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]University of California, 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 general (not particularly text-rich) VQA benchmarks (up to 7.9% in Visual Spatial Reasoning benchmark), and achieved 17.72% overall improvement in a comprehensive multimodal LLM benchmark (MME), 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 11 diverse categories. For researchers interested in further exploration, our code and models are freely accessible at https://github.com/mlpc-ucsd/BLIVA.

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 employ 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 (not particularly text-rich) 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

*These authors contributed equally.

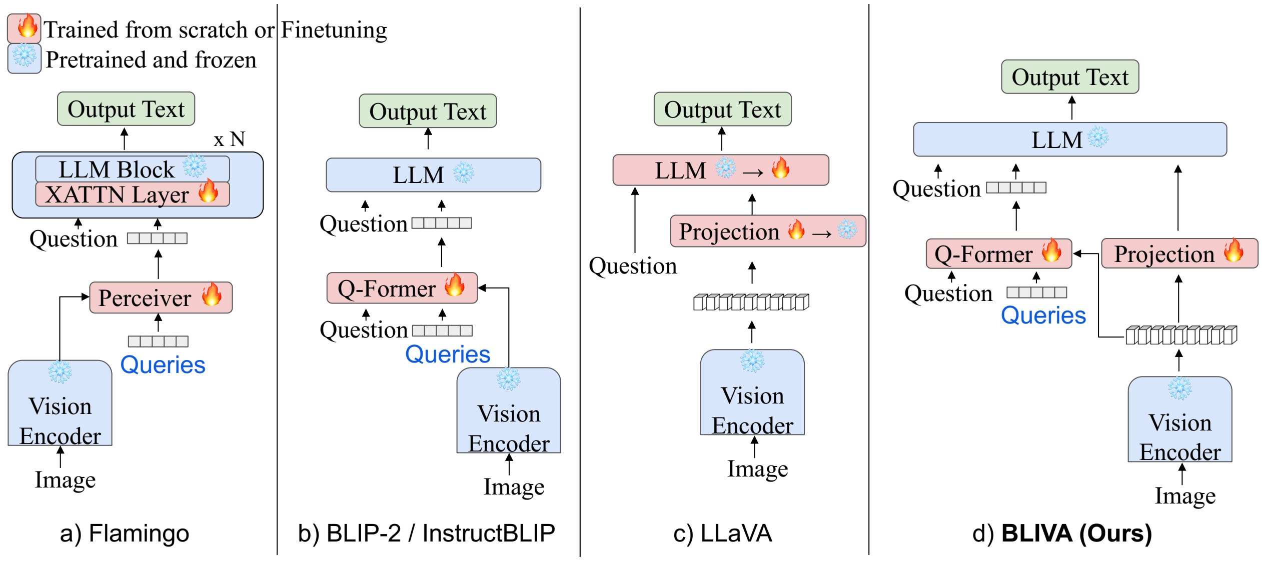

Figure 1: Comparison of various VLM approaches. Both (a) Flamingo (Alayrac et al. 2022) and (b) BLIP-2 / InstructBLIP (Li et al. 2023b; Dai et al. 2023) architecture utilize a fixed, small set of query embeddings. These are used to compress visual information for transfer to the LLM. In contrast, (c) 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.

embeddings and encoded patch embeddings, providing an effective method for interpreting text within images.

- Our experimental results showcase that **BLIVA** provides improvements in the understanding of text within images while maintaining a robust performance in general (not particularly text-rich) VQA benchmarks and achieving the best performance on MME benchmark among previous methods.
- 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. Recent research has 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 which exhibits 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

