

Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution

Peng Wang* Shuai Bai* Sinan Tan* Shijie Wang* Zhihao Fan* Jinze Bai*†
 Keqin Chen Xuejing Liu Jialin Wang Wenbin Ge Yang Fan Kai Dang Mengfei Du
 Xuancheng Ren Rui Men Dayiheng Liu Chang Zhou Jingren Zhou Junyang Lin†
Qwen Team Alibaba Group

Abstract

We present the Qwen2-VL Series, an advanced upgrade of the previous Qwen-VL models that redefines the conventional predetermined-resolution approach in visual processing. Qwen2-VL introduces the Naive Dynamic Resolution mechanism, which enables the model to dynamically process images of varying resolutions into different numbers of visual tokens. This approach allows the model to generate more efficient and accurate visual representations, closely aligning with human perceptual processes. The model also integrates Multimodal Rotary Position Embedding (M-RoPE), facilitating the effective fusion of positional information across text, images, and videos. We employ a unified paradigm for processing both images and videos, enhancing the model's visual perception capabilities. To explore the potential of large multimodal models, Qwen2-VL investigates the scaling laws for large vision-language models (LVLMs). By scaling both the model size-with versions at 2B, 8B, and 72B parameters-and the amount of training data, the Qwen2-VL Series achieves highly competitive performance. Notably, the Qwen2-VL-72B model achieves results comparable to leading models such as GPT-4o and Claude3.5-Sonnet across various multimodal benchmarks, outperforming other generalist models. Code is available at <https://github.com/QwenLM/Qwen2-VL>.

1 Introduction

In the realm of artificial intelligence, Large Vision-Language Models (LVLMs) represent a significant leap forward, building upon the strong textual processing capabilities of traditional large language models. These advanced models now encompass the ability to interpret and analyze a broader spectrum of data, including images, audio, and video. This expansion of capabilities has transformed LVLMs into indispensable tools for tackling a variety of real-world challenges. Recognized for their unique capacity to condense extensive and intricate knowledge into functional representations, LVLMs are paving the way for more comprehensive cognitive systems. By integrating diverse data forms, LVLMs aim to more closely mimic the nuanced ways in which humans perceive and interact with their environment. This allows these models to provide a more accurate representation of how we engage with and perceive our environment.

Recent advancements in large vision-language models (LVLMs) (Li et al., 2023c; Liu et al., 2023b; Dai et al., 2023; Zhu et al., 2023; Huang et al., 2023a; Bai et al., 2023b; Liu et al., 2023a; Wang et al., 2023b; OpenAI., 2023; Team et al., 2023) have led to significant improvements in a short span. These models (OpenAI, 2023; Touvron et al., 2023a,b; Chiang et al., 2023; Bai et al., 2023a) generally follow a common approach of *visual encoder*→*cross-modal connector*→*LLM*. This setup, combined with next-token prediction as the primary training method and the availability of high-quality datasets (Liu et al., 2023a; Zhang et al., 2023; Chen et al., 2023b;

*Equal core contribution, †Corresponding author

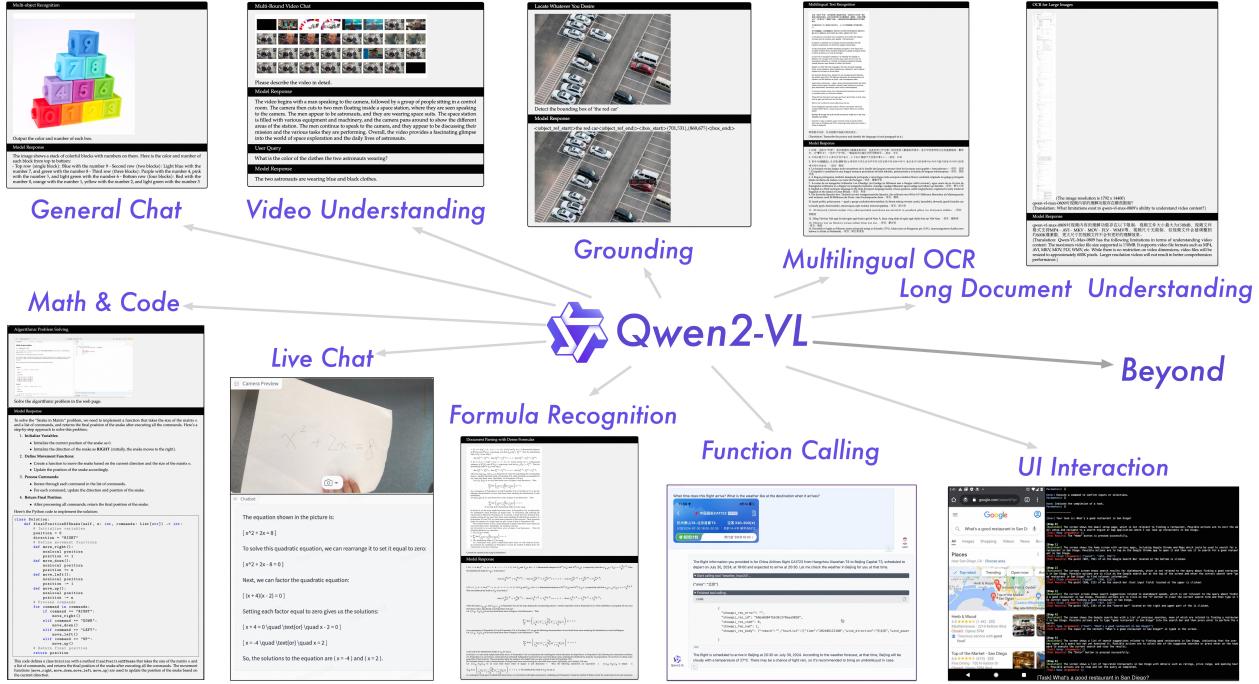


Figure 1: Qwen2-VL capabilities: Multilingual image text understanding, code/math reasoning, video analysis, live chat, agent potential, and more. See Appendix for details.

Li et al., 2023b), has driven much of the progress. Additional factors like larger model architectures (Alayrac et al., 2022), higher-resolution images (Li et al., 2023a,d), and advanced techniques such as mixture-of-expert models (MoE) (Wang et al., 2023b; Ye et al., 2023b), model ensembles (Lin et al., 2023), and more sophisticated connectors (Ye et al., 2023a) between visual and textual modalities have also played a key role in enhancing LVLMs’ ability to process complex visual and textual information more effectively.

However, current large vision-language models (LVLMs) are typically constrained by a fixed image input size. Standard LVLMs encode input images to a fixed resolution (e.g., 224×224), often by either downsampling or upsampling the images (Zhu et al., 2023; Huang et al., 2023a), or by employing a scale-then-padding approach (Liu et al., 2023b,a). While this one-size-fits-all strategy enables processing of images at consistent resolutions, it also limits the model’s ability to capture information at different scales, particularly leading to a significant loss of detailed information in high-resolution images. Consequently, such models fall short of perceiving visual information with the same sensitivity to scale and detail as human vision.

Additionally, most LVLMs rely on a static, frozen CLIP-style (Radford et al., 2021) vision encoder, raising concerns about whether the visual representations produced by such pre-trained models are adequate, particularly for complex reasoning tasks and processing intricate details within images. Recent works (Bai et al., 2023b; Ye et al., 2023a) have attempted to address these limitations by fine-tuning the vision transformer (ViT) during the LVLM training process, which has shown to yield improved results. To further enhance the model’s adaptability to varying resolutions, we introduce dynamic resolution training in the LVLM training process. Specifically, we employ a 2D Rotary Position Embedding (RoPE) in the ViT, thus allowing the model to better capture information across different spatial scales.

When it comes to video content, which is essentially a sequence of frames, many existing models continue to treat it as an independent modality. However, understanding the dynamic nature of reality, as manifested in videos, is crucial for models aiming to grasp the complexities of the real world. Unlike text, which is inherently one-dimensional, the real-world environment exists in three dimensions. The use of one-dimensional position embeddings in current models significantly limits their ability to model three-dimensional space and temporal dynamics effectively. To bridge this gap, we have developed Multimodal Rotary Position Embedding (M-

Table 1: Model descriptions of Qwen2-VL.

Model Name	Vision Encoder	LLM	Model Description
Qwen2-VL-2B	675M	1.5B	The most efficient model, designed to run on-device. It delivers adequate performance for most scenarios with limited resources.
Qwen2-VL-7B	675M	7.6B	The performance-optimized model in terms of cost, significantly upgraded for text recognition and video understanding capabilities. It delivers significant performance across a broad range of visual tasks.
Qwen2-VL-72B	675M	72B	The most capable model, further improvements in visual reasoning, instruction-following, decision-making, and agent capabilities. It delivers optimal performance on most complex tasks.

RoPE), which employs separate components to represent temporal and spatial information. This enables the model to naturally comprehend dynamic content, such as videos or streaming data, improving its ability to understand and interact with the world.

Furthermore, compared to the scaling of large language models (LLMs), current LVLMs are still in the early stages of exploring the impact of scaling in terms of training data and model parameters. The exploration of scaling laws for LVLMs—how increases in model and data size affect performance—remains an open and promising area of research.

In this work, we introduce the newest addition to the large vision-language models of the Qwen family: Qwen2-VL series, which comprises three open-weight models with total parameter counts of 2 billion, 8 billion, and 72 billion. As shown in Figure 1, the key advances in Qwen2-VL include:

- **State-of-the-art understanding across various resolutions and aspect ratios:** Qwen2-VL achieves leading performance on visual benchmarks, including DocVQA, InfoVQA, RealWorldQA, MTVQA, MathVista, and others.
- **Comprehension of extended-duration videos (20 min+):** Qwen2-VL is capable of understanding videos over 20 minutes in length, enhancing its ability to perform high-quality video-based question answering, dialogue, content creation, and more.
- **Robust agent capabilities for device operation:** With advanced reasoning and decision-making abilities, Qwen2-VL can be integrated with devices such as mobile phones, robots, etc., enabling autonomous operation based on visual inputs and text instructions.
- **Multilingual support:** To serve a global audience, beyond English and Chinese, Qwen2-VL now supports multilingual context understanding within images, including most European languages, Japanese, Korean, Arabic, Vietnamese, and others.

2 Approach

The Qwen2-VL series consists of models of 3 sizes, which are Qwen2-VL-2B, Qwen2-VL-7B and Qwen2-VL-72B. Table 1 lists the hyper-parameters and important information. Notably, Qwen2-VL employs a 675M parameter ViT across various-sized LLMs, ensuring that the computational load of the ViT remains constant regardless of the scale of the LLM.

2.1 Model Architecture

Figure 2 illustrates the comprehensive structure of Qwen2-VL. We have retained the Qwen-VL (Bai et al., 2023b) framework, which integrates vision encoders and language models. For various scale adaptations, we

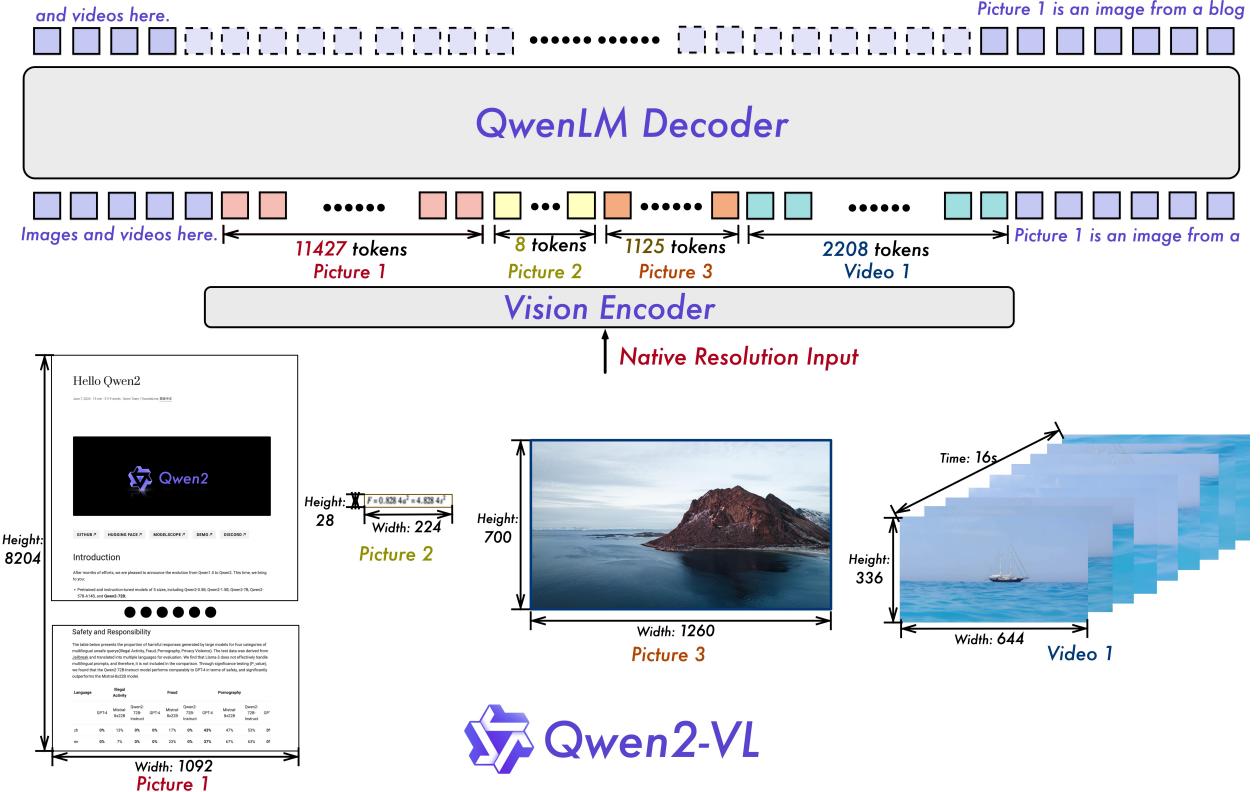


Figure 2: Qwen2-VL is capable of accurately identifying and comprehending the content within images, regardless of their clarity, resolution, or extreme aspect ratios.

have implemented a Vision Transformer (ViT) (Dosovitskiy et al., 2021) with approximately 675 million parameters, adept at handling both image and video inputs. In terms of language processing, we have opted for the more powerful Qwen2 (Yang et al., 2024) series of language models. To further enhance the model’s ability to effectively perceive and comprehend visual information in videos, we introduced several key upgrades:

Naive Dynamic Resolution A key architectural improvement in Qwen2-VL is the introduction of naive dynamic resolution support (Dehghani et al., 2024). Unlike Qwen-VL, Qwen2-VL can now process images of any resolution, dynamically converting them into a variable number of visual tokens.¹ To support this feature, we modified ViT by removing the original absolute position embeddings and introducing 2D-RoPE (Su et al., 2024; Su, 2021) to capture the two-dimensional positional information of images. At the inference stage, images of varying resolutions are packed into a single sequence, with the packed length controlled to limit GPU memory usage. Furthermore, to reduce the visual tokens of each image, a simple MLP layer is employed after the ViT to compress adjacent 2×2 tokens into a single token, with the special <|vision_start> and <|vision_end> tokens placed at the beginning and end of the compressed visual tokens. As a result, an image with a resolution of 224×224 , encoded with a ViT using patch_size=14, will be compressed to 66 tokens before entering LLM.

