BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions

Wenbo Hu[*1], **Yifan Xu**[*2], **Yi Li**[1] **Weiyue Li**[1] **Zeyuan Chen**[1] **Zhuowen Tu**[1]
[1]UC San Diego
[2]Coinbase Global, Inc.
{w1hu, yil115, wel019, zec016, ztu}@ucsd.edu yifan.xu@coinbase.com

Abstract

Vision Language Models (VLMs), which extend Large Language Models (LLM) by incorporating visual understanding capability, have demonstrated significant advancements in addressing open-ended visual question-answering (VQA) tasks. However, these models cannot accurately interpret images infused with text, a common occurrence in real-world scenarios. Standard procedures for extracting information from images often involve learning a fixed set of query embeddings. These embeddings are designed to encapsulate image contexts and are later used as soft prompt inputs in LLMs. Yet, this process is limited to the token count, potentially curtailing the recognition of scenes with text-rich context. To improve upon them, the present study introduces **BLIVA**: an augmented version of Instruct**BLI**P with **V**isual **A**ssistant. BLIVA incorporates the query embeddings from InstructBLIP and also directly projects encoded patch embeddings into the LLM, a technique inspired by LLaVA. This approach assists the model to capture intricate details potentially missed during the query decoding process. Empirical evidence demonstrates that our model, BLIVA, significantly enhances performance in processing text-rich VQA benchmarks (up to 17.76% in OCR-VQA benchmark) and in undertaking typical VQA benchmarks (up to 7.9% in Visual Spatial Reasoning benchmark), comparing to our baseline InstructBLIP. BLIVA demonstrates significant capability in decoding real-world images, irrespective of text presence. To demonstrate the broad industry applications enabled by BLIVA, we evaluate the model using a new dataset comprising YouTube thumbnails paired with question-answer sets across 13 diverse categories. For researchers interested in further exploration, our code and models are freely accessible at https://github.com/mlpc-ucsd/BLIVA.git.

Introduction

Recently, Large Language Models (LLMs) have transformed the field of natural language understanding, exhibiting impressive capabilities in generalizing across a broad array of tasks, both in zero-shot and few-shot settings. This success is mainly contributed by instruction tuning (Wu et al. 2023) which improves generalization to unseen tasks by framing various tasks into instructions. Vision Language Models (VLMs) such as OpenAI's GPT-4 (OpenAI 2023), which incorporates LLM with visual understanding capability, have demonstrated significant advancements in addressing open-ended visual question-answering (VQA) tasks. Several approaches have been proposed for employing LLMs on vision-related tasks by directly aligning with a visual encoder's patch feature (Liu et al. 2023a) or extracting image information through a fixed number of query embeddings. (Li et al. 2023b; Zhu et al. 2023).

However, despite exhibiting considerable abilities for image-based human-agent interactions, these models struggle with interpreting text within images. Images with text are pervasive in our daily lives, and comprehending such content is essential for human visual perception. Previous works utilized an abstraction module with queried embeddings, limiting their capabilities in textual details within images (Li et al. 2023b; Awadalla et al. 2023; Ye et al. 2023).

In our work, we combine learned query embeddings with additional visual assistant branches, utilizing encoded patch embeddings. This approach addresses the constraint image information typically provided to language models, leading to improved text-image visual perception and understanding. Empirically, we report the results of our model in general VQA benchmarks following the evaluation datasets of (Dai et al. 2023) and text-rich image evaluation protocol from (Liu et al. 2023b). Our model is initialized from a pre-trained InstructBLIP and an encoded patch projection layer trained from scratch. Following (Zhu et al. 2023; Liu et al. 2023a), we further demonstrate a two-stage training paradigm. We begin by pre-training the patch embeddings projection layer. Subsequently, with the instruction tuning data, we fine-tune both the Q-former and the patch embeddings projection layer. During this phase, we maintain both the image encoder and LLM in a frozen state. We adopt this approach based on two findings from our experiments: firstly, unfreezing the vision encoder results in catastrophic forgetting of prior knowledge; secondly, training the LLM concurrently didn't bring improvement but brought significant training complexity.

In summary, our study consists of the following highlights:

- We present **BLIVA**, which leverages both learned query embeddings and encoded patch embeddings, providing an effective method for interpreting text within images.
- Our experimental results affirm that **BLIVA** provides im-

[*]These authors contributed equally.

