BRAVE 🦁: Broadening the visual encoding of vision-language models

Oğuzhan Fatih Kar[1,2], Alessio Tonioni[1], Petra Poklukar[1], Achin Kulshrestha[1], Amir Zamir[2], and Federico Tombari[1,3]

[1]Google [2]Swiss Federal Institute of Technology Lausanne (EPFL) [3] TU Munich
https://brave-vlms.epfl.ch

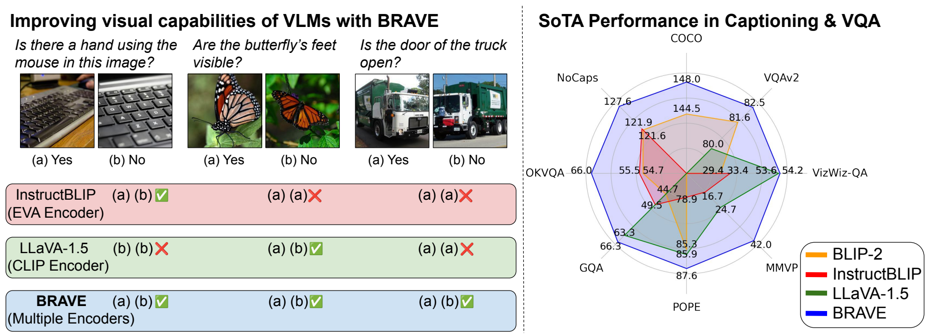

Fig. 1: We propose BRAVE to broaden the visual capabilities of vision-language models (VLMs). **Left:** In contrast to existing methods, e.g. InstructBLIP [25] or LLaVA-1.5 [64], that use a single vision encoder [59, 84], BRAVE combines diverse features from multiple vision encoders into a more versatile and compact representation. The examples are taken from [84] and assess the VLM's ability to differentiate images with visual differences. **Right:** BRAVE leads to state-of-the-art performance on a wide range of captioning and visual question answering tasks. Furthermore, it significantly improves the performance on benchmarks [84] where commonly employed vision encoders fail.

Abstract. Vision-language models (VLMs) are typically composed of a vision encoder, e.g. CLIP, and a language model (LM) that interprets the encoded features to solve downstream tasks. Despite remarkable progress, VLMs are subject to several shortcomings due to the limited capabilities of vision encoders, e.g. "blindness" to certain image features, visual hallucination, etc. To address these issues, we study broadening the visual encoding capabilities of VLMs. We first comprehensively benchmark several vision encoders with different inductive biases for solving VLM tasks. We observe that there is no single encoding configuration that consistently achieves top performance across different tasks, and encoders with different biases can perform surprisingly similarly. Motivated by this, we introduce a method, named BRAVE, that consolidates features from multiple frozen encoders into a more versatile representation that can be directly fed as the input to a frozen LM. `BRAVE` achieves state-of-the-art performance on a broad range of captioning and VQA benchmarks and significantly reduces the aforementioned issues of VLMs, while requiring a smaller number of trainable parameters than existing methods and having a more compressed representation. Our results highlight the potential of incorporating different visual biases for a more broad and contextualized visual understanding of VLMs.

Keywords: Vision-Language · Visual Encoding · Multi-modal Learning

1 Introduction

Vision-language models (VLMs) have recently seen significant improvements on solving tasks requiring both visual and text understanding capabilities such as captioning, visual question answering (VQA), and instruction following. These advancements are fueled by the progress in uni-modal vision encoders [30, 75] and language models (LMs) [22, 24], which are then combined with different "bridging" techniques [3,56,66] to create VLMs. Improved training recipes [25,64] and scalability efforts in terms of training data and model size [6, 19, 77, 87, 99] further boost the performance.

Despite the progress, VLMs are subject to several shortcomings: on the language side, the LMs are known to be susceptible to hallucinations and logical faults [8, 37, 79, 81], while on the vision side, they are limited by the capabilities of the vision encoder. For example, Tong et al. [84] observed that commonly employed vision encoders such as CLIP [75] can exhibit "blindness", i.e. fail to differentiate images with clear visual differences (See Fig. 1, left). Similarly, Li et al. [59] showed that VLMs are prone to visual hallucinations where the models imagine incorrect details about a given image. These limitations create a bottleneck for developing performant and visually grounded VLMs.

To address these, we study broadening the visual encoding capabilities of VLMs. It is known in machine learning that different representations of an input can yield different generalization properties, and each can play a role when ensembled to create a more complete representation [33]. Inspired by this, we first perform a comprehensive evaluation of VLMs which differ only in their vision encoders, i.e. *visual biases*. For this, we consider encoders with different biases, e.g. different objectives, training data, and model sizes. Our results in Table 2 show that different encoders lead to varying performance across vision-language tasks and there is no single encoder achieving a consistent top performance across tasks. Furthermore, we find that encoders with different biases can perform surprisingly similarly, suggesting that there could be multiple cues [35, 45] that can be leveraged to solve the vision-language tasks.

Motivated by these findings, we propose to employ various vision encoders for VLMs and introduce a method to learn how to combine them *efficiently*. We denote the method as `BRAVE`, which stands for ***bro****adening the* ***v****isual* ***e****ncoding of VLMs*. As shown in Fig. 2, `BRAVE` combines features from an *arbitrary* number of vision encoders into a compressed fixed-length visual representation which is then fed as a soft visual prompt to a frozen LM. We achieve the bridging of encoders and the LM by introducing a lightweight ***m**ulti-**e**ncoder **q**uerying trans**former*** (`MEQ-Former`), which is a multi-encoder generalization of the Q-Former proposed in the single encoder BLIP-2 framework [56]. In particular, `MEQ-Former` accepts a textual prompt and a set of learnable queries as inputs and jointly cross-attends to the features from different vision encoders (Fig. 2). We provide a simple yet effective recipe in Sec. 3 for training the `MEQ-Former` in a single stage that bypasses the two-stage pre-training paradigm of [56]. Notably, this is achieved with a smaller number of trainable parameters during pre-training than the existing methods [3, 6, 19, 25, 61, 64, 87] (Sec. 4).

