Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond

Jinze Bai* Shuai Bai* Shusheng Yang* Shijie Wang Sinan Tan
Peng Wang Junyang Lin Chang Zhou† Jingren Zhou
Alibaba Group

Code & Demo & Models: https://github.com/QwenLM/Qwen-VL

Abstract

In this work, we introduce the Qwen-VL series, a set of large-scale vision-language models (LVLMs) designed to perceive and understand both texts and images. Starting from the Qwen-LM as a foundation, we endow it with visual capacity by the meticulously designed (i) visual receptor, (ii) input-output interface, (iii) 3-stage training pipeline, and (iv) multilingual multimodal cleaned corpus. Beyond the conventional image description and question-answering, we implement the grounding and text-reading ability of Qwen-VLs by aligning image-caption-box tuples. The resulting models, including Qwen-VL and Qwen-VL-Chat, set new records for generalist models under similar model scales on a broad range of visual-centric benchmarks (*e.g.*, image captioning, question answering, visual grounding) and different settings (*e.g.*, zero-shot, few-shot). Moreover, on real-world dialog benchmarks, our instruction-tuned Qwen-VL-Chat also demonstrates superiority compared to existing vision-language chatbots. All models are public to facilitate future research.

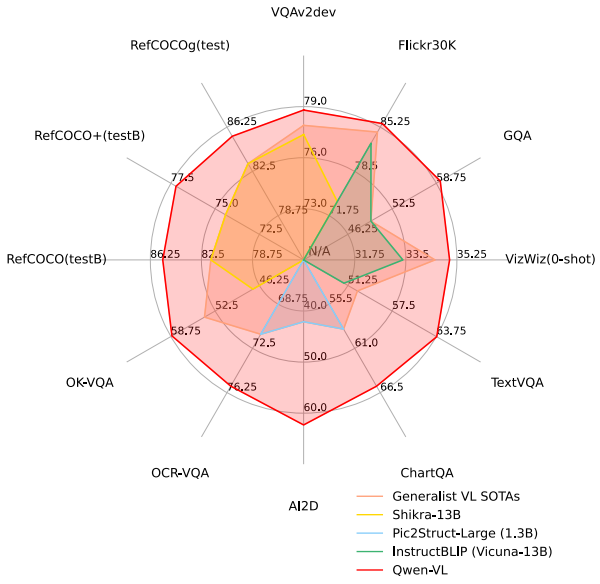

Figure 1: Qwen-VL achieves state-of-the-art performance on a broad range of tasks compared with other generalist models.

*Equal contribution, †Corresponding author

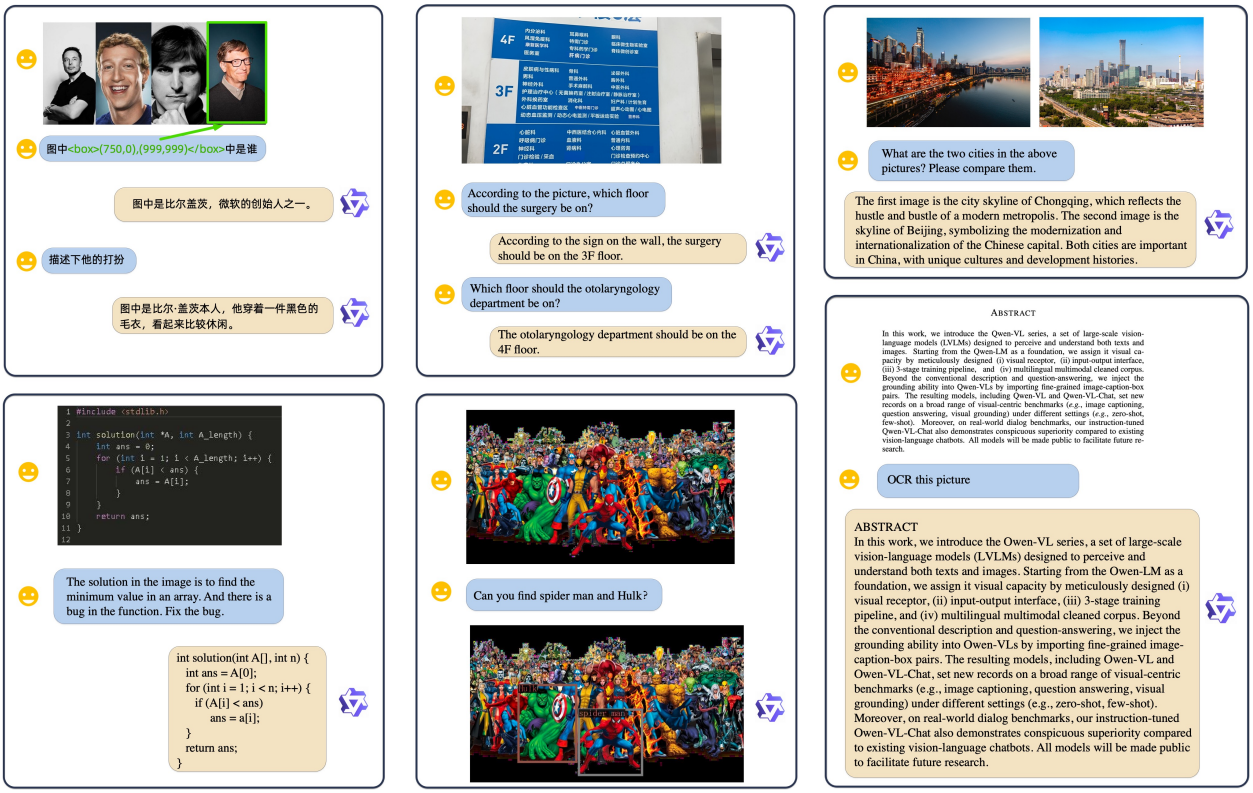

Figure 2: Some qualitative examples generated by our Qwen-VL-Chat. Qwen-VL-Chat supports multiple image inputs, multi-round dialogue, multilingual conversation, text-reading, localization, fine-grained recognition and understanding ability.

1 Introduction

Recently, Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2023; Anil et al., 2023; Gao et al., 2023; Qwen, 2023) have attracted wide attention due to their powerful capabilities in text generation and comprehension. These models can be further aligned with user intent through fine-tuning instructions, showcasing strong interactive capabilities and the potential to enhance productivity as intelligent assistants. However, native large language models only live in the pure-text world, lacking the ability to handle other common modalities (such as images, speech, and videos), resulting in great restrictions on their application scope. Motivated by this, a group of Large Vision Language Models (LVLMs) (Alayrac et al., 2022; Chen et al., 2022; Li et al., 2023c; Dai et al., 2023; Huang et al., 2023; Peng et al., 2023; Zhu et al., 2023; Liu et al., 2023; Ye et al., 2023b,a; Chen et al., 2023a; Li et al., 2023a; Zhang et al., 2023; Sun et al., 2023; OpenAI, 2023) have been developed to enhance large language models with the ability to perceive and understand visual signals. These large-scale vision-language models demonstrate promising potential in solving real-world vision-central problems.

Nevertheless, despite that lots of works have been conducted to explore the limitation and potency of LVLMs, current open-source LVLMs always suffer from inadequate training and optimization, thus lag far behind the proprietary models (Chen et al., 2022, 2023b; OpenAI, 2023), which hinders further exploration and application of LVLMs in open-source community. What's more, as real-world visual scenarios are quite complicated, fine-grained visual understanding plays a crucial role for LVLMs to assist people effectively and precisely. But only a few attempts had been made toward this direction (Peng et al., 2023; Chen et al., 2023a), the majority of open-source LVLMs remain perceiving the image in a coarse-grained approach and lacking the ability to execute fine-grained perception such as object grounding or text reading.