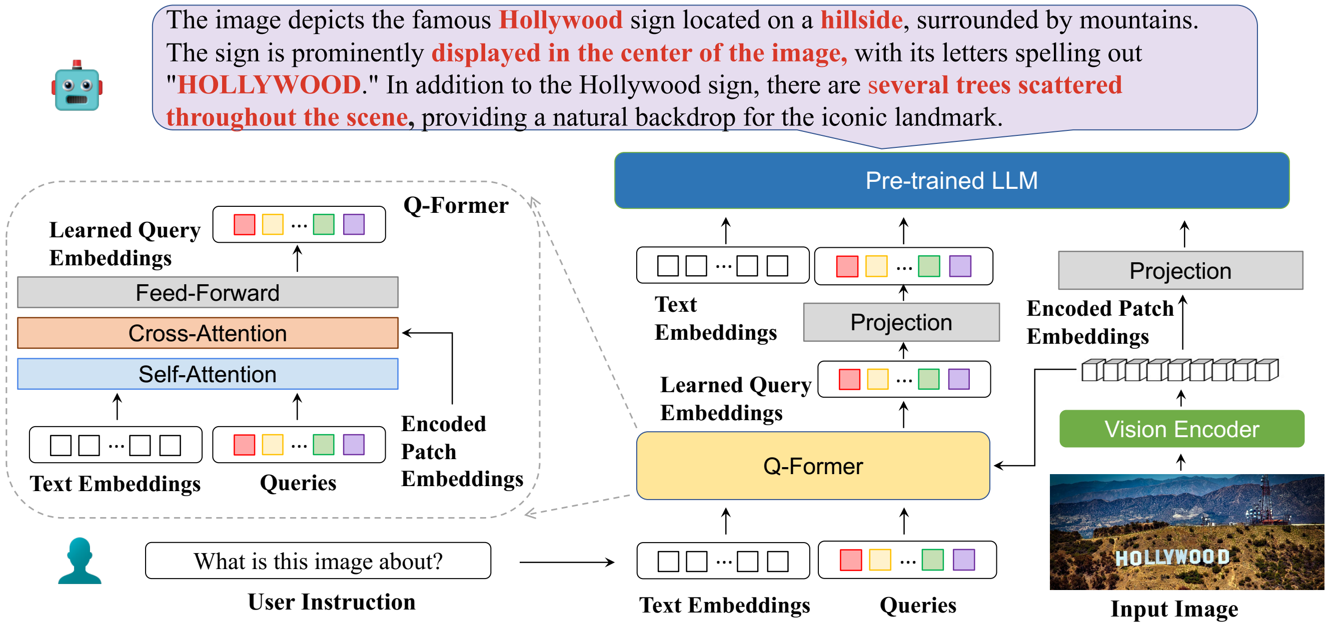

Figure 2: Model architecture of BLIVA. BLIVA uses a Q-Former to draw out instruction-aware visual features from the patch embeddings generated by a frozen image encoder. These learned query embeddings are then fed as soft prompt inputs into the frozen Language-Learning Model (LLM). Additionally, the 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 support another version for commercial usage of our architecture, we also selected FlanT5 XXL as our LLM. This is named as BLIVA ($\text{FLanT5}_{\text{XXL}}$) 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 visual 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 400 **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 11 different categories which is illustrated in the Appendix of arXiv version. 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 11 video categories, "answers" representing the ground truth answer, and "video_link" representing the URL link for each YouTube video.

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 10 OCR-related tasks and 8 general (not particularly text-rich) VQA benchmarks, including image captioning, image question answering, visual reasoning, visual conversational QA, image classification, and video question answering. We also evaluated on a comprehensive multimodal LLM benchmark (MME). We seek to answer the following:

- How does our proposed method compare to alternative single image embeddings approaches in text-rich VQA, general VQA benchmarks and MME benchmark?
- 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. The detailed dataset information can be found at Appendix of arXiv version.

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. We first pre-train our patch embeddings projection layer using LLaVA filtered 558K image-text pairs from LAION (Schuhmann et al. 2021), CC-3M (Sharma et al. 2018), and SBU (Ordonez, Kulkarni, and Berg 2011), captioned by BLIP (Li et al. 2022). Using the pre-training stage leads to slightly better performance. During the vision-language instruction tuning stage, we initialize the Q-Former from InstructBLIP's weight and finetune the parameters of the Q-former and projection layer together while keeping both the image encoder and LLM frozen. We pre-trained the projection layer with 3 epochs with a batch size of 64. During the instruction finetuning stage, we employ a batch size of 24 with a maximum of 200K steps which roughly iterates two epochs of the training data. For both stage training, we used the AdamW (Loshchilov and Hutter 2017) optimizer, with $\beta_1 = 0.9, \beta_2 = 0.999$, and a weight decay of 0.05. Additionally, we apply a linear warmup of the learning rate during the initial 1K steps, increasing from 10^{-8} to 10^{-5}, followed by a cosine decay with a minimum learning rate of 0. The pre-training stage takes 6 hours and the instruction finetuning stage finished within two days on 8 Nvidia A6000 Ada (48G) GPUs.

