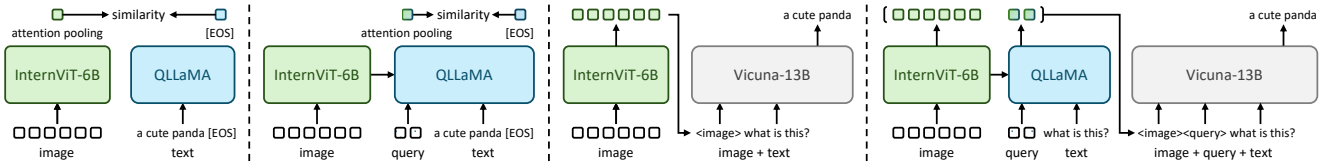
Supplementary Materials for Paper "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks"

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) InternVL-C (b) InternVL-G (c) InternVL-Chat (w/o QLLaMA) (d) InternVL-Chat (w/ QLLaMA)

Figure A. **Different ways to use InternVL.** By flexibly combining the vision encoder and the language middleware, InternVL can support various vision-language tasks, including contrastive tasks, generative tasks, and multi-modal dialogue.

Acknowledgement

We thank Shenglong Zhang, Beitong Zhou, Xinyue Zhang, Dongxing Shi, Weigao Sun, Xingcheng Zhang, and Zhifeng Yue for their contributions to the optimization of the training framework. We thank Zhenhang Huang for his assistance in data preparation.

A. Supplementary Materials

A.1. More Experiments

Zero-Shot Image Classification on 20 Datasets. In this section, we expand our examination to showcase the effectiveness and robustness of InternVL in 20 different zero-shot image classification benchmarks. As indicated in Table B, InternVL registers an average performance of 78.1% across all 20 benchmarks. This performance notably exceeds that of the previously leading method, EVA-02-CLIP-E+ [31], by a margin of 1.0 points. This underscores that, beyond ImageNet [28] and its variants, InternVL possesses robust generalization capabilities across a variety of different domains in zero-shot image classification.

Zero-Shot Image-Text Retrieval on XTD. Table A reports the results of InternVL on the multilingual image-text retrieval dataset XTD [1], spanning eight languages. As can be seen, InternVL-C achieves an average recall@10 score of 95.1% across these languages. The second stage model, InternVL-G, further improves retrieval performance. It attains the highest scores in each individual language and establishes a new record for average performance at 96.6%.

method	EN	ES	FR	ZH	IT	KO	RU	JP	avg.
mUSE m3 [106]	85.3	78.9	78.9	76.7	73.6	67.8	76.1	70.7	76.0
M-CLIP [11]	92.4	91.0	90.0	89.7	91.1	85.2	85.8	81.9	88.4
MURAL [49]	–	92.9	–	89.7	91.8	88.1	87.2	–	–
AltCLIP [17]	95.4	94.1	92.9	95.1	94.2	94.4	91.8	91.7	93.7
OpenCLIP-XLM-R-B [47]	95.8	94.4	92.5	91.8	94.4	86.3	89.9	90.7	92.0
OpenCLIP-XLM-R-H [47]	97.3	96.1	94.5	94.7	96.0	90.2	93.9	94.0	94.6
InternVL-C (ours)	97.3	95.7	95.1	95.6	96.0	92.2	93.3	95.5	95.1
InternVL-G (ours)	**98.6**	**97.7**	**96.5**	**96.7**	**96.9**	**95.1**	**94.8**	**96.1**	**96.6**

Table A. **Comparison of zero-shot multilingual image-text retrieval performance on the XTD dataset.** Multiple languages include English (EN), Spanish (ES), French (FR), Chinese (ZH), Italian (IT), Korean (KO), Russian (RU), and Japanese (JP). We follow M-CLIP [11] to report the recall@10 on Image-to-Text.

Zero-Shot Video Retrieval. In Table C, we present our results of zero-shot video-text retrieval on the MSR-VTT dataset [104] using our InternVL models, *i.e.* InternVL-C and InternVL-G. In the 1-frame setting, we select a single central frame from each video. In the 8-frame setting, we uniformly extract 8 frames from each video, treat them as independent images for encoding, and then average

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

method	CIFAR-10 [53]	CIFAR-100 [53]	MNIST [56]	Caltech-101 [32]	SUN397 [102]	FGVC Aircraft [70]	Country-211 [81]	Stanford Cars [52]	Birdsnap [6]	DTD [19]	Eurosat [42]	FER2013 [35]	Flowers-102 [76]	Food-101 [9]	GTSRB [91]	Pets [79]	Rendered SST2 [81]	Resisc45 [18]	STL10 [21]	VOC2007 [30]	avg. top-1 acc.
OpenAI CLIP-L+ [81]	94.9	74.4	79.0	87.2	68.7	33.4	34.5	79.3	41.0	56.0	61.5	49.1	78.6	93.9	52.4	93.8	70.7	65.4	99.4	78.1	69.6
EVA-01-CLIP-g [92]	98.3	88.7	62.3	87.7	74.2	32.4	28.6	91.7	50.0	61.3	73.6	52.2	74.5	93.5	49.1	94.2	58.4	70.3	98.9	83.2	71.2
OpenCLIP-g [47]	98.2	84.7	71.9	88.1	74.1	44.6	30.9	94.0	51.0	68.7	64.7	55.8	81.0	92.4	49.7	93.9	56.7	69.6	98.9	81.6	72.5
OpenCLIP-H [47]	97.4	84.7	72.9	85.0	75.2	42.8	30.0	93.5	52.9	67.8	72.7	52.0	80.1	92.7	58.4	94.5	64.3	70.5	98.5	77.7	73.2
EVA-02-CLIP-L+ [92]	98.9	89.8	64.3	89.5	74.8	37.5	33.6	91.6	45.8	64.5	71.4	51.0	77.2	94.2	57.6	94.2	64.6	69.8	**99.7**	82.7	72.6
EVA-01-CLIP-g+ [92]	99.1	90.1	71.8	88.1	74.3	39.4	30.8	90.7	52.6	67.3	73.2	56.0	79.7	93.7	66.5	94.8	58.6	71.4	99.5	82.9	74.0
OpenCLIP-G [47]	98.2	87.5	71.6	86.4	74.5	49.7	33.8	94.5	54.5	69.0	70.0	**59.5**	81.5	93.1	62.5	95.2	65.2	72.6	98.5	80.7	74.9
EVA-02-CLIP-E [92]	99.3	92.5	76.7	89.0	**76.5**	47.9	34.7	94.4	56.3	68.2	77.6	55.1	82.5	95.2	67.1	95.6	61.1	73.5	99.2	83.0	76.3
EVA-02-CLIP-E+ [92]	99.3	93.1	74.7	**90.5**	75.1	**54.1**	**35.7**	**94.6**	58.1	68.2	75.8	58.6	84.5	94.9	**67.7**	95.8	61.4	**75.6**	99.2	**85.6**	77.1
InternVL-C (ours)	**99.4**	**93.2**	**80.6**	89.5	76.0	52.7	34.1	94.2	**72.0**	**70.7**	**79.4**	56.2	**86.1**	**95.3**	65.5	**96.0**	**67.9**	74.2	99.5	80.0	**78.1**

