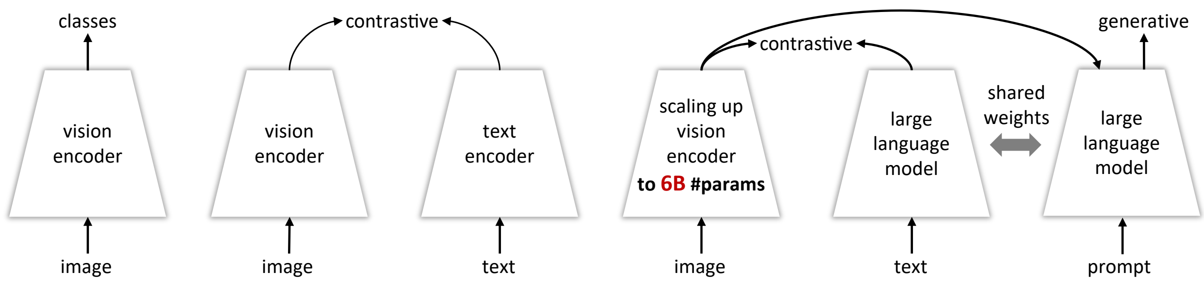
InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks

Zhe Chen[2,1†], Jiannan Wu[3,1†], Wenhai Wang[4,1], Weijie Su[6,1†], Guo Chen[2,1†], Sen Xing[5], Muyan Zhong[5], Qinglong Zhang[1], Xizhou Zhu[5,7,1], Lewei Lu[7,1], Bin Li[6], Ping Luo[3], Tong Lu[2], Yu Qiao[1], Jifeng Dai[5,1✉]
[1]OpenGVLab, Shanghai AI Laboratory [2]Nanjing University
[3]The University of Hong Kong [4]The Chinese University of Hong Kong [5]Tsinghua University
[6]University of Science and Technology of China [7]SenseTime Research

https://github.com/OpenGVLab/InternVL



(a) Supervised pre-training (b) Contrastive pre-training (c) InternVL: Scaling up vision encoder and aligning with LLM (ours)

Figure 1. **Comparisons of different vision and vision-language foundation models.** (a) indicates the traditional vision foundation model, *e.g.* ResNet [43] pre-trained on classification tasks. (b) represents the vision-language foundation models, *e.g.* CLIP [89] pre-trained on image-text pairs. (c) is our InternVL, which presents a workable way to align the large-scale vision foundation model (*i.e.*, InternViT-6B) with the large language model and is versatile for both contrastive and generative tasks.

Abstract

The exponential growth of large language models (LLMs) has opened up numerous possibilities for multi-modal AGI systems. However, the progress in vision and vision-language foundation models, which are also critical elements of multi-modal AGI, has not kept pace with LLMs. In this work, we design a large-scale vision-language foundation model (InternVL), which scales up the vision foundation model to 6 billion parameters and progressively aligns it with the LLM, using web-scale image-text data from various sources. This model can be broadly applied to and achieve state-of-the-art performance on 32 generic visual-linguistic benchmarks including visual perception tasks such as image-level or pixel-level recognition, vision-language tasks such as zero-shot image/video classification, zero-shot image/video-text retrieval, and link with LLMs to create multi-modal dialogue systems. It has powerful visual capabilities and can be a good alternative to the ViT-22B. We hope that our research could contribute to the development of multi-modal large models.

† This work is done when they are interns at Shanghai AI Laboratory; ✉ corresponding author (daijifeng@tsinghua.edu.cn)

1. Introduction

Large language models (LLMs) largely promote the development of artificial general intelligence (AGI) systems with their impressive capabilities in open-world language tasks, and their model scale and performance are still increasing at a fast pace. Vision large language models (VLLMs) [3, 5, 19, 21, 28, 69, 87, 113, 147], which leverage LLMs, have also achieved significant breakthroughs, enabling sophisticated vision-language dialogues and interactions. However, the progress of vision and vision-language foundation models, which are also crucial for VLLMs, has lagged behind the rapid growth of LLMs.

To bridge vision models with LLMs, existing VLLMs [5, 61, 100, 138, 147] commonly employ lightweight "glue" layers, such as QFormer [61] or linear projection [69], to align features of vision and language models. Such alignment contains several limitations: (1) *Disparity in parameter scales.* The large LLMs [38] now boosts up to 1000 billion parameters, while the widely-used vision encoders of VLLMs are still around one billion. This gap may lead to the under-use of LLM's capacity. (2) *Inconsistent representation.* Vision models, trained on pure-vision data or

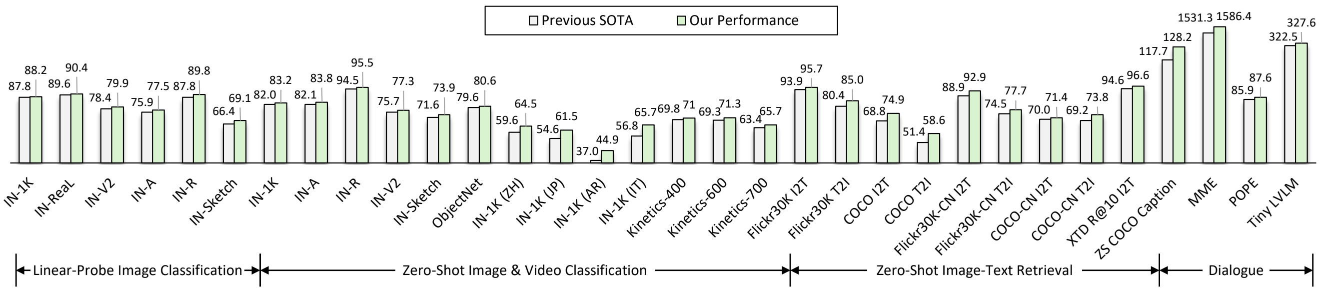


Figure 2. **Comparison results on various generic visual-linguistic tasks**, including image classification, video classification, image-text retrieval, image captioning, and multi-modal dialogue. The proposed InternVL achieves the best performance on all these tasks. Note that only the models trained on public data are included. "IN" is an abbreviation for ImageNet [31].

aligned with the BERT series [52, 54, 70], often exhibit representation inconsistencies with LLMs. (3) *Inefficient connection.* The "glue" layers are usually lightweight and randomly initialized, which may not capture the rich cross-modal interactions and dependencies that are crucial for multi-modal understanding and generation.

These limitations reveal a large gap in both parameter scale and feature representation ability between the vision encoder and the LLM. To bridge this gap, *our inspiration lies in elevating the vision encoder to align with the parameter scale of the LLM and subsequently harmonizing their representations.* However, the training of such large-scale models necessitates a vast amount of image-text data obtained from the Internet. The significant heterogeneity and quality variations within this data pose considerable challenges to the training process. To enhance the efficacy of the training, generative supervision is considered as a complementary approach to contrastive learning, as depicted in Figure 1. This strategy aims to provide additional guidance to the model during training. Yet, the suitability of low-quality data for generative training remains a concern. Besides, how to effectively represent the users' commands and align the representations between the vision encoder and LLM is another open question.