2D-RoPE to
positional embedding
bằng cách di chuyển n
vector của 1 nhánh
góc làm mờ phi và Phi là
1 góc, mà là absolute
position captioning nh
ó

Không rõ lý do tại sao av 1s 1 token trong tinh 2D-RoPEs đang truyền thông tin về thời gian và nh. Khi inference thi image c pack thành 1 sequence

Multimodal Rotary Position Embedding (M-RoPE) Another key architectural enhancement is the innovation of Multimodal Rotary Position Embedding (M-RoPE). Unlike the traditional 1D-RoPE in LLMs, which is limited to encoding one-dimensional positional information, M-RoPE effectively models the positional

¹This technology was previously implemented in the internal iterations, Qwen-VL Plus and Qwen-VL MAX. We have further upgraded it in Qwen2-VL.

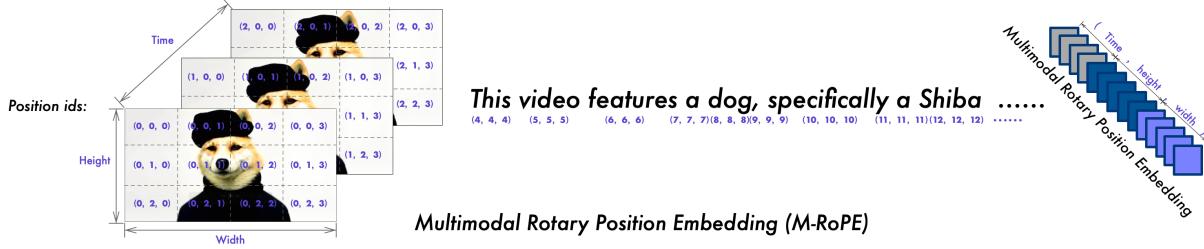


Figure 3: A demonstration of M-RoPE. By decomposing rotary embedding into temporal, height, and width components, M-RoPE can explicitly model the positional information of text, images, and video in LLM.

information of multimodal inputs. This is achieved by deconstructing the original rotary embedding into three components: temporal, height, and width. For text inputs, these components utilize identical position IDs, making M-RoPE functionally equivalent to 1D-RoPE (Su, 2024). When processing images, the temporal IDs of each visual token remain constant, while distinct IDs are assigned to the height and width components based on the token’s position in the image. For videos, which are treated as sequences of frames, the temporal ID increments for each frame, while the height and width components follow the same ID assignment pattern as images. In scenarios where the model’s input encompasses multiple modalities, position numbering for each modality is initialized by incrementing the maximum position ID of the preceding modality by one. An illustration of M-RoPE is shown in Figure 3. M-RoPE not only enhances the modeling of positional information but also reduces the value of position IDs for images and videos, enabling the model to extrapolate to longer sequences during inference.

Unified Image and Video Understanding Qwen2-VL employs a mixed training regimen incorporating both image and video data, ensuring proficiency in image understanding and video comprehension. To preserve video information as completely as possible, we sampled each video at two frames per second. Additionally, we integrated 3D convolutions (Carreira and Zisserman, 2017) with a depth of two to process video inputs, allowing the model to handle 3D tubes instead of 2D patches, thus enabling it to process more video frames without increasing the sequence length (Arnab et al., 2021). For consistency, each image is treated as two identical frames. To balance the computational demands of long video processing with overall training efficiency, we dynamically adjust the resolution of each video frame, limiting the total number of tokens per video to 16384. This training approach strikes a balance between the model’s ability to comprehend long videos and training efficiency.

2.2 Training

Following Qwen-VL (Bai et al., 2023b), we adopt a three-stage training methodology. In the first stage, we focus exclusively on training the Vision Transformer (ViT) component, utilizing a vast corpus of image-text pairs to enhance semantic understanding within the Large Language Model (LLM). In the second stage, we unfreeze all parameters and train with a wider range of data for more comprehensive learning. In the final stage, we lock the ViT parameters and perform exclusive fine-tuning of the LLM using instructional datasets.

The model is pre-trained on a diverse dataset that includes image-text pairs, optical character recognition (OCR) data, interleaved image-text articles, visual question answering datasets, video dialogues, and image knowledge datasets. Our data sources primarily comprise cleaned web pages, open-source datasets, and synthetic data. The cutoff date for our data knowledge is June 2023. This diverse data composition is instrumental in developing a robust multimodal understanding capability.

During the initial pre-training phase, Qwen2-VL is exposed to a corpus of around 600 billion tokens. The LLM component of Qwen2-VL is initialized using the parameters from Qwen2 (Yang et al., 2024), while the vision encoder of Qwen2-VL is initialized with the ViT derived from DFN. However, the fixed position embedding in the original DFN’s ViT (Fang et al., 2023) is replaced by RoPE-2D. This pre-training phase

primarily focuses on learning image-text relationships, textual content recognition within images through OCR, and image classification tasks. Such foundational training is instrumental in enabling the model to develop a robust understanding of core visual-textual correlations and alignments.

The second pre-training phase marks a significant progression, involving an additional 800 billion tokens of image-related data. This stage introduces a higher volume of mixed image-text content, facilitating a more nuanced understanding of the interplay between visual and textual information. The incorporation of visual question answering datasets refines the model's capacity to respond to image-related queries. Moreover, the inclusion of multitasking datasets is pivotal in developing the model's ability to navigate diverse tasks concurrently, a skill of paramount importance when dealing with complex, real-world datasets. Concurrently, purely textual data continues to play a crucial role in maintaining and advancing the model's linguistic proficiency.

Throughout the pre-training stages, Qwen2-VL processes a cumulative total of 1.4 trillion tokens. Specifically, these tokens encompass not only text tokens but also image tokens. During the training process, however, we only provide supervision for the text tokens. This exposure to extensive and diverse linguistic and visual scenarios ensures that the model develops a deep understanding of the intricate relationships between visual and textual information, thereby laying a robust foundation for various multimodal tasks.

During the instruction fine-tuning phase, we employ the ChatML (Openai, 2024) format to construct instruction-following data. This dataset encompasses not only pure text-based dialogue data but also multimodal conversational data. The multimodal components include image question-answering, document parsing, multi-image comparison, video comprehension, video stream dialogue, and agent-based interactions. Our comprehensive approach to data construction aims to enhance the model's capability to understand and execute a wide range of instructions across various modalities. By incorporating diverse data types, we seek to develop a more versatile and robust language model capable of handling complex, multimodal tasks in addition to traditional text-based interactions.

2.2.1 Data Format.

In line with Qwen-VL, Qwen2-VL also employs special tokens to distinguish vision and text inputs. Tokens <|vision_start|> and <|vision_end|> are inserted at the start and end of the image feature sequence to demarcate the image content.

Dialogue Data. In terms of dialogue format, we construct our instruction tuning dataset using the ChatML format, where each interaction's statement is marked with two special tokens (<|im_start|> and <|im_end|>) to facilitate dialogue termination. The sections marked in blue indicate the supervised parts.

The Dataset Format Example of ChatML

```
<|im_start|>user
<|vision_start|>Picture1.jpg<|vision_end|><|vision_start|>Picture2.jpg<|vision_end|>What do the
two pictures have in common?<|im_end|>
<|im_start|>assistant
Both pictures are of SpongeBob SquarePants. <|im_end|>
<|im_start|>user
What is happening in the video?<|vision_start|>video.mp4<|vision_end|><|im_end|>
<|im_start|>assistant
The protagonist in the video is frying an egg.<|im_end|>
```

Visual Grounding. To endow the model with visual grounding capabilities, bounding box coordinates are normalized within [0, 1000] and represented as "($X_{\text{top left}}, Y_{\text{top left}}), (X_{\text{bottom right}}, Y_{\text{bottom right}})$ ". Tokens

<|box_start|> and <|box_end|> are utilized to demarcate bounding box text. To accurately link bounding boxes with their textual descriptions, we introduce tokens <|object_ref_start|> and <|object_ref_end|> to indicate the content that the bounding box references, thereby allowing the model to effectively interpret and generate precise descriptions of specific regions.

Referring Grounding

```
<|vision_start|>Picture1.jpg<|vision_end|>
<|object_ref_start|>the eyes on a giraffe<|object_ref_end|><|box_start|>(176,106),(232,160)
<|box_end|>
```

Visual Agent. To develop Qwen2-VL as a general-purpose VL-Agent, we treat various agent tasks, such as UI Operations, Robotic Control, Games, and Navigation, as sequential decision-making problems, enabling Qwen2-VL to accomplish tasks through multi-step action execution. For each task, we first define a set of permissible actions and keywords pattern (underline) for function call (Qwen Team, 2024). Qwen2-VL then analyzes the observations, performs reasoning and planning, executes the selected actions, and interacts with the environment to acquire new observations. This cycle repeats iteratively until the task is successfully completed. By integrating various tools and leveraging the vision perception capabilities of large vision-language models (LVLMs), Qwen2-VL is able to iteratively execute increasingly complex tasks involving real-world visual interactions.

Visual Agent

```
<|im_start|>system
You are a helpful assistant.
# Actions
## You have the following actions.
### Tap
Tap: A gentle tap that commands, chooses, or navigates through a smartphone's user interface.
Parameters: [{"name": "point", "description": "The specific spot of interest on the monitor, denoted by the coordinates (x, y) where x and y range from 0 to 1000.", "required": True}]
### Home
Home: Go to phone's home screen. Parameters: []
### Other Actions ...
## Continuously take action until the task is completed.
*FUNCTION*: The action to take, should be one of {Actions}.
*ARGS*: The input of the action.
*RESULT*: Action results.
*RETURN*: Reply based on action results. <|im_end|>
<|im_start|>user
Find a pizza restaurant nearby in Map. <|vision_start|>Screenshot_1.jpg<|vision_end|> <|im_end|>
<|im_start|>assistant
Several e-mails are displaying on the phone screen. To open Map, I need go back to the home screen and find the corresponding app icon.
*FUNCTION*: Home *ARGS*: {}
*RESULT*: <|vision_start|>Screenshot_2.jpg<|vision_end|>
*RETURN*: I return to the home screen. Next, I need to find the icon of Map and tap on it.
*FUNCTION*: Tap *ARGS*: {"point": (348,291)}
*RESULT*: <|vision_start|>Screenshot_3.jpg<|vision_end|>
*RETURN*: [Thinking for the next action.]
[Other subsequent actions.] ....
I have found the pizza restaurant nearby in Map. <|im_end|>
```

2.3 Multimodal Model Infrastructure

The Qwen2-VL models were trained on Alibaba Cloud’s PAI-Lingjun Intelligent Computing Service ([Alibaba-Cloud, 2024c](#)) with its scalable computing, auto resuming and straggler detection.

Storage. We use Alibaba Cloud’s ultra-speed CPFS (Cloud Parallel File Storage) ([Alibaba-Cloud, 2024a](#)) to build a storage system of Qwen2-VL pre-training and post-training. We decoupled the text data and vision data storage. We simply store text data on CPFS and use mmap for efficient access. For vision data, we use Alibaba Cloud’s OSS (Object Storage Service) ([Alibaba-Cloud, 2024b](#)) for persistent storage. During training, we accessed vision data through OSS’s python-client concurrently and tuned the concurrency and retrying parameters to avoid reaching the QPS (queries per second) limit. We also found that video data decoding is a main bottleneck, especially for long videos. After several attempts with open-source ([FFmpeg-Developers, 2024](#)) and in-house software failed, we opted for a caching decoding technique. Checkpointing saves each GPU’s optimizer and model states on CPFS.

Parallelism. We use 3D parallelism which combines data parallelism (DP) ([Li et al., 2020](#)), tensor parallelism (TP) ([Krizhevsky et al., 2012; Shoeybi et al., 2019](#)) and pipeline parallelism (PP) ([Huang et al., 2019; Narayanan et al., 2021; Lamy-Poirier, 2023](#)) to scale Qwen2-VL model training. We also leverage deep-speed’s zero-1 redundancy optimizer ([Rajbhandari et al., 2020](#)) to shard states for memory saving. Sequence parallelism (SP) ([Korthikanti et al., 2023](#)) with selective checkpointing activation ([Chen et al., 2016](#)) was leveraged to reduce memory usage. When enabling TP training, we always shard the vision encoder and large language models together but not the vision merger due to its relatively few parameters. We found the TP training would result in different model shared-weights due to the convolution operator’s non-deterministic behavior ². We resolved this issue by performing offline reduction of the shared weights, thereby avoiding an additional **all-reduce** communication step. This approach resulted in only a minimal impact on performance. We leverage 1F1B PP ([Narayanan et al., 2021](#)) for Qwen2-VL 72B training. We combine the vision encoder, vision adapter and several LLM’s decoder layers into one stage, and evenly split the remaining decoder layers. Note that the vision and text sequence lengths are dynamic for each data point. We **broadcast** the dynamic sequence lengths before initiating the 1F1B process and access the shape information using batch indices. We also implemented an interleaved 1F1B PP ([Narayanan et al., 2021](#)) but found it is slower than the standard 1F1B setting.