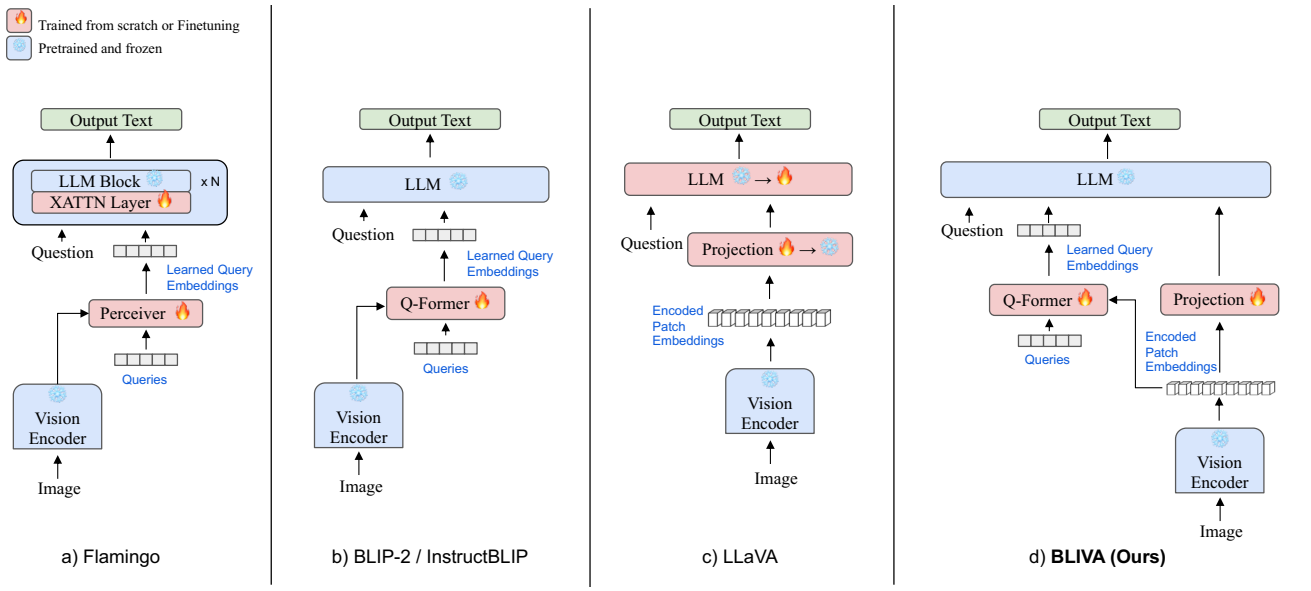

Figure 1: **Comparison of various VLM approaches**. Both (a) Flamingo (Alayrac et al. 2022) and (b) BLIP-2 / InstructBLIP (Li et al. 2023b; Dai et al. 2023) architecture utilize a fixed, small set of query embeddings. These are used to compress visual information for transfer to the LLM. In contrast, (c) LLaVA aligns the encoded patch embeddings directly with the LLM. (d) BLIVA (Ours) builds upon these methods by merging learned query embeddings with additional encoded patch embeddings.

provements in the understanding of text within images while maintaining a satisfactory performance in general VQA benchmarks.

- To underscore the real-world applicability of **BLIVA**, we evaluate the model using a new dataset of YouTube thumbnails with associated question-answer pairs.

Related Work

Multimodal Large Language Model

Large Language Models (LLMs) have demonstrated impressive zero-shot abilities across various open-ended tasks. Despite the successful applications of LLMs in natural language processing, it is still struggling for LLMs to perceive other modalities such as vision and audio. Recent research has also explored the application of LLMs for multimodal generation to understand visual inputs. Some approaches leverage the pre-trained LLM to build unified models for multi-modality. For example, Flamingo (Alayrac et al. 2022) connects the vision encoder and LLM by a Perceiver Resampler, and the gated cross-attention modules exhibit impressive few-shot performance. Additionally, BLIP-2 (Li et al. 2023b) designs a Q-former to align the visual feature with OPT (Zhang et al. 2022) and FLAN-T5 (Wei et al. 2021). MiniGPT-4 (Zhu et al. 2023) employed the same Q-former but changed the LLM to Vicuna (Zheng et al. 2023). Some approaches also finetuned LLM for better alignment with visual features such as LLaVA (Liu et al. 2023a) directly finetuned LLM and mPLUG-Owl (Ye et al. 2023) performs low-rank adaption (LoRA) (Hu et al. 2022) to finetune a LLaMA model (Touvron et al. 2023). PandaGPT (Su et al. 2023) also employed LoRA to finetune a Vicuna model on top of ImageBind (Girdhar et al. 2023), which can take multimodal inputs besides visual. While sharing the same two-stage training paradigm, we focus on developing an end-to-end multimodal model for both text-rich VQA benchmarks and general VQA benchmarks.