To address these issues, we formulate the *InternVL, a large-scale vision-language foundation model, which aligns the representation of the scaled-up vision encoder with the LLM and achieves state-of-the-art performance on various visual and visual-linguistic tasks.* As shown in Figure 1 (c), InternVL has three key designs: (1) *Parameter-balanced vision and language components*: It includes a vision encoder scaled up to 6 billion parameters and an LLM middleware with 8 billion parameters, where the middleware functions as a substantial "glue" layer to reorganize visual features. Unlike previous vision-only (Figure 1 (a)) or dual-tower (Figure 1 (b)) structures, our vision encoder and middleware offer flexible combinations for both contrastive and generative tasks. (2) *Consistent representations*: To maintain the consistency of representations between the vision encoder and LLM, we employ a pre-trained multilingual LLaMA-7B [26], to initialize the middleware and align the vision encoder with it. (3) *Progressive image-text alignment*: We leverage image-text data from diverse sources, ensuring training stability through a progressive alignment strategy. This strategy initiates contrastive learning on large-scale noisy data and subsequently transitions to generative learning on fine-grained data. This approach ensures a consistent enhancement of model performance and task scope.

These designs endow our model with several advantages: (1) *Versatile.* It functions as a standalone vision encoder for perception tasks, or collaborates with the language middleware for vision-language tasks and multi-modal dialogue systems. The language middleware bridges the gap between the vision encoder and the LLM decoder. (2) *Strong.* By leveraging the training strategy, large-scale parameters, and web-scale data, our model has a powerful representation that helps to achieve state-of-the-art results on various vision and vision-language tasks, as shown in Figure 2. (3) *LLM-friendly.* Due to the aligned feature space with LLMs, our model can smoothly integrate with existing LLMs, such as LLaMA series [106, 107], Vicuna [145], and InternLM [104]. These features distinguish our model from the previous approaches and establish a leading vision-language foundation model for various applications.

In summary, our contribution has three folds:

(1) We present a large-scale vision-language foundation model—InternVL, which aligns the large-scale vision encoder with LLMs from scratch for the first time. The model demonstrates strong performance on a wide range of generic visual-linguistic tasks, including visual perception tasks, vision-language tasks, and multi-modal dialogue.

(2) We introduce a progressive image-text alignment strategy for the efficient training of large-scale vision-language foundation models. This strategy maximizes the utilization of web-scale noisy image-text data for contrastive learning and fine-grained, high-quality data for generative learning.

(3) We extensively compare the proposed model with the current state-of-the-art vision foundation models and VLLMs. The results indicate that InternVL achieves

leading performance on a broad range of generic visual-linguistic tasks, including image classification (ImageNet), semantic segmentation (ADE20K), video classification (Kinetics), image-text retrieval (Flickr30K & COCO), video-text retrieval (MSR-VTT), and image captioning (COCO & Flickr30K & NoCaps). Meanwhile, it is also effective for multi-modal dialogue (MME & POPE & Tiny LVLM).

2. Related Work

2.1. Vision Foundation Models

The past decade has witnessed significant development in foundation models within the field of computer vision. Starting with the pioneering AlexNet [55], a variety of convolutional neural networks (CNNs) have emerged, continuously refreshing the ImageNet benchmark [27, 32, 43, 47, 49, 72, 114, 124]. In particular, the introduction of residual connections [43] effectively addressed the problem of vanishing gradients. This breakthrough led to an era of "big & deep" neural networks, signifying that, with adequate training and data, larger and deeper models can achieve better performance. In other words, scaling up matters.

In recent years, ViT [34] has opened up new possibilities for network architectures in the computer vision field. ViT and its variants [13, 23, 30, 71, 89, 111, 112, 125, 139, 140] have significantly increased their capacity and excelled in various important visual tasks. In the LLM era, these vision foundation models often connect with LLMs through some lightweight "glue" layers [60, 69, 147]. However, a gap exists as these models primarily derive from visual-only datasets like ImageNet [31] or JFT [134], or are aligned with the BERT series [52, 54, 70] using image-text pairs, lacking direct alignment with LLMs. Additionally, the prevalent vision models employed to connect with LLMs are still limited to around 1 billion parameters [37, 51], which also constrains the performance of VLLMs.

2.2. Large Language Models

Large language models (LLMs) have revolutionized the field of artificial intelligence, enabling natural language processing tasks that were previously thought exclusive to humans [1, 106, 118]. The emergence of GPT-3 [118] brought a significant leap in capabilities, particularly in few-shot and zero-shot learning, highlighting the immense potential of LLMs. This promise was further realized with the advancements of ChatGPT and GPT-4 [1]. The progress in the field has been further accelerated by the emergence of open-source LLMs, including the LLaMA series [106, 107], Vicuna [145], InternLM [104], MOSS [101], ChatGLM [36], Qwen [4], Baichuan [6], and Falcon [86], among others [26, 103, 119]. However, in real scenarios, interactions are not limited to natural language. The vision modality can bring additional information, which means more possibilities. Therefore, exploring how to utilize the excellent capabilities of LLMs for multi-modal interactions is poised to become the next research trend.

2.3. Vision Large Language Models

Recent advancements have seen the creation of vision large language models (VLLMs) [3, 21, 56, 59, 62, 66, 100, 121, 128, 130, 136, 138, 141, 142, 148], which aim to enhance language models with the capability to process and interpret visual information. Flamingo [3] uses the visual and language inputs as prompts and shows remarkable few-shot performance for visual question answering. Subsequently, GPT-4 [1], LLaVA series [68, 69, 76] and MiniGPT-4 [147] have brought in visual instruction tuning, to improve the instruction-following ability of VLLMs. Concurrently, models such as VisionLLM [113], KOSMOS-2 [87], and Qwen-VL *et al*. [5, 19, 115] have improved VLLMs with visual grounding capabilities, facilitating tasks such as region description and localization. Many API-based methods [73, 74, 95, 102, 120, 127, 129] have also attempted to integrate vision APIs with LLMs for solving vision-centric tasks. Additionally, PaLM-E [35] and EmbodiedGPT [83] represent advanced efforts in adapting VLLMs for embodied applications, significantly expanding their potential applications. These works showcase that VLLMs have achieved significant breakthroughs. However, the progress of vision and vision-language foundation models, equally essential for VLLMs, has not kept pace.

3. Proposed Method

3.1. Overall Architecture

As depicted in Figure 3, unlike traditional vision-only backbones [43, 71, 114] and dual-encoder models [51, 89, 99], the proposed InternVL is designed with a vision encoder InternViT-6B and a language middleware QLLaMA. Specifically, InternViT-6B is a vision transformer with 6 billion parameters, customized to achieve a favorable trade-off between performance and efficiency. QLLaMA is a language middleware with 8 billion parameters, initialized with a pre-trained multilingual LLaMA-7B [26]. It could provide robust multilingual representation for image-text contrastive learning, or serve as a bridge to connect the vision encoder and the off-the-shelf LLM decoder.