Software. We use PyTorch ([Paszke et al., 2019; Ansel et al., 2024](#)) version 2.1.2 with CUDA 11.8 ([Nvidia, 2024b](#)) for training. Additionally, we leverage flash-attention ([Dao et al., 2022; Dao, 2024; Shah et al., 2024](#)) for efficient training in both the vision encoder and the LLM. We also utilize fused operators ([Nvidia, 2024a](#)) such as LayerNorm ([Ba et al., 2016](#)), RMSNorm ([Zhang and Sennrich, 2019](#)), and Adam ([Loshchilov and Hutter, 2019](#)). Besides this, we leverage the overlap of communication and computation during matrix multiplication in our training process.

3 Experiments

In this section, we first evaluate the model’s performance by conducting a comparative analysis across a variety of visual benchmarks, demonstrating the advantages of our approach. Subsequently, we carry out a detailed examination of specific capabilities, including general visual perception, document understanding, multilingual recognition in images, video comprehension, and agent abilities. Finally, we present an ablation study to investigate several key components of our approach.

²<https://pytorch.org/docs/stable/notes/randomness.html>

Table 2: Performance Comparison of Qwen2-VL Models and State-of-the-art.

Benchmark	Previous SoTA	Claude-3.5 Sonnet	GPT-4o	Qwen2-VL-72B	Qwen2-VL-7B	Qwen2-VL-2B
MMMU _{val} (Yue et al., 2023)	66.1 (X.AI, 2024b)	68.3	69.1	64.5	54.1	41.1
DocVQA _{test} (Mathew et al., 2021)	94.1 (Chen et al., 2024c)	95.2	92.8	96.5	94.5	90.1
InfoVQA _{test} (Mathew et al., 2021)	82.0 (Chen et al., 2024c)	-	-	84.5	76.5	65.5
A12D (Kembhavi et al., 2016)	87.6 (Chen et al., 2024c)	80.2(94.7)	84.6(94.2)	88.1	83.0	74.7
ChartQA _{test} (Masry et al., 2022)	88.4 (Chen et al., 2024c)	90.8	85.7	88.3	83.0	73.5
TextVQA _{val} (Singh et al., 2019)	84.4 (Chen et al., 2024c)	-	-	85.5	84.3	79.7
OCRBench (Liu et al., 2023e)	852 (Yao et al., 2024)	788	736	877	866	809
MTVQA (Tang et al., 2024)	23.2 (Team et al., 2023)	25.7	27.8	30.9	25.6	18.1
VCR _{en} easy (Zhang et al., 2024c)	84.7 (Chen et al., 2024c)	63.9	91.6	91.9	89.7	81.5
VCR _{zh} easy (Zhang et al., 2024c)	22.1 (Chen et al., 2024c)	1.0	14.9	65.4	59.9	46.2
RealWorldQA (X.AI, 2024a)	72.2 (Chen et al., 2024c)	60.1	75.4	77.8	70.1	62.9
MME _{sum} (Fu et al., 2023)	2414.7 (Chen et al., 2024c)	1920.0	2328.7	2482.7	2326.8	1872.0
MMBench-EN _{test} (Liu et al., 2023d)	86.5 (Chen et al., 2024c)	79.7	83.4	86.5	83.0	74.9
MMBench-CN _{test} (Liu et al., 2023d)	86.3 (Chen et al., 2024c)	80.7	82.1	86.6	80.5	73.5
MMBench-V1.1 _{test} (Liu et al., 2023d)	85.5 (Chen et al., 2024c)	78.5	82.2	85.9	80.7	72.2
MMT-Bench _{test} (Ying et al., 2024)	63.4 (Chen et al., 2024b)	-	65.5	71.7	63.7	54.5
MMStar (Chen et al., 2024a)	67.1 (Chen et al., 2024c)	62.2	63.9	68.3	60.7	48.0
MMVet _{GPT-4-Turbo} (Yu et al., 2024)	67.5 (OpenAI, 2023)	66.0	69.1	74.0	62.0	49.5
HallBench _{avg} (Guan et al., 2023)	55.2 (Chen et al., 2024c)	49.9	55.0	58.1	50.6	41.7
MathVista _{testmini} (Lu et al., 2024a)	69.0 (X.AI, 2024b)	67.7	63.8	70.5	58.2	43.0
MathVision (Wang et al., 2024)	30.3 (OpenAI, 2023)	-	30.4	25.9	16.3	12.4
MMMU-Pro (Yue et al., 2024)	46.9 (Team et al., 2023)	51.5	51.9	46.2	43.5	37.6

Table 3: Performance of Qwen2-VL and GPT-4o on internal multilingual OCR benchmarks.

Language	Korean	Japanese	French	German	Italian	Russian	Vietnamese	Arabic
GPT-4o	87.8	88.3	89.7	88.3	74.1	96.8	72.0	75.9
Qwen2-VL-72B	94.5	93.4	94.1	91.5	89.8	97.2	73.0	70.7

3.1 Compare to SOTAs

We evaluate the visual capabilities of our model through various visual benchmarks, video tasks, and agent-based assessments. Qwen2-VL demonstrates highly competitive performance at the same scale, achieving new state-of-the-art (SoTA) results. Overall, our 72B model consistently delivers top-tier performance across most evaluation metrics, frequently surpassing even closed-source models such as GPT-4o (OpenAI, 2024) and Claude 3.5-Sonnet (Anthropic, 2024). Notably, it exhibits a significant advantage in document understanding tasks. However, in the MMMU (Yue et al., 2023) benchmark, our model still lags behind GPT-4o to some extent, indicating that Qwen2-VL-72B has room for improvement when handling more complex and challenging problem sets.

3.2 Quantitative Results

In this section, we present an extensive evaluation of the Qwen2-VL series across an array of datasets, offering a comprehensive understanding of the model’s capabilities in various aspects.

3.2.1 General Visual Question Answering

To rigorously assess our models’ capabilities in general visual question answering tasks, we conduct extensive evaluations across a diverse array of state-of-the-art benchmarks: RealWorldQA (X.AI, 2024a), MMStar (Chen et al., 2024a), MMVet (Yu et al., 2024), MMT-Bench (Ying et al., 2024), MMBench (Liu et al., 2023d), MMBench-1.1 (Liu et al., 2023d), MME (Fu et al., 2023), and HallusionBench (Guan et al., 2023). The Qwen2-VL series exhibits exceptional performance across these benchmarks, with the 72B model consistently achieving or surpassing state-of-the-art results, while the 7B and 2B variants also demonstrate robust capabilities. On RealWorldQA, which evaluates real-world spatial comprehension, Qwen2-VL-72B achieves a

Table 4: Performance of Qwen2-VL and other models on video benchmarks.

Benchmark	Previous SoTA	Gemini 1.5-Pro	GPT-4o	Qwen2-VL-72B	Qwen2-VL-7B	Qwen2-VL-2B
MVBench (Li et al., 2024)	69.6	-	-	73.6	67.0	63.2
PerceptionTest _{test} (Patraucean et al., 2024)	66.9	-	-	68.0	62.3	53.9
EgoSchema _{test} (Mangalam et al., 2023)	62.0	63.2	72.2	77.9	66.7	54.9
Video-MME _(wo/w subs) (Fu et al., 2024)	66.3/69.6	75.0/81.3	71.9/77.2	71.2/77.8	63.3/69.0	55.6/60.4

Table 5: Performance Comparison of Qwen2-VL-72B across various agent benchmarks and GPT-4o. SR, GC, TM and EM are short for success rate, goal-condition success, type match and exact match. ALFRED, R2R and REVERIE are performance in valid-unseen.

	Benchmark	Metric	Previous SoTA	GPT-4o	Qwen2-VL-72B
General	FnCall	TM	-	90.2	93.1
		EM	-	50.0	53.2
UI Operations	AITZ (Zhang et al., 2024b)	TM	83.0 (Hong et al., 2023)	70.0	89.6
		EM	47.7 (Zhan and Zhang, 2023)	35.3	72.1
Card Games	Number Line (Zhai et al., 2024)	SR	89.4 (Zhai et al., 2024)	91.5	100.0
	BlackJack (Zhai et al., 2024)	SR	40.2 (Zhai et al., 2024)	34.5	42.6
	EZPoint (Zhai et al., 2024)	SR	50.0 (Zhai et al., 2024)	85.5	100.0
	Point24 (Zhai et al., 2024)	SR	2.6 (Liu et al., 2023b)	3.0	4.5
Robotic Control	ALFRED (Shridhar et al., 2020a)	SR GC	67.7 (Lu et al., 2023) 75.3 (Lu et al., 2023)	- -	67.8 75.8
Navigation	R2R (Anderson et al., 2018)	SR	79.0 (Chen et al., 2022)	43.7	51.7
	REVERIE (Qi et al., 2020)	SR	61.0 (Sigurdsson et al., 2023)	31.6	31.0

score of 77.8, surpassing both the previous state-of-the-art (72.2) and formidable baselines such as GPT-4o (75.4), thus demonstrating superior understanding of physical environments. For MMStar, a benchmark designed to assess genuine multimodal capabilities through visually indispensable samples, Qwen2-VL-72B attains 68.3, outperforming the previous best of 67.1 and highlighting its proficiency in integrating visual and textual information. On MMVet, which evaluates the integration of core vision-language capabilities across 16 complex multimodal tasks, Qwen2-VL-72B achieves a remarkable 74.0, significantly outperforming strong competitors including GPT-4V (67.5) and showcasing its versatility in addressing diverse multimodal challenges. In the MMT-Bench evaluation, which assesses advanced reasoning and instruction following across 32 core meta-tasks and 162 subtasks in multimodal understanding, Qwen2-VL-72B achieves 71.7, markedly surpassing the previous best (63.4) and demonstrating its prowess in applying expert knowledge and executing deliberate visual recognition, localization, reasoning, and planning. On MMBench, which evaluates fine-grained abilities across 20 dimensions, Qwen2-VL-72B exhibits strong performance, achieving 86.5 on the English test set, matching the state-of-the-art, and 86.6 on the Chinese test set, establishing a new benchmark. For MME, which measures a wide spectrum of perception and cognition abilities across 14 subtasks, Qwen2-VL-72B achieves a cumulative score of 2482.7, significantly outperforming the previous best (2414.7), underscoring its advanced capabilities in both visual perception and high-level cognition tasks.

These comprehensive results underscore the Qwen2-VL series' exceptional proficiency in general visual question answering tasks. The models demonstrate advanced capabilities in real-world spatial comprehension, genuine multimodal integration, complex reasoning, instruction following, and a broad range of perception and cognition tasks. The consistent superior performance across diverse benchmarks, particularly the outstanding results of the 72B model, positions the Qwen2-VL series as a leading solution in the field of visual question answering. Our models excel in handling visually indispensable tasks, integrating core vision-language capabilities, and demonstrating expertise across diverse multimodal scenarios, ranging from fundamental perception tasks to complex reasoning and planning. This exhaustive evaluation highlights the Qwen2-VL series' versatility and effectiveness in addressing the multifaceted challenges posed by state-of-the-art multimodal benchmarks, thereby setting a new standard for large vision-language models.

3.2.2 Document and Diagrams Reading

We tested our model’s OCR and document and diagram comprehension on DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2021), TextVQA (Singh et al., 2019), AI2D (Kembhavi et al., 2016) datasets. The DocVQA/InfoVQA/ChartQA dataset focuses on the model’s ability to comprehend text in documents/high-resolution infographics/charts, while the TextVQA dataset examines the ability to comprehend text in naturalistic images. The OCRBench dataset is a dataset of mixed tasks, which focuses on mathematical formula parsing and information extraction in addition to the text-based VQA. The AI2D dataset focuses on multiple-choice questions on scientific diagrams containing text. In addition, we also tested the OCR and formula recognition capabilities of our model on OCRBench (Liu et al., 2023e), as well as the multilingual OCR capabilities of our model on the MTVQA (Tang et al., 2024) dataset.

The experimental results show that our model achieves SoTA level in several metrics, including DocVQA, InfoVQA, TextVQA and OCRBench, demonstrating that our model has good comprehension of textual content in images from multiple domains.

3.2.3 Multilingual Text Recognition and Understanding

In particular, our model surpasses all existing general-purpose LVLMs in multilingual OCR. Our model not only outperforms existing LVLMs (including proprietary models such as GPT-4o, Claude 3.5 Sonnet, etc.) on the public-available MTVQA dataset, it also outperforms GPT-4o on the in-house internal benchmark across all foreign languages except Arabic (Table 3).

3.2.4 Mathematical Reasoning

We’ve conducted experiments on the MathVista (Lu et al., 2024a) and MathVision (Wang et al., 2024) datasets to assess mathematical reasoning capabilities. MathVista is a comprehensive benchmark featuring 6,141 diverse examples of mathematical and visual tasks. The MathVision dataset comprises 3,040 math problems embedded in visual contexts from actual math competitions, covering 16 mathematical disciplines and varying in difficulty across five levels. These challenges underscore the necessity for LVLMs to exhibit strong visual comprehension, a deep understanding of mathematics, and sound logical reasoning skills. The Qwen2-VL series has demonstrated superior performance on MathVista, achieving a 70.5 outperforming other LVLMs. Additionally, it has set a new open-source benchmark on MathVision with 25.9.

3.2.5 Referring Expression Comprehension

Regarding visual localization task, we evaluate Qwen2-VL on RefCOCO, RefCOCO+, and RefCOCOg datasets (Kazemzadeh et al., 2014; Mao et al., 2016). The results, as depicted in Table 6, demonstrate that Qwen2-VL attains top-tier results among generalist models. Benefiting from a more rational structure design, Qwen2-VL is able to perceive details in high-resolution images, leading to significant improvements over Qwen-VL. The superiority of these models in comparison to both generalist and specialized models highlights their potential for advancing the field of visual localization and their capacity for real-world implementation in tasks requiring precise visual understanding.

3.2.6 Video Understanding

We evaluate our models on various video understanding tasks, with related benchmarks covering short videos of a few seconds to long videos of up to one hour. Table 4 presents the performance of Qwen2-VL and baseline models. Overall, Qwen2-VL demonstrates strong results across 2B, 7B, and 72B sizes, with Qwen2-VL-72B achieving the best performance on MVBench (Li et al., 2024), PerceptionTest (Patraucean et al., 2024), and EgoSchema (Mangalam et al., 2023). This showcases Qwen2-VL’s superior capabilities in

Table 6: Performance Comparison on Referring Expression Comprehension Task.