Our evaluation on a wide range of captioning and VQA tasks shows that `BRAVE` effectively consolidates diverse visual signals into a broad and contextual representation, leading to consistently better performance over the state-of-the-art and improved robustness against out-of-distribution inputs, as shown in Fig. 1 and Sec. 4. Furthermore, in contrast to previous works that benefit from scaling the language axis [3, 19, 25, 64] by using larger LMs, our work demonstrates that scaling along the vision axis has a significant potential for VLMs.

Our contributions can be summarized as follows:

- We conduct a systematic analysis of several vision encoders with different inductive biases, e.g. training data, objective, model size, on solving vision-language tasks under a unified training and evaluation framework.
- We introduce `BRAVE`, a method that efficiently consolidates features from any number of vision encoders into a compressed and contextual representation, leading to **1)** state-of-the-art performance on several captioning and VQA tasks and **2)** significantly improved robustness against visual hallucinations and out-of-distribution inputs.
- We perform a comprehensive ablation study of `BRAVE`, highlighting the impact of the design choices, which we hope provide useful insights to the researchers for further advancements in VLMs.

2 Impact of vision encoders for vision-language models

To quantify the impact of visual biases on the performance of VLMs, we compare VLMs with different vision encoders on commonly evaluated VQA and captioning tasks. For this, we develop a pre-training, fine-tuning and evaluation setup, as explained next. To the best of our knowledge, this is the first unified and comprehensive evaluation of different vision encoders for vision-language understanding and generation tasks.

VLM architecture. We use a frozen vision encoder which is connected to a frozen language model by a bridge network with trainable parameters. For the bridge, we adopt the Q-Former module from [56]. This choice is based on the following reasons: **1)** the Q-Former resamples visual features to a fixed-length

Table 1: Overview of vision encoders we benchmarked. All encoders use a ViT [27] backbone, yet they differ in terms of objective, training data, and model size. **ITC:** Image-text contrastive learning [75], **MIM:** Masked image modelling [41], **LGC:** Local-to-global correspondence learning [11]. See Sec. 2 for details.

Vision encoder	Parameters	Training data	Objective
CLIP-L/14 [75]	0.3B	OpenAI WIT	ITC
OpenCLIP-G/14 [21]	1.8B	LAION-2B	ITC
EVA-CLIP-g [30]	1B	LAION-400M	MIM (CLIP features)
SIGLIP-G/14 [100]	1.8B	WebLI	ITC (Sigmoid loss)
SILC-G/16 [73]	1.8B	WebLI	ITC + LGC
ViT-e [19]	3.8B	JFT-3B	Classification
ViT-G [99]	1.8B	JFT-3B	Classification
DINOv2-L/14 [74]	0.3B	LVD-142M	LGC + iBOT [105]

output before feeding them to the LM which enables a fair comparison of vision encoders with different output dimensionalities, **2)** the Q-Former is efficient to train and evaluate as it is compatible with frozen pre-trained vision encoders and LMs. For the vision encoder, we consider a diverse set of options as described in the next paragraph. For the LM, we follow [56] and use FlanT5-XL [24]. The Q-Former has 110M parameters consisting of a series of cross- and self-attention layers and a fixed number of learnable queries. Unless specified otherwise, we use 32 queries of dimension 768 which is also the hidden dimension of the Q-Former. The queries interact with visual features through cross-attention layers, enabling a significant reduction in their dimensionality. For example, CLIP-L/14 [75] visual features are reduced from 257×1024 to 32×768. The output of the Q-Former is then linearly projected to the textual embedding space of the LM to create a *soft visual prompt*, and then prepended to the textual prompt that specifies the task. Both visual and textual prompts are fed as inputs to the frozen LM. For captioning, we use the textual prompt "*A photo of* " and for VQA, we directly use the question as the prompt. Similar to InstructBLIP [25], we give the textual prompt as additional input to the Q-Former together with the learnable queries to extract more task-aligned visual prompts.

Vision encoders. As summarized in Table 1, we consider eight recently introduced vision encoders: CLIP [75], OpenCLIP [21], EVA-CLIP [30], SIGLIP [100], SILC [73], ViT-e [19], ViT-G [99], and DINOv2 [74]. While they all use ViT-based backbones [27], they differ in terms of **1)** training data, e.g. LAION [77], WebLI [19], LVD-142M [74], etc., **2)** training objective, e.g. image-text contrastive learning, masked image modelling, classification, etc., and **3)** model size, e.g. 300M to 4B parameters. Due to this diversity, they incorporate different biases and thus potentially capture different aspects of the underlying depicted scene.

Pre-training data and objectives. We pre-train the Q-Former using WebLI [19] dataset at 224×224 resolution. We use the filtered and de-duplicated

Table 2: Benchmarking vision encoders on different vision-language tasks. For COCO captioning, we report CIDEr [86] score. For VQA tasks, we report top-1 accuracy. For MMVP, we report average pair accuracy [84]. Top-3 best results are shown with Blue, Green, and Orange, respectively. All VLMs use FlanT5-XL [24] as the language model and a Q-Former [56] to bridge vision and language modalities. See Sec. 2 for details.

Vision encoder	COCO cap. ↑ Karpathy val	VQAv2 ↑ Karpathy val	OKVQA ↑ val	GQA ↑ test-dev	MMVP ↑ test
CLIP-L/14	133.0	74.4	61.0	48.7	15.3
OpenCLIP-G	128.3	73.3	60.6	48.0	22.0
EVA-CLIP-g	140.9	77.0	63.0	50.1	27.3
SIGLIP-G/14	133.0	74.7	62.5	48.6	24.0
SILC-G/16	141.1	77.0	63.4	49.7	24.0
ViT-e	137.8	75.6	61.9	49.1	25.3
ViT-G	133.8	74.2	61.2	48.3	20.7
DINOv2-L/14	127.6	71.3	59.0	48.0	22.0

English-only subset with 100 million image-text pairs [19,73]. We keep both the vision encoder and the LM frozen and only train the parameters of the Q-Former. The model is trained with captioning objective using the alt-text in WebLI as target, as it has been shown to be effective [92]. Similar to [56], we remove the last Transformer layer of the encoder. We refer the reader to the supplementary material for additional training details and hyperparameters.