Table B. **Comparison of zero-shot image classification performance on 20 other datasets.** These results indicate that, in addition to ImageNet [28], InternVL also possesses good generalization capabilities in zero-shot image classification across various domains.

method	#F	MSR-VTT (1K test set) [104]						avg.
		Video → Text			Text → Video			
		R@1	R@5	R@10	R@1	R@5	R@10	
OpenAI CLIP-L [81]	1	27.8	49.4	58.0	29.0	50.5	59.2	45.7
InternVL-C (ours)	1	35.3	56.6	66.6	37.5	60.9	**70.9**	54.6
InternVL-G (ours)	1	**36.6**	**58.3**	**67.7**	**39.1**	**61.7**	70.7	**55.7**
OpenAI CLIP-L [81]	8	26.6	50.8	61.8	30.7	54.4	64.0	48.1
Florence [109]	8	–	–	–	37.6	63.8	72.6	–
InternVideo† [101]	8	39.6	–	–	40.7	–	–	–
UMT-L† [60]	8	38.6	59.8	69.6	42.6	64.4	73.1	58.0
LanguageBind† [116]	8	40.9	66.4	75.7	44.8	70.0	78.7	62.8
InternVL-C (ours)	8	40.2	63.1	74.1	44.7	68.2	78.4	61.5
InternVL-G (ours)	8	**42.4**	**65.9**	**75.4**	**46.3**	**70.5**	**79.6**	**63.4**

Table C. **Comparison of zero-shot video-text retrieval performance on MSR-VTT.** "#F" denotes the number of frames. † These models are trained with temporal attention layers.

method	Flickr30K (English, 1K test set) [80]						avg.
	Image → Text			Text → Image			
	R@1	R@5	R@10	R@1	R@5	R@10	
ALIGN [50]	95.3	99.8	100.0	84.9	97.4	98.6	96.0
FILIP [107]	96.6	100.0	100.0	87.1	97.7	99.1	96.8
Florence [109]	97.2	99.9	–	87.9	98.1	–	–
BLIP [58]	97.4	99.8	99.9	87.6	97.7	99.0	96.9
OmniVL [97]	97.3	99.9	100.0	87.9	97.8	99.1	97.0
BEiT-3 [99]	97.5	99.9	100.0	89.1	98.6	**99.3**	97.4
ONE-PEACE [98]	97.6	100.0	100.0	89.6	98.0	99.1	97.4
InternVL-C-FT (ours)	97.2	100.0	100.0	88.5	98.4	99.2	97.2
InternVL-G-FT (ours)	**97.9**	**100.0**	**100.0**	**89.6**	**98.6**	99.2	**97.6**

method	Flickr30K-CN (Chinese, 1K test set) [55]						avg.
Wukong-ViT-L [38]	92.7	99.1	99.6	77.4	94.5	97.0	93.4
CN-CLIP-ViT-H [105]	95.3	99.7	100.0	83.8	96.9	98.6	95.7
R2D2-ViT-L [103]	95.6	99.8	100.0	84.4	96.7	98.4	95.8
InternVL-C-FT (ours)	96.5	99.9	100.0	85.2	97.0	98.5	96.2
InternVL-G-FT (ours)	**96.9**	**99.9**	**100.0**	**85.9**	**97.1**	**98.7**	**96.4**

Table D. **Comparison of fine-tuned image-text retrieval performance.** We evaluate English and Chinese image-text retrieval using Flickr30K [80] and Flickr30K-CN [55], with separate fine-tuning for each to prevent data leakage.

the embeddings. The results showcase consistent improvement across various metrics such as R@1, R@5, R@10, and the average score. Importantly, both models exhibit promising outcomes in single-frame and multi-frame configurations, with InternVL-G achieving slightly higher performance than InternVL-C, especially in the multi-frame setting. These results underscore the effectiveness of QLLaMA in harmonizing visual and linguistic features.

Fine-tuned Image-Text Retrieval. In Table D, we report the fine-tuned image-text retrieval results of InternVL, on both the English and Chinese versions of the Flickr30K dataset [55, 80]. The specific hyperparameters for fine-tuning are shown in Table I. As can be seen, our models obtain competitive performance, with InternVL-G-FT marginally surpassing InternVL-C-FT in both datasets. Notably, in the highly challenging Flickr30K-CN, both models show a promising ability to handle cross-lingual retrieval tasks. These results demonstrate the effectiveness of our language middleware, especially in the retrieval tasks.

Tiny LVLM. Tiny LVLM [87] is an ability-level benchmark for evaluating the performance of multimodal dialogue models. It provides a systematic assessment of five categories of multimodal capabilities, including visual perception, visual knowledge acquisition, visual reasoning, visual commonsense, and object hallucination. We report our results on Tiny LVLM in Table E.

A.2. More Ablation Studies

Compatibility with Other LLM. In this experiment, we test the compatibility of InternVL with LLMs other than Vicuna [114]. The experimental setup used here is the same as in Table 9 of the main paper. As shown in Table F, InternLM-7B [93] achieves slightly better performance than Vicuna-7B [114]. This indicates that our InternVL exhibits promising compatibility with various LLMs.