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, general VQA benchmarks and MME benchmark?

	ST VQA ↑	OCR VQA ↑	Text VQA ↑	Doc VQA ↑	Info VQA ↑	Chart QA ↑	EST VQA ↑	FUN SD ↑	SRO IE ↑	PO IE ↑	Average ↑
OpenFlamingo (Awadalla et al. 2023)	19.32	27.82	29.08	5.05	14.99	9.12	28.20	0.85	0.12	2.12	13.67
BLIP2-OPT$_{6.7b}$ (Li et al. 2023b)	13.36	10.58	21.18	0.82	8.82	7.44	27.02	0.00	0.00	0.02	8.92
BLIP2-FLanT5$_{XXL}$ (Li et al. 2023b)	21.38	30.28	30.62	4.00	10.17	7.20	42.46	1.19	0.20	2.52	15.00
MiniGPT4 (Zhu et al. 2023)	14.02	11.52	18.72	2.97	13.32	4.32	28.36	1.19	0.04	1.31	9.58
LLaVA (Liu et al. 2023a)	22.93	15.02	28.30	4.40	13.78	7.28	33.48	1.02	0.12	2.09	12.84
mPLUG-Owl (Ye et al. 2023)	26.32	35.00	37.44	6.17	**16.46**	**9.52**	**49.68**	1.02	0.64	**3.26**	18.56
InstructBLIP (FlanT5$_{XXL}$) (Dai et al. 2023)	26.22	55.04	36.86	4.94	10.14	8.16	43.84	1.36	0.50	1.91	18.90
InstructBLIP (Vicuna-7B) (Dai et al. 2023)	28.64	47.62	39.60	5.89	13.10	5.52	47.66	0.85	0.64	2.66	19.22
BLIVA (FlanT5$_{XXL}$)	28.24	61.34	39.36	5.22	10.82	9.28	45.66	**1.53**	0.50	2.39	20.43
BLIVA (Vicuna-7B)	**29.08**	**65.38**	**42.18**	**6.24**	13.50	8.16	48.14	1.02	**0.88**	2.91	**21.75**

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 making OCR-VQA evaluation not zero-shot.

Models	VSR ↑	IconQA ↑	TextVQA ↑	Visdial ↑	Flickr30K ↑	HM ↑ (val)	VizWiz ↑ (val-dev)	MSRVTT ↑ (val-dev)
Flamingo-3B (Alayrac et al. 2022)	-	-	30.1	-	60.6	-	-	-
Flamingo-9B (Alayrac et al. 2022)	-	-	31.8	-	61.5	-	-	-
Flamingo-80B (Alayrac et al. 2022)	-	-	35.0	-	67.2	-	-	-
MiniGPT-4 (Zhu et al. 2023)	50.65	-	18.56	-	-	29.0	34.78	-
LLaVA (Liu et al. 2023a)	56.3	-	37.98	-	-	9.2	36.74	-
BLIP-2 (Vicuna-7B) (Dai et al. 2023)	50.0	39.7	40.1	44.9	74.9	50.2	**49.34**	4.17
InstructBLIP (Vicuna-7B) (Dai et al. 2023)	54.3	43.1	50.1	45.2	82.4	54.8	43.3	18.7
InstructBLIP Baseline (Vicuna-7B)	58.67	44.34	37.58	40.58	84.61	50.6	44.10	20.97
BLIVA (Vicuna-7B)	**62.2**	**44.88**	**57.96**	**45.63**	**87.1**	**55.6**	42.9	**23.81**

Table 2: Zero-shot results on general *(not particularly text-rich)* VQA benchmarks. Our baseline is obtained by directly fine-tuning InstructBLIP (Dai et al. 2023). For the three datasets on the right, due to the unavailability of test-set answers, we have evaluated them using validation dev. Here, Visdial and HM denote the Visual Dialog and Hateful Memes datasets, respectively. Following previous works (Alayrac et al. 2022; Yang et al. 2021; Murahari et al. 2020), we report the CIDEr score (Vedantam, Zitnick, and Parikh 2015) for Flickr30K, AUC score for Hateful Memes, and Mean Reciprocal Rank (MRR) for Visual Dialog. For all remaining datasets, we report the top-1 accuracy (%). Notably, for Text-VQA, we have followed InstructBLIP's method of using OCR-tokens for comparison. While InstructBLIP also included GQA, iVQA, and MSVDQA, we were unable to access these datasets due to either unresponsive authors or the datasets being removed from their websites. For ScienceQA and Nocaps, we were unable to reproduce the results of InstructBLIP, hence their results are not reported here.