Evaluation tasks. We evaluate the obtained VLMs on standard captioning and VQA tasks. For captioning, we use the COCO captioning benchmark [20] and fine-tune the pre-trained VLM on the commonly employed Karpathy training split [53]. Similar to the pre-training, only the Q-Former parameters are updated. For VQA, we again follow the standard practices [19, 25, 56, 64] and fine-tune on a mixture of VQA data that enforces the Q-Former to extract more task-aligned visual features. Our mixture includes the training sets of VQAv2 [36] and OKVQA [70] as well as the synthetically generated VQA data from $\mathrm{VQ^2A}$ [12], resulting in total 17M training examples. We then evaluate the fine-tuned models on VQAv2 and OKVQA validation sets. Similar to [25, 56], we also evaluate the VLMs' zero-shot VQA capabilities on GQA [48]. Finally, we evaluate the zero-shot performance on MMVP [84] that includes images with semantic differences that are challenging for the current VLMs to handle. Unlike captioning, we fine-tune both Q-Former and the LM parameters while keeping the vision encoder frozen, as we find this to improve the performance, similar to [64]. For both captioning and VQA tasks, we use 224×224 image resolution. See the supplementary for more details and results.

Results. We report the findings in Table 2 and make the following observations:

- **Encoders with different biases can obtain similar performance.** While we observe some variation in VLMs' performance across different encoders, the gap remains small, especially among the top-performing ones. For example, EVA-CLIP-g, SILC-G/16, and ViT-e achieve the best results for VQAv2, GQA and COCO captioning despite their differences in training objective and data, as shown in Table 1. Similar observations can be made for OpenCLIP and DINOv2 encoders that have a small performance gap.
- **MMVP stays challenging for all encoders.** While it is originally curated with "CLIP-blind" pairs in [84], i.e. images that CLIP perceives as similar despite their visual differences (See Fig. 1 for examples), most of the encoders stay below the random guess accuracy of 25%. Exceptions are EVA-CLIP and ViT-e which slightly surpass the random guess level (27.3% and 25.3%).
- **For tasks requiring compositional reasoning and open-world knowledge, the performance decreases and the gap among VLMs diminishes,** as shown by the results on GQA and OKVQA. More precisely, the standard deviation across all the VLMs' performance on GQA and OKVQA is 0.8 and 1.36, respectively, but 1.74 on VQAv2 and 4.91 on COCO.
- **Scaling the size of vision encoder can improve the performance** as seen by the consistent improvements obtained by ViT-e over ViT-G when scaling from 1.8B to 3.8B parameters.
- **Pre-training data distribution can play a critical role.** OpenCLIP-G/14 model is larger than CLIP-L/14 and uses more training data, yet it noticeably underperforms on most of the evaluated VQA and captioning tasks except the MMVP benchmark, where OpenCLIP significantly outperforms the CLIP model. Since both models are trained using the same training objective, this indicates that the impact of the training objective and dataset to the performance of the VLM can be disentangled, and both are important.
- **Performance correlates between VQA and captioning tasks.** Encoders with better performance on VQA tasks generally also perform better on captioning, and vice versa. Spearman's rank correlation between VQA tasks and COCO captioning is 0.90, 0.80, 0.79, 0.56 for VQAv2, OK-VQA, GQA, and MMVP, respectively. Thus, improving the visual component in VLMs is expected to improve performance for a broad range of tasks.

The above results highlight the importance of using the right vision encoder for VLMs, motivating us to focus on the following question: ***Can we broaden the vision capabilities of VLMs through combining vision encoders with different visual biases?*** We answer positively in the next section.

3 Combining vision encoders with `BRAVE`

Motivated by the findings in Sec. 2, we propose using multiple vision encoders with different, and potentially complementary, visual biases to create more capable VLMs. To do this, we introduce `BRAVE`, a method that combines the strengths of different vision encoders while staying efficient in terms of number of trainable parameters. Below, we provide details of `BRAVE`.

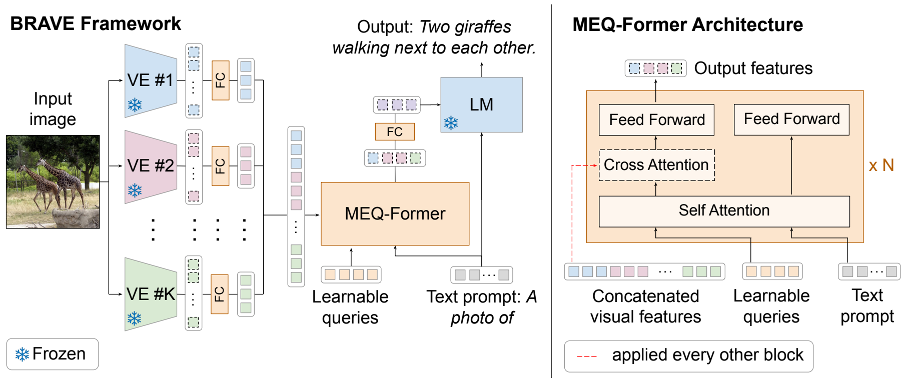

Fig. 2: Overview of BRAVE. Left: We keep all the vision encoders (VEs) and the language model (LM) frozen. The linear projection layers are used to concatenate features from K different VEs, e.g. $K = 5$, sequence-wise. These are then resampled by the `MEQ-Former` which accepts a set of learnable queries and a text prompt describing the task as inputs. The output of `MEQ-Former` is projected to the input space of the LM using fully-connected (FC) layers. The total number of trainable parameters is 116M ($\approx$ 1% of the total parameters). **Right:** Architecture of the `MEQ-Former` with $N = 12$ transformer layers. It interacts with the concatenated features through cross-attention layers and produces a fixed-length output to be fed as input to the LM.