Multimodal instruction tuning

Instruction tuning has been shown to improve the generalization performance of language models to unseen tasks. In the natural language processing (NLP) community, some approaches collect instruction-tuning data by converting existing NLP datasets into instruction format (Wang et al. 2022b; Wei et al. 2021; Sanh et al. 2022; Chung et al. 2022) others use LLMs to generate instruction data (Taori et al. 2023; Zheng et al. 2023; Wang et al. 2023; Honovich et al. 2022). Recent research expanded instruction tuning to multimodal settings. In particular, for image-based instruction tuning, MiniGPT-4 (Zhu et al. 2023) employs human-curated instruction data during the finetuning stage. LLaVA (Liu et al. 2023a) generates 156K multimodal instruction-following data by prompting GPT-4 (OpenAI 2023) with image captions and bounding boxes coordinates. mPLUG-Owl (Ye et al. 2023) also employs 400K mixed text only and multimodal instruction data for finetuning. Instruction tuning also enhanced the previous vision language foundation model's performance. For example, MultimodalGPT (Gong et al. 2023) designed various instruction templates that incorporate vision and language data for multi-modality instruction tuning OpenFlamingo (Awadalla et al. 2023). (Xu, Shen, and Huang 2023) built a multimodal instruction tuning benchmark dataset that consists of 62 diverse multimodal tasks in a unified seq-to-seq format and finetuned OFA (Wang

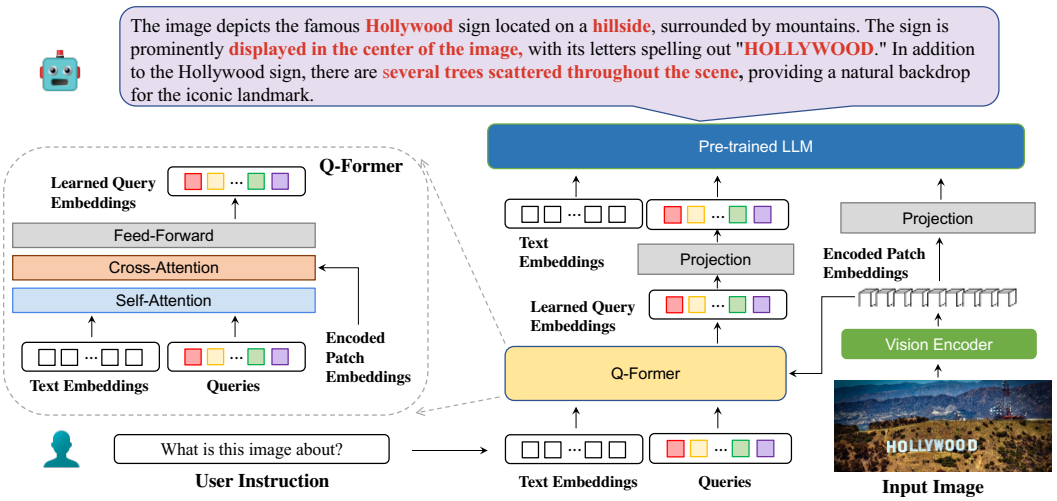

Figure 2: **Model architecture of BLIVA**. BLIVA uses a Q-Former to draw out instruction-aware visual features from the patch embeddings generated by a frozen image encoder. These learned query embeddings are then fed as soft prompt inputs into the frozen Language-Learning Model (LLM). Additionally, the system repurposes the originally encoded patch embeddings through a fully-connected projection layer, serving as a supplementary source of visual information for the frozen LLM.

et al. 2022a). MIMIC-IT (Li et al. 2023a) built a bigger dataset comprising 2.8 million multimodal instruction-response pairs to train a stronger model Otter (Li et al. 2023a). We also employed instruction tuning data following the same prompt as InstructBLIP(Dai et al. 2023) to demonstrate the effectiveness of utilizing additional encoded patch embeddings.