Type	Model	RefCOCO			RefCOCO+			RefCOCOg	
		val	test-A	test-B	val	test-A	test-B	val	test
Generalist	OFA-L (Wang et al., 2022)	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
	Shikra (Chen et al., 2023a)	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
	Qwen-VL (Bai et al., 2023b)	89.4	92.3	85.3	83.1	88.3	77.2	85.6	85.5
	Ferretv2 (Zhang et al., 2024a)	92.6	95.0	88.9	87.4	92.1	81.4	89.4	90.0
	CogVLM (Wang et al., 2023b)	92.8	94.8	89.0	88.7	92.9	83.4	89.8	90.8
	InternVL2 _{2b} (Chen et al., 2024c)	82.3	88.2	75.9	73.5	82.8	63.3	77.6	78.3
	InternVL2 _{8b} (Chen et al., 2024c)	87.1	91.1	80.7	79.8	87.9	71.4	82.7	82.7
	InternVL2 _{76b} (Chen et al., 2024c)	92.2	94.8	88.4	88.8	93.1	82.8	89.5	90.3
	Qwen2-VL _{2b}	87.6	90.6	82.3	79.0	84.9	71.0	81.2	80.3
	Qwen2-VL _{7b}	91.7	93.6	87.3	85.8	90.5	79.5	87.3	87.8
	Qwen2-VL _{72b}	93.2	95.3	90.7	90.1	93.8	85.6	89.9	90.4
Specialist	G-DINO-L (Liu et al., 2023c)	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
	UNINEXT-H (Yan et al., 2023)	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
	ONE-PEACE (Wang et al., 2023a)	92.6	94.2	89.3	88.8	92.2	83.2	89.2	89.3

video understanding tasks, and scaling up Qwen2-VL yields significant improvements. For the challenging Video-MME benchmark ([Fu et al., 2024](#)), which includes videos up to one hour, it is noteworthy that we limited the maximum number of frames extracted per video to 768 during evaluation, potentially impacting performance on longer videos. Future work will focus on extending Qwen2-VL to support longer sequences, thereby accommodating longer videos.

3.2.7 Visual Agent

Qwen2-VL is evaluated first for its ability to interact with the environment via function calls and then for its capacity to complete complex sequential decision tasks through multiple rounds of interaction. The implementation is based on the Qwen-Agent framework ([Qwen Team, 2024](#)).

Function Calling Unlike function calling in LLMs ([Yan et al., 2024; Srinivasan et al., 2023; Chen et al., 2023c](#)), function calling in LVLMs often involves extracting information from visual cues. Due to the absence of public benchmarks for evaluating the capabilities of LVLMs in function calling, we constructed our internal evaluation dataset.

To construct the evaluation dataset, we undertook the following procedures ([Chen et al., 2023c](#)): Scene Categorization, Image Collection, Image Content Extraction, and Question/Functions/Arguments Generation. Firstly, we classified scenes into categories based on different visual applications. Subsequently, we downloaded and meticulously selected high-quality, representative images from the internet for each category. Thereafter, utilizing an advanced LVLM ([Bai et al., 2023b](#)), we analyzed each image to extract key visual elements and textual information. Finally, based on the content information from the images, we used an advanced LLM ([Yang et al., 2024](#)) to generate a series of questions that required specific functions to answer, along with specifying the input parameters needed for these function calls.

Similar to the function calling evaluation method in LLMs ([Yan et al., 2024](#)), we designed two metrics to evaluate the accuracy of the function selection and the correctness of the arguments input. Specifically, Type Match(TM), is calculated as the ratio of times the model successfully invoked the correct function to the total number of calls attempted. Exact Match(EM), for each function calling, we checked whether the arguments passed to the function exactly matched those recorded in the image's content information, calculating this correctness ratio.

As shown in Table 5, the performance of Qwen2-VL in both Type Match(93.1 vs. 90.2) and Exact Match(53.2 vs. 50.0) over GPT-4o substantiates the efficacy of Qwen2-VL's capability in function calling, thereby underscoring

its significant potential for application expansion through external tool integration.

The evaluation results demonstrated that GPT-4o underperformed, primarily due to two factors: in scenarios where uncertainty arises, GPT-4o demonstrates a conservative approach by avoiding using external tools. The Optical Character Recognition (OCR) capability of GPT-4o is outperformed by Qwen2-VL, particularly in the context of Chinese characters.

UI Operations/Games/Robotics/Navigation To assess Qwen2-VL’s ability to generally handle complex tasks, we conduct evaluations across multiple VL agent tasks, including mobile operations (Zhang et al., 2024b; Rawles et al., 2024b; Lu et al., 2024b; Rawles et al., 2024a), robotic control (Kolve et al., 2017; Shridhar et al., 2020a; Inoue and Ohashi, 2022; Lu et al., 2023; Jiang et al., 2022; Huang et al., 2023b), card games (Zhai et al., 2024), and vision-language navigation (Anderson et al., 2018; Qi et al., 2020). As these tasks need multiple actions to complete tasks, we keep the history (observation, action) through Qwen2-VL supports a 32K context length, then append each new observation image after every action, enabling continuous reasoning about subsequent steps.

UI Operations: we evaluate Qwen2-VL using the AITZ task (Zhang et al., 2024b), which constructs a core clean test set derived from AITW (Rawles et al., 2024b). Based on common operation patterns of phone, we define actions such as tap, input and swipe (Rawles et al., 2024b) for Qwen2-VL to interact with on-screen icons for task completion. For example, when Qwen2-VL is tasked with finding a pizza restaurant nearby by Google Maps, it should input "pizza" in the search term, swipe to select the appropriate restaurant, and tap the corresponding link. Following the AITZ setting, we report both type match (correctness of tap, input, or swipe) and exact match (correctness of tap location, input text, or swipe direction). With the support of grounding capability on UI, Qwen2-VL surpasses GPT-4 and previous SoTA (Zhang et al., 2024b; Zhan and Zhang, 2023).

Robotic Control: we evaluate Qwen2-VL on the ALFRED task (Shridhar et al., 2020a) in AI2THOR (Kolve et al., 2017). The task requires agent to perform complex household tasks, such as toasting bread and slicing an apple to prepare a meal. To work in the virtual environment, we define high-level actions (GotoLocation, Pickup, PutDown, Open, Close, Clean, Heat, Cool, Slice) (Shridhar et al., 2020b) as the action set. Moreover, agent needs to localize objects for manipulation (e.g., it can only pick up an apple if the apple is recognized). To improve the accuracy of manipulation, we integrate SAM (Kirillov et al., 2023). ALFRED task reports task success rate (SR) (e.g., preparing dinner) and sub-goal completion metrics (GC) (e.g., whether the bread is toasted or the apple is sliced). Qwen2-VL slightly outperforms the previously specialized model ThinkBot (Lu et al., 2023) on the valid-unseen set.

Card Games: we leverage the card game environment from RL4VLM (Zhai et al., 2024) to assess Qwen2-VL’s performance in a series of card-based games: Number Line, BlackJack, EZPoint, and Point24. Each game presents distinct challenges: (1) reaching a target number using +1 or -1 operations, (2) drawing or holding cards to compete against the dealer, (3) applying basic arithmetic operations to reach a total of 12, and (4) using arithmetic operations to achieve a total of 24. We report the success rate of the tasks. They not only evaluate agent capabilities but also require strong OCR skills to recognize these cards and understand the progression of the game. Qwen2-VL demonstrates superior performance across all tasks.

Vision-Language Navigation: we evaluate Qwen2-VL on the Vision-and-Language Navigation (VLN) task using the R2R (Anderson et al., 2018) and REVERIE (Qi et al., 2020). In VLN, the model must autonomously determine the next location based on instruction, current observations. We report the success rate (SR) of VLM in reaching the predetermined destination for this task. The performance of Qwen2-VL is comparable to that of GPT-4o, but both models fall significantly behind current specialized VLN models (Chen et al., 2022; Sigurdsson et al., 2023). We attribute this gap to the incomplete and unstructured map information generated by the model from multiple images. Accurately modeling maps and locations in a 3D environment remains a major challenge for multimodal models.

Table 7: Qwen2-VL-7B under fixed/dynamic image tokens. Adjusting image sizes only results in small perturbations in performance, demonstrating the robustness to varying image sizes. Moreover, the dynamic resolution strategy achieves top-tier performance while consuming fewer tokens on average, demonstrating the efficiency of our model.

Strategy	Average Image Tokens	InfoVQA _{val}	RealWorldQA	OCRBench	MMMU
Fixed Image Tokens	64	28.85	56.47	572	53.33
	576	65.72	65.88	828	52.78
	1600	74.99	69.54	824	52.89
	3136	77.27	70.59	786	53.44
Dynamic Image Tokens	1924	75.89	70.07	866	53.44

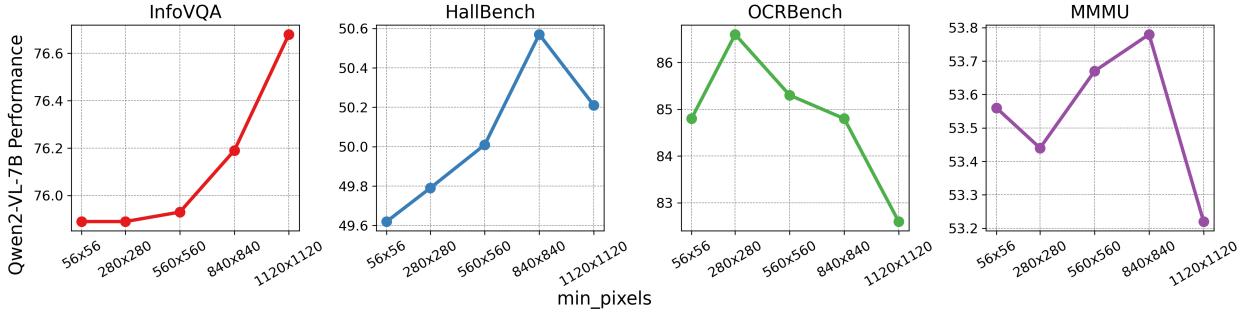


Figure 4: Qwen2-VL-7B with different min_pixels. Small images are upscaled to surpass a specified min_pixels threshold before input into the model. Increasing the image size within a reasonable range shows enhanced performance on perceptual tasks like InfoVQA, HallusionBench, and OCRBench.

3.3 Ablation Study

In this section, we present ablation studies on image dynamic resolution, M-RoPE, and model scale. These experiments aim to provide insights into the impact of these key components on our model’s performance.

3.3.1 Dynamic Resolution

As shown in Table 7, we compare the performance between dynamic resolution and fixed resolution. For fixed resolution, we resize the images to ensure a constant number of image tokens being input to the model, rather than resizing to a specific height and width, as this would distort the original aspect ratio. For dynamic resolution, we only set $\text{min_pixels} = 100 \times 28 \times 28$ and $\text{max_pixels} = 16384 \times 28 \times 28$, allowing the number of image tokens depend primarily on the image’s native resolution. It can be observed that adjusting image sizes only results in small perturbations in performance, demonstrating the model robustness to varying image sizes. Moreover, dynamic resolution approach is more efficient. We can observe that no single fixed resolution achieves optimal performance across all benchmarks. In contrast, the dynamic resolution approach consistently achieves top-tier performance while consuming fewer tokens on average.

Additionally, we observe that merely increasing the image size does not always lead to improved performance. It is more important to choose an appropriate resolution for different images. As detailed in Figure 4, we upscale small images to surpass a specified min_pixels threshold. Evaluations on upsampled images shows enhanced performance on perceptual tasks like InfoVQA, HallusionBench, and OCRBench. We attribute these gains to increased computational load. However, for OCRBench, a too-high min_pixels value leads to a severe performance decline. This is likely because OCRBench contains numerous extremely small images, and excessive enlargement causes these images to deviate from the training data distribution, turning them into out-of-distribution samples. In contrast, the effect of increasing min_pixels on the MMMU benchmark is negligible. We hypothesize that the performance bottleneck in MMMU is more related to the model’s

Table 8: Ablation studies of M-RoPE. Compared to 1D-RoPE, using M-RoPE achieves better performance in downstream tasks, particularly in video benchmarks. RWQ means RealworldQA.

	Image Benchmarks								Video Benchmarks		
	MathVista	MMB	MMStar	RWQ	DocVQA	ChartQA	InfoVQA	TextVQA	PerceptionTest	NextQA	STAR
1D-RoPE	39.2	58.6	36.7	54.5	82.5	68.0	50.8	71.3	46.6	43.9	55.5
M-RoPE	43.4	60.6	36.7	53.7	82.8	68.4	50.3	71.8	47.4	46.0	57.9

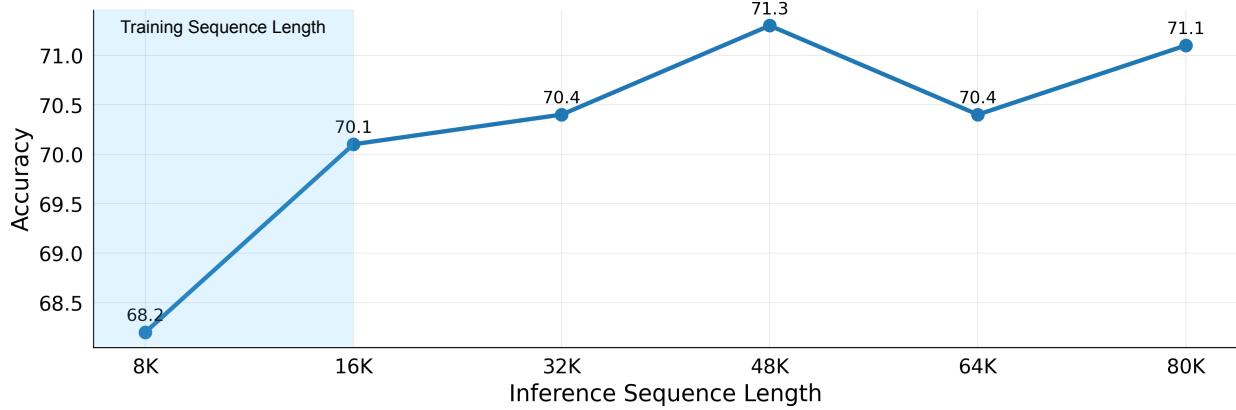


Figure 5: Evaluate the length extrapolation capability of Qwen2-VL-72B on Video-MME Medium Video. With the help of M-RoPE, the model demonstrated robust performance when the inference length exceeded the maximum training length of 16384 tokens.

reasoning capability rather than image resolution.