Overview. As depicted in Fig. 2, we introduce the ***m**ulti-**e**ncoder **q**uerying trans**former***, or `MEQ-Former`, that combines visual features from an *arbitrary* set of encoders and outputs a *fixed-length* visual representation that is given as a soft prompt to a frozen LM. It takes as an input a sequence comprised of a fixed number of learnable queries and embedded text tokens which describe the task – for captioning, they are the textual prompt *"A photo of"*, while for VQA, they are the question prompt for a given image (See supplementary for examples). `MEQ-Former` interacts with visual features through cross-attention layers. The visual features from different encoders are first linearly projected to the same dimension and then concatenated sequence-wise. The obtained concatenated feature is given as a key and value pair to the cross-attention layers in `MEQ-Former`, and is cross-attended to by the `MEQ-Former`'s query sequence (Fig. 2). This resampling enables to efficiently processing a large number of visual features since it bypasses the quadratic complexity of self-attention. It also acts as a "bottleneck" that keeps the total number of VLM's parameters low compared to a naive ensembling of VLMs that is prohibitively expensive (See a comparison in Sec. 4.3). Moreover, the visual features are not "marked" with any encoder-specific embedding, thus `MEQ-Former` is not biased to differentiate between encoders, allowing for a simpler design.

Differences with previous work. Our work differs from other popular approaches like LLaVa [64] and PaLI [19] that require significantly more trainable parameters ($\approx$10B) and do not have a resampling mechanism. Resampling visual features from a single encoder has been successfully demonstrated in previous work for single- [25,56] and multiple-frames [3,55,97]. To the best of our knowledge, our work is the first one demonstrating an effective resampling mechanism to consolidate visual features from *several arbitrarily different* encoders while keeping the number of trainable parameters small (our `MEQ-Former` having 116M parameters vs BLIP-2's Q-Former is 188M). See Sec. 5 for more comparisons.

Pre-training details. We use the same data and objective as in Sec. 2 and only train the `MEQ-Former` while keeping the vision encoders and the LM frozen. During pre-training, we randomly mask features from each encoder with 20% probability as we found that this can act as a regularizer and enforces `MEQ-Former` to avoid local minima by attending only to the features of a single encoder (Ablated in Sec. 4.3). After pre-training, we fine-tune the `MEQ-Former` and (optionally) the LM on the downstream tasks, as explained in Sec. 4.

Architecture. For the main results, we combine five vision encoders, namely *EVA-CLIP-g* [30], *CLIP-L/14* [75], *SILC-G/16* [73], *ViT-e* [19], and *DINOv2-L/14* [74] to cover all training datasets and objectives from Table 1. Please see Sec. 4.3 and the supplementary for analysis on the contributions of encoders. For the LM, we employ FlanT5-XL [24] as before. We set the visual feature dimension after linear projection as 1408 for concatenation. For `MEQ-Former`, we use $32 \times 5 = 160$ learnable queries to proportionally scale the model capacity to the number of vision encoders and set the hidden dimension as 768. The `MEQ-Former` resampling results in $14\times$ reduction in total feature size (from 1223×1408 to 160×768). The total number of trainable parameters during pre-training is 116M which is about 1% of the total number of parameters in the VLM (10.3B).

Implementation. We use FLAX [42] with the Scenic [26] library to implement our models and training pipeline. For evaluations, we use both Scenic and EvalAI API [93]. We pre-train with batch size of 1024 on 64 TPUv5 chips.

4 Experiments

We perform evaluations on a broad range of captioning (Sec. 4.1) and VQA tasks (Sec. 4.2) to show the effectiveness of `BRAVE`. Please see Fig. 1 right for a summary of the results and supplementary for an overview of the tasks. We then provide a comprehensive analysis of `BRAVE` in Sec. 4.3.

For captioning, we evaluate on COCO [20] and NoCaps [2] benchmarks. The latter includes out-of-distribution samples with novel classes that do not exist in the COCO benchmark, hence it serves as a challenging testbed to assess robustness. The results (Table 3) show that `BRAVE` pushes the state-of-the-art for both benchmarks while staying efficient in terms of trainable parameters.

For VQA, we evaluate on VQAv2 [36], OKVQA [70], GQA [48], and VizWiz-QA [39] benchmarks. To assess the robustness of VLMs, we additionally include

Table 3: Captioning. We compare BRAVE with state-of-the-art captioning methods on COCO and NoCaps (zero-shot). CIDEr [86] score is reported (Best in **bold**, second best is underlined). Numbers of other methods are taken from the corresponding publications. Dashed line means no result was reported. BRAVE achieves the best results in NoCaps evaluation sets (both out-domain and overall in validation & test) and second best for COCO, while having much fewer trainable parameters than other methods. For example, compared to PaLI-17B [19], BRAVE uses significantly fewer trainable parameters, training data, and image pixels. Yet, it outperforms the baselines on NoCaps and stays competitive on COCO, demonstrating the effectiveness of having different visual biases for generalization. See Sec. 4.1 for more details.