Method

Architecture Overview

As illustrated in Figure 1, there are mainly two types of end-to-end multimodal LLMs: 1) Models that utilize learned query embeddings for LLM. For instance, MiniGPT-4 (Zhu et al. 2023) used the frozen Q-former module from BLIP-2 (Li et al. 2023b) to extract image features by querying the CLIP vision encoder. Flamingo (Alayrac et al. 2022), employed a Perceiver Resampler, which reduced image features to a fixed number of visual outputs for LLM. 2) Models that directly employed image-encoded patch embeddings, such as LLaVA (Liu et al. 2023a), which connect its vision encoder to the LLM using an MLP. Nevertheless, these models exhibit certain constraints. Some models employ learned query embeddings for LLM, which help in better understanding the vision encoder but may miss crucial information from encoded patch embeddings. On the other hand, some models directly use encoded image patch embeddings through a linear projection layer, which might have limited capability in capturing all the information required for LLM.

To address this, we introduce BLIVA, a multimodal LLM designed to incorporate both learned query embeddings — which are more closely aligned with the LLM — and image-encoded patch embeddings that carry richer image information. In particular, Figure 2 illustrates that our model incorporates a vision tower, which encodes visual representations from the input image into encoded patch embeddings. Subsequently, it is sent separately to the Q-former to extract refined learned query embeddings, and to the projection layer, allowing the LLM to grasp the rich visual knowledge. We concatenate the two types of embeddings and feed them directly to the LLM. These combined visual embeddings are appended immediately after the question text embedding to serve as the final input to the LLM. During inference, we employed beam search to select the best-generated output. Conversely, for classification and multi-choice VQA benchmarks, we adopted the vocabulary ranking method as outlined in InstructBLIP (Dai et al. 2023). Given our prior knowledge of a list of candidates, we calculated the log-likelihood for each and chose the one with the highest value as the final prediction.

To make another version for commercial usage of our architecture, we also select FlanT5 XXL as our LLM. This is named as BLIVA (FLanT5$_{\text{XXL}}$) in this paper.

Two stages Training Scheme

We adopted the typical two-stage training scheme: 1) In the pre-training stage, the goal is to align the LLM with visual information using image-text pairs from image captioning datasets that provide global descriptions of images. 2) After pre-training, the LLM becomes familiar with the visual embedding space and can generate descriptions of images. However, it still lacks the capability to discern the finer details of images and respond to human questions. In the second stage, we use instruction tuning data to enhance performance and further align the visual embeddings with the LLM and human values. Recent methods have predominantly adopted a two-stage training approach (Zhu et al. 2023; Liu et al. 2023a; Ye et al. 2023) except PandaGPT (Su et al. 2023), which utilizes a one-stage training method, has also demonstrated commendable results. In BLIVA, our vi-

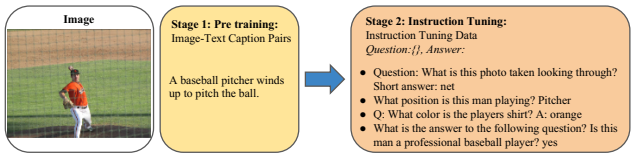

Figure 3: **A typical multi-stage VLM training paradigm.** We follow the paradigm established by InstructBLIP (Dai et al. 2023). The training process involves two key stages. For Q-former, the first stage is done by (Li et al. 2023b) where image and text caption pairs are pre-trained to accomplish a raw alignment between visual and language modalities. As for the patch feature, we followed (Liu et al. 2023a) to use the same pre-training dataset. In the second stage, the alignment is further refined using instruction tuning VQA data, which facilitates a more detailed understanding of visual input based on language instructions.

sual assistant branch, specifically the encoded patch embeddings, diverges from the approach of BLIP-2 (Li et al. 2023b), which uses a 129M pre-training dataset. Instead, it leverages a more compact 0.5M pre-training caption data following (Liu et al. 2023a). This presents a more efficient strategy for aligning the visual encoder and LLM at the first stage.

We employed language model loss as our training objective. The model learns to generate subsequent tokens based on the preceding context.

Thumbnails Dataset

To showcase the wide-ranging industry applications made feasible by BLIVA, we assess the model by introducing a new evaluation dataset, named **YTTB-VQA** which consists of 100 **Y**ou**T**ube **T**hum**b**nail **V**isual **Q**uestion-**A**nswer pairs to evaluate the visual perception abilities of in-text images. It covers 13 different categories which is illustrated in the Appendix Figure 7. During the data collection, we randomly selected YouTube videos with text-rich thumbnails from different categories. We recorded the unique video ID for each YouTube video and obtained the high-resolution thumbnail from the URL "http://img.youtube.com/vi/<YouTube-Video-ID>/maxresdefault.jpg". After retrieving all the YouTube thumbnails, we created the annotation file with the following fields: "video_id" representing the unique identification for a specific YouTube video, "question" representing the human-made question based on the text and image in the thumbnail, "video_classes" representing the 13 video categories, "answers" representing the ground truth answer, and "video_link" representing the URL link for each YouTube video. Our Youtube thumbnail datasets are available at https://huggingface.co/datasets/mlpc-lab/YTTB-VQA.