Efficiency Analysis. In this study, we analyze the computational efficiency of InternVL in encoding image-text pairs. The entire encoding process consists of two parts: image encoding and text encoding. The analysis covered two models (InternVL-C and InternVL-G) and their performance across three different image sizes (224, 336, and

method	LLM	VR	VP	VKA	VC	OH	Overall
MiniGPT-4 [117]	Vicuna-7B	37.6	37.8	17.6	49.0	50.7	192.6
LLaVA [67]	Vicuna-7B	41.6	38.3	18.7	49.4	49.0	197.0
VisualGLM [29]	ChatGLM-6B	37.3	36.3	46.9	37.6	54.0	211.9
Otter [57]	Otter-9B	41.6	37.0	15.1	52.4	74.0	216.4
LLaMA-Adapter-V2 [34]	LLaMA-7B	43.5	46.8	22.3	56.0	60.7	229.2
Lynx [110]	Vicuna-7B	52.2	65.8	17.6	57.4	86.3	279.2
BLIP-2 [59]	FlanT5xl	44.9	49.0	64.1	44.0	82.7	284.7
InstructBLIP [24]	Vicuna-7B	46.7	48.0	61.7	59.2	85.0	300.6
LLaVA-1.5 [66]	Vicuna-7B	55.6	49.0	57.0	57.2	88.3	307.2
Qwen-VL-Chat [3]	Qwen-7B	62.4	54.5	55.1	54.8	90.0	316.8
Bard [36]	Bard	64.2	57.0	68.1	59.6	70.7	319.6
InternLM-XComposer [111]	InternLM-7B	55.8	53.8	64.1	61.8	87.0	322.5
InternVL-Chat (ours)	Vicuna-13B	56.4	52.3	68.0	62.0	89.0	**327.6**

Table E. **Evaluation of Tiny LVLM test set.** Here we report five categories of multimodal capabilities, including visual reasoning (VR), visual perception (VP), visual knowledge acquisition (VKA), visual commonsense (VC), and object hallucination (OH).

visual encoder	glue layer	LLM	visual question answering				dialogue	
			VQA^{v2}	GQA	VizWiz	VQA^{T}	MME	POPE
IViT-6B	MLP	Vicuna-7B	79.3	62.9	52.5	57.0	1525.1	86.4
IViT-6B	MLP	InternLM-7B	79.7	63.2	53.1	58.0	1532.8	86.4

Table F. **Compatibility with other LLM.** Here we use InternLM [93] as an example to verify the compatibility of InternVL with LLMs other than Vicuna [114]. The experimental settings used here are the same as in Table 9 of the main paper.

method	image size	encode image (ms)		encode text (ms)	total time	FPS
		InternViT-6B	QLLaMA	QLLaMA		
InternVL-C	224	15.5	–	4.9	20.4	48.9
InternVL-C	336	35.2	–	4.9	40.1	24.9
InternVL-C	448	66.9	–	4.9	71.8	13.9
InternVL-G	224	15.5	8.2	4.9	28.6	35.0
InternVL-G	336	35.2	10.3	4.9	50.4	19.8
InternVL-G	448	66.9	12.8	4.9	84.6	11.8

Table G. **Efficiency analysis of InternVL for encoding image-text pairs.** The total time to encode an image-text pair includes both the image encoding part and the text encoding part. We measure the time cost with a batch size of 128 on a single A100 GPU. Flash Attention [25] and bf16 precision are used during testing.

448). The results are shown in Table G.

From these results, we find that: (1) As the image size increases, the encoding time also significantly increases, leading directly to a decrease in frame rate; (2) InternVL-G slightly increased the encoding time due to the introduction of QLLaMA for secondary image encoding, but it still maintains a reasonable frame rate across all image sizes; (3) Even though we scale up the text encoder, the additional cost of text encoding is not significant, as the main time expenditure lies in image encoding. In summary, when choosing between InternVL-C and InternVL-G, one should weigh the trade-off between computational efficiency and potential performance improvements based on specific requirements. Additionally, these results were measured using PyTorch with Flash Attention [25] and bf16 precision, and there is still considerable room for optimization, such as using model quantization and TensorRT.

config	stage 1	stage 2
image enc. weight init.	random init. [4]	from stage 1
text enc. weight init.	from [23]	from stage 1
image enc. peak learning rate	1e-3	frozen
text enc. peak learning rate	1e-4	frozen
cross attn peak learning rate	–	5e-5
learning rate schedule	cosine decay	cosine decay
optimizer	AdamW [68]	AdamW [68]
optimizer hyper-parameters	β_1, β_2 = 0.9, 0.95	β_1, β_2 = 0.9, 0.98
weight decay	0.1	0.05
input resolution	$196^2 \rightarrow 224^2$	224^2
patch size	14	14
total batch size	164K	20K
warm-up iterations	5K	2K
total iterations	175K	80K
samples seen	28.7B	1.6B
drop path rate [45]	uniform (0.2)	0.0
data augmentation	random resized crop	random resized crop
numerical precision	DeepSpeed bf16 [82]	DeepSpeed bf16 [82]
trainable / total parameters	13B / 13B	1B / 14B
GPUs for training	640×A100 (80G)	160×A100 (80G)

Table H. **Training settings of InternVL's stage 1 and stage 2.** "$196^2 \rightarrow 224^2$" means we initially train at a 196×196 resolution, and later switch to 224×224 resolution for the final 0.5 billion samples, for higher training efficiency.

A.3. Detailed Training Settings

Settings of Stage 1. As shown in Table H, in this stage, the image encoder InternViT-6B is randomly initialized using the BEiT's initialization method [4], and the text encoder LLaMA-7B is initialized with the pre-trained weights from [23], a multilingual LLaMA-7B. All parameters are fully trainable. We employ the AdamW optimizer [68] with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay at 0.1, and a cosine learning rate schedule starting at 1e-3 and 1e-4 for the image and text encoders, respectively. We adopt a uniform drop path rate of 0.2. The training involves a total batch size of 164K across 640 A100 GPUs, extending over 175K iterations to process about 28.7 billion samples. To enhance efficiency, we initially train at a 196×196 resolution, masking 50% of image tokens [63], and later switch to 224×224 resolution without masking for the final 0.5 billion samples.