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 embedding. 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, InstructBLIP. The results are illustrated in Table 1. Our model consistently shows significant improvement across all the text-rich VQA datasets compared to InstructBLIP. Note that since InstructBLIP utilized OCR-VQA as its training dataset, the comparison for this specific dataset isn't zero-shot. We evaluated both InstructBLIP and our model using the OCR-VQA validation set. BLIVA achieved state-of-the-art results among 6 text-rich datasets while mPlug-Owl performed the best in 4 datasets. Compared to mPlug-Owl, which employs about 1104M image captioning data in the Pre-training stage, BLIVA only employs 558K image caption pairs which could explain why BLIVA is not performing the best in information-based VQA tasks such as InfoVQA, ChartQA and ESTVQA. BLIVA demonstrates the best performance on average compared to all previous methods, underscoring our design choice to employ learned query embeddings, further aided by encoded patch embeddings.

Zero-shot evaluation for general *(not particularly text-rich)* VQA benchmarks Next, we compared BLIVA with models that employ single image features. Results are given in Table 2 and in Table 3 for LLMs available for commercial use. Our model consistently and significantly outperformed the original InstructBLIP model in VSR, IconQA, TextVQA, Visual Dialog, Hateful Memes, MSRVTT, and Flickr30K. For VizWiz, our model nearly matched InstructBLIP's performance. This naturally raises the question: why didn't additional visual assistance improve all the benchmarks? We speculate that the additional visual information didn't aid VizWiz task. We continue to investigate this phenomenon in the next ablation study section. Overall, our de-

Models	VSR ↑	IconQA ↑	TextVQA ↑	Visdial ↑	Flickr30K ↑	HM ↑	VizWiz ↑	MSRVTT ↑
						(val)	(val-dev)	(val-dev)
BLIP-2 ($FlanT5_{XXL}$) (Li et al. 2023b)	68.2	45.4	44.1	46.9	73.7	52.0	29.4	17.4
InstructBLIP ($FlanT5_{XXL}$) (Dai et al. 2023)	65.6	51.2	46.6	**48.5**	83.5	**53.6**	41.35	20.79
BLIVA ($FlanT5_{XXL}$)	**68.82**	**52.42**	**57.2**	36.18	**87.66**	50.0	**43.97**	**23.78**

Table 3: Zero-shot results on general *(not particularly text-rich)* VQA benchmarks for models with open LLM eligible for commercial use. Here, the commercial use applicable LLM we reported is $FlanT5_{XXL}$. Same as Table 2, we report the same evaluation datasets with the same evaluation metrics.

Model	Overall ↑	Perception ↑										Cognition ↑			
		Exist.	Count	Pos.	Color	OCR	Poster	Cele.	Scene	Land.	Art.	Comm.	Num.	Trans.	Code
LLaVA(Liu et al. 2023a)	712.5	50.0	50.0	50.0	50.0	50.0	50.0	48.8	50.0	50.0	49.0	57.1	50.0	57.5	50.0
MiniGPT-4(Zhu et al. 2023)	694.3	68.3	55.0	43.3	43.3	57.5	41.8	54.4	71.8	54.0	60.5	59.3	45.0	0.0	40.0
mPLUG-Owl(Ye et al. 2023)	1238.4	120.0	50.0	50.0	50.0	65.0	136.1	100.3	135.5	159.3	96.3	78.6	60.0	80.0	57.5
InstructBLIP(Dai et al. 2023)	1417.9	185.0	143.3	66.7	66.7	72.5	123.8	101.2	153.0	79.8	134.3	129.3	40.0	65.0	57.5
BLIP-2(Li et al. 2023b)	1508.8	160.0	135.0	73.3	73.3	110.0	141.8	105.6	145.3	138.0	136.5	110.0	40.0	65.0	75.0
BLIVA	**1669.2**	180.0	138.3	81.7	180.0	87.5	155.1	140.9	151.5	89.5	133.3	136.4	57.5	77.5	60.0