3.3.2 M-RoPE

In this subsection, we demonstrate the effectiveness of M-RoPE. First, we validate its capability on various downstream tasks. We employ Qwen2-1.5B and ViT-L as the backbone and report the results of the pre-trained models. As shown in Table 8, compared to 1D-RoPE, using M-RoPE achieves better performance in downstream tasks, particularly in video benchmarks. Furthermore, we assess the length extrapolation capability of M-RoPE on Video-MME medium-length videos. Figure 5 illustrates the performance of Qwen2-VL-72B at different inference lengths. Leveraging M-RoPE, the model demonstrates robust results across various inference lengths. Notably, despite limiting the maximum tokens per video to 16K during training, the model still exhibits exceptional performance at a maximum inference length of 80K tokens.

3.3.3 Model Scaling

We evaluate the performance of models of varying scales across multiple capability dimensions. Specifically, we categorize these dimensions into complex college-level problem-solving, mathematical abilities, document and table comprehension, general scenario question-answering, and video comprehension. The overall capability of a model is assessed by averaging its scores across different benchmarks associated with each dimension.

In particular, we use the MMMU (Yue et al., 2023) benchmark to represent college-level problem-solving ability, while the average scores from MathVista (Lu et al., 2024a) and MathVision (Wang et al., 2024) serve as indicators of mathematical ability. For general scenario question-answering, we compute the average score across the RealWorldQA (X.AI, 2024a), MMBench-V1.1 (Liu et al., 2023d), MMT-Bench (Ying et al., 2024), HallBench (Guan et al., 2023), MMVet (Yu et al., 2024), and MMStar (Chen et al., 2024a)

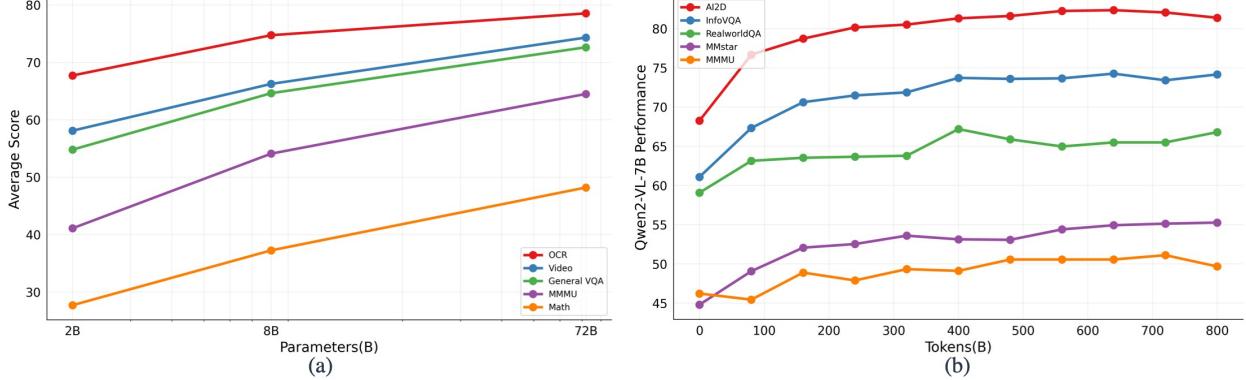


Figure 6: Model Performance Scaling Across Capabilities and Training Progress. As model size and the volume of training data increase, performance consistently improves across a range of capabilities and benchmarks.

benchmarks. Document and table comprehension capability is reflected through the average score from benchmarks like DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), OCRBench (Liu et al., 2023e), and MTVQA (Tang et al., 2024). Lastly, video comprehension ability is measured by averaging scores across MVbench (Li et al., 2024), Perception-Test (Patraucean et al., 2024), EgoSchema (Mangalam et al., 2023), and Video-MME (Fu et al., 2024).

As illustrated in Figure 6(a), there is a consistent improvement in performance with increasing model size, particularly with respect to mathematical abilities, which show a positive correlation with the number of model parameters. On the other hand, for optical character recognition (OCR)-related tasks, even smaller-scale models exhibit relatively strong performance.

As shown in Figure 6(b), we visualize the relationship between model performance and the number of training tokens during the second stage of pretraining for Qwen2-VL-7B. As the number of training tokens increases, the model performance improves; however, performance on vision question answering (VQA) tasks exhibits some fluctuation. In contrast, for tasks such as AI2D (Kembhavi et al., 2016) and InfoVQA (Mathew et al., 2021)—both of which involve understanding textual and graphical information in images—the model performance shows steady improvement as training data is augmented.

4 Conclusion

We have presented the Qwen2-VL series, the versatile large vision-language models, including three open-weight models with total parameter counts of 2, 8, and 72 billion. Qwen2-VL matches the performance of top-tier models like GPT-4o and Claude3.5-Sonnet in a range of multimodal scenarios, surpassing all other open-weight LVLM models. Qwen2-VL series introduces naive dynamic resolution and multimodal rotary position embedding (M-RoPE) to fuse information across modals effectively and be capable of understanding videos over 20 minutes in length. With advanced reasoning and decision-making abilities, Qwen2-VL can be integrated with devices such as mobile phones, robots, etc. Furthermore, Qwen2-VL now supports understanding multilingual texts within images, including most European languages, Japanese, Korean, Arabic, Vietnamese, and others.

We have made the Qwen2-VL model weights openly accessible, which enables researchers and developers to harness the full potential in a variety of applications and research projects. We aim to advance AI technologies and enhance their beneficial effects on society by dedicating ourselves to these endeavors.

Acknowledgements

We express our gratitude to Juan Zhu, Fan Hong, Jie Zhang, Yong Li of Alibaba Cloud’s PAI team ([Alibaba-Cloud, 2024c](#)) for supporting the training infrastructure of Qwen2-VL. This work was also supported by Qwen LLM team ([Yang et al., 2024](#)), and we especially thank Na Ni, Yichang Zhang, Jianxin Ma, Bowen Yu, Zheren Fu for their data contribution and insightful discussion.

References

- 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. [2](#)
- Alibaba-Cloud. Cloud parallel file storage (cpfs), 2024a. URL <https://www.alibabacloud.com/en/product/cpfs>. [8](#)
- Alibaba-Cloud. Object storage service (oss), 2024b. URL <https://www.alibabacloud.com/en/product/object-storage-service>. [8](#)
- Alibaba-Cloud. Pai-lingjun intelligent computing service, 2024c. URL <https://www.alibabacloud.com/en/product/pai-lingjun>. [8, 17](#)
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. [10, 13](#)
- Jason Ansel, Edward Z. Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshitij Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsche, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *ASPLOS*, 2024. [8](#)
- Anthropic. Claude 3.5 sonnet, 2024. URL <https://www.anthropic.com/news/clause-3-5-sonnet>. [9](#)
- Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. [5](#)
- Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv:1607.06450*, 2016. [8](#)
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv:2309.16609*, 2023a. [1](#)
- 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:2308.12966*, 2023b. [1, 2, 3, 5, 12](#)
- Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. [5](#)
- 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. [12](#)

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*, 2023b. 1

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024a. 9, 15

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022. 10, 13

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv:1604.06174*, 2016. 8

Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, et al. T-eval: Evaluating the tool utilization capability step by step. *arXiv:2312.14033*, 2023c. 12

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821*, 2024b. 9

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024c. URL <https://internvl.github.io/blog/2024-07-02-InternVL-2.0>. 9, 12

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>. 1

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. 1

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024. 8

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, 2022. 8

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. In *NeurIPS*, 2024. 4

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. 4

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 46, 48, 49

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv:2309.17425*, 2023. 5

FFmpeg-Developers. ffmpeg tool, 2024. URL <http://ffmpeg.org/>. 8

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. 9

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv:2405.21075*, 2024. 10, 12, 16

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv:2310.14566*, 2023. 9, 15

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv:2312.08914*, 2023. 10

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*, 2023a. 1, 2

Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv:2305.11176*, 2023b. 13

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Xu Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *NeurIPS*, 2019. 8

Yuki Inoue and Hiroki Ohashi. Prompter: Utilizing large language model prompting for a data efficient embodied instruction following. *arXiv:2211.03267*, 2022. 13

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv:2210.03094*, 2022. 13

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 11

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 9, 11, 16

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 13

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv:1712.05474*, 2017. 13

Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. In *MLSys*, 2023. 8

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 8

Joel Lamy-Poirier. Breadth-first pipeline parallelism. In *MLSys*, 2023. 8

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model. *arXiv:2311.04219*, 2023a. 2

Chen Li, Yixiao Ge, Dian Li, and Ying Shan. Vision-language instruction tuning: A review and analysis. *arXiv:2311.08172*, 2023b. 2

- 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. 1
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 10, 11, 16
- Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. Pytorch distributed: Experiences on accelerating data parallel training. In *VLDB*, 2020. 8
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv:2311.06607*, 2023d. 2
- Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Jiao Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv:2311.07575*, 2023. 2
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023a. 1, 2
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023b. 1, 2, 10
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023c. 12
- Yuan Liu, Haodong Duan, Bo Li Yuanhan Zhang, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahu Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023d. 9, 15
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895*, 2023e. 9, 11, 16
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 8
- Guanxing Lu, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. Thinkbot: Embodied instruction following with thought chain reasoning. *arXiv:2312.07062*, 2023. 10, 13
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, 2021. 32
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024a. 9, 11, 15
- Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv:2406.08451*, 2024b. 13
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023. 10, 11, 16
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 11

- 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. 9, 11, 16
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 9, 11, 16
- Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. Efficient large-scale language model training on GPU clusters using megatron-lm. In *SC*, 2021. 8
- Nvidia. Apex, 2024a. URL <https://github.com/NVIDIA/apex>. 8
- Nvidia. Cuda, 2024b. URL <https://developer.nvidia.com/cuda-toolkit>. 8
- OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 1, 9
- OpenAI. Gpt-4v(ision) system card, 2023. URL <https://openai.com/research/gpt-4v-system-card>. 1, 9
- Openai. Chatml documents, 2024. URL <https://github.com/openai/openai-python/blob/main/chatml.md>. 6
- OpenAI. Hello gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o>. 9
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 8
- Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2024. 10, 11, 16
- Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020. 10, 13
- Alibaba Group Qwen Team. Qwen-agent framework, 2024. URL <https://github.com/QwenLM/Qwen-Agent>. 7, 12
- 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. 2
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: memory optimizations toward training trillion parameter models. In *SC*, 2020. 8
- Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv:2405.14573*, 2024a. 13
- Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Androidinthewild: A large-scale dataset for android device control. In *NeurIPS*, 2024b. 13
- Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *arXiv:2407.08608*, 2024. 8
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv:1909.08053*, 2019. 8

- Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020a. 10, 13
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv:2010.03768*, 2020b. 13
- Gunnar A Sigurdsson, Jesse Thomason, Gaurav S Sukhatme, and Robinson Piramuthu. Rrex-bot: Remote referring expressions with a bag of tricks. In *IROS*, 2023. 10, 13
- 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*, 2019. 9, 11, 16
- Venkat Krishna Srinivasan, Zhen Dong, Banghua Zhu, Brian Yu, Damon Mosk-Aoyama, Kurt Keutzer, Jiantao Jiao, and Jian Zhang. Nexusraven: a commercially-permissive language model for function calling. In *NeurIPS Workshop*, 2023. 12
- Jianlin Su. Transformer upgrade path: 4. rotary position encoding for two-dimensional positions, 2021. URL <https://www.spaces.ac.cn/archives/8397>. 4
- Jianlin Su. Transformer upgrade path: 17. insights into multimodal positional encoding, 2024. URL <https://spaces.ac.cn/archives/10040>. 5
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. In *Neurocomputing*, 2024. 4
- Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv:2405.11985*, 2024. 9, 11, 16
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 1, 9
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023a. 1
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023b. 1
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv:2402.14804*, 2024. 9, 11, 15
- 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*, 2022. 12
- 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*, 2023a. 12
- Weihan Wang, Qingsong Lv, Wemeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogylm: Visual expert for pretrained language models. *arXiv:2311.03079*, 2023b. 1, 2, 12
- X.AI. Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>, 2024a. 9, 15
- X.AI. Grok-2 beta release. <https://x.ai/blog/grok-2>, 2024b. 9

- B. Yan, Yi Jiang, Jiannan Wu, D. Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 12
- Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Berkeley function calling leaderboard, 2024. URL https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html. 12
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv:2407.10671*, 2024. 4, 5, 12, 17
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv:2309.17421*, 2023. 30, 44
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv:2408.01800*, 2024. 9
- 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*, 2023a. 2
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv:2311.04257*, 2023b. 2
- Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv:2404.16006*, 2024. 9, 15
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024. 9, 15
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv:2311.16502*, 2023. 9, 15
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 9
- Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *arXiv:2405.10292*, 2024. 10, 13
- Zhuosheng Zhan and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. *arXiv:2309.11436*, 2023. 10, 13
- Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019. 8
- Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv:2404.07973*, 2024a. 12
- Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv:2403.02713*, 2024b. 10, 13
- 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:2309.15112*, 2023. 1

Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. Vcr: Visual caption restoration. *arXiv:2406.06462*, 2024c. 9

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. 1, 2

A Model Capabilities and Qualitative Examples

In this section, we present some practical examples of our Qwen2-VL.

A.1 General Chat and OCR

The Qwen2-VL models are now more adept at accurately describing and identifying complex information within images, as well as providing detailed background and answering related questions. Besides, the text processing capabilities of the Qwen2-VL models have seen significant improvements, particularly concerning the recognition of Chinese and English text within images.

Multi-object Recognition



The image shows a stack of colorful blocks arranged in four rows. The top row has one blue block with the number 9. The second row has two blocks: a light blue block with the number 7 and a green block with the number 8. The third row has three blocks: a purple block with the number 4, a pink block with the number 5, and a light green block with the number 6. The bottom row has four blocks: a red block with the number 0, an orange block with the number 1, a yellow block with the number 2, and a light green block with the number 3.

Output the color and number of each box.

Model Response

The image shows a stack of colorful blocks with numbers on them. Here is the color and number of each block from top to bottom:

- Top row (single block): Blue with the number 9 - Second row (two blocks): Light blue with the number 7, and green with the number 8 - Third row (three blocks): Purple with the number 4, pink with the number 5, and light green with the number 6 - Bottom row (four blocks): Red with the number 0, orange with the number 1, yellow with the number 2, and light green with the number 3

Figure 7: When presented with an image of cubes of different colors, the models identify their layout and the color of each cube.