Settings of Stage 2. In this stage, InternViT-6B and QLLaMA inherit their weights from the first stage, while the learnable queries and cross-attention layers in QLLaMA are randomly initialized. Benefiting from the powerful encoding capabilities learned in the first stage, we keep both InternViT-6B and QLLaMA frozen and only train the newly added parameters. The input images are processed at a resolution of 224×224. For optimization, the AdamW optimizer [68] is employed with $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay set at 0.05, and a total batch size of 20K. The training extends over 80K steps across 160 A100 GPUs, inclusive of 2K warm-up steps, and is governed by a cosine learning rate schedule with a peak learning rate of 5e-5. More detailed training settings are listed in Table H.

Settings of Stage 3. At this stage, we have two different configurations. One is to use InternViT-6B separately, as

config	retrieval fine-tuning
image-text data	Flickr30K [80] / Flickr30K-CN [55]
peak learning rate	1e-6
layer-wise lr decay rate	InternViT-6B (0.9), QLLaMA (0.9)
learning rate schedule	cosine decay
optimizer	AdamW [68]
optimizer hyper-parameters	β_1, β_2 = 0.9, 0.999
weight decay	0.05
input resolution	364^2
patch size	14
total batch size	1024
warm-up iterations	100
training epochs	10
drop path rate [45]	0.3
data augmentation	random resized crop & flip
numerical precision	DeepSpeed bf16 [82]
trainable / total parameters	14B / 14B
GPUs for training	32×A100 (80G)

Table I. **Training settings of retrieval fine-tuning.** We fine-tune InternVL on Flickr30K and Flickr30K-CN separately.

config	ImageNet linear probing
peak learning rate	0.2
learning rate schedule	cosine decay
optimizer	SGD
optimizer momentum	0.9
weight decay	0.0
input resolution	224^2
patch size	14
total batch size	1024
warm-up epochs	1
training epochs	10
data augmentation	random resized crop & flip
GPUs for training	8×A100 (80G)

Table J. **Training settings of ImageNet linear probing.**

shown in Figure A (c). The other is to use the entire InternVL model simultaneously, as shown in Figure A (d).

(1) InternVL-Chat (w/o QLLaMA): For this setup, we follow the training recipes of LLaVA-1.5 [66]. We use the same hyperparameters and datasets for supervised fine-tuning, *i.e.* we first train the MLP layers with the LGS-558K [67] dataset, and then train the LLM with the LLaVA-Mix-665K [66] dataset, both for one epoch.

(2) InternVL-Chat (w/ QLLaMA): For this more advanced setup, we also conducted the training in two steps. We first train the MLP layers with our custom SFT dataset and then fine-tune the LLM with it. Due to the expansion of the dataset, we increased the batch size to 512.

Settings of Retrieval Fine-tuning. In this experiment, all parameters of InternVL are set to be trainable. We conduct separate fine-tuning on the Flickr30K [80] and Flickr30K-CN [55]. Following common practice [59], a 364×364 resolution is adopted for fine-tuning. To avoid over-fitting, we apply a layer-wise learning rate decay of 0.9 to both InternViT-6B and QLLaMA, along with a drop path rate of 0.3 for InternViT-6B. The AdamW optimizer [68] is utilized, with a total batch size of 1024, for fine-tuning the InternVL model across 10 epochs. For more detailed training settings, please refer to Table I.

Settings of ImageNet Linear Probing. We follow the common practices of linear probing in previous methods [27, 41, 77]. Specifically, we employ an additional BatchNorm [48] to normalize the pre-trained backbone features during training. Besides, we concatenate the average-pooled patch token features with the class token. The linear head is trained using the SGD optimizer for 10 epochs on ImageNet-1K [28], with a total batch size of 1024, a peak learning rate of 0.2, 1 epoch warm-up, and no weight decay. Data augmentation involves random-resized-crop and flip. For more training details, please see Table J.

config	linear probing / head tuning / full tuning
peak learning rate	4e-5
layer-wise lr decay rate	– / – / 0.95
learning rate schedule	polynomial decay
optimizer	AdamW [68]
optimizer hyper-parameters	β_1, β_2 = 0.9, 0.999
weight decay	0.0 / 0.05 / 0.05
input resolution	504^2
patch size	14
total batch size	16
warm-up iterations	1.5K
total iterations	80K
drop path rate [45]	0.0 / 0.0 / 0.4
data augmentation	default augmentation in MMSeg [22]
numerical precision	DeepSpeed bf16 [82]
GPUs for training	8×A100 (80G)

Table K. **Training settings of ADE20K semantic segmentation.** We list the hyperparameters for three different configurations, including linear probing, head tuning, and full-parameter tuning.

Settings of ADE20K Semantic Segmentation. In Table K, we have listed the hyperparameters for three different configurations in ADE20K semantic segmentation, including linear probing, head tuning, and full-parameter tuning.

A.4. Data Preparation for Pre-training

Training Data for Stage 1 & Stage 2. During the first and second stages, we employed a vast collection of image-text pair data (see Figure B (a)), such as LAION-en [84], LAION-multi [84], LAION-COCO [85], COYO [10], Wukong [38], among others [15, 78, 88]. A detailed introduction to these datasets is provided in Table L.

Training Data Cleaning for Stage 1 & Stage 2. To fully utilize web-scale image-text data, we adopted different data filtering strategies in stage 1 and stage 2.

(1) Stage 1: In the first stage, we applied only minor data filtering, thus retaining the vast majority of the data. We considered six factors: CLIP similarity, watermark probability, unsafe probability, aesthetic score, image resolution, and caption length, to remove extreme data points and avoid disrupting training stability. Additionally, we removed data that was duplicated with ImageNet-1K/22K [28], Flickr30K [80], and COCO [64] to ensure the reliability of our zero-shot evaluations. Due to download failures and the use of our data filtering pipeline, the total amount of data retained in the first stage was 4.98 billion.

