface.co/openai/clip-vit-large-patch14
[5]https://huggingface.co/lmsys/vicuna-7b-v1.5, https://huggingface.co/lmsys/vicuna-13b-v1.5
[6]https://huggingface.co/openai/clip-vit-large-patch14-336
[7]https://huggingface.co/liuhaotian/llava-v1.5-mlp2x-336px-pretrain-vicuna-7b-v1.5, https://huggingface.co/liuhaotian/llava-v1.5-mlp2x-336px-pretrain-vicuna-13b-v1.5
[8]https://dl.fbaipublicfiles.com/imagebind/imagebind_huge.pth
[9]All the experiments are trained under Deepspeed Zero-3 FP16 configuration.

Method (shots)	A-OKVQA Acc	RefCOCO Acc	RefCOCO+ Acc	RefCOCOg Acc
CaMML-7B (3)	81.1	66.6	60.3	57.6
CaMML-13B (3)	82.0	70.6	65.9	60.5

Table 5: CaMML multimodal performance on A-OKVQA, Refcoco/+/g.

A.2 CaMML Retrieval Setup

We built CaMML retriever in following steps:

- Utilize ImageBind (Girdhar et al., 2023) model to inference upon images and get corresponding visual embedding.
- Utilize Faiss (Johnson et al., 2019b) to build datastore index for visual embedding, and bind each data source (image and text) to the index, ensuring the right one-to-one retrieval.
- For each query, we forward with ImageBind model and compute similarity between query embedding and datastore index. We select top-k similarity samples as our contexts.

Note that, in our quantitative experiments, we compute query (text) embedding's similarity with (vision) datastore, while in qualitative analyses, we compute query (vision) embedding's similarity with (vision) datastore, to obtain relevant image information.

A.3 Experimental Results

A.3.1 Multimodal Task Performance

In addition to the results presented in the main paper tables, CaMML demonstrates versatility in handling various multimodal tasks without requiring further fine-tuning. Table 5 shows that CaMML achieves exceptional performance on the Augmented Outside-Knowledge VQA (A-OKVQA) task with an accuracy of 82.0, and also exhibits good ability in localizing visual grounding tasks such as RefCOCO/+/g.

A.3.2 Finegrained Evaluation

CaMML is tested on MMBench and MMBench-CN to showcase each fine-grained ability such as Logic Reasoning (LR), Attribute Recognition (AR), RR (Relation Reasoning), Instance-Level Fine-Grained Perception (FP-S), Cross-Instance Fine-Grained Perception (FP-C), Coarse Perception (CP). According to Table 6, CaMML-13B has achieved 8 out of 12 state-of-the-art results among large multimodal models.

B Additional Cases

Image Generation CaMML demonstrates its ability to generate reliable content based on a wealth of multimodal contextual sources. In Figure 8 9, we present an example of CaMML's capability in prompting image generation. In the case, CaMML is known for its ability to incorporate contextual elements into generating descriptions. For example, it can generate descriptions of traditional Chinese-style buildings in cyber-punk style. When using CaMML, the generated images are likely to include a mix of old buildings with a modern style, based on the provided contextual samples. On the other hand, when using simple prompt without any contextual samples, the generated image is more likely to be limited to modern style skyscrapers only.

Implicit-Description Localization CaMML also highlights its strong localization capabilities, particularly in terms of reasoning the concept behind and high-level understanding. We utilize CaMML to detect objects by providing implicit descriptions and allowing CaMML to reason about the nature and location of these objects. As shown in Figure 10, CaMML is able to identify the objects with implicit descriptions (such as airplane: "The thing that carries hundreds of souls to the sky" or mirror: "The thing that makes two same dogs existing in this image", *etc.*).

Figure 8: CaMML capability of Image generation: Create a city view via DALL·E3 API.