Plant identification



What kind of flower is this?

Model Response

The flowers in the picture are forget-me-nots (*Myosotis*). These small, delicate flowers are known for their vibrant colors, typically blue, pink, or white, and are often used in bouquets and gardens for their charming appearance.

Figure 8: The model displays an adeptness in recognizing flowers in photographs.

Document Parsing with Dense Formulas

1. If $i > 0$, let $\underline{d}'' := (i - 1, d - i - l + 1, l)$. Let $V_{\underline{d}}$ and $V_{\underline{d}''}$ be $r + 1$ -dimensional subspaces of $H^0(\mathcal{L}_{\underline{d}})$ and $H^0(\mathcal{L}_{\underline{d}''})$, respectively, such that $\varphi_{\underline{d}, \underline{d}''}(V_{\underline{d}}) = V_{\underline{d}''}^{X_1, 0}$. Then the distributivity holds in $V_{\underline{d}''}$ if and only if

$$\dim(V_{\underline{d}''}^{X_2, 0} + V_{\underline{d}''}^{X_3, 0}) - \dim(V_{\underline{d}''}^{X_1, 0} + V_{\underline{d}''}^{X_3, 0}) = r + 1 - \dim(V_{\underline{d}''}^{X_1, 0} + V_{\underline{d}''}^{X_2, 0} + V_{\underline{d}''}^{X_3, 0})$$

2. If $i + l \leq d - 1$, let $\underline{d}''' := (i, d - i - l - 1, l + 1)$. Let $V_{\underline{d}}$ and $V_{\underline{d}'''}$ be $r + 1$ -dimensional subspaces of $H^0(\mathcal{L}_{\underline{d}})$ and $H^0(\mathcal{L}_{\underline{d}'''})$, respectively, such that $\varphi_{\underline{d}, \underline{d}'''}(V_{\underline{d}}) = V_{\underline{d}'''}^{X_3, 0}$. Then the distributivity holds in $V_{\underline{d}}$ if and only if

$$\dim(V_{\underline{d}'''}^{X_1, 0} + V_{\underline{d}'''}^{X_2, 0}) - \dim(V_{\underline{d}'''}^{X_1, 0} + V_{\underline{d}'''}^{X_3, 0}) = r + 1 - \dim(V_{\underline{d}'''}^{X_1, 0} + V_{\underline{d}'''}^{X_2, 0} + V_{\underline{d}'''}^{X_3, 0}),$$

where the maps $\varphi_{\underline{d}, \underline{d}''}$ and $\varphi_{\underline{d}''' \rightarrow \underline{d}}$ in Proposition 3.14 are the maps linking the corresponding sheaves. Another important result is Proposition 3.16, which establishes an inequality for any exact limit linear series. Specifically, our Proposition 3.16 says:

Let $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ be an exact limit linear series of degree d and dimension r . Then

$$\sum_{\underline{d}} \dim\left(\frac{V_{\underline{d}}}{V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}}\right) \geq r + 1.$$

As a consequence of Proposition 3.14 and Proposition 3.16, in Corollary 3.17, we get the following characterization of exact limit linear series satisfying the distributivity at each multidegree:

Let $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ be an exact limit linear series of degree d and dimension r . Then

$$\sum_{\underline{d}} \dim\left(\frac{V_{\underline{d}}}{V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}}\right) = r + 1$$

if and only if the distributivity holds in $V_{\underline{d}}$ for any \underline{d} .

In Section 3, we also study simple limit linear series. In Proposition 3.18, we characterize the multidegrees which determine all simple bases. In Proposition 3.20, following the construction in Muñoz [5], Proposition 4.4, we present a certain structure of linearly independent sections for any exact limit linear series satisfying the distributivity property. In propositions 3.21 and 3.22, we obtain some properties of that structure. These properties imply the existence of a simple basis (we give a proof of that in Proposition 3.23).

We conclude Section 3 with a characterization of simple limit linear series among the exact limit linear series. Specifically, our Corollary 3.24 says:

Let $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ be an exact limit linear series of degree d and dimension r . Then the following statements are equivalent:

1. $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ is simple.

$$2. \sum_{\underline{d}} \dim\left(\frac{V_{\underline{d}}}{V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}}\right) = r + 1.$$

3. The distributivity holds in $V_{\underline{d}}$ for any \underline{d} .

In a subsequent work, given a refined limit linear series, we will construct all simple extensions by combining our Proposition 3.14 and the method of Muñoz [6] for the construction of all exact extensions.

Convert the content in the image to Markdown

Model Response

1. If $i > 0$, let $\underline{d}'' := (i - 1, d - i - l + 1, l)$. Let $V_{\underline{d}}$ and $V_{\underline{d}''}$ be $r + 1$ -dimensional subspaces of $H^0(\mathcal{L}_{\underline{d}})$ and $H^0(\mathcal{L}_{\underline{d}''})$, respectively, such that $\varphi_{\underline{d}, \underline{d}''}(V_{\underline{d}}) = V_{\underline{d}''}^{X_1, 0}$. Then the distributivity holds in $V_{\underline{d}''}$ if and only if

$$\dim(V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}) - \dim(V_{\underline{d}''}^{X_2, 0} + V_{\underline{d}''}^{X_3, 0}) = r + 1 - \dim(V_{\underline{d}''}^{X_1, 0} + V_{\underline{d}''}^{X_2, 0} + V_{\underline{d}''}^{X_3, 0})$$

2. If $i + l \leq d - 1$, let $\underline{d}''' := (i, d - i - l - 1, l + 1)$. Let $V_{\underline{d}}$ and $V_{\underline{d}'''}$ be $r + 1$ -dimensional subspaces of $H^0(\mathcal{L}_{\underline{d}})$ and $H^0(\mathcal{L}_{\underline{d}'''})$, respectively, such that $\varphi_{\underline{d}''' \rightarrow \underline{d}}(V_{\underline{d}}) = V_{\underline{d}'''}^{X_3, 0}$. Then the distributivity holds in $V_{\underline{d}}$ if and only if

$$\dim(V_{\underline{d}'''}^{X_1, 0} + V_{\underline{d}'''}^{X_2, 0}) - \dim(V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0}) = r + 1 - \dim(V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}),$$

where the maps $\varphi_{\underline{d}, \underline{d}''}$ and $\varphi_{\underline{d}''' \rightarrow \underline{d}}$ in Proposition 3.14 are the maps linking the corresponding sheaves. Another important result is Proposition 3.16, which establishes an inequality for any exact limit linear series. Specifically, our Proposition 3.16 says:

Let $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ be an exact limit linear series of degree d and dimension r . Then

$$\sum_{\underline{d}} \dim\left(\frac{V_{\underline{d}}}{V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}}\right) \geq r + 1.$$

As a consequence of Proposition 3.14 and Proposition 3.16, in Corollary 3.17, we get the following characterization of exact limit linear series satisfying the distributivity at each multidegree:

Let $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ be an exact limit linear series of degree d and dimension r . Then

$$\sum_{\underline{d}} \dim\left(\frac{V_{\underline{d}}}{V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}}\right) = r + 1$$

if and only if the distributivity holds in $V_{\underline{d}}$ for any \underline{d} .

In Section 3, we also study simple limit linear series. In Proposition 3.18, we characterize the multidegrees which determine all simple bases. In Proposition 3.20, following the construction in Muñoz [5], Proposition 4.4, we present a certain structure of linearly independent sections for any exact limit linear series satisfying the distributivity property. In propositions 3.21 and 3.22, we obtain some properties of that structure. These properties imply the existence of a simple basis (we give a proof of that in Proposition 3.23).

We conclude Section 3 with a characterization of simple limit linear series among the exact limit linear series. Specifically, our Corollary 3.24 says:

Let $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ be an exact limit linear series of degree d and dimension r . Then the following statements are equivalent:

1. $\{\mathcal{L}_{\underline{d}}, V_{\underline{d}}\}_{\underline{d}}$ is simple.

$$\sum_{\underline{d}} \dim\left(\frac{V_{\underline{d}}}{V_{\underline{d}}^{X_1, 0} + V_{\underline{d}}^{X_2, 0} + V_{\underline{d}}^{X_3, 0}}\right) = r + 1.$$

3. The distributivity holds in $V_{\underline{d}}$ for any \underline{d} .

In a subsequent work, given a refined limit linear series, we will construct all simple extensions by combining our Proposition 3.14 and the method of Muñoz [6] for the construction of all exact extensions.

Figure 9: Literary writing in multiple languages based on visual stimuli.

Multilingual Text Recognition

汉语，也称为“华语”。是中国使用人数最多的语言，也是世界上作为第一语言使用人数最多的语言。是中华优秀传统文化的重要载体。繁体字，又称为“繁体中文”，与“简化字”/“简体字”相对。一般是指汉字简化运动被简化字所代替的汉字。

日本語は地方ごとに多様な方言があり、とりわけ琉球諸島で方言差が著しい。

한국어(韓國語), 조선말(朝鮮 말)은 대한민국과 조선민주주의인민공화국의 공용어이다. 둘은 표기나 문법에서는 차이가 있지만 동사 어미나 표현에서 차이가 있다.

Le français est une langue indo-européenne de la famille des langues romanes dont les locuteurs sont appelés « francophones ».

El español o castellano es una lengua romance procedente del latín hablado, perteneciente a la familia de lenguas indoeuropeas.

A língua portuguesa, também designada português, é uma língua indo-europeia romântica flexiva ocidental originada no galego-português falado no Reino da Galiza e no norte de Portugal.

Is ceann de na teangacha Ceilteacha i an Ghaeilge (nó Gaeilge na hÉireann mar a thugtar uirthi corrúair), agus ceann de na trí cinn de theangacha Ceilteacha ar a dtugtar na teangacha Gaelacha (Gaeilge, Gaeilge Mhanann agus Gaeilge na hAlban) go háirthe.

English is a West Germanic language in the Indo-European language family, whose speakers, called Anglophones, originated in early medieval England on the island of Great Britain.

Die deutsche Sprache bzw. Deutsch ist eine westgermanische Sprache, die weltweit etwa 90 bis 105 Millionen Menschen als Muttersprache und weiteren rund 80 Millionen als Zweit- oder Fremdsprache dient.

Język polski, polszczyzna – język z grupy zachodniośląskiej (do której należą również czeski, kaszubski, słowacki, języki Łużyckie czy wymarły język drzewiński), stanowiącej część rodzin indoeuropejskich.

Η ελληνική γλώσσα ανήκει στην ινδοευρωπαϊκή οικογένεια και αποτελεί το μοναδικό μέλος του ελληνικού κλαδού.

Tiếng Việt hay Việt ngữ là một ngôn ngữ thuộc ngữ hệ Nam Á, được công nhận là ngôn ngữ chính thức tại Việt Nam.

Монгол хэл нь Монгол улсын албан ёсны хэл юм.

Число владеющих русским языком в России составляет 137,5 млн человек (2010). Всего в мире на русском говорят 258,2 млн человек (2022).

हिन्दी विश्व की एक प्रमुख भाषा है और भारत की एक राजभाषा है। केंद्रीय सरकार भारत में सह-आधिकारिक भाषा घोषणी है।

Kiswahili ni lugha ya Kibantu yenye misamati mingi ya Kiarabu (35%), lakini sasa ya Kiingereza pia (10%), inayozungumzwa katika eneo kubwa la Afrika ya Mashariki.

转录图中内容，并识别图中每段内容的语言。

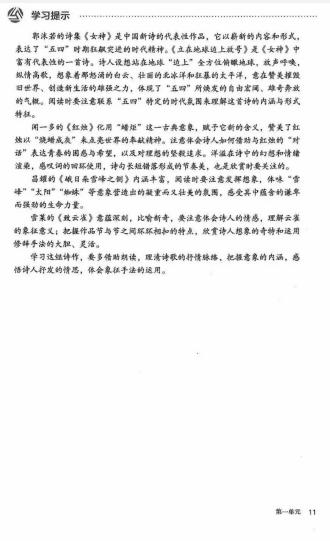
(Translation: Transcribe the picture and identify the language of each paragraph in it.)

Model Response

- 汉语，也称为“华语”。是中国使用人数最多的语言，也是世界上作为第一语言使用人数最多的语言。是中华优秀传统文化的重要载体。繁体字，又称为“繁中文”，“化字”/“字”相。一般是指数字化被化字所代替的字。 - 语言: 中文
- 日本は地方ごとに多様な方言があり、とりわけ琉球で方言差が著しい。 - 语言: 日语
- 한국어(韓國語), 조선말(朝鮮 말)은 대한민국과 조선민주주의인민공화국의 공용어이다. 둘은 표기나 문법에서는 차이가 없지만 동사 어미나 표현에서 차이가 있다。 - 语言: 韩语
- Le français est une langue indo-européenne de la famille des langues romanes dont les locuteurs sont appelés « francophones ». - 语言: 法语
- El español o castellano es una lengua romance procedente del latín hablado, perteneciente a la familia de lenguas indoeuropeas. - 语言: 西班牙语
- A língua portuguesa, também designada português, é uma língua indo-europeia romântica flexiva ocidental originada no galego-português falado no Reino da Galiza e no norte de Portugal. - 语言: 葡萄牙语
- Is ceann de na teangacha Ceilteacha i an Ghaeilge (nó Gaeilge na hÉireann mar a thugtar uirthi corrúair), agus ceann de na trí cinn de theangacha Ceilteacha ar a dtugtar na teangacha Gaelacha (Gaeilge, Gaeilge Mhanann agus Gaeilge na hAlban) go háirthe. - 语言: 爱尔兰语
- English is a West Germanic language in the Indo-European language family, whose speakers, called Anglophones, originated in early medieval England on the island of Great Britain. - 语言: 英语
- Die deutsche Sprache bzw. Deutsch ist eine westgermanische Sprache, die weltweit etwa 90 bis 105 Millionen Menschen als Muttersprache und weiteren rund 80 Millionen als Zweit- oder Fremdsprache dient. - 语言: 德语
- Język polski, polszczyzna – język z grupy zachodniośląskiej (do której należą również czeski, kaszubski, słowacki, języki Łużyckie czy wymarły język drzewiński), stanowiącej część rodzin indoeuropejskich. - 语言: 波兰语
- Η ελληνική γλώσσα ανήκει στην ινδοευρωπαϊκή οικογένεια και αποτελεί το μοναδικό μέλος του ελληνικού κλαδού。 - 语言: 希腊语
- Tiếng Việt hay Việt ngữ là một ngôn ngữ thuộc ngữ hệ Nam Á, được công nhận là ngôn ngữ chính thức tại Việt Nam. - 语言: 越南语
- Монгол хэл нь Монгол улсын албан ёсны хэл юм。 - 语言: 蒙古语
- Kiswahili ni lugha ya Kibantu yenye misamati mingi ya Kiarabu (35%), lakini sasa ya Kiingereza pia (10%), inayozungumzwa katika eneo kubwa la Afrika ya Mashariki。 - 语言: 斯瓦希里语

Figure 10: The model displays an adeptness in recognizing multilingual texts in images.