To align the two large-scale components with substantial gaps in modalities and structures, we introduce a progressive alignment training strategy. The training strategy is conducted progressively, beginning with contrastive learning on large-scale noisy data, and gradually moving towards generative learning on exquisite and high-quality data. In this way, we ensure the effective organization and full utilization of web-scale image-text data from a variety of sources. Then, equipped with the aligned vision encoder

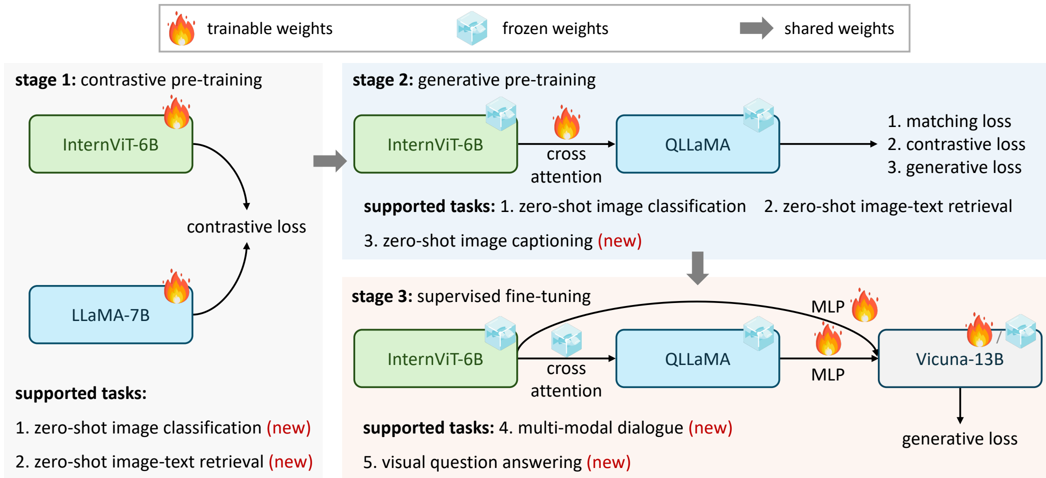


Figure 3. **The training strategy of the proposed InternVL model.** It consists of three progressive stages, including vision-language contrastive training, vision-language generative training, and supervised fine-tuning. These stages effectively leverage public data from diverse sources, ranging from noisy image-text pairs on the web to high-quality caption, VQA, and multi-modal dialogue datasets.

name	width	depth	MLP	#heads	#param (M)
ViT-G [134]	1664	48	8192	16	1843
ViT-e [21]	1792	56	15360	16	3926
EVA-02-ViT-E [99]	1792	64	15360	16	4400
ViT-6.5B [98]	4096	32	16384	32	6440
ViT-22B [30]	6144	48	24576	48	21743
InternViT-6B (ours)	3200	48	12800	25	5903

Table 1. **Architecture details of the InternViT-6B model.**

and language middleware, our model functions like a Swiss Army knife. It boasts a flexible composition that can be adapted for a wide array of generic visual-linguistic tasks. These tasks range from visual perception and image/video-text retrieval to image captioning, visual question answering, and multi-modal dialogue, among others.

3.2. Model Design

Large-Scale Vision Encoder: InternViT-6B. We implement the vision encoder of InternVL with vanilla vision transformer (ViT) [34]. To match the scale of LLMs, we scale up the vision encoder to 6 billion parameters, resulting in the InternViT-6B model. To obtain a good trade-off between accuracy, speed, and stability, we conduct a hyperparameter search for InternViT-6B. We vary the model depth within $\{32, 48, 64, 80\}$, the head dimension within $\{64, 128\}$, and the MLP ratio within $\{4, 8\}$. The model width and the head number are calculated based on the given model scale and other hyperparameters.

We employ contrastive learning on a 100M subset of the LAION-en dataset [91] to measure the accuracy, speed, and stability of InternViT-6B variants with different configurations. We report the following findings: (1) *Speed.* For different model settings, when computation is not saturated, the models with smaller depths exhibit faster speed per image. However, as the GPU computation is fully utilized, the speed difference becomes negligible; (2) *Accuracy.* With the same number of parameters, the depth, head dimension, and MLP ratio have little impact on the performance. Based on these findings, we identified the most stable configuration for our final model, as shown in Table 1.

Language Middleware: QLLaMA. The language middleware QLLaMA is proposed to align visual and linguistic features. As shown in Figure 3, QLLaMA is developed based on the pre-trained multilingual LLaMA [26], and newly added 96 learnable queries and cross-attention layers (1 billion parameters) that are randomly initialized. This manner allows QLLaMA to smoothly integrate visual elements into the language model, thereby enhancing the coherence and effectiveness of the combined features.

Compared to recently popular approaches [61, 69] that use lightweight "glue" layers, such as QFormer [61] and linear layers [69] to connect vision encoder and LLMs, our method has three advantages: (1) By initializing with the pre-trained weights of [26], QLLaMA can transform image tokens generated by InternViT-6B into the representation that is aligned with the LLMs; (2) QLLaMA has 8 billion parameters for vision-language alignment, which are 42 times larger than the QFormer. Therefore, even with a frozen LLM decoder, InternVL can achieve promising performance on multi-modal dialogue tasks. (3) It can also be applied to contrastive learning, providing a powerful text representation for image-text alignment tasks, such as zero-shot image classification and image-text retrieval.

"Swiss Army Knife" Model: InternVL. By flexibly combining the vision encoder and the language middleware, InternVL can support various vision or vision-language tasks. (1) *For visual perception tasks*, the vision encoder of InternVL, *i.e.* InternViT-6B, can be used as the backbone for vision tasks. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, our

dataset	characteristics		stage 1		stage 2	
	language	original	cleaned	remain	cleaned	remain
LAION-en [91]	English	2.3B	1.94B	84.3%	91M	4.0%
LAION-COCO [92]		663M	550M	83.0%	550M	83.0%
COYO [12]		747M	535M	71.6%	200M	26.8%
CC12M [18]		12.4M	11.1M	89.5%	11.1M	89.5%
CC3M [94]		3.0M	2.6M	86.7%	2.6M	86.7%
SBU [85]		1.0M	1.0M	100%	1.0M	100%
Wukong [41]	Chinese	100M	69.4M	69.4%	69.4M	69.4%
LAION-multi [91]	Multi	2.2B	1.87B	85.0%	100M	4.5%
Total	Multi	6.03B	4.98B	82.6%	1.03B	17.0%

Table 2. **Details of the training data for InternVL in stage 1 and stage 2.** Among them, LAION-en [91], LAION-multi [91], COYO [12], and Wukong [41] are web-scale image-text pairs data. LAION-COCO [92] is a synthetic dataset with high-quality captions from LAION-en. CC12M [18], CC3M [94], SBU [85] are academic caption datasets. "Multi" means multilingual.

task	#samples	dataset
Captioning	588K	COCO Caption [20], TextCaps [96]
VQA	1.1M	VQAv2 [40], OKVQA [79], A-OKVQA [93], IconQA [75], AI2D [53], GQA [48]
OCR	294K	OCR-VQA [82], ChartQA [80], DocVQA [25], ST-VQA [11], EST-VQA [116], InfoVQA [81], LLaVAR [143]
Grounding	323K	RefCOCO/+/g [78, 132], Toloka [108]
Grounded Cap.	284K	RefCOCO/+/g [78, 132]
Conversation	1.4M	LLaVA-150K [69], SVIT [144], VisDial [29], LRV-Instruction [67], LLaVA-Mix-665K [68]

Table 3. **Details of the training data for InternVL in stage 3.** We collect a wide range of high-quality instruction data, totaling approximately 4 million samples. For a fair comparison, we only use the training split of these datasets.

model can generate a feature map $F \in \mathbb{R}^{H/14 \times W/14 \times D}$ for dense prediction tasks, or work with global average pooling and linear projection to make image classification.