(2) Stage 2: In the second stage, we implemented a more

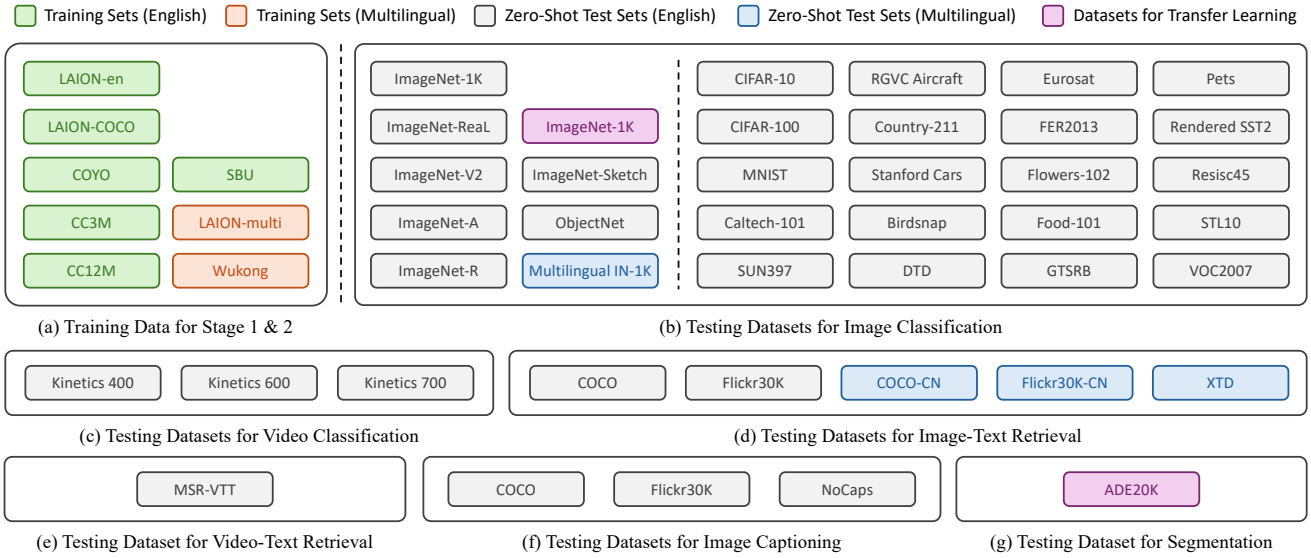


Figure B. **Panoramic overview of the datasets used in InternVL's stage 1 and stage 2.** During the training of stage 1 and stage 2, we utilize web-scale image-text data from a variety of sources to train our InternVL model, as shown in (a). To assess InternVL's capabilities in handling generic visual-linguistic tasks, we conducted extensive validations across a range of tasks and datasets, including (b) image classification, (c) video classification, (d) image-text retrieval, (e) video-text retrieval, (f) image captioning, and (g) semantic segmentation.

stringent data filtering strategy. With generative supervision included, we deleted most of the low-quality data based on the captions, mainly considering the length, completeness, readability, and whether they were gibberish or boilerplate (like menus, error messages, or duplicate text), contained offensive language, placeholder text, or source code. We retained only 1.03 billion entries.

Testing Datasets for Image Classification. We conducted extensive validation on image classification tasks (see Figure B (b)), including the linear probing performance of InternViT-6B and the zero-shot performance of InternVL-C. These datasets used are listed in Table L.

Testing Datasets for Video Classification. As shown in Figure B (c), to evaluate the capabilities of video classification, we utilize the following Kinetics datasets: Kinetics 400 [12], Kinetics 600 [13], and Kinetics 700 [14].

Testing Datasets for Image-Text Retrieval. We use five datasets (see Figure B (d)) to evaluate InternVL's zero-shot, multilingual image-text retrieval capabilities. A detailed introduction to these datasets is provided in Table M.

Testing Dataset for Video-Text Retrieval. As shown in Figure B (e), we use the MSR-VTT [104] dataset to evaluate our InternVL in zero-shot video-text retrieval.

Testing Dataset for Image Captioning. As illustrated in Figure B (f), we use three image captioning datasets to test our InternVL model. A detailed introduction to these datasets is provided in Table N.

Testing Dataset for Semantic Segmentation. We use the ADE20K [115] dataset to study the pixel-level perceptual capacity of InternViT-6B, as shown in Figure B (g). A detailed introduction to this dataset is provided in Table N.

A.5. Data Preparation for SFT

Training Data for SFT. In this stage, we collect a wide range of high-quality instruction data. For non-dialogue datasets, we follow the method described in [66] for conversion. A detailed introduction is provided in Table O.

Testing Datasets for SFT. We validate the effectiveness of our supervised fine-tuned InternVL-Chat models on three tasks, including image captioning, visual question answering, and multi-modal dialogue. There datasets are listed in Table P. For most of these datasets, we employ the same response formatting prompt as for LLaVA-1.5 [66].