Table 4: Evaluation of MME-Benchmark. Here we report the results on all the sub tasks, including Existence(Exist.), Count, Position(Pos.), Color, OCR, Poster, Celebrity(Cele.), Scene, Landmark(Land.), Artwork(Art.), Commonsense Reasoning(Comm.), Numerical Calculation(Num.), Text Translation(Trans.), and Code Reasoning(Code). We bold the highest overall score and highlight the Top-2 model of each sub task with underline.

Instruct-BLIP	Baseline (Instruction Tuning Qformer)	Patch Embed-dings	Pre-Training	Fine-tuning LLM	ST-VQA	OCR-VQA	Text–VQA	Doc-VQA	Info-VQA	Chart-QA	EST-VQA	FUNSD	SROIE	POIE	Improvement
✓					28.64	47.62	39.60	5.89	13.10	5.52	47.66	0.85	0.64	2.66	+ 0 %
✓	✓				**30.08**	65.8	40.5	6.13	12.03	8.08	47.02	0.85	0.57	2.62	+ 7.40%
✓	✓	✓			28.86	65.04	40.7	**6.65**	**14.28**	**8.24**	47.72	**1.19**	**1.66**	2.83	**+ 31.72%**
✓	✓	✓	✓		29.08	65.38	**42.18**	6.24	13.50	8.16	**48.14**	1.02	0.88	**2.91**	+ 17.01%
✓	✓	✓	✓	✓	29.94	**66.48**	41.9	6.47	12.51	7.52	46.76	1.02	0.51	2.85	+ 9.65%

Table 5: Results of adding individual techniques of our framework in text-rich VQA benchmarks. We include four ablations that accumulate each technique (i) baseline: instruction tuning InstructBLIP's Qformer. (ii) instruction tuning patch embeddings (iii) pre-training stage of patch embeddings (iv) Finetuning LLM with LORA during the instruction tuning stage.

Instruct-BLIP	Baseline (Instruction Tuning Qformer)	Patch Embed-dings	Pre-Training	Fine-tuning LLM	VSR	IconQA	TextVQA	Visdial	Flickr 30K	HM (val)	VizWiz (val-dev)	MSRVTT (val-dev)	Improvement
✓					54.3	43.1	50.1	45.2	82.4	54.8	43.3	18.7	+ 0%
✓	✓				58.67	44.34	37.58	40.58	84.61	50.6	44.1	20.97	- 1.91%
✓	✓	✓			58.85	**44.91**	**58.8**	41.67	**87.4**	49.1	42.83	23.70	+ 5.43%
✓	✓	✓	✓		**62.2**	44.88	57.96	**45.63**	87.1	**55.6**	42.9	**23.81**	**+ 8.61%**
✓	✓	✓	✓	✓	51.39	41.34	57.82	42.32	82.7	46.2	**44.91**	22.67	+ 1.15%

Table 6: Results of adding individual techniques of our framework in general *(not particularly text-rich)* VQA benchmarks. We include four ablations that accumulate each technique same as in Table 5.

sign not only achieved significant improvements in understanding text-rich images but also improves 7 out of 8 general VQA benchmarks.

MME Benchmark We further evaluated BLIVA on a comprehensive Multlimodal LLM benchmark (MME) (Fu et al. 2023). As illustrated in Table 4, BLIVA demonstrates the best performance among the current methods for both the perception and cognition tasks overall. For all text-rich tasks such as OCR, Poster, Numerical Calculation, Text Translation, and code, BLIVA outperforms InstructBLIP. BLIVA achieved top 2 performance across all the tasks except artwork and landmark which demand extensive informational knowledge. This is consistent with our findings from informational VQA, indicating that our light-weight pre-training stage and the missing LAION-115M web image caption dataset during instruction tuning stage both likely contribute to a degradation in BLIVA's internet knowledge base.

2. How do the individual components of our method influence its success?

To investigate the impact of image-encoded patch embeddings, the pre-training stage, and fine-tuning the LLM, we conducted ablation studies, incorporating each element respectively. For simplicity, here we only conduct ablation on the BLIVA (Vicuna-7B) model. Since our baseline is the InstructBLIP, we report the results of using baseline alone as directly finetuning the InstructBLIP model with our data and implementation.