We also provide two sample scenarios from the YTTB-VQA dataset. Figure 4 illustrates BLIVA's capability to provide detailed captions and answer users' visual questions.

Experiment

In this section, we conduct extensive experiments and analyses to show the efficacy of our model. We evaluate our model, baseline, and other SOTA models on four OCR-related tasks and eight general VQA benchmarks, including image captioning, image question answering, visual reasoning, visual conversational QA, image classification, and video question answering. We seek to answer the following:

- How does our proposed method compare to alternative single-image embeddings approaches in text-rich VQA and general VQA benchmarks?
- How do the individual components of our method influence its success?
- How does BLIVA enhance the recognition of YouTube thumbnails?

Datasets

To demonstrate the effectiveness of patch embeddings, we followed (Dai et al. 2023) to use the same training and evaluation data unless mentioned explicitly. Due to the illegal contents involved in LAION-115M dataset (Schuhmann et al. 2021), we cannot download it securely through the university internet. Besides lacking a subset of samples of image captioning, we keep all other training data the same. It includes MSCOCO (Lin et al. 2015) for image captioning, TextCaps (Sidorov et al. 2020), VQAv2 (Goyal et al. 2017), OKVQA (Marino et al. 2019), A-OKVQA (Schwenk et al. 2022), OCR-VQA (Mishra et al. 2019) and LLaVA-Instruct-150K (Liu et al. 2023a). For evaluation datasets, we also follow (Dai et al. 2023) but only keep Flickr30K (Young et al. 2014), VSR (Liu, Emerson, and Collier 2023), IconQA (Lu et al. 2022), TextVQA (Singh et al. 2019), Visual Dialog (Das et al. 2017), Hateful Memes (Kiela et al. 2020), VizWiz (Gurari et al. 2018), and MSRVTT QA (Xu et al. 2017) datasets. Here, for Vizwiz, since there's no ground truth answer for the test split, we choose to use a validation split. For Hateful Memes, the test split also misses answers, so we picked the same number of examples from the training set as our evaluation data. InstructBLIP originally also had GQA (Hudson and Manning 2019) and iVQA (Yang et al. 2021); we contacted the authors for access to their datasets but received no reply yet. As for MSVDQA (Xu et al. 2017), the authors completely removed this dataset from their competition website. For OCR task datasets, we followed (Liu et al. 2023b) to select OCR-VQA (Mishra et al. 2019), Text-VQA (Singh et al. 2019), ST-VQA (Biten et al. 2022), and DOC-VQA (Mathew, Karatzas, and Jawahar 2021).

Implementation Details

We selected the ViT-G/14 from EVA-CLIP (Sun et al. 2023) as our visual encoder. The pre-trained weights are initialized and remain frozen during training. We removed the last layer from ViT (Dosovitskiy et al. 2020) and opted to use the output features of the second last layer, which yielded slightly better performance. In line with InstructBLIP, we employed Vicuna-7B which is a recently released, decoder-only Transformer. It has been instruction-tuned from LLaMA (Touvron