Text Information Recognition & Processing



Translate all the text in the picture into English and standardize the format

Model Response

The text in the image reads:

Learning Prompt

Guo Moruo's collection of poems, "The Goddess", is a representative work of modern Chinese poetry. It expresses new content and forms, embodying the spirit of the May Fourth Movement. "Standing on the Edge of the Earth" is a representative poem from "The Goddess". The poet imagines standing on the edge of the earth to observe it from all angles, shouting and singing, imagining the surging clouds, magnificent Arctic ice, and violent Pacific Ocean, praising the power to destroy the old world and create a new life. This poem embodies the free, and bold spirit of the May Fourth Movement. When reading, pay attention to connecting with the specific context of the May Fourth Movement to understand the content and form of this poem.

Wen Yiduo's "Red Candle" uses the classical image of "candle" to give it a new meaning, praising the red candle for its dedication to lighting up the world. Pay attention to how the poet uses the dialogue between "red candle" and "red candle" to express the confusion and hope of youth, as well as the firm pursuit of ideals. The use of fantasy and emotional rhetoric, the use of rhetorical devices such as alliteration and assonance, and the formation of rhythmic beauty through the arrangement of long and short sentences in the poem are also important aspects to appreciate.

Chang Yao's "Snow Capped Peak on the Side of the Sun" is rich in content. When reading, pay attention to the development of imagination and taste the heavy and majestic atmosphere created by the images of "snow-capped peak", "sun", and "spider". Feel the powerful contained in it.

Shelley's "To a Skylark" has profound implications and novel imagery. Pay attention to the poet's emotions and understand the symbolic meaning of the skylark. Grasp the characteristics of the interlocking between sections of the work and appreciate the poet's use of bold and flexible imagery.

When learning this group of poems, rely more on reading aloud to clarify the emotional context of the poem and grasp the connotation of the image. Feel the emotions expressed by the poet and understand the use of symbolism.

Figure 11: Recognition of dense Chinese text. The model is capable of directly converting lengthy, closely packed Chinese characters into standard English with accuracy and fluency.

Multilingual Text Recognition



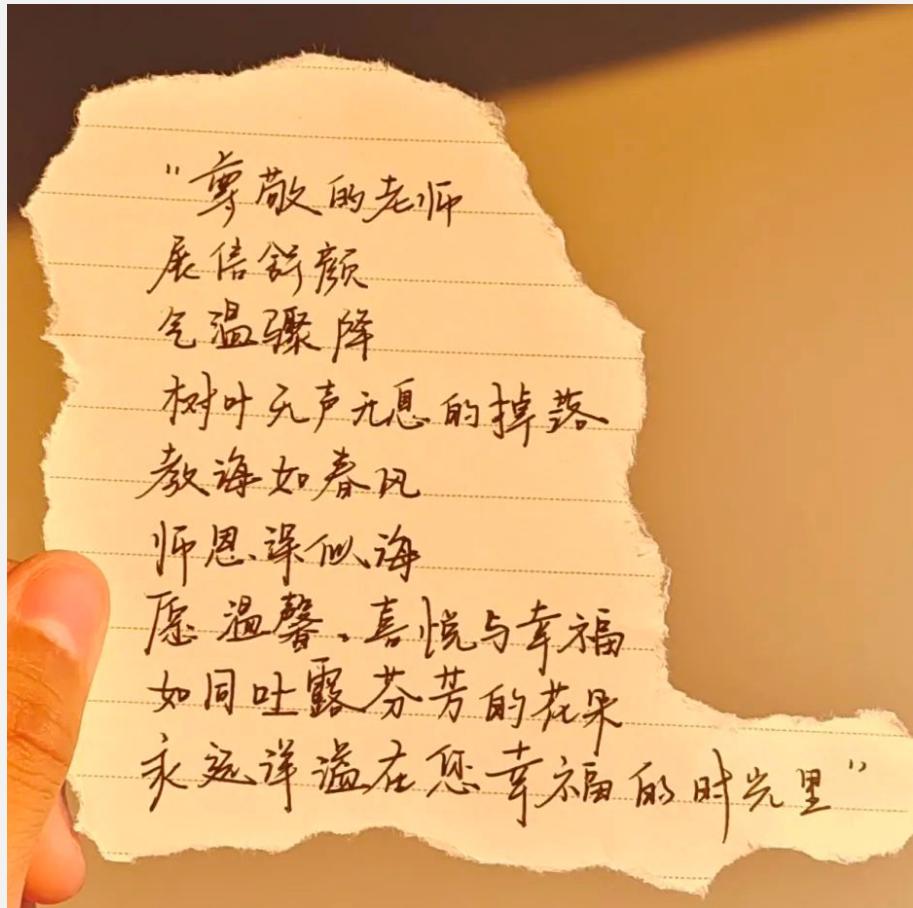
What is the text?

Model Response

The text on the wall is in Portuguese and reads: “não é normal ter medo de andar sozinha na rua!!!”
This translates to: “It is not normal to be afraid to walk alone on the street!!!”

Figure 12: The model displays an adeptness in recognizing multilingual texts in images. Image source: ([Yang et al., 2023](#))

Multilingual Text Recognition



识别图中所有文字，无需添加标点。

(Translation: Recognize all the text in the image without adding punctuation.)

Model Response

尊敬的老师
展信舒颜
气温骤降
树叶无声无息的掉落
教诲如春风
师恩深似海
愿温馨、喜悦与幸福
如同吐露芬芳的花朵
永远洋溢在您幸福的时光里

Figure 13: The model displays an adeptness in recognizing multilingual texts in images.

A.2 Information extraction and Visual Reasoning

A notable advancement in the Qwen2-VL models is their enhanced visual reasoning capability. This advancement allows the models to interpret and comprehend complex representations such as flowcharts, diagrams, and other symbolic systems.

Mathematical Problem Solving

Find the length of AC in the isosceles triangle ABC .

Model Response

Since triangle ABC is isosceles, we know that $AC = BC$. Therefore, we can set the expressions for AC and BC equal to each other and solve for x

$$4x + 1 = 5x - 0.5$$

Subtracting $4x$ from both sides, we get:

$$1 = x - 0.5$$

Adding 0.5 to both sides, we get:

$$1.5 = x$$

Now that we know the value of x , we can substitute it back into the expression for AC to find its length:

$$AC = 4x + 1$$
$$AC = 4(1.5) + 1$$
$$AC = 6 + 1$$
$$AC = 7$$

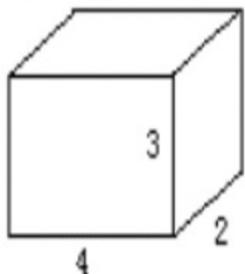
Therefore, the length of AC is 7.

Figure 14: The model displays an adeptness in solving mathematical problems. Image source: [Lu et al. \(2021\)](#)

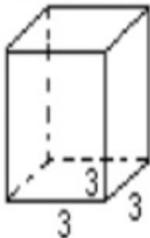
Mathematical Problem Solving

3、计算下面图形的表面积和体积(8分)

①(单位:厘米)



②(单位:分米)



Read the image and solve it step by step.

Model Response

The image shows two geometric figures, a rectangular prism and a cube, with their dimensions labeled. The problem asks to calculate the surface area and volume of these figures.

1. For the rectangular prism:

$$\text{Length (l)} = 4 \text{ cm}$$

$$\text{Width (w)} = 2 \text{ cm}$$

$$\text{Height (h)} = 3 \text{ cm}$$

$$\text{Surface Area (SA) of a rectangular prism} = 2lw + 2lh + 2wh$$

$$\text{Volume (V) of a rectangular prism} = lwh$$

Substituting the given values:

$$SA = 2(4\text{cm})(2\text{cm}) + 2(4\text{cm})(3\text{cm}) + 2(2\text{cm})(3\text{cm})$$

$$SA = 16\text{cm}^2 + 24\text{cm}^2 + 12\text{cm}^2 \quad SA = 52\text{cm}^2$$

$$V = (4\text{cm})(2\text{cm})(3\text{cm}) \quad V = 24\text{cm}^3$$

2. For the cube:

$$\text{Side length (a)} = 3 \text{ dm}$$

$$\text{Surface Area (SA) of a cube} = 6a^2 \quad \text{Volume (V) of a cube} = a^3$$

Substituting the given value:

$$SA = 6(3\text{dm})^2 \quad SA = 6(9\text{dm}^2) \quad SA = 54\text{dm}^2$$

$$V = (3\text{dm})^3 \quad V = 27\text{dm}^3$$

So, the surface area and volume for each figure are as follows:

For the rectangular prism:

$$\text{Surface Area} = 52 \text{ cm}^2$$

$$\text{Volume} = 24 \text{ cm}^3$$

For the cube:

$$\text{Surface Area} = 54 \text{ dm}^2$$

$$\text{Volume} = 27 \text{ dm}^3$$

Figure 15: Solving a math problem. The model calculates the surface area and volume of these figures with step-by-step explanation

Algorithmic Problem Solving

Weekly Contest 410 < > Run Submit Description Editorial Solutions Submissions Register or Sign in Premium

3248. Snake in Matrix

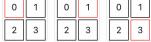
Easy Companies Hint

There is a snake in an $n \times n$ matrix grid and can move in **four possible directions**. Each cell in the grid is identified by the position: $\text{grid}[i][j] = (i * n) + j$.

The snake starts at cell 0 and follows a sequence of commands.

You are given an integer n representing the size of the grid and an array of strings `commands` where each `command[i]` is either "UP", "RIGHT", "DOWN", and "LEFT". It's guaranteed that the snake will remain within the grid boundaries throughout its movement.

Return the position of the final cell where the snake ends up after executing `commands`.

Example 1:
Input: $n = 2$, `commands` = ["RIGHT", "DOWN"]
Output: 3
Explanation:


Example 2:
Input: $n = 3$, `commands` = ["DOWN", "RIGHT", "UP"]
Output: 1
Explanation:


Code

```
1 class Solution(object):
2     def finalPositionOfSnake(self, n, commands):
3         """
4             :type n: int
5             :type commands: List[str]
6             :rtype: int
7         """
8
```

Python Auto

Saved You need to [Login / Sign up](#) to run or submit Ln 1, Col 1

Testcase Test Result

Solve the algorithmic problem in the web page.

Model Response

To solve the “Snake in Matrix” problem, we need to implement a function that takes the size of the matrix n and a list of commands, and returns the final position of the snake after executing all the commands. Here’s a step-by-step approach to solve this problem:

1. Initialize Variables:

- Initialize the current position of the snake as 0.
- Initialize the direction of the snake as **RIGHT** (initially, the snake moves to the right).

2. Define Movement Functions:

- Create a function to move the snake based on the current direction and the size of the matrix n .
- Update the position of the snake accordingly.

3. Process Commands:

- Iterate through each command in the list of commands.
- For each command, update the direction and position of the snake.

4. Return Final Position:

- After processing all commands, return the final position of the snake.

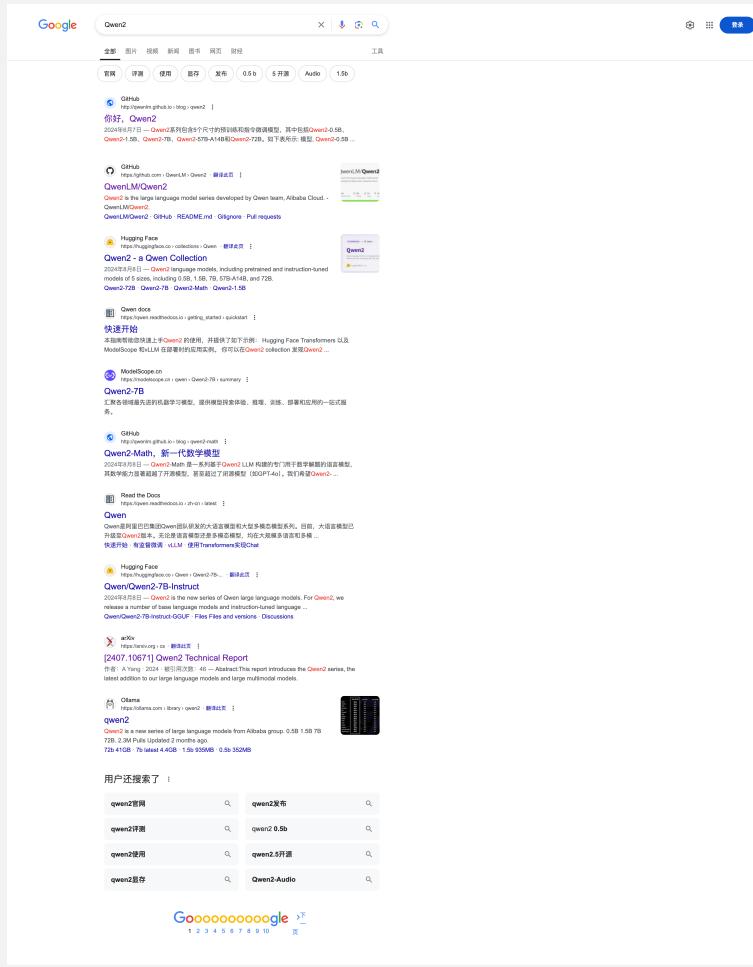
Here’s the Python code to implement the solution:

```
class Solution:  
    def finalPositionOfSnake(self, n: int, commands: List[str]) -> int:  
        # Initialize variables  
        position = 0  
        direction = "RIGHT"  
        # Define movement functions  
        def move_right():  
            nonlocal position  
            position += 1  
        def move_down():  
            nonlocal position  
            position += n  
        def move_left():  
            nonlocal position  
            position -= 1  
        def move_up():  
            nonlocal position  
            position -= n  
        # Process commands  
        for command in commands:  
            if command == "RIGHT":  
                move_right()  
            elif command == "DOWN":  
                move_down()  
            elif command == "LEFT":  
                move_left()  
            elif command == "UP":  
                move_up()  
        # Return final position  
        return position
```

This code defines a class `Solution` with a method `finalPositionOfSnake` that takes the size of the matrix n and a list of commands, and returns the final position of the snake after executing all the commands. The movement functions (`move_right`, `move_down`, `move_left`, `move_up`) are used to update the position of the snake based on the current direction.