(2) *For contrastive tasks*, we introduce two inference modes: **InternVL-C** and **InternVL-G**, using the vision encoder InternViT or the combination of InternViT and QLLaMA to encode visual features. Specifically, we apply attention pooling to the visual features of InternViT or the query features of QLLaMA, to calculate the global visual feature I_f. Besides, we encode text as T_f by extracting the feature from the `[EOS]` token of QLLaMA. By computing similarity scores between I_f and T_f, we support various contrastive tasks such as image-text retrieval.

(3) *For generative tasks*, unlike QFormer [60], QLLaMA inherently has promising image captioning abilities thanks to its scaled-up parameters. The queries of QLLaMA reorganize the visual representations from InternViT-6B and play as the prefix texts for QLLaMA. The subsequent text tokens are generated one by one sequentially.

(4) *For multi-modal dialogue*, we introduce **InternVL-Chat**, leveraging InternVL as the visual component to connect with LLMs. For this purpose, we have two distinct configurations. One option is to employ the InternViT-6B independently, akin to the approach in LLaVA-1.5 [68]. The alternative is to employ the complete InternVL model concurrently, as illustrated in Figure 3.

3.3. Alignment Strategy

As shown in Figure 3, the training of InternVL consists of three progressive stages. These stages effectively leverage public data from diverse sources, ranging from noisy image-text pairs on the web to high-quality caption, VQA, and multi-modal dialogue datasets.

Vision-Language Contrastive Training. In the first stage, we conduct contrastive learning to align InternViT-6B with a multilingual LLaMA-7B [26] on web-scale, noisy image-text pairs. The data are all publicly available and comprise multilingual content, including LAION-en [91], LAION-multi [91], LAION-COCO [92], COYO [12], Wukong [41], etc. We use the combination of these datasets and filter out some extremely low-quality data to train our model. As summarized in Table 2, the original dataset contains 6.03 billion image-text pairs, and 4.98 billion remains after cleaning. More details about data preparation will be provided in the supplementary materials.

During training, we adopt the LLaMA-7B to encode the text as T_f, and use InternViT-6B to extract the visual feature I_f. Following the objective function of CLIP [89], we minimize a symmetric cross-entropy loss on the similarity scores of image-text pairs in a batch. This stage allows InternVL to excel on contrastive tasks like zero-shot image classification and image-text retrieval, and the vision encoder of this stage can also perform well on visual perception tasks.

Vision-Language Generative Training. In the second stage of training, we connect InternViT-6B with QLLaMA and adopt a generative training strategy. Specifically, QLLaMA inherits the weights of LLaMA-7B in the first stage. We keep both InternViT-6B and QLLaMA frozen and only train the newly added learnable queries and cross-attention layers with filtered, high-quality data. Table 2 summarizes the datasets for the second stage. It can be seen that we further filtered out data with low-quality captions, reducing it from 4.98 billion in the first stage to 1.03 billion.

Following the loss function of BLIP-2 [61], the loss in this stage is computed as the sum of three components: image-text contrastive (ITC) loss, image-text matching (ITM) loss, and image-grounded text generation (ITG) loss. This enables the queries to extract powerful visual representations, and further align feature space with LLMs, attributable to the effective training objectives and the utilization of our large-scale, LLM-initialized QLLaMA.

Supervised Fine-tuning. To demonstrate the benefits of InternVL in creating multi-modal dialogue systems, we connect it with an off-the-shelf LLM decoder (*e.g.*, Vicuna [145] or InternLM [104]) through an MLP layer, and conduct supervised fine-tuning (SFT). As detailed in Table 3, we collect a wide range of high-quality instruction data, totaling approximately 4 million samples. For non-dialogue datasets, we follow the format described in [68] for conversion. Owing to the similar feature space of QLLaMA

method	#param	IN-1K	IN-ReaL	IN-V2	IN-A	IN-R	IN-Ske	avg.
OpenCLIP-H [51]	0.6B	84.4	88.4	75.5	–	–	–	–
OpenCLIP-G [51]	1.8B	86.2	89.4	77.2	63.8	87.8	66.4	78.5
DINOv2-g [84]	1.1B	86.5	89.6	78.4	75.9	78.8	62.5	78.6
EVA-01-CLIP-g [37]	1.1B	86.5	89.3	77.4	70.5	87.7	63.1	79.1
MAWS-ViT-6.5B [98]	6.5B	87.8	–	–	–	–	–	–
ViT-22B* [30]	21.7B	89.5	90.9	83.2	83.8	87.4	–	–
InternViT-6B (ours)	5.9B	**88.2**	**90.4**	**79.9**	**77.5**	**89.8**	**69.1**	**82.5**

Table 4. **Linear evaluation on image classification.** We report the top-1 accuracy on ImageNet-1K [31] and its variants [9, 45, 46, 90, 109]. *ViT-22B [30] uses the private JFT-3B dataset [134].

method	#param	crop size	1/16	1/8	1/4	1/2	1
ViT-L [105]	0.3B	504^2	36.1	41.3	45.6	48.4	51.9
ViT-G [134]	1.8B	504^2	42.4	47.0	50.2	52.4	55.6
ViT-22B [30]	21.7B	504^2	44.7	47.2	50.6	52.5	54.9
InternViT-6B (ours)	5.9B	504^2	**46.5**	**50.0**	**53.3**	**55.8**	**57.2**

(a) Few-shot semantic segmentation with limited training data. Following ViT-22B [30], we fine-tune the InternViT-6B with a linear classifier.

method	decoder	#param (train/total)	crop size	mIoU
OpenCLIP-G$_{\text{frozen}}$ [51]	Linear	0.3M / 1.8B	512^2	39.3
ViT-22B$_{\text{frozen}}$ [30]	Linear	0.9M / 21.7B	504^2	34.6
InternViT-6B$_{\text{frozen}}$ (ours)	Linear	0.5M / 5.9B	504^2	**47.2**
ViT-22B$_{\text{frozen}}$ [30]	UperNet	0.8B / 22.5B	504^2	52.7
InternViT-6B$_{\text{frozen}}$ (ours)	UperNet	0.4B / 6.3B	504^2	**54.9**
ViT-22B [30]	UperNet	22.5B / 22.5B	504^2	55.3
InternViT-6B (ours)	UperNet	6.3B / 6.3B	504^2	**58.9**

(b) Semantic segmentation performance in three different settings, from top to bottom: linear probing, head tuning, and full-parameter tuning.