In this paper, we explore a way out and present the newest members of the open-sourced Qwen families: Qwen-VL series. Qwen-VLs are a series of highly performant and versatile vision-language foundation models based on Qwen-7B (Qwen, 2023) language model. We empower the LLM basement with visual capacity by introducing a new visual receptor including a language-aligned visual encoder and a position-aware adapter. The overall model architecture as well as the input-output interface are quite concise and we elaboratedly design a 3-stage training pipeline to optimize the whole model upon a vast collection of image-text corpus.

Our pre-trained checkpoint, termed Qwen-VL, is capable of perceiving and understanding visual inputs, generating desired responses according to given prompts, and accomplishing various vision-language tasks such as image captioning, question answering, text-oriented question answering, and visual grounding. Qwen-VL-Chat is the instruction-tuned vision-language chatbot based on Qwen-VL. As shown in Fig. 2, Qwen-VL-Chat is able to interact with users and perceive the input images following the intention of users.

Specifically, the features of the Qwen-VL series models include:

- Leading performance: Qwen-VLs achieve top-tier accuracy on a vast of vision-centric understanding benchmarks compared to counterparts with similar scales. Besides, Qwen-VL's stuning performance covers not only the conventional benchmarks *e.g.*, captioning, question-answering, grounding), but also some recently introduced dialogue benchmarks.
- Multi-lingual: Similar to Qwen-LM, Qwen-VLs are trained upon multilingual image-text data with a considerable amount of corpus being in English and Chinese. In this way, Qwen-VLs naturally support English, Chinese, and multilingual instructions.
- Multi-image: In the training phase, we allow arbitrary interleaved image-text data as Qwen-VL's inputs. This feature allows our Qwen-Chat-VL to compare, understand, and analyze the context when multiple images are given.
- Fine-grained visual understanding: Thanks to the higher-resolution input size and fine-grained corpus we used in training, Qwen-VLs exhibit highly competitive fine-grained visual understanding ability. Compared to existing vision-language generalists, our Qwen-VLs possess much better grounding, text-reading, text-oriented question answering, and fine-grained dialog performance.

2 Methodology

2.1 Model Architecture

The overall network architecture of Qwen-VL consists of three components and the details of model parameters are shown in Table 1:

Large Language Model: Qwen-VL adopts a large language model as its foundation component. The model is initialized with pre-trained weights from Qwen-7B (Qwen, 2023).

Visual Encoder: The visual encoder of Qwen-VL uses the Vision Transformer (ViT) (Dosovitskiy et al., 2021) architecture, initialized with pre-trained weights from Openclip's ViT-bigG (Ilharco et al., 2021). During both training and inference, input images are resized to a specific resolution. The visual encoder processes images by splitting them into patches with a stride of 14, generating a set of image features.

Position-aware Vision-Language Adapter: To alleviate the efficiency issues arising from long image feature sequences, Qwen-VL introduces a vision-language adapter that compresses the image features. This adapter comprises a single-layer cross-attention module initialized randomly. The module uses a group of trainable vectors (Embeddings) as query vectors and the image features from the visual encoder as keys for cross-attention operations. This mechanism compresses the visual feature sequence to a fixed length of 256. The ablation about the number of queries is shown in Appendix E.2. Additionally, considering the significance

of positional information for fine-grained image comprehension, 2D absolute positional encodings are incorporated into the cross-attention mechanism's query-key pairs to mitigate the potential loss of positional details during compression. The compressed image feature sequence of length 256 is subsequently fed into the large language model.

Table 1: Details of Qwen-VL model parameters.

Vision Encoder	VL Adapter	LLM	Total
1.9B	0.08B	7.7B	9.6B

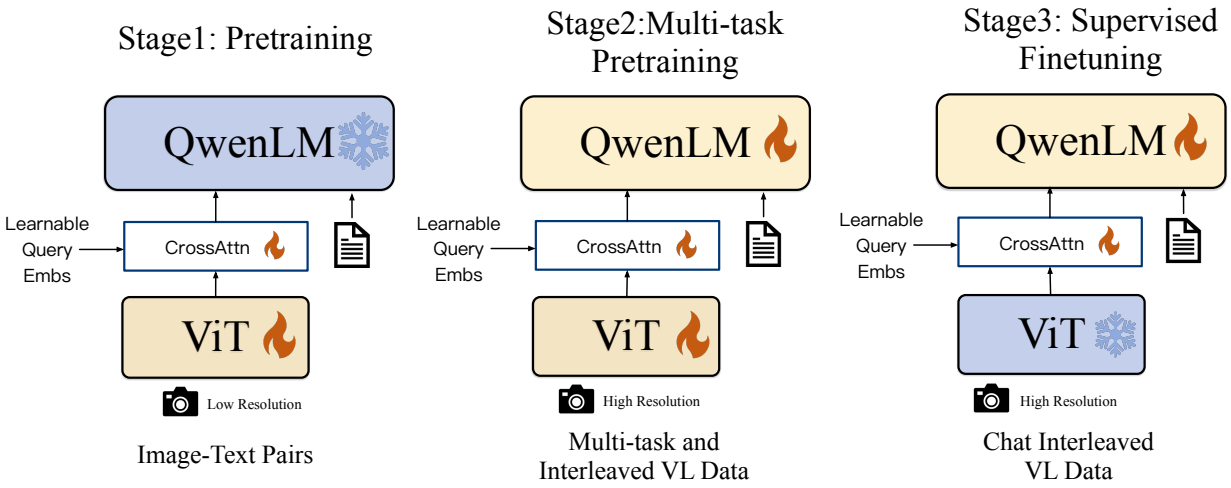

Figure 3: The training pipeline of the Qwen-VL series.

2.2 Inputs and Outputs

Image Input: Images are processed through the visual encoder and adapter, yielding fixed-length sequences of image features. To differentiate between image feature input and text feature input, two special tokens (<img> and </img>) are appended to the beginning and end of the image feature sequence respectively, signifying the start and end of image content.

Bounding Box Input and Output: To enhance the model's capacity for fine-grained visual understanding and grounding, Qwen-VL's training involves data in the form of region descriptions, questions, and detections. Differing from conventional tasks involving image-text descriptions or questions, this task necessitates the model's accurate understanding and generation of region descriptions in a designated format. For any given bounding box, a normalization process is applied (within the range $[0, 1000)$) and transformed into a specified string format: "$(X_{topleft}, Y_{topleft}), (X_{bottomright}, Y_{bottomright})$". The string is tokenized as text and does not require an additional positional vocabulary. To distinguish between detection strings and regular text strings, two special tokens (<box> and </box> are added at the beginning and end of the bounding box string. Additionally, to appropriately associate bounding boxes with their corresponding descriptive words or sentences, another set of special tokens (<ref> and </ref>) is introduced, marking the content referred to by the bounding box.

3 Training

As illustrated in Fig. 3, the training process of the Qwen-VL model consists of three stages: two stages of pre-training and a final stage of instruction fine-tuning training.

3.1 Pre-training

In the first stage of pre-training, we mainly utilize a large-scale, weakly labeled, web-crawled set of image-text pairs. Our pre-training dataset is composed of several publicly accessible sources and some in-house data. We made an effort to clean the dataset of certain patterns. As summarized in Table 2, the original dataset contains a total of 5 billion image-text pairs, and after cleaning, 1.4 billion data remain, with 77.3% English (text) data and 22.7% Chinese (text) data.