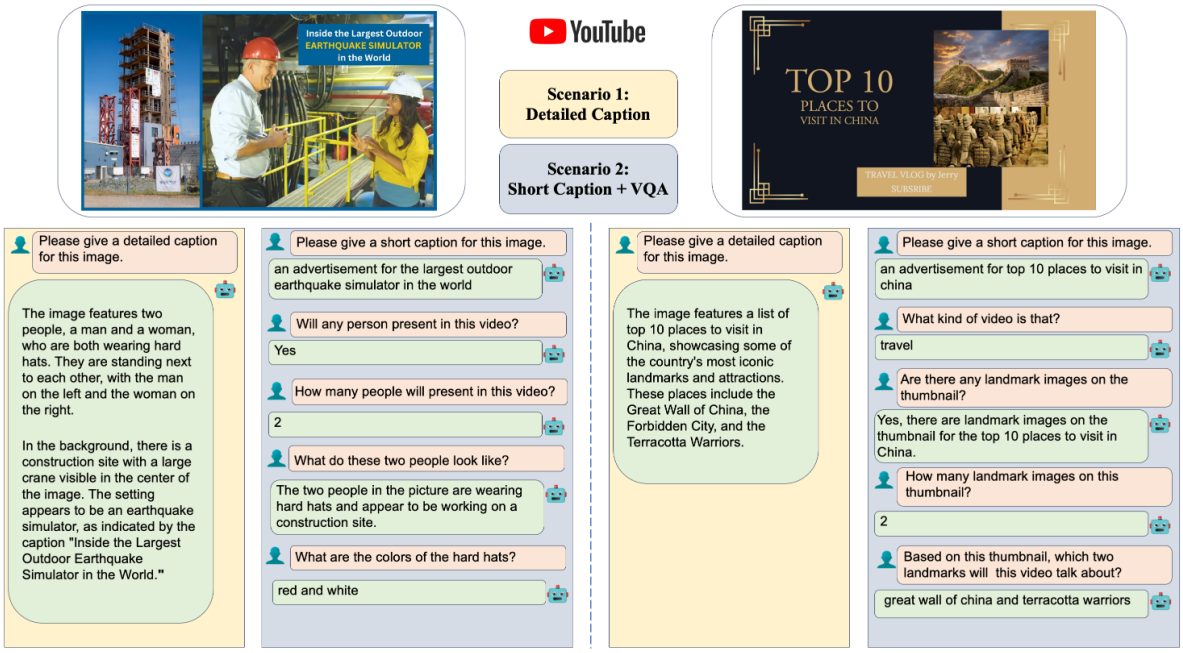

Figure 4: **Two Sample Scenarios from the YTTB-VQA Dataset**: This dataset demonstrates the dual application of BLIVA. The first scenario highlights BLIVA's capability to provide detailed captions that encompass all visual information within an image. The second scenario showcases BLIVA's utility in summarizing visual data into concise captions, followed by its ability to field more detailed visual queries posed by users.

Models	ST-VQA	OCR-VQA	TextVQA	DocVQA
OpenFlamingo (Awadalla et al. 2023)	19.32	27.82	29.08	5.05
BLIP2-OPT$_{6.7b}$ (Li et al. 2023b)	13.36	10.58	21.18	0.82
BLIP2-FLanT5$_{XXL}$ (Li et al. 2023b)	21.38	30.28	30.62	4.00
MiniGPT4 (Zhu et al. 2023)	14.02	11.52	18.72	2.97
LLaVA (Liu et al. 2023a)	22.93	15.02	28.30	4.4
LLaVAR (224^2) (Zhang et al. 2023)	**30.2**	23.4	39.5	6.2
mPLUG-Owl (Ye et al. 2023)	26.38	35.00	37.44	6.17
InstructBLIP (FlanT5$_{XXL}$) (Dai et al. 2023)	26.22	55.04	36.86	4.94
InstructBLIP (Vicuna-7B) (Dai et al. 2023)	28.64	47.62	39.60	5.89
BLIVA (FlanT5$_{XXL}$)	28.24	61.34	39.36	5.22
BLIVA (Vicuna-7B)	29.08	**65.38**	**42.18**	**6.24**

Table 1: **Zero-Shot OCR-Free Results on Text-Rich VQA benchmarks**: This table presents the accuracy (%) results for OCR-free methods, implying no OCR-tokens were used. Note that our work follows InstructBLIP which incorporated OCR-VQA in its training dataset, thus inevitably makes OCR-VQA evaluatoion not Zero-Shot. Except for InstructBLIP, all other results are sourced from (Liu et al. 2023b) on May 20th, 2023, with which we ensure consistency in our results.

et al. 2023) and serves as our LLM. Additional details can be found in Appendix .

Results & Discussions

We introduce our results in the context of each of our three questions and discuss our main findings.

1. How does our proposed method compare to alternative single image embeddings approaches in text-rich VQA and general VQA benchmarks?

Zero-shot evaluation for text-rich VQA benchmarks We compared our data with state-of-the-art Multimodality LLMs. This includes LLaVA, which showcases robust OCR capabilities using only patch embeddings, and its recent OCR-centric version, LLaVAR, which utilizes additional OCR training data. We also considered BLIP2's previous best version, BLIP-FLanT5xxL, the state-of-the-art vision-language model mPlug-Owl (trained on a vast amount of both text and vision-text data), and our baseline, Instruct-