Table 5. **Semantic segmentation on ADE20K.** Results show that InternViT-6B has better pixel-level perceptual capacity.

and LLMs, we can achieve robust performance even when freezing the LLM, choosing to train just the MLP layer or both the MLP layer and QLLaMA. This approach not only expedites the SFT process but also maintains the original language capabilities of the LLMs.

4. Experiments

4.1. Implementation Details

Stage 1. In this stage, the image encoder InternViT-6B is randomly initialized [7], and the text encoder LLaMA-7B is initialized with the pre-trained weights from [26]. All parameters are fully trainable.

Stage 2. In this stage, InternViT-6B and QLLaMA inherit their weights from the first stage, while the new learnable queries and cross-attention layers in QLLaMA are randomly initialized. We keep both InternViT-6B and QLLaMA frozen and only train the new parameters.

Stage 3. At this stage, we have two different configurations. One is to use InternViT-6B separately, similar to LLaVA-1.5 [68]. The other is to use the entire InternVL model simultaneously, as shown in Figure 3. More details will be provided in the supplementary materials.

4.2. Visual Perception Benchmarks

First of all, we validate the visual perception capabilities of InternViT-6B, the most core component of InternVL.

Transfer to Image Classification. We evaluate the quality of visual representation produced by InternViT-6B using the ImageNet-1K [31] dataset. Following common practices [30, 44, 84], we adopt the linear probing evaluation, *i.e.* training a linear classifier while keeping the backbone frozen. In addition to the ImageNet-1K validation set, we also report performance metrics on several ImageNet variants [9, 45, 46, 90, 109], to benchmark the domain generalization capability. As shown in Table 4, InternViT-6B achieves a very significant improvement over previous state-of-the-art methods [37, 51, 84] on linear probing. To our knowledge, this represents the currently best linear evaluation results without the JFT dataset [134].

Transfer to Semantic Segmentation. To investigate the pixel-level perceptual capacity of InternViT-6B, we conduct extensive experiments of semantic segmentation on the ADE20K [146] dataset. Following ViT-22B [30], we begin with few-shot learning experiments, *i.e.* fine-tuning the backbone with a linear head on a limited dataset. As indicated in Table 5a, InternViT-6B consistently outperforms ViT-22B across five experiments with varying proportions of training data. Additionally, Table 5b presents our further verification in three distinct settings, including linear probing, head tuning [122], and full-parameter tuning. Notably, in the case of linear probing, InternViT-6B attains 47.2 mIoU, a substantial +12.6 mIoU improvement over ViT-22B. These results underscore the strong out-of-the-box pixel-level perceptual capacity of our InternViT-6B.

4.3. Vision-Language Benchmarks

In this section, we evaluate the inherent capabilities of InternVL on various vision-language tasks.

Zero-Shot Image Classification. We conduct thorough validation of the zero-shot image classification capability of InternVL-C. As depicted in Table 6a, InternVL-C attains leading performance on various ImageNet variants [31, 45, 46, 90, 109] and ObjectNet [8]. Compared to EVA-02-CLIP-E+ [99], it exhibits stronger robustness to distribution shift, manifesting in a more consistent accuracy across ImageNet variants. Additionally, as shown in Table 6b, our model showcases robust multilingual capabilities, outperforming competing models [14, 24, 51, 126] on the multilingual ImageNet-1K benchmark.

Zero-Shot Video Classification. Following previous methods [89, 99, 117], we report the top-1 accuracy and the mean of top-1 and top-5 accuracy on Kinetics-400/600/700 [15–17]. As shown in Table 8, when sampling only a single center frame in each video, our method achieves an average accuracy of 71.0%, 71.3%, and 65.7% on the three datasets, surpassing EVA-02-CLIP-E+ [99] by +1.2, +2.0, and +2.3 points, respectively. Additionally, when uniformly sampling 8 frames in each video, InternVL-C is even better than ViCLIP [117] that trained using web-scale video data.

method	IN-1K	IN-A	IN-R	IN-V2	IN-Sketch	ObjectNet	Δ↓	avg.
OpenCLIP-g [51]	78.5	60.8	90.2	71.7	67.5	69.2	5.5	73.0
OpenAI CLIP-L+ [89]	76.6	77.5	89.0	70.9	61.0	72.0	2.1	74.5
EVA-01-CLIP-g [99]	78.5	73.6	92.5	71.5	67.3	72.3	2.5	76.0
OpenCLIP-G [51]	80.1	69.3	92.1	73.6	68.9	73.0	3.9	76.2
EVA-01-CLIP-g+ [99]	79.3	74.1	92.5	72.1	68.1	75.3	2.4	76.9
MAWS-ViT-2B [98]	81.9	–	–	–	–	–	–	–
EVA-02-CLIP-E+ [99]	82.0	82.1	94.5	75.7	71.6	79.6	1.1	80.9
CoCa* [131]	86.3	90.2	96.5	80.7	77.6	82.7	0.6	85.7
LiT-22B* [30, 135]	85.9	90.1	96.0	80.9	–	87.6	–	–
InternVL-C (ours)	**83.2**	**83.8**	**95.5**	**77.3**	**73.9**	**80.6**	**0.8**	**82.4**

(a) ImageNet variants [31, 45, 46, 90, 109] and ObjectNet [8].

method	EN	ZH	JP	AR	IT	avg.
M-CLIP [14]	–	–	–	–	20.2	–
CLIP-Italian [10]	–	–	–	–	22.1	–
Japanese-CLIP-ViT-B [77]	–	–	54.6	–	–	–
Taiyi-CLIP-ViT-H [137]	–	54.4	–	–	–	–
WuKong-ViT-L-G [41]	–	57.5	–	–	–	–
CN-CLIP-ViT-H [126]	–	59.6	–	–	–	–
AltCLIP-ViT-L [24]	74.5	59.6	–	–	–	–
EVA-02-CLIP-E+ [99]	82.0	3.6	5.0	0.2	41.2	–
OpenCLIP-XLM-R-H [51]	77.0	55.7	53.1	37.0	56.8	55.9
InternVL-C (ours)	**83.2**	**64.5**	**61.5**	**44.9**	**65.7**	**64.0**

(b) Multilingual ImageNet-1K [31, 57].