dataset	introduction
Training Data for Stage 1 & Stage 2.	
LAION-en [84]	LAION-en is a part of the LAION-5B dataset, containing 2.32 billion English-only image-text pairs.
LAION-multi [84]	LAION-multi is another segment of LAION-5B, featuring 2.26 billion image-text pairs across more than 100 languages, and is ideal for multilingual studies.
Laion-COCO [85]	Laion-COCO comprises 663 million synthetic captions for web images, generated using a blend of BLIP-L/14 [58] and CLIP models [81].
COYO [10]	COYO-700M is a large-scale dataset that contains 747 million image-text pairs as well as many other meta-attributes to increase the usability to train various models. It follows a similar strategy to previous vision-language datasets, collecting many informative pairs of alt-text and its associated image in HTML documents.
Wukong [38]	Wukong is a large-scale Chinese image-text dataset for benchmarking different multi-modal pre-training methods. It contains 100 million Chinese image-text pairs from the web.
CC3M [88]	This dataset consists of approximately 3 million images, each annotated with a caption.
CC12M [15]	CC12M is a dataset with 12 million image-text pairs. It is larger and covers a much more diverse set of visual concepts than the CC3M [88].
SBU [78]	The SBU Captioned Photo Dataset is a collection of over 1 million images with associated text descriptions extracted from Flicker.
Testing Datasets for Image Classification.	
ImageNet-1K [28]	A large-scale dataset commonly used in image classification, consisting of over 1 million images across 1K different classes.
ImageNet-ReaL [7]	It contains ImageNet val images augmented with a new set of "re-assessed" labels. These labels are collected using an enhanced protocol, resulting in multi-label and more accurate annotations.
ImageNet-V2 [83]	A dataset created to test the robustness of models trained on ImageNet-1K, containing new test images collected following the original methodology.
ImageNet-A [44]	It consists of real-world, unmodified, and naturally occurring examples that are misclassified by ResNet models [40]. It's designed to highlight the challenges of adversarial examples in natural settings.
ImageNet-R [43]	A set of images labeled with ImageNet labels obtained by collecting art, cartoons, deviantart, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video game renditions of ImageNet classes. It has renditions of 200 ImageNet classes resulting in 30K images.
ImageNet-Sketch [96]	It consists of 51K images, approximately 50 images for each of the ImageNet classes. It is constructed using Google Image queries with the standard class name followed by "sketch of".
ObjectNet [5]	ObjectNet is a crowd-sourced test set of 50K images featuring objects in unusual poses and cluttered scenes, designed to challenge recognition performance. It includes controls for rotation, background, and viewpoint, and covers 313 object classes, with 113 overlapping with ImageNet [28].
Multilingual IN-1K [54]	An adaptation of ImageNet-1K supporting multilingual annotations, facilitating research in cross-lingual image classification.
CIFAR-10/100 [53]	It comprises 60K 32×32 images in 10 classes (CIFAR-10) or 100 classes (CIFAR-100).
MNIST [56]	A classic dataset containing 70K 28×28 gray-scale images of handwritten digits.
Caltech-101 [32]	The dataset comprises images of objects from 101 classes and a background clutter class, each labeled with a single object. It contains about 40 to 800 images per class, totaling approximately 9K images.
SUN397 [102]	The SUN397 or Scene UNderstanding (SUN) is a dataset for scene recognition consisting of 397 categories with 109K images.
FGVC Aircraft [70]	The dataset contains 10K images of aircraft, with 100 images for each of 102 different aircraft model variants, most of which are airplanes.
Country-211 [81]	It is a dataset released by OpenAI, designed to assess the geolocation capability of visual representations. It filters the YFCC100M [94] dataset to find 211 countries that have at least 300 photos with GPS coordinates. OpenAI built a balanced dataset with 211 categories, by sampling 200 photos for training and 100 photos for testing, for each country.
Stanford Cars [52]	This dataset consists of 196 classes of cars with a total of 16K images, taken from the rear. The data is divided into almost a 50-50 train/test split with 8K training images and 8K testing images.

Table L. **Introduction of datasets used in InternVL's stage 1 and stage 2.** In summary, we utilize a vast amount of image-text data for pre-training and conduct comprehensive evaluation across a wide range of generic visual-linguistic tasks.

dataset	introduction
Testing Datasets for Image Classification.	
Birdsnap [6]	Birdsnap is a large bird dataset consisting of 49,829 images from 500 bird species with 47,386 images used for training and 2,443 images used for testing. Due to broken links, we are only able to download 1,845 out of the 2,443 testing images.
DTD [19]	The Describable Textures Dataset (DTD) contains 5,640 texture images in the wild. They are annotated with human-centric attributes inspired by the perceptual properties of textures.
Eurosat [42]	This dataset is based on Sentinel-2 satellite images covering 13 spectral bands and consisting of 10 classes with 27K labeled and geo-referenced samples.
FER2013 [35]	This dataset includes around 30K RGB facial images, categorized into seven expressions: angry, disgust, fear, happy, sad, surprise, and neutral.
Flowers-102 [76]	It is consistent with 102 flower categories commonly occurring in the United Kingdom. Each class consists of between 40 and 258 images.
Food-101 [9]	The Food-101 dataset consists of 101 food categories with 750 training and 250 test images per category, making a total of 101K images.
GTSRB [91]	The German Traffic Sign Recognition Benchmark (GTSRB) contains 43 classes of traffic signs, split into 39,209 training images and 12,630 test images.
Pets [79]	The Oxford-IIIT Pet Dataset is a 37-category pet dataset with roughly 200 images for each class created by the Visual Geometry Group at Oxford.
Rendered SST2 [81]	This dataset is used to evaluate the model's capability on optical character recognition. It was generated by rendering sentences in the Standford Sentiment Treebank v2 dataset.
Resisc45 [21]	This is a dataset for remote sensing scene classification. It contains 31,500 RGB images divided into 45 scene classes, each class containing 700 images.
STL10 [76]	The STL-10 dataset, inspired by CIFAR-10 [53], includes 10 classes with 500 training and 800 test color images each, sized 96×96 pixels.
VOC2007 [30]	The Pascal VOC 2007 dataset focuses on recognizing objects in realistic scenarios and contains 20 object classes across 9,963 images with 24,640 labeled objects. The data has been divided into 50% for training/validation and 50% for testing. Following common practice, we conduct zero-shot image classification by cropping images to isolate objects using bounding boxes.
Testing Datasets for Video Classification.	
Kinetics 400 [12]	A large-scale dataset containing around 400 human action classes with at least 400 video clips for each class, sourced from YouTube.
Kinetics 600 [13]	An expansion of Kinetics 400, this dataset includes 600 action classes and provides an increased diversity in video representation.
Kinetics 700 [14]	The latest in the series, Kinetics 700 offers an even broader range with 700 action categories, further challenging the robustness of retrieval models.
Testing Datasets for Image-Text Retrieval.	
COCO [16]	The COCO Caption dataset contains diverse images with detailed captions, widely used for image-text retrieval and image captioning tasks.
COCO-CN [61]	COCO-CN is a bilingual image description dataset enriching COCO with manually written Chinese sentences and tags. The new dataset can be used for multiple tasks including image tagging, captioning, and retrieval, all in a cross-lingual setting.
Flickr30K [80]	This dataset comprises 31,000 images sourced from Flickr, each annotated with five captions, making it suitable for image-text retrieval.
Flickr30K-CN [55]	Flickr30K-CN offers Chinese captions for the images, enabling studies in cross-lingual and multi-modal retrieval tasks.
XTD [1]	A newly developed 1K multilingual test set, featuring COCO images annotated in various languages.
Testing Dataset for Video-Text Retrieval.	
MSR-VTT [104]	This is a large-scale dataset for open-domain video captioning and video-text retrieval, comprising 10,000 video clips across 20 categories. Each clip is annotated with 20 English sentences, totaling about 29,000 distinct words in all captions. The standard division of the dataset allocates 6,513 clips for training, 497 for validation, and 2,990 for testing purposes.