Method	# params		COCO (fine-tuned)	NoCaps (zero-shot, val)		NoCaps (zero-shot, test)	
	Trainable	Total	Karpathy test	out-domain	overall	out-domain	overall
Flamingo [3]	10.6B	80B	138.1	-	-	-	-
SimVLM [91]	632M	632M	143.3	113.7	112.2	-	110.3
Qwen-VL [6]	9.6B	9.6B	-	-	121.4	-	-
BLIP-2 [56]	1.1B	4.1B	144.5	124.8	121.6	-	-
InstructBLIP [25]	188M	14.2B	-	-	121.9	-	-
CoCa [96]	2.1B	2.1B	143.6	-	122.4	-	120.6
GiT2 [87]	5.1B	5.1B	145.0	<u>130.6</u>	126.9	122.3	<u>124.8</u>
PaLI-17B [19]	16.9B	16.9B	**149.1**	-	<u>127.0</u>	<u>126.7</u>	124.4
BRAVE	**116M**	10.3B	<u>148.0</u>	**133.3**	**127.6**	**127.1**	**125.6**

POPE [59] that measures object hallucinations, and MMVP [84] which exposes the shortcomings of CLIP-based vision encoders that are widely used in VLMs. Our results (Table 4) show that BRAVE achieves consistent improvements over recent methods for all the tasks by leveraging diverse visual features.

4.1 Image captioning

Fine-tuning details. Similar to Sec. 2, we only fine-tune the MEQ-Former on COCO captioning and keep the vision encoders and the LM frozen. During fine-tuning, we set a higher input resolution as 336×336 to further boost the performance, similar to previous works [19, 56, 64] (See Sec. 4.3 for an ablation). Following [19, 25, 56, 87, 96], we also zero-shot evaluated the COCO fine-tuned model on NoCaps [2]. Please see the supplementary for more training details.

Results. We summarize the captioning evaluations in Table 3. BRAVE uses the least number of trainable parameters (116M) yet achieves strong results for both COCO and NoCaps benchmarks. For NoCaps, BRAVE is the best performing method with significant gains over recent methods. This is especially the case for out-domain samples with novel classes, demonstrating the usefulness of diversity in visual features for robustness. For COCO, BRAVE stays competitive with the best performing model, PaLI-17B [19]. Notably, BRAVE achieves this while using 150× fewer trainable parameters (116M vs 16.9B), 16× less pre-training data (100M vs 1.6B) and 3× fewer image pixels (336×336 vs 588×588) than PaLI-17B, suggesting that having different visual biases is effective for generalization, while keeping the sample complexity low.

Table 4: Visual question answering. We compare BRAVE with state-of-the-art VQA methods on VQAv2, OKVQA, GQA, VizWiz-QA, MMVP, and POPE. Top-1 accuracies are reported (Best in **bold**, second best is underlined). For MMVP, we report average pair accuracy [84]. Numbers of other methods are taken from the corresponding publications. Dashed line means no result was reported. BRAVE achieves the best results for six out of seven benchmarks, while staying efficient in terms of trainable parameters. Similar to the COCO captioning results in Table 3, BRAVE stays competitive with PaLI-17B [19] for VQAv2, while using significantly less pre-training data, image resolution, and model parameters (both trainable and total). See Sec. 4.2 for details.

Method	# params		Fine-tuned			Zero-shot			
	Trainable	Total	VQAv2 test-dev	OKVQA val	GQA test-dev	VizWiz-QA test-dev	GQA test-dev	MMVP test	POPE test
SimVLM [91]	632M	632M	80.0	-	-	-	-	-	-
Flamingo [3]	10.2B	80B	82.0	57.8	-	31.6	-	-	-
MiniGPT-v2 [14]	7B	8B	-	57.8	60.1	<u>53.6</u>	-	-	-
GiT2 [87]	5.1B	5.1B	81.7	-	-	-	-	-	-
Qwen-VL [6]	9.6B	9.6B	79.5	58.6	59.3	35.2	-	-	-
SPHINX-2k [61]	13B	16.5B	80.7	62.6	63.1	44.9	-	-	<u>87.2</u>
PaLI-17B [19]	16.9B	16.9B	**84.3**	<u>64.5</u>	-	-	-	-	-
BLIP-2 [56]	1.2B	12.1B	81.6	54.7	-	29.4	44.7	-	85.3
InstructBLIP [25]	188M	14.2B	-	55.5	-	33.4	<u>49.5</u>	16.7	78.9
ShareGPT4V [16]	13.4B	13.4B	81.0	-	64.8	-	-	-	-
LLaVa$^{1.5}$ [64]	13B	13.4B	80.0	-	63.3	<u>53.6</u>	-	24.7	85.9
LLaVa$^{1.6}$ [65]	13B	13.4B	-	46.3	<u>65.4</u>	-	-	-	86.3
LLaVA$^{1.5}$ (I-MoF) [84]	13B	13.6B	79.3	-	-	-	-	<u>31.3</u>	86.7
BRAVE	3B	10.3B	<u>82.5</u>	**66.0**	**66.3**	**54.2**	**52.7**	**42.0**	**87.6**

4.2 Visual question answering

Fine-tuning details. We use the same VQA mixture as in Sec. 2 and fine-tune both the MEQ-Former and the LM at 224×224 resolution. Vision encoders are kept frozen. This is followed by a high-resolution fine-tuning stage on VQAv2 and OKVQA training sets at 336×336 resolution, similar to previous work [19,56,64]. The resulting model is evaluated on VQAv2 and OKVQA benchmarks as well as zero-shot evaluated on GQA [48], VizWiz-QA [39], MMVP [84] and POPE [59], following [25,56,64,84]. To be able to compare to the methods that fine-tune on GQA training set, e.g. [6,61,64], we also perform an additional training stage on top of the mixture-finetuned model, and evaluate the model on the GQA evaluation set (test-dev). Please see the supplementary for the details.

Results. We summarize the VQA evaluations in Table 4. BRAVE achieves strong results for all the benchmarks in both fine-tuned and zero-shot evaluation scenarios. For OKVQA, GQA, VizWiz-QA, MMVP, and POPE benchmarks, BRAVE achieves the best results, and for VQAv2, it stays competitive with PaLI-17B:

- For MMVP, BRAVE significantly improves the performance over previous approaches, making it less susceptible to failure modes of CLIP-based encoders. This can be further seen from the qualitatives in Figures 1 and 3.
- For VizWiz-QA, BRAVE achieves the best results without using any scene text understanding or OCR data during pre-training, as opposed to [6,14,64].