Table 6. **Comparison of zero-shot image classification performance.** "Δ↓": The gap between the averaged top-1 accuracy and the IN-1K top-1 accuracy. *CoCa [131] and LiT-22B [30] use the private JFT-3B dataset [134] during training. Multilingual evaluation involves 5 languages, including English (EN), Chinese (ZH), Japanese (JP), Arabic (AR), and Italian (IT).

method	multi-lingual	Flickr30K (English, 1K test set) [88]						COCO (English, 5K test set) [20]						avg.
		Image → Text			Text → Image			Image → Text			Text → Image			
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
Florence [133]	×	90.9	99.1	–	76.7	93.6	–	64.7	85.9	–	47.2	71.4	–	–
ONE-PEACE [110]	×	90.9	98.8	99.8	77.2	93.5	96.2	64.7	86.0	91.9	48.0	71.5	79.6	83.2
OpenCLIP-g [51]	×	91.4	99.2	99.6	77.7	94.1	96.9	66.4	86.0	91.8	48.8	73.3	81.5	83.9
EVA-01-CLIP-g+ [99]	×	91.6	99.3	99.8	78.9	94.5	96.9	68.2	87.5	92.5	50.3	74.0	82.1	84.6
CoCa [131]	×	92.5	99.5	99.9	80.4	95.7	97.7	66.3	86.2	91.8	51.2	74.2	82.0	84.8
OpenCLIP-G [51]	×	92.9	99.3	99.8	79.5	95.0	97.1	67.3	86.9	92.6	51.4	74.9	83.0	85.0
EVA-02-CLIP-E+ [99]	×	93.9	99.4	99.8	78.8	94.2	96.8	68.8	87.8	92.8	51.1	75.0	82.7	85.1
BLIP-2† [61]	×	97.6	100.0	100.0	89.7	98.1	98.9	–	–	–	–	–	–	–
InternVL-C (ours)	✓	94.7	99.6	99.9	81.7	96.0	98.2	70.6	89.0	93.5	54.1	77.3	84.6	86.6
InternVL-G (ours)	✓	**95.7**	**99.7**	**99.9**	**85.0**	**97.0**	**98.6**	**74.9**	**91.3**	**95.2**	**58.6**	**81.3**	**88.0**	**88.8**
method		Flickr30K-CN (Chinese, 1K test set) [58]						COCO-CN (Chinese, 1K test set) [63]						avg.
WuKong-ViT-L [41]	×	76.1	94.8	97.5	51.7	78.9	86.3	55.2	81.0	90.6	53.4	80.2	90.1	78.0
R2D2-ViT-L [123]	×	77.6	96.7	98.9	60.9	86.8	92.7	63.3	89.3	95.7	56.4	85.0	93.1	83.0
Taiyi-CLIP-ViT-H [137]	×	–	–	–	–	–	–	–	–	–	60.0	84.0	93.3	–
AltCLIP-ViT-H [24]	✓	88.9	98.5	99.5	74.5	92.0	95.5	–	–	–	–	–	–	–
CN-CLIP-ViT-H [126]	×	81.6	97.5	98.8	71.2	91.4	95.5	63.0	86.6	92.9	69.2	89.9	96.1	86.1
OpenCLIP-XLM-R-H [51]	✓	86.1	97.5	99.2	71.0	90.5	94.9	70.0	91.5	97.0	66.1	90.8	96.0	87.6
InternVL-C (ours)	✓	90.3	98.8	99.7	75.1	92.9	96.4	68.8	92.0	96.7	68.9	91.9	96.5	89.0
InternVL-G (ours)	✓	**92.9**	**99.4**	**99.8**	**77.7**	**94.8**	**97.3**	**71.4**	**93.9**	**97.7**	**73.8**	**94.4**	**98.1**	**90.9**

Table 7. **Comparison of zero-shot image-text retrieval performance.** We evaluate the retrieval capability in English using the Flickr30K [88] and COCO [20], as well as in Chinese using Flickr30K-CN [58] and COCO-CN [63]. †BLIP-2 [61] is finetuned on COCO and zero-shot transferred to Flickr30K, contributing to the enhanced zero-shot performance on Flickr30K.

method	#F	K400 [15]		K600 [16]		K700 [17]	
		top-1	avg.	top-1	avg.	top-1	avg.
OpenCLIP-g [51]	1	–	63.9	–	64.1	–	56.9
OpenCLIP-G [51]	1	–	65.9	–	66.1	–	59.2
EVA-01-CLIP-g+ [99]	1	–	66.7	–	67.0	–	60.9
EVA-02-CLIP-E+ [99]	1	–	69.8	–	69.3	–	63.4
InternVL-C (ours)	1	–	**71.0**	–	**71.3**	–	**65.7**
ViCLIP [117]	8	64.8	75.7	62.2	73.5	54.3	66.4
InternVL-C (ours)	8	**69.1**	**79.4**	**68.9**	**78.8**	**60.6**	**71.5**

Table 8. **Comparison of zero-shot video classification results on Kinetics 400/600/700.** We report the top-1 accuracy and the mean of top-1 and top-5 accuracy. "#F" denotes the number of frames.

Zero-Shot Image-Text Retrieval. InternVL exhibits a powerful multilingual image-text retrieval capability. In Table 7, we evaluate these capabilities in English using the Flickr30K [88] and COCO [20] datasets, as well as in Chinese using the Flickr30K-CN [58] and COCO-CN [63]. In summary, InternVL-C achieves state-of-the-art performance across most retrieval metrics, and with the second stage of pre-training, InternVL-G further enhances zero-shot image-text retrieval performance. These improvements indicate a more effective alignment between visual and linguistic features by using the QLLaMA.

Zero-Shot Image Captioning. Benefiting from vision-language generative training on a vast collection of high-quality image-text pairs, our QLLaMA possesses promising capability in zero-shot image captioning. As shown in Table 10, QLLaMA surpasses other models in zero-shot performance on the COCO Karpathy test set [20]. When InternVL is linked with an LLM (*e.g.*, Vicuna-7B/13B [145]) and subjected to SFT, a notable enhancement in zero-shot performance is observed for both Flickr30K [88] and NoCaps [2] datasets, as shown in Table 9.

4.4. Multi-Modal Dialogue Benchmarks

Beyond the traditional multi-modal tasks, the emergence of ChatGPT [1] has led to a growing focus on evaluating the performance of multi-modal models in real usage scenarios, specifically within the realm of multi-modal dialogue. We conducted testing of InternVL-Chat models on two prominent multi-modal dialogue benchmarks, including MME [39] and POPE [65]. As shown in Table 9, it clearly demonstrates that our models exhibit superior performance compared with previous methods.