Table 2: Details of Qwen-VL pre-training data. LAION-en and LAION-zh are the English and Chinese language subset of LAION-5B (Schuhmann et al., 2022a). LAION-COCO (Schuhmann et al., 2022b) is a synthetic dataset generated from LAION-en. DataComp (Gadre et al., 2023) and Coyo (Byeon et al., 2022) are collections of image-text pairs. CC12M (Changpinyo et al., 2021), CC3M (Sharma et al., 2018), SBU (Ordonez et al., 2011) and COCO Caption (Chen et al., 2015) are academic caption datasets.

Language	Dataset	Original	Cleaned	Remaining%
English	LAION-en	2B	280M	14%
	LAION-COCO	600M	300M	50%
	DataComp	1.4B	300M	21%
	Coyo	700M	200M	28%
	CC12M	12M	8M	66%
	CC3M	3M	3M	100%
	SBU	1M	0.8M	80%
	COCO Caption	0.6M	0.6M	100%
Chinese	LAION-zh	108M	105M	97%
	In-house Data	220M	220M	100%
	Total	5B	1.4B	28%

We freeze the large language model and only optimize the vision encoder and VL adapter in this stage. The input images are resized to 224×224. The training objective is to minimize the cross-entropy of the text tokens. The maximum learning rate is $2e^{-4}$ and the training process uses a batch size of 30720 for the image-text pairs, and the entire first stage of pre-training lasts for 50,000 steps, consuming approximately 1.5 billion image-text samples. More hyperparameters are detailed in Appendix C and the convergence curve of this stage is shown in Figure 6.

3.2 Multi-task Pre-training

In the second stage of multi-task pre-training, we introduce high-quality and fine-grained VL annotation data with a larger input resolution and interleaved image-text data. As summarized in Table 3, we trained Qwen-VL on 7 tasks simultaneously. For text generation, we use the in-house collected corpus to maintain the LLM's ability. Captioning data is the same with Table 2 except for far fewer samples and excluding LAION-COCO. We use a mixture of publicly available data for the VQA task which includes GQA (Hudson and Manning, 2019), VGQA (Krishna et al., 2017), VQAv2 (Goyal et al., 2017), DVQA (Kafle et al., 2018), OCR-VQA (Mishra et al., 2019) and DocVQA (Mathew et al., 2021). We follow Kosmos-2 to use the GRIT (Peng et al., 2023) dataset for the grounding task with minor modifications. For the reference grounding and grounded captioning duality tasks, we construct training samples from GRIT (Peng et al., 2023), Visual Genome (Krishna et al., 2017), RefCOCO (Kazemzadeh et al., 2014), RefCOCO+, and RefCOCOg (Mao et al.,

2016). In order to improve the text-oriented tasks, we collect pdf and HTML format data from Common Crawl[1] and generate synthetic OCR data in English and Chinese language with natural scenery background, following (Kim et al., 2022). Finally, we simply construct interleaved image-text data by packing the same task data into sequences of length 2048.

Table 3: Details of Qwen-VL multi-task pre-training data.

Task	# Samples	Dataset
Captioning	19.7M	LAION-en & zh, DataComp, Coyo, CC12M & 3M, SBU, COCO, In-house Data
VQA	3.6M	GQA, VGQA, VQAv2, DVQA, OCR-VQA, DocVQA, TextVQA, ChartQA, AI2D
Grounding[2]	3.5M	GRIT
Ref Grounding	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
Grounded Cap.	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
OCR	24.8M	SynthDoG-en & zh, Common Crawl pdf & HTML
Pure-text Autoregression	7.8M	In-house Data

We increase the input resolution of the visual encoder from 224×224 to 448×448, reducing the information loss caused by image down-sampling. Besides, we ablate the window attention and global attention for higher resolutions of the vision transformer in Appendix E.3. We unlocked the large language model and trained the whole model. The training objective is the same as the pre-training stage.

3.3 Supervised Fine-tuning

During this stage, we finetuned the Qwen-VL pre-trained model through instruction fine-tuning to enhance its instruction following and dialogue capabilities, resulting in the interactive Qwen-VL-Chat model. The multi-modal instruction tuning data primarily comes from caption data or dialogue data generated through LLM self-instruction, which often only addresses single-image dialogue and reasoning and is limited to image content comprehension. We construct an additional set of dialogue data through manual annotation, model generation, and strategy concatenation to incorporate localization and multi-image comprehension abilities into the Qwen-VL model. We confirm that the model effectively transfers these capabilities to a wider range of languages and question types. Additionally, we mix multi-modal and pure text dialogue data during training to ensure the model's universality in dialogue capabilities. The instruction tuning data amounts to 350k. In this stage, we freeze the visual encoder and optimize the language model and adapter module. We demonstrate the data format of this stage in Appendix B.2.

4 Evaluation

In this section, we conduct an overall evaluation on various multi-modal tasks to comprehensively assess our models' visual understanding ability. In the following, Qwen-VL denotes the model after the multi-task training, and Qwen-VL-Chat denotes the model after supervised fine-tuning (SFT) stage.

Table 9 provides a detailed summary of the used evaluation benchmarks and corresponding metrics.

4.1 Image Caption and General Visual Question Answering

Image caption and general visual question answering (VQA) are two conventional tasks for vision-language models. Specifically, image caption requires the model to generate a description for a given image and general VQA requires the model to generate an answer for a given image-question pair.

[1] https://digitalcorpora.org/corpora/file-corpora/cc-main-2021-31-pdf-untruncated
[2] This task is to generate noun/phrase grounded captions (Peng et al., 2023).

Table 4: Results on Image Captioning and General VQA.

Model Type	Model	Image Caption		General VQA				
		Nocaps (0-shot)	Flickr30K (0-shot)	VQAv2	OKVQA	GQA	SciQA-Img (0-shot)	VizWiz (0-shot)
Generalist Models	Flamingo-9B	-	61.5	51.8	44.7	-	-	28.8
	Flamingo-80B	-	67.2	56.3	50.6	-	-	31.6
	Unified-IO-XL	100.0	-	77.9	54.0	-	-	-
	Kosmos-1	-	67.1	51.0	-	-	-	29.2
	Kosmos-2	-	80.5	51.1	-	-	-	-
	BLIP-2 (Vicuna-13B)	103.9	71.6	65.0	45.9	32.3	61.0	19.6
	InstructBLIP (Vicuna-13B)	**121.9**	82.8	-	-	49.5	63.1	33.4
	Shikra (Vicuna-13B)	-	73.9	77.36	47.16	-	-	-
	Qwen-VL (Qwen-7B)	121.4	**85.8**	**79.5**	**58.6**	**59.3**	67.1	35.2
	Qwen-VL-Chat	120.2	81.0	78.2	56.6	57.5	**68.2**	**38.9**
Specialist SOTAs	-	127.0 (PALI-17B)	84.5 (InstructBLIP -FlanT5-XL)	86.1 (PALI-X -55B)	66.1 (PALI-X -55B)	72.1 (CFR)	92.53 (LLaVa+ GPT-4)	70.9 (PALI-X -55B)

For the image caption task, we choose Nocaps (Agrawal et al., 2019) and Flickr30K (Young et al., 2014) as benchmarks and report CIDEr score (Vedantam et al., 2015) as metric. We utilize greedy search for caption generation with a prompt of *"Descripe the image in English:"*.

For general VQA, we utilize five benchmarks including VQAv2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), GQA (Hudson and Manning, 2019), ScienceQA (Image Set) (Lu et al., 2022b) and VizWiz VQA (Gurari et al., 2018). For VQAv2, OKVQA, GQA and VizWiz VQA, we employ open-ended answer generation with greedy decoding strategy and a prompt of *"{question} Answer:"*, without any constrain on model's output space. However, for ScienceQA, we constrain the model's output to possible options (instead of open-ended), choose the option with highest confidence as model's prediction, and report the Top-1 accuracy.