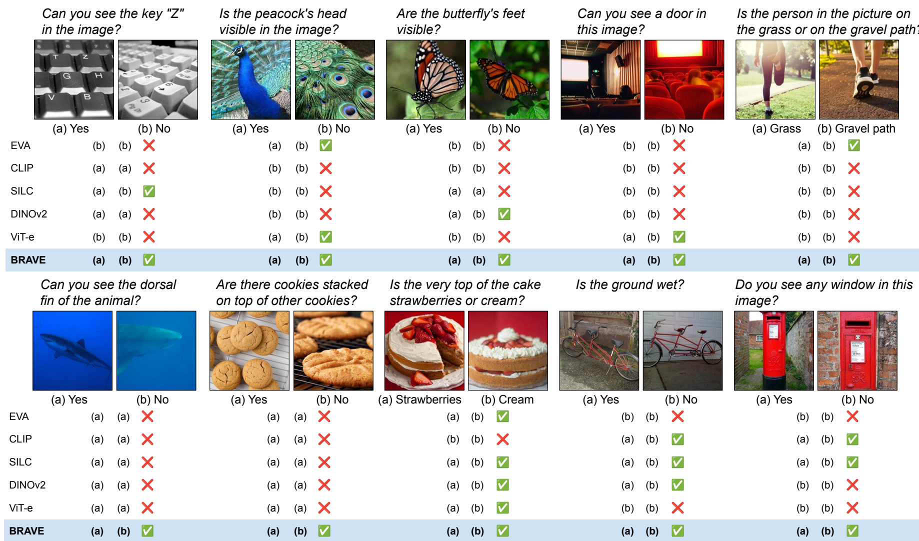

Fig. 3: Qualitative results. We compare predictions of `BRAVE` and the VLMs with different vision encoders, e.g. CLIP, on samples from the MMVP benchmark. Following [84], a model is considered correct only if it answers to both images in a pair correctly, i.e. if it can successfully differentiate between images with semantic differences. Note that the images in a pair are seen independently, i.e. neither of the images is provided as context for the other one. All encoders output some correct predictions, yet none of them performs consistently well on a broad range of inputs. `BRAVE` alleviates this by combining diverse visual features, leading to a more consistent performance. The quantitative difference is indeed stark: 42% for `BRAVE` vs 27.3% for best single encoder (Tables 2 and 4). See supplementary for more qualitative results.

- For GQA and OKVQA that requires different capabilities, i.e. spatial reasoning and open-world knowledge, `BRAVE` noticeably improves the performance.
- For POPE, `BRAVE` further reduces the visual hallucinations over dual encoder methods [61,84], confirming that using multiple encoders can be effective for visual grounding (Also see Sec. 5 for a discussion).
- For VQAv2, `BRAVE` is the second best after PaLI-17B, while staying more efficient in terms of training data, model size, and input resolution, similar to the COCO captioning result in Sec. 4.1. We expect `BRAVE` to benefit from such scaling efforts as well, which is a future direction to investigate.

4.3 Analysis of `BRAVE`

We study different aspects of `BRAVE` including an ablation study, contribution of vision encoders to the final performance, comparing `MEQ-Former` to an ensembling strategy, and the role of pre-training data. Due to the space limit, please see the supplementary for more qualitative and quantitative results.

Table 5: Ablation study of `BRAVE`. We start from the base pre-training setup explained in Sec. 3, denoted as A0. Each row corresponds to a particular deviation from A0, while keeping the rest fixed. For example, A5 evaluates the performance when high-resolution fine-tuning is disabled. N/A means the ablation is not applicable for that evaluation. See Sec. 4.3 for the discussions and supplementary for more results.

Ablation ID	Study Subject	Original	Change	COCO Captioning (Karpathy val)	VQAv2 (Karpathy val)	OKVQA (val)
A0	Base	-	-	147.0	81.8	65.7
A1	LM fine-tuning for VQA	Full fine-tuning	No fine-tuning LoRA (r=64) LoRA (r=128)	N/A N/A N/A	78.6 80.7 81.0	57.5 62.8 62.9
A2	Synthetic VQA data [12]	✓	✗	N/A	81.1	64.0
A3	Encoder dropout	✓	✗	145.3	81.3	66.0
A4	`MEQ-Former` text input	✓	✗	145.9	81.4	64.9
A5	High-res fine-tuning	✓	✗	145.2	79.6	65.0
A6	A3 + A5	✓	✗	144.0	79.0	64.5
A7	A4 + A5	✓	✗	145.1	78.3	63.4
A8	Language model	FlanT5-XL	FlanT5-L	142.5	79.9	65.5

Ablation study. We perform a comprehensive study in Table 5 where we ablate different design choices, e.g. training data and choice of language model. The ablations are labeled from A1 to A8, each showing a particular deviation from the original pre-training setup A0 (Sec. 3). We summarize the key findings:
(A1): LM fine-tuning for VQA significantly boosts the performance compared to frozen LM, similar to the observation in [64]. To minimize the performance loss, we perform a LoRA [46] fine-tuning by inserting LoRA layers to the LM. With a rank of 128, we compensated for 70% of the performance gap while using 10× fewer trainable parameters than full fine-tuning (324M vs 3B). This underscores the development of parameter-efficient fine-tuning methods as an important direction for more efficient VLMs.
(A2): Using synthetic VQA data [12] noticeably enhances our performance, while being much cheaper to collect than human annotated data. This can potentially be further improved by using more diverse templates [38, 83].
(A3,A4,A5): Encoder dropout during training improves the performance for COCO captioning and VQAv2 but slightly degrades for OKVQA. Using text prompt as an additional input to `MEQ-Former` helps with extracting better task-aligned features. Similarly, high-resolution fine-tuning stage gives a boost to the performance at the expense of processing more visual tokens.
(A6,A7): These ablations combine the impact of previously ablated design choices, further demonstrating their contribution to the final performance.
(A8): A stronger LM significantly helps for all the tasks, underscoring that scaling the VLM along the vision and language axes simultaneously have complementary benefits, and both are important.