method	visual encoder	glue layer	LLM	Res.	PT	SFT	train. param	image captioning			visual question answering				dialogue	
								COCO	Flickr	NoCaps	VQA^{v2}	GQA	VizWiz	VQA^{T}	MME	POPE
InstructBLIP [28]	EVA-g	QFormer	Vicuna-7B	224	129M	1.2M	188M	–	82.4	123.1	–	49.2	34.5	50.1	–	–
BLIP-2 [61]	EVA-g	QFormer	Vicuna-13B	224	129M	–	188M	–	71.6	103.9	41.0	41.0	19.6	42.5	1293.8	85.3
InstructBLIP [28]	EVA-g	QFormer	Vicuna-13B	224	129M	1.2M	188M	–	82.8	121.9	–	49.5	33.4	50.7	1212.8	78.9
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-7B	224	1.0B	4.0M	64M	141.4^{*}	89.7	120.5	72.3^{*}	57.7^{*}	44.5	42.1	1298.5	85.2
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-13B	224	1.0B	4.0M	90M	142.4^{*}	89.9	123.1	71.7^{*}	59.5^{*}	54.0	49.1	1317.2	85.4
Shikra [19]	CLIP-L	Linear	Vicuna-13B	224	600K	5.5M	7B	117.5^{*}	73.9	–	77.4^{*}	–	–	–	–	–
IDEFICS-80B [50]	CLIP-H	Cross-Attn	LLaMA-65B	224	1.6B	–	15B	91.8^{*}	53.7	65.0	60.0	45.2	36.0	30.9	–	–
Qwen-VL [5]	CLIP-G	VL-Adapter	Qwen-7B	448	$1.4B^{\dagger}$	$50M^{\dagger}$	9.6B	–	85.8	121.4	78.8^{*}	59.3^{*}	35.2	63.8	–	–
Qwen-VL-Chat [5]	CLIP-G	VL-Adapter	Qwen-7B	448	$1.4B^{\dagger}$	$50M^{\dagger}$	9.6B	–	81.0	120.2	78.2^{*}	57.5^{*}	38.9	**61.5**	1487.5	–
LLaVA-1.5 [68]	$CLIP-L_{336}$	MLP	Vicuna-7B	336	558K	665K	7B	–	–	–	78.5^{*}	62.0^{*}	50.0	58.2	1510.7	85.9
LLaVA-1.5 [68]	$CLIP-L_{336}$	MLP	Vicuna-13B	336	558K	665K	13B	–	–	–	80.0^{*}	63.3^{*}	53.6	61.3	1531.3	85.9
InternVL-Chat (ours)	IViT-6B	MLP	Vicuna-7B	336	558K	665K	7B	–	–	–	79.3^{*}	62.9^{*}	52.5	57.0	1525.1	86.4
InternVL-Chat (ours)	IViT-6B	MLP	Vicuna-13B	336	558K	665K	13B	–	–	–	80.2^{*}	63.9^{*}	54.6	58.7	1546.9	87.1
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-13B	336	1.0B	4.0M	13B	**146.2^{*}**	**92.2**	**126.2**	**81.2^{*}**	**66.6^{*}**	**58.5**	**61.5**	**1586.4**	**87.6**

Table 9. **Comparison with SoTA methods on 9 benchmarks.** Image captioning datasets include: COCO Karpathy test [20], Flickr30K Karpathy test [88], NoCaps val [2]. VQA datasets include: VQAv2 test-dev [40], GQA test-balanced [48], VizWiz test-dev [42], and TextVQA val [97]. *The training annotations of the datasets are observed during training. "IViT-6B" represents our InternViT-6B.

method	glue layer	LLM	COCO	Flickr30K	NoCaps
Flamingo-80B [3]	Cross-Attn	Chinchilla-70B	84.3	67.2	–
KOSMOS-2 [87]	Linear	KOSMOS-1	–	66.7	–
PaLI-X-55B [22]	Linear	UL2-32B	–	–	**126.3**
BLIP-2 [61]	QFormer	Vicuna-13B	–	71.6	103.9
InstructBLIP [28]	QFormer	Vicuna-13B	–	82.8	121.9
Shikra-13B [19]	Linear	Vicuna-13B	–	73.9	–
ASM [115]	QFormer	Husky-7B	–	**88.0**	116.9
Qwen-VL [5]	VL-Adapter	Qwen-7B	–	85.8	121.4
Emu-I [100]	QFormer	LLaMA-13B	117.7	–	–
DreamLLM [33]	Linear	Vicuna-7B	115.4	–	–
InternVL-G (ours)	Cross-Attn	QLLaMA	**128.2**	79.2	113.7

Table 10. **Comparison of zero-shot image captioning.**

name	width	depth	MLP	#heads	#param	FLOPs	throughput	zs IN
variant 1	3968	32	15872	62	6051M	1571G	35.5 / 66.0	65.8
variant 2	3200	48	12800	50	5903M	1536G	28.1 / 64.9	66.1
variant 3	3200	48	12800	25	5903M	1536G	28.0 / 64.6	66.2
variant 4	2496	48	19968	39	5985M	1553G	28.3 / 65.3	65.9
variant 5	2816	64	11264	44	6095M	1589G	21.6 / 61.4	66.2
variant 6	2496	80	9984	39	5985M	1564G	16.9 / 60.1	66.2

Table 11. **Comparison of hyperparameters in InternViT-6B.** The throughput (img/s) and GFLOPs are measured at 224×224 input resolution, with a batch size of 1 or 128 on an A100 GPU.

4.5. Ablation Study

Hyperparameters of InternViT-6B. As discussed in Section 3.2, we explored variations in model depth {32, 48, 64, 80}, head dimension {64, 128}, and MLP ratio {4, 8}, resulting in 16 distinct models. In selecting the optimal model, we initially narrowed down our focus to 6 models, chosen based on their throughput, as listed in Table 11. These models underwent further evaluation using contrastive learning on a 100M subset of LAION-en [91] over 10K iterations. For the experimental setup, the primary difference was the use of a randomly initialized text encoder from CLIP-L [89], in order to speed up the training. For the sake of accuracy, inference speed, and training stability, we ultimately chose variant 3 as the final InternViT-6B.

Consistency of Feature Representation. In this study, we validate the consistency of the feature representation of InternVL with LLMs. We adopt a minimalist setting, *i.e.* conducting a single-stage SFT using only the LLaVA-Mix-665K [64]. Moreover, only the MLP layers are trainable, thereby confirming the inherent alignment level among features from various vision foundation models and LLMs. The results are shown in Table 12. These significant improvements clearly delineate that *the feature representation of InternVL is more consistent with the off-the-shelf LLM.*

visual encoder	glue layer	LLM	dataset	dialogue MME	caption NoCaps	visual question answering OKVQA	$VizWiz_{val}$	GQA
EVA-E	MLP	V-7B	665K [68]	970.5	75.1	40.1	25.5	41.3
IViT-6B	MLP	V-7B	665K [68]	1022.3	80.8	42.9	28.3	45.8
IViT-6B	QLLaMA	V-7B	665K [68]	1227.5	94.5	51.0	38.4	57.4
IViT-6B	QLLaMA	V-7B	Ours	1298.5	120.5	51.8	44.9	57.7
IViT-6B	QLLaMA	V-13B	Ours	1317.2	123.1	55.5	55.7	59.5

Table 12. **Ablation of InternVL's feature representations.** V-7B and V-13B denote Vicuna-7B and Vicuna-13B [145], respectively.