The overall performance on image caption and general VQA tasks are reported in Table 4. As the results shown, our Qwen-VL and Qwen-VL-Chat both achieve obviously better results compared to previous generalist models in terms of both two tasks. Specifically, on zero-shot image caption task, Qwen-VL achieves state-of-the-art performance (*i.e.*, 85.8 CIDEr score) on the Flickr30K karpathy-test split, even outperforms previous generalist models with much more parameters (*e.g.*, Flamingo-80B with 80B parameters).

On general VQA benchmarks, our models also exhibit distinct advantages compared to others. On VQAv2, OKVQA and GQA benchmarks, Qwen-VL achieves 79.5, 58.6 and 59.3 accuracy respectively, which surpasses recent proposed LVLMs by a large margin. It's worth noting that Qwen-VL also shows strong zero-shot performance on ScienceQA and VizWiz datasets.

4.2 Text-oriented Visual Question Answering

Text-oriented visual understanding has a broad application prospect in real-world scenarios. We assess our models' ability toward text-oriented visual question answering on several benchmarks including TextVQA (Sidorov et al., 2020), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), AI2Diagram (Kembhavi et al., 2016), and OCR-VQA (Mishra et al., 2019). Similarly, the results are shown in Table 5. Compared to previous generalist models and recent LVLMs, our models show better performance on most benchmarks, frequently by a large margin.

4.3 Refer Expression Comprehension

We show our models' fine-grained image understanding and localization ability by evaluating on a sort of refer expression comprehension benchmarks such as RefCOCO (Kazemzadeh et al., 2014), RefCOCOg (Mao et al., 2016), RefCOCO+ (Mao et al., 2016) and GRIT (Gupta et al., 2022). Specifically, the refer expression comprehension task requires the model to localize the target object under the guidance of a description. The

Table 5: Results on Text-oriented VQA.

Model type	Model	TextVQA	DocVQA	ChartQA	AI2D	OCR-VQA
Generalist Models	BLIP-2 (Vicuna-13B)	42.4	-	-	-	-
	InstructBLIP (Vicuna-13B)	50.7	-	-	-	-
	mPLUG-DocOwl (LLaMA-7B)	52.6	62.2	57.4	-	-
	Pix2Struct-Large (1.3B)	-	**76.6**	58.6	42.1	71.3
	Qwen-VL (Qwen-7B)	**63.8**	65.1	65.7	**62.3**	**75.7**
	Qwen-VL-Chat	61.5	62.6	**66.3**	57.7	70.5
Specialist SOTAs	PALI-X-55B (Single-task fine-tuning, without OCR Pipeline)	71.44	80.0	70.0	81.2	75.0

Table 6: Results on Referring Expression Comprehension task.

Model type	Model	RefCOCO			RefCOCO+			RefCOCOg		GRIT
		val	test-A	test-B	val	test-A	test-B	val	test	refexp
Generalist Models	GPV-2	-	-	-	-	-	-	-	-	51.50
	OFA-L*	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	61.70
	Unified-IO	-	-	-	-	-	-	-	-	**78.61**
	VisionLLM-H		86.70	-	-	-	-	-	-	-
	Shikra-7B	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	69.34
	Shikra-13B	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	69.03
	Qwen-VL-7B	**89.36**	92.26	**85.34**	**83.12**	88.25	**77.21**	85.58	85.48	78.22
	Qwen-VL-7B-Chat	88.55	**92.27**	84.51	82.82	**88.59**	76.79	**85.96**	**86.32**	-
Specialist SOTAs	G-DINO-L	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02	-
	UNINEXT-H	92.64	94.33	91.46	85.24	89.63	79.79	88.73	89.37	-
	ONE-PEACE	92.58	94.18	89.26	88.77	92.21	83.23	89.22	89.27	-

results are shown in Table 6. Compared to previous generalist models or recent LVLMs, our models obtain top-tier results on all benchmarks.

4.4 Few-shot Learning on Vision-Language Tasks

Our model also exhibits satisfactory in-context learning (*a.k.a.*, few-shot learning) ability. As shown in Figure 4, Qwen-VL achieves better performance through in-context few-shot learning on OKVQA (Marino et al., 2019), Vizwiz (Gurari et al., 2018), TextVQA (Sidorov et al., 2020), and Flickr30k (Young et al., 2014) when compared with models with similar number of parameters (Flamingo-9B(Alayrac et al., 2022), OpenFlamingo-9B(**?**) and IDEFICS-9B**?**). Qwen-VL's performance is even comparable with much larger models (Flamingo-80B and IDEFICS-80B). Note that we adopt naïve random sample to construct the few-shot exemplars, sophisticated few-shot exemplar construction methods such as RICES (Yang et al., 2022b) are not used despite better results would be achieved.

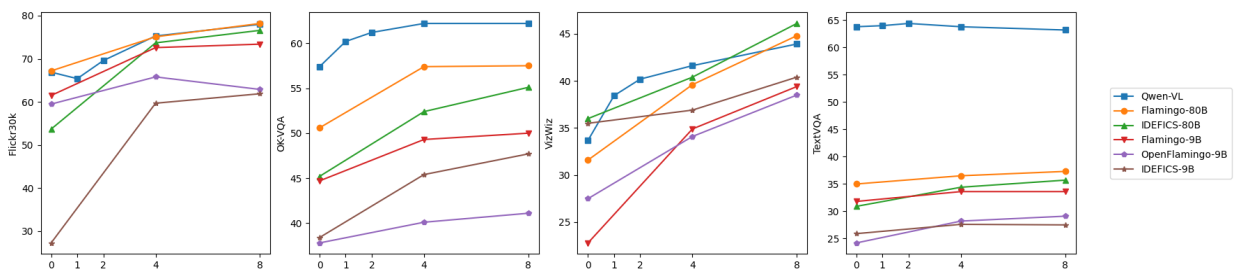

Figure 4: Few-shot learning results of Qwen-VL in comparison with other models.

Table 7: Results on Instruction-following benchmarks.

Model	TouchStone		SEED-Bench			MME	
	En	Cn	All	Img	Video	Perception	Cognition
VisualGLM	-	247.1	-	-	-	705.31	181.79
PandaGPT	488.5	-	-	-	-	642.59	228.57
MiniGPT4	531.7	-	42.8	47.4	29.9	581.67	144.29
InstructBLIP	552.4	-	53.4	58.8	38.1	1212.82	291.79
LLaMA-AdapterV2	590.1	-	32.7	35.2	25.8	972.67	248.93
LLaVA	602.7	-	33.5	37.0	23.8	502.82	214.64
mPLUG-Owl	605.4	-	34.0	37.9	23.0	967.34	276.07
Qwen-VL	-	-	56.3	62.3	**39.1**	-	-
Qwen-VL-Chat	**645.2**	**401.2**	**58.2**	**65.4**	37.8	**1487.58**	**360.71**

4.5 Instruction Following in Real-world User Behavior

In addition to previous conventional vision-language evaluations, to evaluate our Qwen-VL-Chat model's capacity under real-world user behavior, we further conduct the evaluations on the TouchStone (Bai et al., 2023), SEED-Bench (Li et al., 2023b), and MME (Fu et al., 2023). TouchStone is an open-ended vision-language instruction-following benchmark. We compare the instruction-following ability of Qwen-VL-Chat with other instruction-tuned LVLMs in both English and Chinese on the TouchStone benchmark. SEED-Bench consists of 19K multiple-choice questions with accurate human annotations for evaluating Multimodal LLMs, covering 12 evaluation dimensions including both the spatial and temporal understanding. MME measures both perception and cognition abilities on a total of 14 subtasks.