Table M. **Introduction of datasets used in InternVL's stage 1 and stage 2.** In summary, we utilize a vast amount of image-text data for pre-training and conduct comprehensive evaluation across a wide range of generic visual-linguistic tasks.

dataset	introduction
Testing Datasets for Image Captioning.	
COCO [16]	We use the Karpathy test set for testing.
Flickr30K [80]	We use the Karpathy test set for testing.
NoCaps [2]	NoCaps stands out for testing models' capabilities in open-ended caption generation, using images that go beyond the training data's domain. We report the performance on the NoCaps val set.
Testing Dataset for Semantic Segmentation.	
ADE20K [115]	ADE20K contains more than 20K scene-centric images exhaustively annotated with pixel-level objects and object parts labels. There are a total of 150 semantic categories, which include stuffs like sky, road, grass, and discrete objects like person, car, bed. We report the performance on the ADE20K val set.

Table N. **Introduction of datasets used in InternVL's stage 1 and stage 2.** In summary, we utilize a vast amount of image-text data for pre-training and conduct comprehensive evaluation across a wide range of generic visual-linguistic tasks.

dataset	introduction
Training Data for SFT.	
COCO Caption [16]	It contains over 0.5 million captions describing over 110K images. Following common practice, we use the Karpathy training set for training. We transform it into a dialogue dataset using the response formatting prompt: "Provide a one-sentence caption for the provided image."
TextCaps [89]	TextCaps contains 145K captions for 28K images. It challenges a model to recognize text, relate it to its visual context, and decide what part of the text to copy or paraphrase. OCR tokens are used during training. We transform it into a dialogue dataset using the response formatting prompt: "Provide a one-sentence caption for the provided image."
VQAv2 [37]	VQAv2, the second version of the VQA dataset, features open-ended questions related to images. Answering these questions demands a grasp of vision, language, and common sense. We convert it into a dialogue dataset using the prompt: "Answer the question using a single word or phrase."
OKVQA [72]	A dataset with over 14K questions requiring external knowledge for answers, focusing on knowledge-based visual question answering. We transform it into a dialogue dataset using the response formatting prompt: "Answer the question using a single word or phrase."
A-OKVQA [86]	An augmented successor of OKVQA [72] and contains 25K questions requiring a broad base of commonsense and world knowledge to answer. We transform it into a dialogue dataset using the response formatting prompt: "Answer with the option's letter from the given choices directly."
IconQA [69]	A dataset with 107K questions across three sub-tasks, focusing on abstract diagram recognition and comprehensive visual reasoning. We convert it into a dialogue dataset using these prompts: "Answer with the option's letter from the given choices directly." and "Answer the question using a single word or phrase."
AI2D [51]	AI2D features over 5K grade school science diagrams with rich annotations and 15K multiple-choice questions for diagram understanding research. We convert it into a dialogue dataset using the prompt: "Please answer the question based on the options mentioned before."
GQA [46]	GQA is a large-scale dataset with more than 110K images and 22 million questions, combining real images with balanced question-answer pairs for visual reasoning. We transform it into a dialogue dataset using the prompt: "Answer the question using a single word or phrase."
OCR-VQA [75]	The OCR-VQA dataset contains 207,572 images of book covers and more than 1 million question-answer pairs about these images. We convert it into a dialogue dataset using the response formatting prompt: "Answer the question using a single word or phrase."
ChartQA [73]	ChartQA is a dataset for question answering about charts, focusing on visual and logical reasoning. It comprises 9.6K human-written questions and 23.1K questions generated from human-written chart summaries. We convert it using the prompt: "Answer the question using a single word or phrase."
DocVQA [20]	The DocVQA dataset consists of 50,000 questions defined on over 12,000 document images. We convert it into a dialogue dataset using the prompt: "Answer the question using a single word or phrase."
ST-VQA [8]	The ST-VQA dataset contains a total of 31,791 questions over 23,038 images. The training set alone consists of 26,308 questions based on 19,027 images. We convert it into a dialogue dataset using the response formatting prompt: "Answer the question using a single word or phrase."

Table O. **Introduction of datasets used in InternVL's stage 3.** We collect a wide range of high-quality instruction data. For non-dialogue datasets, we follow the response formatting prompts described in [66] for conversion. Note that only the training set is used for training.

dataset	introduction
Training Data for SFT.	
EST-VQA [100]	The EST-VQA dataset provides questions, images, and answers, but also a bounding box for each question that indicates the area of the image that informs the answer. We convert it into a dialogue dataset using the response formatting prompt: "Answer the question using a single word or phrase."
InfoVQA [74]	This dataset includes a diverse collection of infographics with natural language questions and answers. It focuses on reasoning over document layout, textual content, graphical elements, and data visualizations. We convert it into a dialogue dataset using the prompt: "Answer the question using a single word or phrase."
LLaVAR [112]	The LLaVAR dataset advances visual instruction tuning for Large Language Models by focusing on text-rich images. It incorporates 422K images processed with OCR and 16K GPT-4 generated conversations, enhancing text-based VQA performance and human interaction capabilities in diverse scenarios. Note that, we only use the 20K high-quality data for fine-tuning of LLaVAR.
RefCOCO [71, 108]	A mixed dataset of RefCOCO [108], RefCOCO+[108], and RefCOCO-g [71]. We convert it into a dialogue dataset following LLaVA-1.5 [66].
Toloka [95]	The TolokaVQA dataset comprises images with associated textual questions, each marked with a bounding box indicating the visual answer. It's sourced from a licensed subset of the COCO dataset and labeled on the Toloka platform. We convert it into a dialogue dataset following LLaVA-1.5 [66].
LLaVA-150K [67]	This is a set of GPT-generated multi-modal instruction-following data, constructed for visual instruction tuning and building large multi-modal models towards GPT-4 vision/language capability. It includes 158K unique language-image instruction-following samples.
SVIT [113]	This dataset includes 3.2 million visual instruction tuning data, with 1.6M conversation QA pairs, 1.6M complex reasoning QA pairs, and 106K detailed image descriptions. It is designed to improve multi-modal performance in visual perception, reasoning, and planning. For this dataset, we merge the QA pairs from the same training image into a single conversation.
VisDial [26]	A dataset based on the COCO images, featuring dialogues created by two Amazon Mechanical Turk workers. One plays the 'questioner', seeing only an image's text description, and the other, the 'answerer', sees the image. They engage in a 10-round Q&A session about the image.
LRV-Instruction [65]	The LRV-Instruction dataset is designed to combat hallucination in large multi-modal models. It comprises 120K GPT-4-generated visual instructions for 16 vision-and-language tasks, including both positive and negative instructions for robust tuning. Negative instructions focus on Nonexistent and Existent Element Manipulation. This dataset helps improve accuracy and consistency in multi-modal tasks.
LLaVA-Mix-665K [66]	LLaVA-Mix-665K is an instruction-following dataset mixed from 10 academically oriented datasets.
Testing Dataset for SFT (Image Captioning).	
COCO [16]	Karpathy test set is used for testing. The prompt is: "Provide a one-sentence caption for the provided image."
Flickr30K [80]	Karpathy test set is used for testing. The prompt is: "Provide a one-sentence caption for the provided image."
NoCaps [2]	NoCaps val set is used for testing. The prompt is: "Provide a one-sentence caption for the provided image."
Testing Dataset for SFT (Visual Question Answering).	
VQAv2 [37]	VQAv2 test-dev set is used for testing. The prompt is: "Answer the question using a single word or phrase."
GQA [46]	GQA test-balanced set is used. The prompt is: "Answer the question using a single word or phrase."
VizWiz [39]	VizWiz test-dev set is used for testing. The prompt is: "When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase."
TextVQA [90]	TextVQA val set is used for testing. The prompt is: "Answer the question using a single word or phrase."
Testing Dataset for SFT (Multi-Modal Dialogue).	
MME [33]	MME is a comprehensive evaluation benchmark for multi-modal large language models. It measures both perception and cognition abilities on a total of 14 subtasks, including existence, count, position, color, poster, celebrity, scene, landmark, artwork, OCR, commonsense reasoning, numerical calculation, text translation, and code reasoning. The prompt for this dataset is: "Answer the question using a single word or phrase."
POPE [62]	POPE is a popular dataset used to evaluate object hallucination. The response formatting prompt used for this dataset is: "Answer the question using a single word or phrase."