Ablation in text-rich VQA benchmarks For text-rich image related tasks, Table 5 illustrates the results of adding each technique separately. Compared to the baseline, adding patch embeddings improved performance across all tasks with the exception of ST-VQA and OCR-VQA. This can stem from data contamination, as STVQA includes data already present in InstructBLIP's Qformer training set but not included in patch embedding's training set. Without the pre-training stage, the performance of ST-VQA, OCR-VQA, TextVQA, ESTVQA, and POIE decreased, while the rest are benefited. Since the pre-training stage employs image caption pairs, we observed that it didn't benefit BLIVA's performance in text-rich VQA tasks as consistently as in the general VQA tasks. Considering the improvement of all tasks, pre-training is still adopted. BLIVA on average outperforms InstructBLIP by 31.72% without pre-training and 17.01% with it, both outpacing the 7.40% improvement from instruction tuning Qformer. These studies indicate that our design of employing patch embeddings provides more detailed visual information. It also supports our hypothesis that an additional visual assistant improves visual knowledge in areas where the query embeddings either neglect or have limited extraction capabilities.

Ablation in general *(not particularly text-rich)* VQA benchmarks As illustrated in Table 6, the presence of encoded patch embeddings improves performance in all benchmarks significantly except HM and VizWiz. For tasks where we observed a drop in performance, such as HM, which focuses on interpreting the feeling of hatefulness, and VizWiz, which predicts whether a visual question can be answered. We conjecture these tasks can be fulfilled by utilizing global-level query embeddings information such as feeling the hatefulness in the image or if the image's object is unrelated to the question asking. When adding the first pre-training stage, the performance for VSR, VisDial, HM, and MSRVTT tasks improves substantially while others are kept roughly the same. These ablation results confirmed the necessity of two-stage training. During the instruction tuning stage, we also experimented with fine-tuning the LLM using LoRA in conjunction with Q-former and encoded patch embeddings. However, this approach didn't yield as much improvement as our best model and even reduced performance in many tasks. Nonetheless, we have included these results in the ablation study for completeness. We conjecture that frozen LLM has a satisfactory understanding of visual information after our two-stage alignment. The visual embeddings are interpreted as a ”foreign language” to LLM and thus finetuning LLM together is not needed in this case.

Models	Accuracy (%)
MiniGPT4 (Zhu et al. 2023)	47.75
LLaVA (Liu et al. 2023a)	41.75
InstructBLIP (Dai et al. 2023)	82.2
BLIVA (Vicuna-7B)	**83.5**

Table 7: Evaluation results of our collected Youtube thumbnails dataset. We report the top-1 accuracy (%).

3. How does BLIVA enhance the recognition of YouTube thumbnails?

Youtube Thumbnails Evaluation Table 7 illustrates the results of the youtube thumbnail dataset with BLIVA achieving the best performance. From an application perspective, BLIVA has the ability to extract extra visual information from images besides extracting information from YouTube captions alone like LLMs. Our success in this use case can be further expanded to large-scale thumbnail images.

Qualitative Analysis

We use real-life scene images, movie posters, webpages, and memes to demonstrate our model's performance regarding interaction with humans based on text-rich images. The examples are in Appendix of arXiv version. BLIVA showcases exceptional OCR capabilities, paired with a robust localization ability that accurately identifies texts and objects within images.

Conclusion

In this paper, we illustrate the effectiveness of assisting learned query embeddings with encoded image patch embeddings as a visual assistant. This straightforward yet innovative design bolsters performance in both general VQA benchmarks and text-rich VQA benchmarks. Our model, BLIVA, demonstrates superior performance in both academic benchmarks and qualitative real-world examples. Moreover, human evaluation of the model's performance reveals that BLIVA struggles with deciphering numerical symbols in images. This could be attributed to the reduced pixel representation often used for these symbols and needs future work to develop valuable insights. Our work also demonstrates the effectiveness of mixing different types of visual embeddings. We encourage more future work to explore how to scale more visual embeddings to LLM which can be the key to the next stage of Large Vision-Language Models.

Acknowledgements

Zhuowen Tu is funded by NSF Award IIS-2127544.