The results on three benchmarks are shown in Table 7. Qwen-VL-Chat has achieved obvious advantages over other LVLMs on all three datasets, indicating that our model performs better in understanding and answering diverse user instructions. In SEED-Bench, we have found that our model's visual capabilities can be effectively transferred to video tasks by simply sampling four frames. In terms of the overall scores presented in TouchStone, our model demonstrates a clear advantage compared to other LVLMs, especially in terms of its Chinese capabilities. In terms of the broad categories of abilities, our model exhibits a more pronounced advantage in understanding and recognition, particularly in areas such as text recognition and chart analysis. For more detailed information, please refer to the TouchStone dataset.

5 Related Work

In recent years, researchers have shown considerable interest in vision-language learning (Su et al., 2019; Chen et al., 2020; Li et al., 2020; Zhang et al., 2021; Li et al., 2021b; Lin et al., 2021; Kim et al., 2021; Dou et al., 2022; Zeng et al., 2021; Li et al., 2021a, 2022), especially in the development of multi-task generalist models (Hu and Singh, 2021; Singh et al., 2022; Zhu et al., 2022; Yu et al., 2022; Wang et al., 2022a; Lu et al., 2022a; Bai et al., 2022). CoCa (Yu et al., 2022) proposes an encoder-decoder structure to address image-text retrieval and vision-language generation tasks simultaneously. OFA (Wang et al., 2022a) transforms specific vision-language tasks into sequence-to-sequence tasks using customized task instructions. Unified I/O (Lu et al., 2022a) further introduces more tasks like segmentation and depth estimation into a unified framework. Another category of research focuses on building vision-language representation models (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; Yuan et al., 2021; Yang et al., 2022a). CLIP (Radford et al., 2021) leverages contrastive learning and large amounts of data to align images and language in a semantic space, resulting in strong generalization capabilities across a wide range of downstream tasks. BEIT-3 (Wang et al., 2022b) employs a mixture-of-experts (MOE) structure and unified masked token prediction objective, achieving state-of-the-art results on various visual-language tasks. In addition to vision-language learning, ImageBind (Girdhar et al., 2023) and ONE-PEACE (Wang et al., 2023) align more modalities such as speech into a unified semantic space, thus creating more general representation models.

Despite achieving significant progress, previous vision-language models still have several limitations such

as poor robustness in instruction following, limited generalization capabilities in unseen tasks, and a lack of in-context abilities. With the rapid development of large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; Anil et al., 2023; Gao et al., 2023; Qwen, 2023), researchers have started building more powerful large vision-language models (LVLMs) based on LLMs (Alayrac et al., 2022; Chen et al., 2022; Li et al., 2023c; Dai et al., 2023; Huang et al., 2023; Peng et al., 2023; Zhu et al., 2023; Liu et al., 2023; Ye et al., 2023b,a; Chen et al., 2023a; Li et al., 2023a; Zhang et al., 2023; Sun et al., 2023). BLIP-2 (Li et al., 2023c) proposes Q-Former to align the frozen vision foundation models and LLMs. Meanwhile, LLAVA (Liu et al., 2023) and MiniGPT4 (Zhu et al., 2023) introduce visual instruction tuning to enhance instruction following capabilities in LVLMs. Additionally, mPLUG-DocOwl (Ye et al., 2023a) incorporates document understanding capabilities into LVLMs by introducing digital documents data. Kosmos2 (Peng et al., 2023), Shikra (Chen et al., 2023a), and BuboGPT (Zhao et al., 2023) further enhance LVLMs with visual grounding abilities, enabling region description and localization. In this work, we integrate image captioning, visual question answering, OCR, document understanding, and visual grounding capabilities into Qwen-VL. The resulting model achieves outstanding performance on these diverse style tasks.

6 Conclusion and Future Work

We release the Qwen-VL series, a set of large-scale multilingual vision-language models that aims to facilitate multimodal research. Qwen-VL outperforms similar models across various benchmarks, supporting multilingual conversations, multi-image interleaved conversations, grounding in Chinese, and fine-grained recognition. Moving forward, we are dedicated to further enhancing Qwen-VL's capabilities in several key dimensions:

- Integrating Qwen-VL with more modalities, such as speech and video.
- Augmenting Qwen-VL by scaling up the model size, training data and higher resolution, enabling it to handle more complex and intricate relationships within multimodal data.
- Expanding Qwen-VL's prowess in multi-modal generation, specifically in generating high-fidelity images and fluent speech.

References

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*, 2019.

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. In *NeurIPS*, 2022.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv:2305.10403*, 2023.

Jinze Bai, Rui Men, Hao Yang, Xuancheng Ren, Kai Dang, Yichang Zhang, Xiaohuan Zhou, Peng Wang, Sinan Tan, An Yang, et al. Ofasys: A multi-modal multi-task learning system for building generalist models. *arXiv:2212.04408*, 2022.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv:2308.16890*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022. URL https://github.com/kakaobrain/coyo-dataset.

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*, 2021.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023a.

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. *arXiv:2209.06794*, 2022.

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*, 2023b.

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:1504.00325*, 2015.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Zi-Yi* Dou, Aishwarya* Kamath, Zhe* Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*, 2022.

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:2306.13394*, 2023.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv:2304.14108*, 2023.

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:2304.15010*, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Google. Puppeteer, 2023. URL https://github.com/puppeteer/puppeteer.

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*, 2017.

Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark. *arXiv:2204.13653*, 2022.

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*, 2018.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *ICCV*, 2021.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv:2302.14045*, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

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, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv:2102.05918*, 2021.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, 2022.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2017.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*, 2023b.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023c.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*, 2021b.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. In *KDD*, 2021.

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*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv:2206.08916*, 2022a.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022b.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

Openai. Chatml documents. URL https://github.com/openai/openai-python/blob/main/chatml.md.

OpenAI. Gpt-4 technical report, 2023.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011.

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:2306.14824*, 2023.

Qwen. Introducing qwen-7b: Open foundation and human-aligned models (of the state-of-the-arts), 2023. URL https://github.com/QwenLM/Qwen-7B.

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*, 2021.

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. *arXiv:2210.08402*, 2022a.

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/*, 2022b.

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.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022.

Artifex Software. Pymupdf, 2015. URL https://github.com/pymupdf/PyMuPDF.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019.

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. *arXiv:2307.05222*, 2023.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022a.

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:2305.11172*, 2023.

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 all vision and vision-language tasks. *arXiv:2208.10442*, 2022b.

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

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, 2022b.

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:2307.02499*, 2023a.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178*, 2023b.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *ACL*, 2014.

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

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv:2111.11432*, 2021.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv:2111.08276*, 2021.

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*, 2022.

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

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv:2307.08581*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023.

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, 2022.

A Dataset details

A.1 Image-text pairs

We use web-crawled image-text pairs dataset for pre-training, which includes LAION-en (Schuhmann et al., 2022a), LAION-zh (Schuhmann et al., 2022a), LAION-COCO (Schuhmann et al., 2022b), DataComp (Gadre et al., 2023) and Coyo (Byeon et al., 2022). We clean these noisy data by several steps:

1. Removing pairs with too large aspect ratio of the image
2. Removing pairs with too small image
3. Removing pairs with a harsh CLIP score (dataset-specific)
4. Removing pairs with text containing non-English or non-Chinese characters
5. Removing pairs with text containing emoji characters
6. Removing pairs with text length too short or too long
7. Cleaning the text's HTML-tagged part
8. Cleaning the text with certain unregular patterns

For academic caption datasets, we remove pairs whose text contains the special tags in CC12M (Changpinyo et al., 2021) and SBU (Ordonez et al., 2011). If there is more than one text matching the same image, we select the longest one.