Contribution of vision encoders. To understand the impact of vision encoders to the final performance of `BRAVE`, we perform additional evaluations:

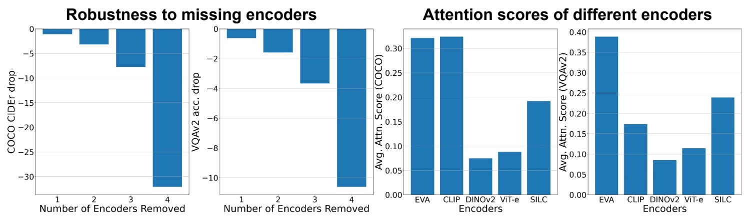

Fig. 4: Contribution of vision encoders to BRAVE. Left: We analyze the robustness of BRAVE when a subset of encoders are removed during evaluation. We report the average drop in CIDEr for COCO and accuracy for VQAv2. **Right:** We compute average attention scores for different vision encoders cross-attended by the MEQ-Former for COCO and VQAv2. See Sec. 4.3 for the discussions.

- In Fig. 4-left, we perform a robustness analysis by removing a subset of vision encoders during inference. This is performed for all the combinations and we report the average drop in performance for COCO captioning and VQAv2 tasks (Full results in supplementary). The results show that the performance of BRAVE degrades gracefully up to two encoder removals for both tasks, indicating some redundancy among the encoders. Because of this, even if some encoders are missing, BRAVE can make use of the remaining set of encoders to compensate. Beyond the two encoder removal, the drop in the performance is more severe, suggesting that some features are more unique and consequently harder to replace using the remaining set.
- In Fig. 4-right, we compute attention scores of features from different encoders that are cross-attended by the MEQ-Former queries. These are averaged for randomly sampled 1k samples from COCO and VQAv2 validation sets. The results show that the MEQ-Former can cross-attend, or "pay attention to", some encoders more than the others depending on the downstream task adaptively, further confirming the usefulness of having different biases.

5 Related work

Several works have focused on training VLMs to solve tasks requiring both vision and text understanding capabilities. Most of these works make use of pre-trained vision encoders [9,27,31,49,75,99] and LMs [6,22,24,44,76,85,103]. Architecture-wise, different approaches were developed to project visual features to the textual embedding space, e.g. Q-Former [25,56,106], P-Former [51], cross-attention [3,6,19,95], linear projector [14,15,66] and MLP [64,89]. Different objectives were also considered depending on the downstream tasks, e.g. image-text contrastive learning [50,73,75,94,101], auto-regressive text generation [17,19,23,87,88,91] or both [56–58,96] (see [102] for a survey). In contrast to ours, these works consider a single vision encoder which has inherent limitations [47,59,82,84] (Sec. 2).

Concurrent to our work, LLaVa-MoF [84] proposes combining CLIP [75] and DINOv2 [74] encoders using separate adapters. SPHINX [61] mixes CLIP, DINOv2, and Q-Former outputs with linear projection layers. Both of the methods input a concatenation of the extracted visual embeddings to the LM, which is not scalable to multiple encoders. We differ from these works by proposing a flexible and unified resampling mechanism that combines features from several vision encoders while staying efficient and improving the performance (Tables 3, 4).

Our work focuses on scaling the VLMs along the vision axis, unlike other works that explored scaling homogenously model size across the visual and/or language component [1,3,4,18,19,28,34,80,90], training data [6,10,13,49,50,60, 71,78,87,89] and number of tasks and modalities [5,7,40,43,67–69,72,98]. BRAVE can be combined with these efforts to further improve performance. Moreover, several studies investigated the visual aspect of VLMs by evaluating them against hallucinations and other robustness issues [32, 47, 59, 62, 63, 82, 104]. BRAVE improves the robustness to those issues by broadening visual capabilities of VLMs.

6 Conclusions and Limitations

We focus on broadening the visual encoding aspect of VLMs. Our evaluation of vision encoders revealed that there is no encoder achieving a consistent top performance across tasks and encoders with different biases can perform similarly. Motivated by this, we introduce BRAVE that combines diverse features from several vision encoders using a small number of trainable parameters. Extensive evaluations show that BRAVE achieves state-of-the-art performance and improves robustness of VLMs against out-of-distribution inputs and visual hallucinations. We discuss limitations of our work and future directions below:

Adaptive mechanisms. While MEQ-Former can resample from the useful features and discard the irrelevant ones, it still needs forward passes from all the encoders. A future direction could be to explore adaptive mechanisms to *preselect* which encoders to resample from, reducing the inference cost.

Improving sample efficiency. BRAVE uses around 100M image-text pairs during pre-training which is an order of magnitude less than recent methods [6,19, 87], yet can benefit from further lowering the sample complexity similar to [51].

More broad vision encoders. The set of encoders we investigated do not completely cover all types of visual biases, e.g. including the ones with strong 3D [29, 52] or semantic [54] priors could further improve the performance. Our results show that the current set is a good starting point to see the benefits of consolidation. Another interesting direction is to distill from the MEQ-Former-combined features back into a single encoder for a more unified solution.

Different modalities and multiple frames. While we focus on image and text modalities, BRAVE is a general method that could be potentially used to fuse more modalities, e.g. audio or 3D, or extended to incorporate multiple frames, e.g. for video understanding or in-context learning.

Biases of text-generative models. As we employ pre-trained LMs in our setup, the resulting VLMs are susceptible to the longstanding bias and fairness issues of these models, thus they should be used with caution.