5. Conclusion

In this paper, we present InternVL, a large-scale vision-language foundation model that scales up the vision foundation model to 6 billion parameters and is aligned for generic visual-linguistic tasks. Specifically, we design a large-scale vision foundation model InternViT-6B, progressively align it with an LLM-initialized language middleware QLLaMA, and leverage web-scale image-text data from various sources for efficient training. It bridges the gap between vision foundation models and LLMs, and demonstrates proficiency in a wide range of generic visual-linguistic tasks, such as image/video classification, image/video-text retrieval, image captioning, visual question answering, and multi-modal dialogue. We hope this work could contribute to the development of the VLLM community.

Acknowledgement

The work is supported by the National Key R&D Program of China (No. 2022ZD0161300), and the National Natural Science Foundation of China (No. 62376134, 62372223). Tong Lu and Zhe Chen are supported by the China Mobile Zijin Innovation Insititute (No. NR2310J7M).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 7

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *ICCV*, pages 8948–8957, 2019. 7, 8

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736, 2022. 1, 3, 8

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3, 8

[6] Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. 3

[7] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 6

[8] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *NeurIPS*, 32, 2019. 6, 7

[9] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 6

[10] Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*, 2021. 7

[11] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 5

[12] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022. 5

[13] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. Reversible column networks. In *ICLR*, 2022. 3

[14] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *LREC*, pages 6848–6854, 2022. 6, 7

[15] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 6, 7

[16] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 7

[17] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 6, 7

[18] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021. 5

[19] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 3, 8

[20] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5, 7, 8

[21] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2022. 1, 3, 4

[22] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 8

[23] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2022. 3

[24] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. In *ACL*, pages 8666–8682, 2023. 6, 7

[25] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018. 5

[26] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023. 2, 3, 4, 5, 6

[27] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 3

[28] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36, 2024. 1, 8

[29] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 326–335, 2017. 5

[30] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, pages 7480–7512, 2023. 3, 4, 6, 7

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2, 3, 6, 7

[32] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, pages 13733–13742, 2021. 3

[33] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024. 8

[34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 3, 4

[35] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[36] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *ACL*, pages 320–335, 2022. 3

[37] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023. 3, 6

[38] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(1):5232–5270, 2022. 1

[39] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 7

[40] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. 5, 8

[41] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *NeurIPS*, 35: 26418–26431, 2022. 5, 7

[42] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 8

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 3

[44] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 6

[45] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 6, 7

[46] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 6, 7

[47] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 3

[48] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 5, 8

[49] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 3

[50] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023. 8

[51] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. Zenodo. Version 0.1. https://doi.org/10.5281/zenodo.5143773, 2021. DOI: 10.5281/zenodo.5143773. 3, 6, 7

[52] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2, 3

[53] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016. 5

[54] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 2, 3

[55] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012. 3

[56] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

[57] LAION-AI. Clip benchmark: Clip-like model evaluation. https://github.com/LAION-AI/CLIP_benchmark, 2023. 7

[58] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. In *ACM MM*, pages 1549–1557, 2017. 7

[59] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 3

[60] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 3, 5

[61] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 1, 4, 5, 7, 8

[62] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3

[63] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *TMM*, 21 (9):2347–2360, 2019. 7

[64] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4804–4814, 2022. 8

[65] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023. 7

[66] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023. 3

[67] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 5

[68] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3, 5, 6, 8

[69] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023. 1, 3, 4, 5

[70] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 3

[71] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[72] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 3

[73] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3

[74] Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, and Wenhai Wang. Controlllm: Augment language models with tools by searching on graphs. *arXiv preprint arXiv:2310.17796*, 2023. 3

[75] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 5

[76] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023. 3

[77] Kei Sawada Makoto Shiin, Tianyu Zhao. Construction and public release of language image pretraining models in japanese. In *MIRU*, 2022. 7

[78] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 5

[79] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019. 5

[80] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pages 2263–2279, 2022. 5

[81] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 5

[82] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952. IEEE, 2019. 5

[83] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *NeurIPS*, 36, 2024. 3

[84] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. 6

[85] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24, 2011. 5

[86] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. 3

[87] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 3, 8

[88] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 7, 8

[89] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 3, 5, 6, 7, 8

[90] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 6, 7

[91] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35: 25278–25294, 2022. 4, 5, 8

[92] Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en. *https://laion.ai/blog/laion-coco/*, 2022. 5

[93] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162, 2022. 5

[94] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 5

[95] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *NeurIPS*, 36, 2024. 3

[96] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020. 5

[97] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 8

[98] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *ICCV*, pages 5484–5494, 2023. 4, 6, 7

[99] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3, 4, 6, 7

[100] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. In *ICLR*, 2024. 1, 3, 8

[101] Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, et al. Moss: Training conversational language models from synthetic data. *arXiv preprint arXiv:2307.15020*, 7, 2023. 3

[102] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, pages 11888–11898, 2023. 3

[103] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023. 3

[104] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`, 2023. 2, 3, 5

[105] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, pages 516–533, 2022. 6

[106] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 3

[107] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3

[108] Dmitry Ustalov, Nikita Pavlichenko, Sergey Koshelev, Daniil Likhobaba, and Alisa Smirnova. Toloka visual question answering benchmark. *arXiv preprint arXiv:2309.16511*, 2023. 5

[109] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 6, 7

[110] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 7

[111] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021. 3

[112] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt

v2: Improved baselines with pyramid vision transformer. *CVMJ*, 8(3):415–424, 2022. 3

[113] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *NeurIPS*, 36, 2023. 1, 3

[114] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419, 2023. 3

[115] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *ICLR*, 2024. 3, 8

[116] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pages 10126–10135, 2020. 5

[117] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024. 6, 7

[118] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022. 3

[119] Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023. 3

[120] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3

[121] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 3

[122] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 6

[123] Chunyu Xie, Jincheng Li, Heng Cai, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, et al. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework. *arXiv preprint arXiv:2205.03860*, 2022. 7

[124] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 3

[125] Wanli Xue, Jingze Liu, Siyi Yan, Yuxi Zhou, Tiantian Yuan, and Qing Guo. Alleviating data insufficiency for chinese sign language recognition. *Visual Intelligence*, 1(1):26, 2023. 3

[126] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022. 6, 7

[127] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *NeurIPS*, 36, 2024. 3

[128] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 3

[129] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3

[130] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 3

[131] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 7

[132] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 5

[133] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 7

[134] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, pages 12104–12113, 2022. 3, 4, 6, 7

[135] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022. 7

[136] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[137] Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, et al. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv preprint arXiv:2209.02970*, 2022. 7

[138] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 1, 3

[139] Qinglong Zhang and Yu-Bin Yang. Rest: An efficient transformer for visual recognition. *NeurIPS*, 34:15475–15485, 2021. 3

[140] Qinglong Zhang and Yu-Bin Yang. Rest v2: simpler, faster and stronger. *NeurIPS*, 35:36440–36452, 2022. 3

[141] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2024. 3

[142] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[143] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 5

[144] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 5

[145] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36, 2024. 2, 3, 5, 7, 8

[146] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 6

[147] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024. 1, 3

[148] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023. 3