A.2 VQA

For the VQAv2 (Goyal et al., 2017) dataset, we select the answer annotation based on the maximum confidence. For other VQA datasets, we didn't do anything special.

A.3 Grounding

For the GRIT (Peng et al., 2023) dataset, we found that there are many recursive grounding box labels in one caption. We use the greedy algorithm to clean the caption to make sure each image contains the most box labels with no recursive box labels. For other grounding datasets, we simply concatenate the noun/phrase with respective bounding box coordinates.

A.4 OCR

We generated the synthetic OCR dataset using Synthdog (Kim et al., 2022). Specifically, we use the COCO (Lin et al., 2014) train2017 and unlabeld2017 dataset split as the natural scenery background. Then we selected 41 English fonts and 11 Chinese fonts to generate text. We use the default hyperparameters as in Synthdog. We track the generated text locations in the image and convert them to quadrilateral coordinates and we also use these coordinates as training labels. The visualization example is illustrated in the second row of Fig 5.

For all the PDF data we collected, we follow the steps below to pre-process the data using PyMuPDF (Software, 2015) to get the rendering results of each page in a PDF file as well as all the text annotations with their bounding boxes.

1. Extracting all texts and their bounding boxes for each page.

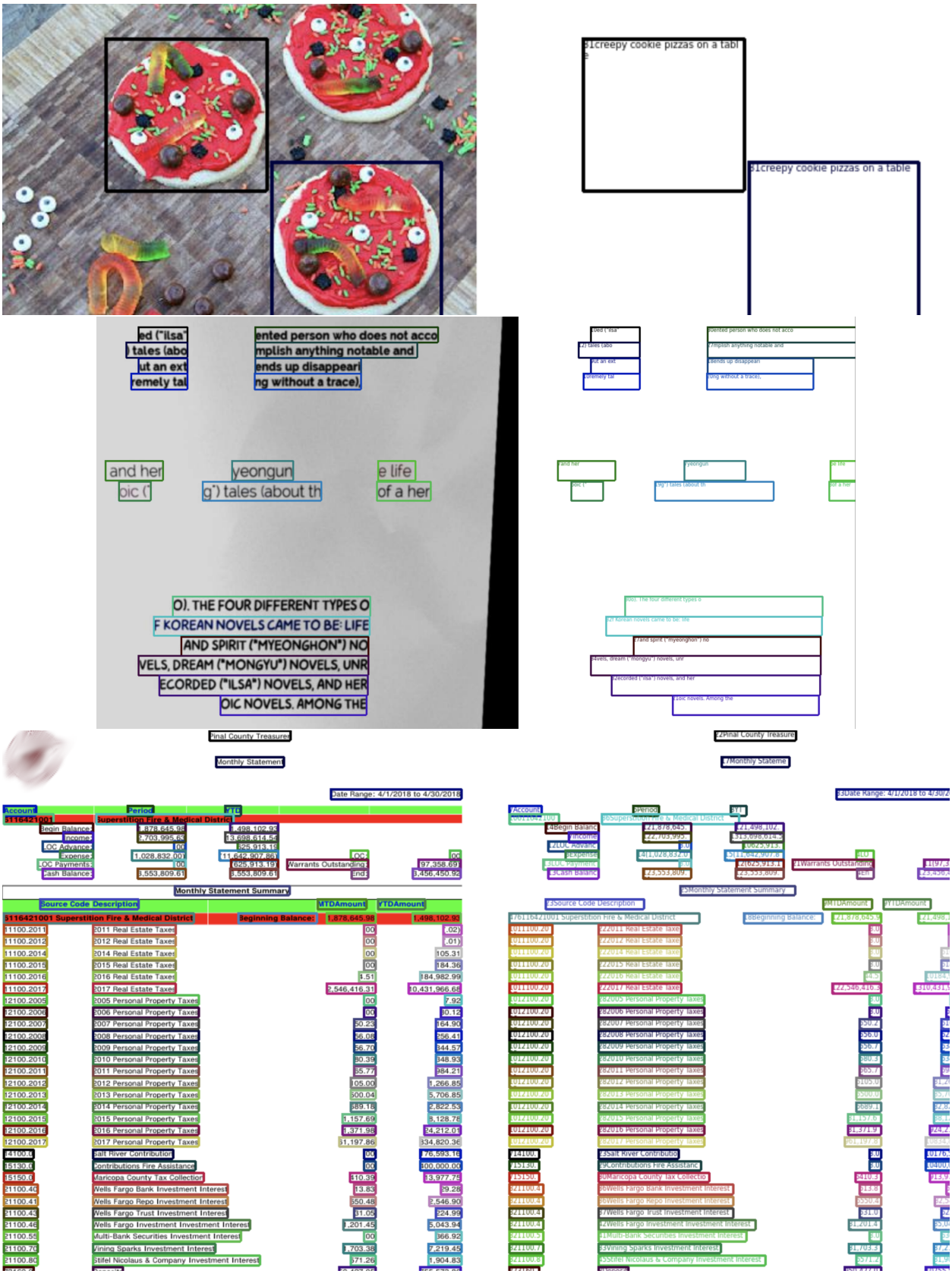

Figure 5: Visualization of the Grounding and OCR data used for training Qwen-VL

2. Rendering each page and save them as an image file.
3. Removing too small image.
4. Removing images with too many or too few characters.
5. Removing images containing Unicode characters in the "Latin Extended-A" and "Latin Extended-B" blocks.
6. Removing images containing Unicode characters in the "Private Use Area (PUA)" block.

For all HTML web pages we collected, we pre-process them in a similar approach to all the PDF data we collected, but we use Puppeteer (Google, 2023) instead of PyMuPDF to render these HTML pages and get the ground truth annotation. We follow the steps below to pre-process the data.

1. Extracting all texts for each webpage.
2. Rendering each page and save them as an image file.
3. Removing too small image.
4. Removing images with too many or too few characters.
5. Removing images containing Unicode characters in the "Private Use Area (PUA)" block.

B Data Format Details of Training

B.1 Data Format of Multi-Task Pre-training

We visualize the Multi-Task Pre-training data format in Box B.1. The Box contains all 7 tasks with the black-colored text as the prefix sequence without loss and blue-colored text as the ground truth labels with loss.

Image Captioning

<img>cc3m/01581435.jpg</img>Generate the caption in English: the beautiful flowers for design.<eos>

Vision Question Answering

<img>VG_100K_2/1.jpg</img> Does the bandage have a different color than the wrist band? Answer: No, both the bandage and the wrist band are white.<eos>

OCR VQA

<img>ocr_vqa/1.jpg</img> What is the title of this book? Answer: Asi Se Dice!, Volume 2: Workbook And Audio Activities (Glencoe Spanish) (Spanish Edition)<eos>

Caption with Grounding

<img>coyo700m/1.jpg</img>Generate the caption in English with grounding: Beautiful shot of <ref>bees</ref><box>(661,612),(833,812)</box><box>(120,555),(265,770) </box> gathering nectars from <ref>an apricot flower</ref><box>(224,13),(399,313) </box><eos>

Referring Grounding

<img>VG_100K_2/3.jpg</img><ref>the ear on a giraffe</ref><box>(176,106),(232,160) </box><eos>

Grounded Captioning

<img>VG_100K_2/4.jpg</img><ref>This</ref><box>(360,542),(476,705)</box> is Yellow cross country ski racing gloves<eos>

OCR

<img>synthdog/1.jpg</img>OCR with grounding: <ref>It is managed</ref> <quad> (568,121), (625,131), (624,182), (567,172)</quad>...<eos>

B.2 Data Format of Supervised Fine-tuning

To better accommodate multi-image dialogue and multiple image inputs, we add the string "Picture id:" before different images, where the id corresponds to the order of image input dialogue. In terms of dialogue format, we construct our instruction tuning dataset using the ChatML (Openai) format, where each interaction's statement is marked with two special tokens (<im_start> and <im_end>) to facilitate dialogue termination.