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 (NeurIPS)*, 35: 23716–23736.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Jitsev, J.; et al. 2023. Openflamingo.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; and Ji, R. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind: One Embedding Space To Bind Them All. In *Computer Vision and Pattern Recognition Conference (CVPR)*.

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. arXiv:2305.04790.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 (CVPR)*, 3608–3617.

Honovich, O.; Scialom, T.; Levy, O.; and Schick, T. 2022. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. arXiv:2212.09689.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

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 (NeurIPS)*, 33: 2611–2624.

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, D.; Li, J.; Le, H.; Wang, G.; Savarese, S.; and Hoi, S. C. 2023b. LAVIS: A One-stop Library for Language-Vision Intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 31–41. Toronto, Canada: Association for Computational Linguistics.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, F.; Emerson, G.; and Collier, N. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics (TACL)*, 11: 635–651.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning.

Liu, Y.; Li, Z.; Li, H.; Yu, W.; Liu, Y.; Yang, B.; Huang, M.; Peng, D.; Liu, M.; Chen, M.; Li, C.; Yin, X.; lin Liu, C.; Jin, L.; and Bai, X. 2023b. On the Hidden Mystery of OCR in Large Multimodal Models. arXiv:2305.07895.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, P.; Qiu, L.; Chen, J.; Xia, T.; Zhao, Y.; Zhang, W.; Yu, Z.; Liang, X.; and Zhu, S.-C. 2022. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. arXiv:2110.13214.

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 (CVPR)*, 3195–3204.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. OCR-VQA: Visual Question Answering by Reading Text in Images. In *ICDAR*.

Murahari, V.; Batra, D.; Parikh, D.; and Das, A. 2020. Large-Scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *ECCV*.

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

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; Dey, M.; Bari, M. S.; Xu, C.; Thakker, U.; Sharma, S. S.; Szczechla, E.; Kim, T.; Chhablani, G.; Nayak, N.; Datta, D.; Chang, J.; Jiang, M. T.-J.; Wang, H.; Manica, M.; Shen, S.; Yong, Z. X.; Pandey, H.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Fevry, T.; Fries, J. A.; Teehan, R.; Bers, T.; Biderman, S.; Gao, L.; Wolf, T.; and Rush, A. M. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. arXiv:2110.08207.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. *arXiv*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of ACL*.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. TextCaps: a Dataset for Image Captioning with Reading Comprehension. arXiv:2003.12462.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 8317–8326.

Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; and Cai, D. 2023. PandaGPT: One Model To Instruction-Follow Them All. *arXiv preprint arXiv:2305.16355*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. arXiv:2303.15389.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca. Accessed:2023-06-04.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022a. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *CoRR*, abs/2202.03052.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560.

Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Ashok, A.; Dhanasekaran, A. S.; Naik, A.; Stap, D.; Pathak, E.; Karamanolakis, G.; Lai, H. G.; Purohit, I.; Mondal, I.; Anderson, J.; Kuznia, K.; Doshi, K.; Patel, M.; Pal, K. K.; Moradshahi, M.; Parmar, M.; Purohit, M.; Varshney, N.; Kaza, P. R.; Verma, P.; Puri, R. S.; Karia, R.; Sampat, S. K.; Doshi, S.; Mishra, S.; Reddy, S.; Patro, S.; Dixit, T.; Shen, X.; Baral, C.; Choi, Y.; Smith, N. A.; Hajishirzi, H.; and Khashabi, D. 2022b. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. arXiv:2204.07705.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wu, Y.; Zhao, Y.; Li, Z.; Qin, B.; and Xiong, K. 2023. Improving Cross-Task Generalization with Step-by-Step Instructions. *arXiv preprint arXiv:2305.04429*.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, 1645–1653.

Xu, Z.; Shen, Y.; and Huang, L. 2023. MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning. arXiv:2212.10773.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *International Conference on Computer Vision (ICCV)*, 1686–1697.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Jiang, C.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qi, Q.; Zhang, J.; and Huang, F. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Nlp.cs.illinois.edu*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; Ott, M.; Shleifer, S.; Shuster, K.; Simig, D.; Koura, P. S.; Sridhar, A.; Wang, T.; and Zettlemoyer, L. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.