Table P. **Introduction of datasets used in InternVL's stage 3.** We collect a wide range of high-quality instruction data. For non-dialogue datasets, we follow the response formatting prompts described in [66] for conversion. Note that only the training set is used for training. We evaluate our InternVL-Chat models on three tasks, including image captioning, VQA, and multi-modal dialogue. For these datasets, we employ the same response formatting prompts as for LLaVA-1.5 [66].

References

[1] Pranav Aggarwal and Ajinkya Kale. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*, 2020. 1, 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. 8, 9

[3] 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. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 3

[5] 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

[6] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, pages 2011–2018, 2014. 2, 7

[7] 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

[8] 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. 8

[9] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 2, 7

[10] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022. 4, 6

[11] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *LREC*, pages 6848–6854, 2022. 1

[12] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 5, 7

[13] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 5, 7

[14] 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. 5, 7

[15] 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. 4, 6

[16] 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. 7, 8, 9

[17] 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. 1

[18] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2

[19] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 2, 7

[20] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018. 8

[21] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTAT*, pages 215–223, 2011. 2, 7

[22] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020. 4

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

[24] 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. 3

[25] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 35:16344–16359, 2022. 3

[26] 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. 9

[27] 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. 4

[28] 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. 1, 2, 4, 6

[29] 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

[30] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015. 2, 7

[31] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 1

[32] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pages 178–178, 2004. 2, 6

[33] 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. 9

[34] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 3

[35] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *ICONIP*, pages 117–124, 2013. 2, 7

[36] Google. Google bard. https://bard.google.com/, 2023. 3

[37] 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. 8, 9

[38] 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. 2, 4, 6

[39] 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. 9

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[41] 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. 4

[42] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2, 7

[43] 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

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

[45] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016. 3, 4

[46] 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. 8, 9

[47] 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. 1, 2

[48] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 4

[49] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*, 2021. 1

[50] 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

[51] 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. 8

[52] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. 2, 6

[53] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 2, 6, 7

[54] LAION-AI. Clip benchmark: Clip-like model evaluation. https://github.com/LAION-AI/CLIP_benchmark, 2023. 6

[55] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. In *ACM MM*, pages 1549–1557, 2017. 2, 4, 7

[56] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2, 6

[57] 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

[58] 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. 2, 6

[59] 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. 3, 4

[60] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, pages 19948–19960, 2023. 2

[61] 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

[62] 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. 9

[63] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, pages 23390–23400, 2023. 3

[64] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 4

[65] 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. 9

[66] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3, 4, 5, 8, 9

[67] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023. 3, 4, 9

[68] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3, 4

[69] 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. 8

[70] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2, 6

[71] 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. 9

[72] 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. 8

[73] 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. 8

[74] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 9

[75] 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. 8

[76] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008. 2, 7

[77] 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. 4

[78] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24, 2011. 4, 6

[79] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. 2, 7

[80] 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. 2, 4, 7, 8, 9

[81] 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. 2, 6, 7

[82] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*, pages 3505–3506, 2020. 3, 4

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

[84] 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, 6

[85] 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. 4, 6

[86] 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. 8

[87] Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. Tiny lvlm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*, 2023. 2

[88] 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. 4, 6

[89] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020. 8

[90] 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. 9

[91] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 2, 7

[92] 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. 2

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

[94] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 6

[95] Dmitry Ustalov, Nikita Pavlichenko, Sergey Koshelev, Daniil Likhobaba, and Alisa Smirnova. Toloka visual question answering benchmark. *arXiv preprint arXiv:2309.16511*, 2023. 9

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

[97] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *NeurIPS*, 35: 5696–5710, 2022. 2

[98] 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. 2

[99] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023. 2

[100] 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. 9

[101] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2

[102] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPRW*, pages 3485–3492, 2010. 2, 6

[103] 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. 2

[104] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 1, 2, 5, 7

[105] 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. 2

[106] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. Multilingual universal sentence encoder for semantic retrieval. In *ACL*, pages 87–94, 2020. 1

[107] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021. 2

[108] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 9

[109] 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. 2

[110] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023. 3

[111] 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. 3

[112] 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. 9

[113] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 9

[114] 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

[115] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 5, 8

[116] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2023. 2

[117] 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. 3