The Dataset Format Example of ChatML

<im_start>user
Picture 1: <img>vg/VG_100K_2/649.jpg</img>What is the sign in the picture?<im_end>
<im_start>assistant
The sign is a road closure with an orange rhombus.<im_end>
<im_start>user
How is the weather in the picture?<im_end>
<im_start>assistant
The shape of the road closure sign is an orange rhombus.<im_end>

During training, we ensure the consistency between prediction and training distributions by only supervising answers and special tokens (blue in the example), and not supervising role names or question prompts.

C Hyperparameters

We report the detailed training hyperparameter settings of Qwen-VL in Table 8.

Table 8: Training hyperparameters of Qwen-VL

Configuration	Pre-training	Multi-task Pre-training	Supervised Fine-tuning
ViT init.	Open-CLIP-bigG	Qwen-VL 1st-stage	Qwen-VL 2nd-stage
LLM init.	Qwen-7B	Qwen-7B	Qwen-VL 2nd-stage
VL Adapter init.	random	Qwen-VL 1st-stage	Qwen-VL 2nd-stage
Image resolution	224^2	448^2	448^2
ViT sequence length	256	1024	1024
LLM sequence length	512	2048	2048
Learnable query numbers	256	256	256
Optimizer		AdamW	
Optimizer hyperparameter		$\beta_1 = 0.9, \beta_2 = 0.98, eps = 1e^{-6}$	
Peak learning rate	$2e^{-4}$	$5e^{-5}$	$1e^{-5}$
Minimum learning rate	$1e^{-6}$	$1e^{-5}$	$1e^{-6}$
ViT learning rate decay	0.95	0.95	0
ViT Drop path rate		0	
Learning rate schedule		cosine decay	
Weight decay		0.05	
Gradient clip		1.0	
Training steps	50k	19k	8k
Warm-up steps	500	400	3k
Global batch size	30720	4096	128
Gradient Acc.	6	8	8
Numerical precision		`bfloat16`	
Optimizer sharding		✓	
Activation checkpointing		✗	
Model parallelism	✗	2	2
Pipeline parallelism		✗	

In the first pre-training stage, the model is trained using AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98, eps = 1e^{-6}$. We use the cosine learning rate schedule and set the maximum learning rate of $2e^{-4}$ and minimum of $1e^{-6}$ with a linear warm-up of 500 steps. We use a weight decay of $5e^{-2}$ and a gradient clipping of 1.0. For the ViT image encoder, we apply a layer-wise learning rate decay strategy with a decay factor of 0.95. The training process uses a batch size of 30720 for the image-text pairs, and the entire first stage of pre-training lasts for 50,000 steps, consuming approximately 1.5 billion image-text samples and 500 billion image-text tokens.

In the second multi-task training stage, we increase the input resolution of the visual encoder from 224×224 to 448×448, reducing the information loss caused by image down-sampling. We unlocked the large language model and trained the whole model. The training objective is the same as the pre-training stage. We use AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98, eps = 1e^{-6}$. We trained for 19000 steps with 400 warm-up steps and a cosine learning rate schedule. Specifically, we use the model parallelism techniques for ViT and LLM.

D Summary of the evaluation benchmarks

We provide a detailed summary of the used evaluation benchmarks and corresponding metrics in Table 9.

Table 9: Summary of the evaluation benchmarks.

Task	Dataset	Description	Split	Metric
Image Caption	Nocaps	Captioning of natural images	val	CIDEr(↑)
	Flickr30K	Captioning of natural images	karpathy-test	CIDEr(↑)
General VQA	VQAv2	VQA on natural images	test-dev	VQA Score(↑)
	OKVQA	VQA on natural images requiring outside knowledge	val	VQA Score(↑)
	GQA	VQA on scene understanding and reasoning	test-balanced	EM(↑)
	ScienceQA-Img	Multi-choice VQA on a diverse set of science topics	test	Accuracy(↑)
	VizWiz	VQA on photos taken by people who are blind	test-dev	VQA Score(↑)
Text-oriented VQA	TextVQA	VQA on natural images containing text	val	VQA Score(↑)
	DocVQA	VQA on images of scanned documents	test	ANLS(↑)
	ChartQA	VQA on images of charts	test	Relaxed EM(↑)
	OCRVQA	VQA on images of book covers	test	EM(↑)
	AI2Diagram	VQA on images of scientific diagrams	test	EM(↑)
Refer Expression Comprehension	RefCOCO	Refer grounding on natural images	val & testA & testB	Accuracy(↑)
	RefCOCO+	Refer grounding on natural images	val & testA & testB	Accuracy(↑)
	RefCOCOg	Refer grounding on natural images	val & test	Accuracy(↑)
	GRiT	Refer grounding on natural images	test	Accuracy(↑)
Instruction Following	TouchStone	Open-ended VL instruction following benchmark	English & Chinese	GPT-4 Score (↑)
	MME	Open-ended VL Benchmark by yes/no questions	Perception & Cognition	Accuracy (↑)
	Seed-Bench	Open-ended VL Benchmark by Multi-choice VQA	Image & Video	Accuracy (↑)

E Additional experimental details

E.1 Convergence of the Pre-training Stage

In Figure 6, we show the convergence of the Pre-training Stage (stage one). The whole models are trained using BFloat16 mixed precision, the batch size is 30720, and the learning rate is $2e^{-4}$. All images are only trained once (one epoch). The training loss decreases steadily with the increase of the number of training pictures. Note that, the pre-training stage (Stage one) has no VQA data being added, but the Zero-shot VQA score increases amidst fluctuations.

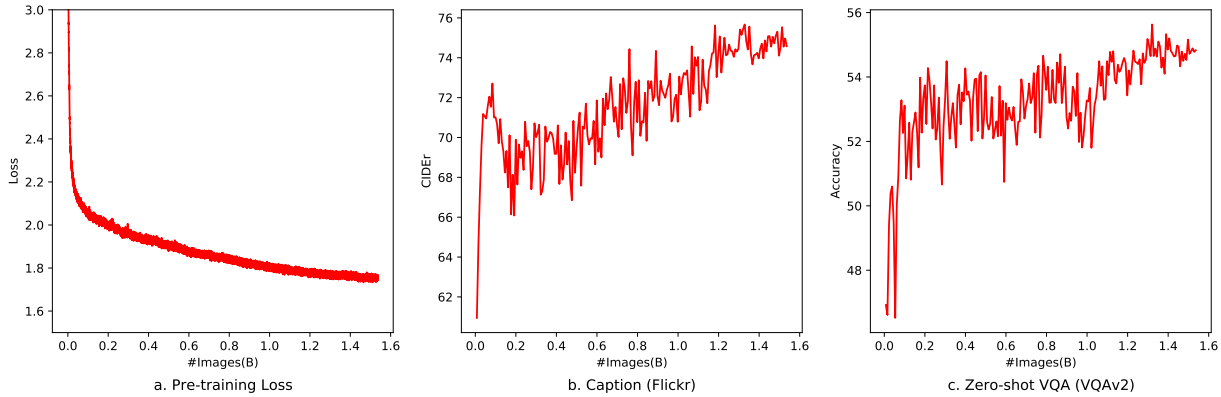

Figure 6: Visualization of the Convergence of the Pre-training Stage

E.2 Number of Learnable Queries in the Vision-Language Adapter

The vision-language adapter uses cross-attention to compress the visual feature sequence by a set of learning queries of length. Too few queries can lead to the loss of some visual information, while too many queries may reduce in greater convergence difficulty and computational cost.