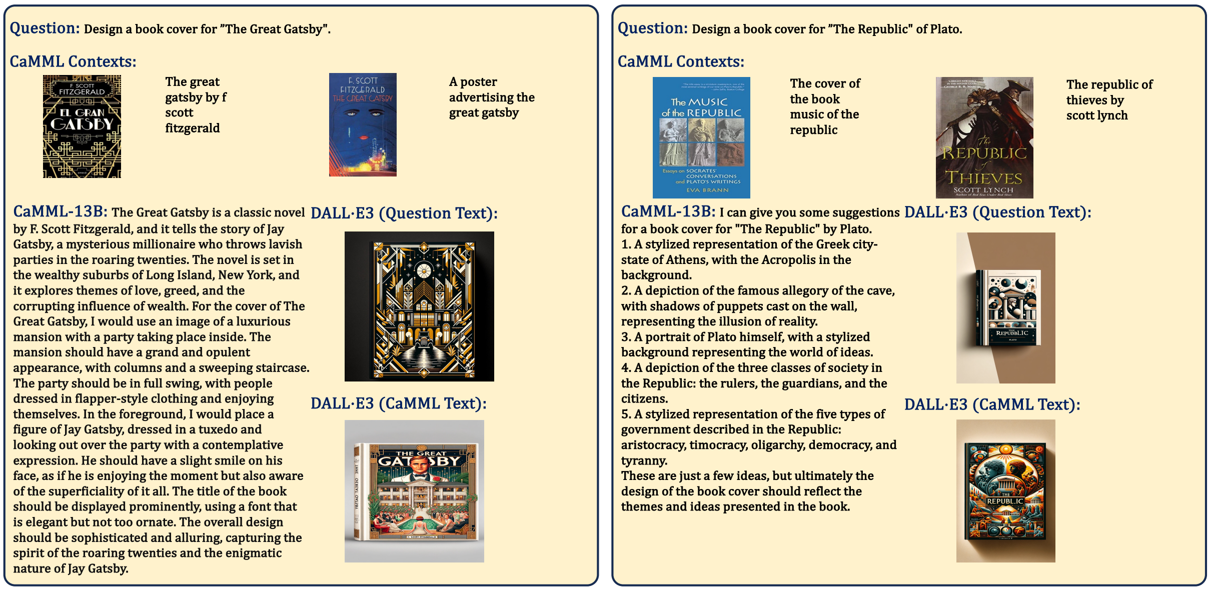

Figure 9: CaMML capability of Image generation: Design a book cover via DALL·E3 API.

Method	LLM	MMBench-dev							MMBenchCN-dev						
		Overall	LR	AR	RR	FP-S	FP-C	CP	**Overall**	LR	AR	RR	FP-S	FP-C	CP
MMICL (Zhao et al., 2023)	FLANT5-XXL	67.9	49.2	71.6	73.0	66.7	57.2	77.2	-	-	-	-	-	-	-
mPLUG-Owl2 (Ye et al., 2023)	LLaMA2-7B	66.5	32.2	72.4	60.9	68.6	60.1	79.4	59.5	28.8	64.8	48.7	60.1	50.3	76.0
SPHINX (Lin et al., 2023)	LLaMA2-13B	67.2	33.1	67.3	58.3	74.4	59.4	80.7	58.6	21.2	61.8	43.5	62.1	58.7	73.6
LLaVA-1.5 (Liu et al., 2023a)-7B	Vicuna-7B	63.0	26.3	68.8	53.0	67.2	56.6	76.4	57.4	25.4	58.8	55.7	55.3	49.7	75.7
LLaVA-1.5 (Liu et al., 2023a)-13B	Vicuna-13B	68.2	44.1	67.3	60.0	72.0	59.4	82.1	61.9	36.4	65.8	49.6	62.1	59.4	75.0
CaMML-7B	Vicuna-7B	66.9	34.2	71.6	66.1	70.4	56.5	78.9	60.6	27.1	87.0	55.7	58.0	57.3	77.0
CaMML-13B	Vicuna-13B	**70.2**	44.2	**87.5**	60.9	72.4	**67.8**	**84.6**	**63.5**	**37.3**	**88.9**	52.2	**63.1**	**59.4**	**77.0**

Table 6: Experiments on fine-grained multimodal reasoning ability: CaMML evaluated on MMBench and MMBenchCN, compared with other state-of-the-art methods. The categories include Logic Reasoning (LR), Attribute Recognition (AR), RR (Relation Reasoning), Instance-Level Fine-Grained Perception (FP-S), Cross-Instance Fine-Grained Perception (FP-C), Coarse Perception (CP).

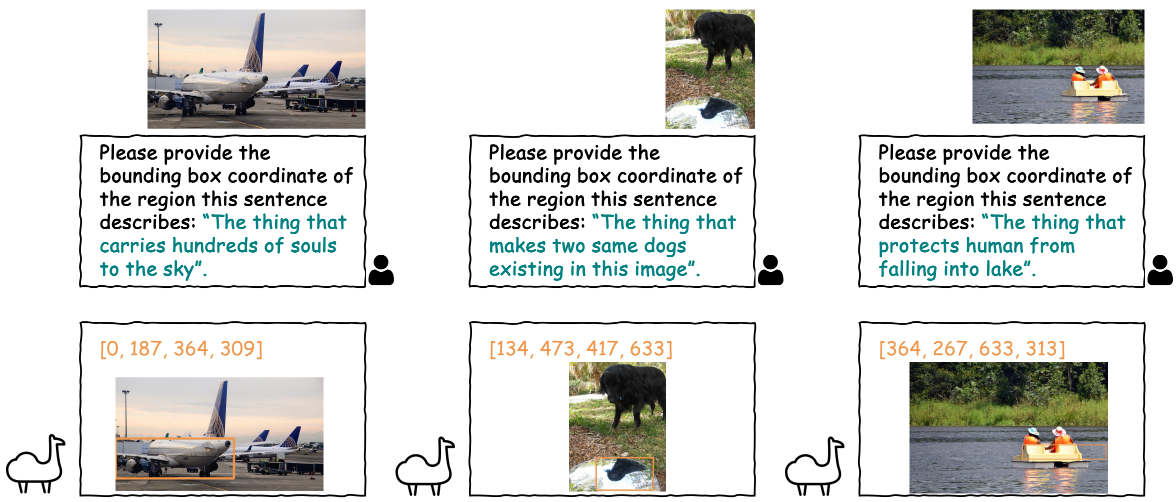

Figure 10: CaMML capability of localization on implicit description.