Acknowledgements. We thank Diego Martin Arroyo, Ferjad Naeem, Xingyi Zhou, Yannick Strümpler and Yongqin Xian for their help with the project.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
4. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
5. Bachmann, R., Kar, O.F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., Zamir, A.: 4m-21: An any-to-any vision model for tens of tasks and modalities. arXiv preprint arXiv:2406.09406 (2024)
6. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
7. Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., Efros, A.A.: Sequential modeling enables scalable learning for large vision models. arXiv preprint arXiv:2312.00785 (2023)
8. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al.: A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023 (2023)
9. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization. In: International Conference on Machine Learning. pp. 1059–1071. PMLR (2021)
10. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
12. Changpinyo, S., Kukliansky, D., Szpektor, I., Chen, X., Ding, N., Soricut, R.: All you may need for vqa are image captions. arXiv preprint arXiv:2205.01883 (2022)
13. Changpinyo, S., Sharma, P.K., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3557–3567 (2021)
14. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
15. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)

16. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
17. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. In: International Conference on Learning Representations (2022)
18. Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., et al.: Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565 (2023)
19. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
20. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
21. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
22. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023)
23. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: International Conference on Machine Learning. pp. 1931–1942. PMLR (2021)
24. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
25. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=vvoWPYqZJA
26. Dehghani, M., Gritsenko, A., Arnab, A., Minderer, M., Tay, Y.: Scenic: A jax library for computer vision research and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21393–21398 (2022)
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
28. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
29. Eftekhar, A., Sax, A., Bachmann, R., Malik, J., Zamir, A.R.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10766–10776 (2021)
30. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)

31. Fang, Y., Wang, W., Xie, B., Sun, Q.S., Wu, L.Y., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: Exploring the limits of masked visual representation learning at scale. ArXiv **abs/2211.07636** (2022)
32. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
33. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural computation **4**(1), 1–58 (1992)
34. Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790 (2023)
35. Goodwin, C.: Seeing in depth. Social studies of science **25**(2), 237–274 (1995)
36. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
37. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y.: How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597 (2023)
38. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10867–10877 (2023)
39. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)
40. Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al.: Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:2309.03905 (2023)
41. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15979–15988 (2021)
42. Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., van Zee, M.: Flax: A neural network library and ecosystem for jax, 2020. URL http://github. com/google/flax **1** (2020)
43. Herzig, R., Abramovich, O., Ben Avraham, E., Arbelle, A., Karlinsky, L., Shamir, A., Darrell, T., Globerson, A.: Promptonomyvit: Multi-task prompt learning improves video transformers using synthetic scene data. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6803–6815 (2024)
44. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)
45. Howard, I.P., Rogers, B.J.: Binocular vision and stereopsis. Oxford University Press, USA (1995)
46. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: LoRA: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021), https://api.semanticscholar.org/CorpusID:235458009
47. Huang, W., Liu, H., Guo, M., Gong, N.Z.: Visual hallucinations of multi-modal large language models. arXiv preprint arXiv:2402.14683 (2024)

48. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
49. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below.
50. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
51. Jian, Y., Gao, C., Vosoughi, S.: Bootstrapping vision-language learning with decoupled language pre-training. Advances in Neural Information Processing Systems **36** (2024)
52. Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3d common corruptions and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18963–18974 (2022)
53. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
54. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
55. Korbar, B., Xian, Y., Tonioni, A., Zisserman, A., Tombari, F.: Text-conditioned resampler for long form video understanding. arXiv preprint arXiv:2312.11897 (2023)
56. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
57. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022)
58. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
59. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
60. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533 (2023)
61. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
62. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations (2023)
63. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)

64. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
65. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (2024)
66. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
67. Liu, S., Fan, L.J., Johns, E., Yu, Z., Xiao, C., Anandkumar, A.: Prismer: A vision-language model with an ensemble of experts. ArXiv **abs/2303.02506** (2023)
68. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. arXiv preprint arXiv:2312.17172 (2023)
69. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-IO: A unified model for vision, language, and multi-modal tasks. In: The Eleventh International Conference on Learning Representations (2023)
70. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
71. McKinzie, B., Gan, Z., Fauconnier, J.P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Weers, F., et al.: Mm1: Methods, analysis & insights from multimodal llm pre-training. arXiv preprint arXiv:2403.09611 (2024)
72. Mizrahi, D., Bachmann, R., Kar, O.F., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4M: Massively multimodal masked modeling. In: Advances in Neural Information Processing Systems (2023)
73. Naeem, M.F., Xian, Y., Zhai, X., Hoyer, L., Van Gool, L., Tombari, F.: Silc: Improving vision language pretraining with self-distillation. arXiv preprint arXiv:2310.13355 (2023)
74. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. ArXiv **abs/2304.07193** (2023)
75. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
76. Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
77. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
78. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)

79. Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih, G., Moy, L.: Chatgpt and other large language models are double-edged swords (2023)
80. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
81. Thorp, H.H.: Chatgpt is fun, but not an author (2023)
82. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
83. Tiong, A.M.H., Li, J., Li, B., Savarese, S., Hoi, S.C.: Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. arXiv preprint arXiv:2210.08773 (2022)
84. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. arXiv preprint arXiv:2401.06209 (2024)
85. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
86. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
87. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)
88. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022)
89. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)
90. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. ArXiv **abs/2208.10442** (2022)
91. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021)
92. Xu, J., Zhou, X., Yan, S., Gu, X., Arnab, A., Sun, C., Wang, X., Schmid, C.: Pixel Aligned Language Models. arXiv preprint arXiv: 2312.09237 (2023)
93. Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., Singh, S.B., Lee, S., Batra, D.: Evalai: Towards better evaluation systems for ai agents (2019)
94. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
95. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)

96. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022)
97. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. arXiv preprint arXiv:2305.06988 (2023)
98. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
99. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12104–12113 (2022)
100. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. arXiv preprint arXiv:2303.15343 (2023)
101. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133 (2022)
102. Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., Yu, D.: Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601 (2024)
103. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
104. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems **36** (2024)
105. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBoT: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (2022)
106. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)