An ablation experiment is conducted on the number of learnable queries in the vision-language adapter. We

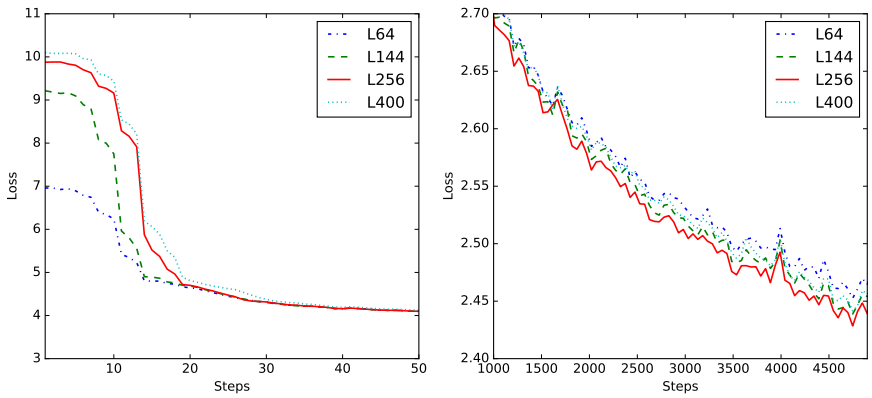

Figure 7: Visualization of the training loss when using different compressed feature lengths of the vision-language adapter. The left depicts the initial training loss (within 50 steps), and the right depicts the loss in convergence (1k-5k steps). In the legend, L64 denotes that the adapter uses 64 queries to compress the visual feature sequence to a fixed length of 64, and so on. The loss curves have been smoothed to avoid shading owing to fluctuations.

used ViT-L/14 as the visual encoder and the 224×224 resolution picture as input, so the sequence length of ViT's output is $(224/14)^2 = 256$. As shown in the left part of Figure 7, the fewer queries used at the beginning of training, the lower the initial loss. However, with convergence, too many or too few queries will cause convergence to slow down, as shown in the right part of Figure 7. Considering that the second training stage (Multi-task Pre-train) applies 448*448 resolution, where the sequence length of ViT's output is $(448/14)^2 = 1024$. Too few queries can result in more information being lost. We finally chose to use 256 queries for the vision-language adapter in Qwen-VL.

E.3 Window Attention vs Global Attention for Vision Transformer

Using a high-resolution Vision Transformer in the model will significantly increase the computational cost. One possible solution to reduce the computational cost of the model is to use Window Attention in the Vision Transformer, i.e., to perform Attention only in a window of 224×224 in most layers of the ViT part of the model, and to perform Attention for the full 448×448 or 896×896 image in a small number of layers (e.g. 1 out of every 4 layers) of the ViT part of the model.

To this end, we conducted ablation experiments to compare the performance of the model when using Global Attention and Window Attention for ViT. We compare the experimental results for analysing the trade-off between computational efficiency and convergence of the model.

Table 10: Training speed of Window Attention vs Global Attention for different input image resolutions

Model input resolution & Attention type	Training speed
448×448, Global Attention	10s / iter
448×448, Window Attention	9s / iter
896×896, Global Attention	60s / iter
896×896, Window Attention	25s / iter

As shown in Figure 8 and Table 10, the loss of the model is significantly higher when Window Attention instead of Vanilla Attention is used. And the training speeds for both of them are similar. Therefore, we decided to use Vanilla Attention instead of Window Attention for the Vision Transformer when training Qwen-VL.

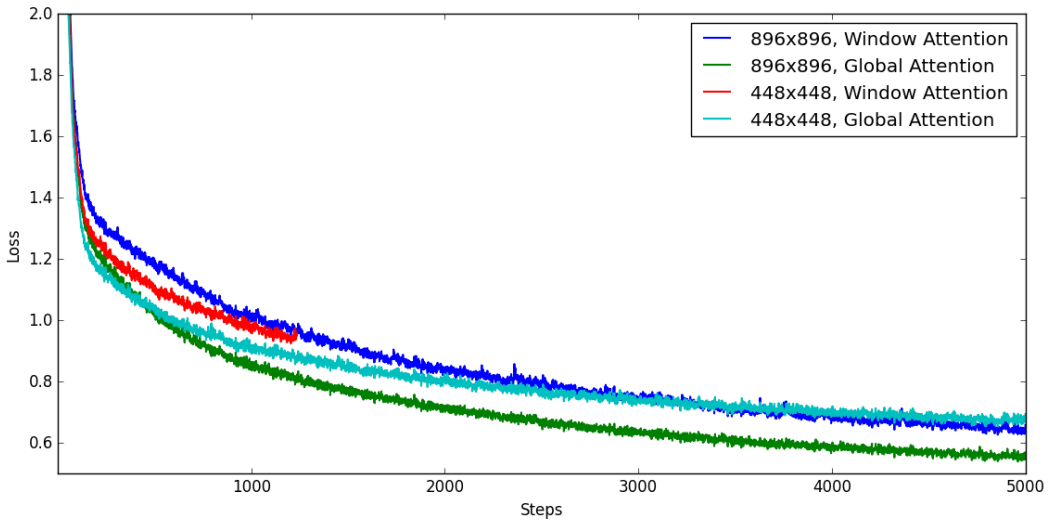

Figure 8: Visualization of the Loss when using Window Attention vs Global Attention

The reason we don't use Window Attention with 896×896 resolution is that its training speed is too slow for us. Although it reaches a loss value similar to model with 448×448 resolution input at 5000 steps. It takes almost 2.5 times longer to train than the model with 448×448 resolution input.

E.4 Performance on Pure-text Tasks

In order to study the effect of multi-modal training on pure-text ability, we show the performance of pure-text tasks of Qwen-VL compared to open-source LLM in Table 11.

Qwen-VL uses an intermediate checkpoint of Qwen-7B as the LLM initialization. The reason why we did not use the final released checkpoint of Qwen-7B is that Qwen-VL and Qwen-7B were developed at a very similar period. Because Qwen-VL has a good initialization on LLM by Qwen-7B, it is comparable to many text-only LLMs on pure-text tasks.

Table 11: Performance on Pure-text Benchmarks of Qwen-VL compared to open-source LLM. Due to the introduction of pure-text data in the multi-task training and SFT stage, Qwen-VL do not compromise any pure-text ability.

Model	MMLU	CMMLU	C-Eval
LLaMA-7B	35.1	26.8	-
LLaMA2-7B	46.8	31.8	32.5
Baichuan-7B	42.3	44.4	42.8
Baichuan2-7B	54.2	57.1	54.0
ChatGLM2-6B	47.9	48.8	51.7
InternLM-7B	51.0	51.8	52.8
Qwen-7B (final released)	58.2	62.2	63.5
Qwen-7B (intermediate, use as Qwen-VL's LLM initialization)	49.9	-	48.5
Qwen-VL	50.7	49.5	51.1

Furthermore, in the multi-task training and SFT stages, Qwen-VL not only utilizes visual and language-related data but also incorporates pure-text data for training. The purpose of this is to prevent the catastrophic

forgetting of text comprehension by leveraging the information from pure-text data. The results in Table 11 indicate that the Qwen-VL model does not exhibit any degradation in terms of its pure text capability and even demonstrates improvement after multi-task training.