Figure 16: The model displays an adeptness in solving algorithmic problems.

Web Content Recognition



Extract all page titles in the screenshot of search result.

Model Response

The page titles in the screenshot of the search result are:

1. 你好, Qwen2
2. QwenLM/Qwen2
3. Qwen2 - a Qwen Collection
4. 快速开始
5. Qwen2-7B
6. Qwen2-Math, 新一代数学模型
7. Qwen
8. Qwen/Qwen2-7B-Instruct
9. [2407.10671] Qwen2 Technical Report
10. qwen2

Figure 17: The model displays an adeptness in recognizing content from the web page.

OCR and Mathematical Problem Solving

Ruka Kayamori		Yuki Izumi	
A	SS	A	SS
			
Personal information		Personal information	
Japanese Name	茅森 月歌	Japanese Name	和泉ユキ
Addresses self as	atashi (あたし)	Addresses self as	atashi(あたし)
Height	159cm	Height	160cm
Birthday	24 June	Birthday	17 September
Seraphim Information		Seraphim Information	
Squad	 31A	Squad	 31A
Seraphim Weapon	Brave Blue 	Seraphim Weapon	Rapid Fire 
Seraphim Code	あたしの伝説はこれから始まる My legend is starting now	Seraphim Code	Hello World

Please tell me how many days are between Ruka's birthday and Yuki's birthday? Think carefully step by step.

Model Response

To determine the number of days between Ruka's and Yuki's birthdays, we need to calculate the difference between their respective birth dates.

1. Ruka's birthday is on June 24.
2. Yuki's birthday is on September 17.

Let's break it down step by step:

1. **From June 24 to June 30:**
 - June 24 to June 30 is 6 days.
2. **From July 1 to August 31:**
 - July has 31 days.
 - August has 31 days.
 - Total days in July and August: $31 + 31 = 62$ days.
3. **From September 1 to September 17:**
 - September 1 to September 17 is 17 days.

Now, we add all these days together:

- June 24 to June 30: 6 days
- July 1 to August 31: 62 days
- September 1 to September 17: 17 days

Total days = $6 + 62 + 17 = 85$ days

Therefore, there are 85 days between Ruka's and Yuki's birthdays.

Figure 18: The model displays an adeptness in OCR and mathematical reasoning.

OCR and Format Following

The screenshot shows the homepage of The Linux Kernel Archives. At the top right, there is a yellow button labeled "Latest Release" with the text "6.10.4" and a small icon. Below this, there is a table with columns "Protocol" and "Location". The table includes rows for HTTP, GIT, and RSYNC. The main content area displays a list of kernel versions with their release dates and download links. On the left, there is a sidebar with "Other resources" and "Social" sections. The bottom of the page has a footer with copyright information.

Protocol Location

HTTP	https://www.kernel.org/pub/
GIT	https://git.kernel.org/
RSYNC	rsync://rsync.kernel.org/pub/

Latest Release
6.10.4

mainline:	version	release date	tarball	patch	diff	browse
stable:	6.10.4	2024-08-11	[tarball]	[patch]	[view diff]	[browse]
stable:	6.9.12 [EOL]	2024-08-11	[tarball]	[patch]	[view diff]	[browse]
longterm:	6.6.45	2024-08-11	[tarball]	[patch]	[view diff]	[browse]
longterm:	6.1.104	2024-08-11	[tarball]	[patch]	[view diff]	[browse]
longterm:	5.15.164	2024-07-27	[tarball]	[patch]	[view diff]	[browse]
longterm:	5.10.223	2024-07-27	[tarball]	[patch]	[view diff]	[browse]
longterm:	5.4.281	2024-07-27	[tarball]	[patch]	[view diff]	[browse]
longterm:	4.19.319	2024-07-27	[tarball]	[patch]	[view diff]	[browse]
linux-next:	next-20240812	2024-08-12				[browse]

linux-next: next-20240812 2024-08-12

Other resources

Git Trees	Documentation	Kernel Mailing Lists
Patchwork	Wikis	Bugzilla
Mirrors	Linux.com	Linux Foundation

Social

Site Atom Feed	Releases Atom Feed
	Kernel Planet

This site is operated by the Linux Kernel Organization, Inc., a 501(c)3 nonprofit corporation, with support from the following sponsors.

What are the Linux versions and their release dates in the picture? Return results as a JSON list.

Model Response

```
[{"version": "6.11-rc3", "release_date": "2024-08-11"}, {"version": "6.10.4", "release_date": "2024-08-11"}, {"version": "6.9.12 [EOL]", "release_date": "2024-07-27"}, {"version": "6.6.45", "release_date": "2024-08-11"}, {"version": "6.1.104", "release_date": "2024-08-11"}, {"version": "5.15.164", "release_date": "2024-07-27"}, {"version": "5.10.223", "release_date": "2024-07-27"}, {"version": "5.4.281", "release_date": "2024-07-27"}, {"version": "4.19.319", "release_date": "2024-07-27"}, {"version": "next-20240812", "release_date": "2024-08-12"}]
```

Figure 19: The model displays an adeptness in OCR and following formats.

OCR and Format Following



将图中的每日的最高气温和最低气温整理成表格。(Translation: Organize the daily highest and lowest temperatures from the image into a table.)

Model Response

以下是图中每日的最高气温和最低气温整理成的表格:

日期	最高气温	最低气温
08/12	31°C	23°C
08/13	30°C	22°C
08/14	30°C	23°C
08/15	31°C	24°C
08/16	31°C	24°C
08/17	30°C	24°C
08/18	30°C	23°C
08/19	30°C	22°C

希望这对你有帮助！

Figure 20: The model displays an adeptness in OCR and following formats.

OCR for Large Images



Original Image
(1792 x 14400)

视频理解

qwen-vl-max-0809支持对视频内容的理解功能。您可以直接传入视频文件。请参考以下限制条件：

- 如果传入视频文件：
 - 视频文件大小：最大 150MB。
 - 视频文件格式：MP4、AVI、MKV、MOV、FLV、WMV 等。
 - 视频尺寸：无限制，但是视频文件会被调整到约 600K 像素数，更大尺寸的视频文件不会有更好的理解效果。

Cropped region related to the question.
Note: The cropping effect is only for visualization. **The original image is directly used as the input of the model, without any crop.**

qwen-vl-max-0809对视频内容的理解功能存在哪些限制？

(Translation: What limitations exist in qwen-vl-max-0809's ability to understand video content?)

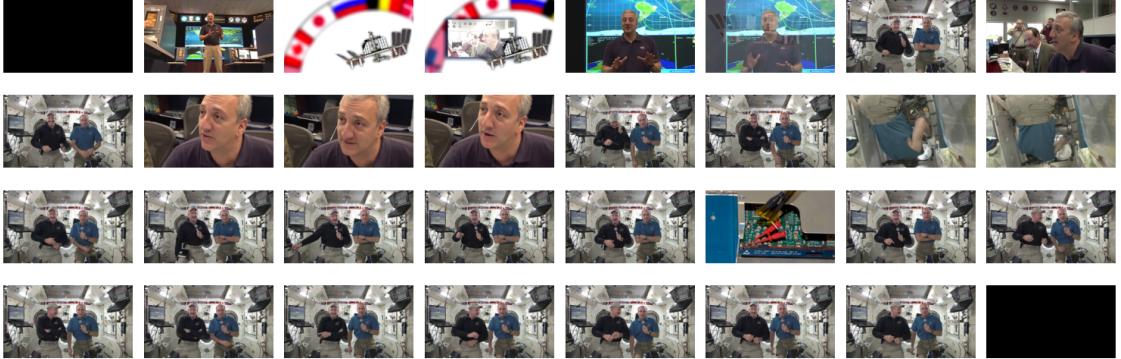
Model Response

qwen-vl-max-0809对视频内容的理解功能存在以下限制：视频文件大小最大为150MB，视频文件格式支持MP4、AVI、MKV、MOV、FLV、WMV等，视频尺寸无限制，但视频文件会被调整到约600K像素数，更大尺寸的视频文件不会有更好的理解效果。(Translation: qwen-vl-max-0809 has the following limitations on the video content understanding: the maximum video file size is 150MB, the video file formats supported are MP4, AVI, MKV, MOV, FLV, WMV, etc. There is no limitation on the video size, but the video file will be resized to about 600K pixels, and video files of larger sizes won't have better video understanding results.)

Figure 21: The model displays an adeptness in large images OCR.

A.3 Video Understanding

Multi-Round Video Chat



Please describe the video in detail.

Model Response

The video begins with a man speaking to the camera, followed by a group of people sitting in a control room. The camera then cuts to two men floating inside a space station, where they are seen speaking to the camera. The men appear to be astronauts, and they are wearing space suits. The space station is filled with various equipment and machinery, and the camera pans around to show the different areas of the station. The men continue to speak to the camera, and they appear to be discussing their mission and the various tasks they are performing. Overall, the video provides a fascinating glimpse into the world of space exploration and the daily lives of astronauts.

User Query

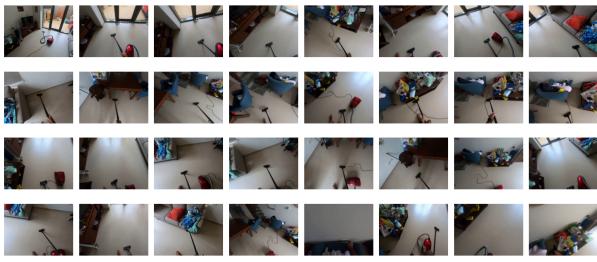
What is the color of the clothes the two astronauts wearing?

Model Response

The two astronauts are wearing blue and black clothes.

Figure 22: The model displays an adeptness in recognizing multi-round video chat.

Multi-Video Understanding



视频中的人在做什么?

(Translation: What are the people in the video doing?)

Model Response

视频中的人正在使用吸尘器清理地板。

(Translation: The man in the video is using a vacuum cleaner to clean the floor.)

User Query

如果要进一步清洁地面的话，接下来该怎么做?

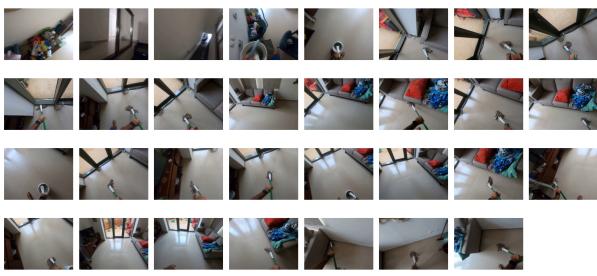
(Translation: What should I do next if I want to clean the floor further?)

Model Response

如果要进一步清洁地面，可以使用拖把或湿布擦拭地面，以去除吸尘器无法清除的污渍和灰尘。

(Translation: For further cleaning of the floor, wipe the floor with a mop or damp cloth to remove stains and dirt that the vacuum cleaner cannot remove.)

User Query



已经用拖把清洁了地面，评价一下清洁工作?

(Translation: Already cleaned the floor with a mop and please comment on the cleaning?)

Model Response

从视频中可以看出，清洁工作已经完成，地面已经变得干净整洁。

(Translation: As you can see from the video, the cleaning has been completed and the floor has been made clean and tidy.)

User Query

通过你看到的，接下来还能做些什么?

(Translation: By what you see, what else can you do next?)

Model Response

接下来可以将拖把清洗干净，晾干备用。

(Translation: Next you can clean the mop, dry it and set it aside.)

Figure 23: The model displays an adeptness in recognizing multi-video understanding.

A.4 Visual Agent Capability

The Qwen2-VL also excels in location and agent tasks.

Locate Whatever You Desire



Detect the bounding box of 'the red car'

Model Response

<|object_ref_start|>the red car<|object_ref_end|><|box_start|>(701,531),(869,675)<|box_end|>



Figure 24: Our models were able to locate specific elements within images, such as identifying the red car accurately.

Visual Referring Prompting

Maison Pour toi Suivant Vitrine de l'actualité | NOUS Monde Local Entreprise Technologie Divertissement Des sports Science Santé

9to5Mac

Disponibilité limitée de l'iPhone 15 : laquelle serait la moins frustrante ?

il y a 4 heures

IGN

L'acquisition d'Activision Blizzard par Microsoft est de bon augure pour Ubisoft...

il y a 2 heures

Le bord

OpenAI lance des instructions personnalisées pour ChatGPT

il y a 18 heures

Couverture complète

Couverture complète

Couverture complète

What does the pointed news talk about?

Model Response

The pointed news with a green logo of OpenAI and a title "Le chatbot ChatGPT d'OpenAI dispose désormais d'instructions personnalisées pour stocker vos préférences" is talking about the new feature of ChatGPT that can now store personalized instructions to better serve users.

Figure 25: Our model is capable of annotating and posing questions regarding the content present in webpage screenshots, demonstrating its potential as a visual agent. Image source: ([Yang et al., 2023](#))

Function Calling - Basic

几点到达，到的时候目的地天气怎么样
(Translation: What time does this flight arrive? What is the weather like at the destination when it arrives?)

Function Input

北京

Function Result

时间: 2024-07-30 14:00, 风向: 西南风, 风力: 0-3级微风<5.4m/s, 温度: 29°C, 天气: 小雨
时间: 2024-07-30 15:00, 风向: 西南风, 风力: 0-3级微风<5.4m/s, 温度: 28°C, 天气: 小雨
时间: 2024-07-30 16:00, 风向: 西南风, 风力: 0-3级微风<5.4m/s, 温度: 28°C, 天气: 小雨
时间: 2024-07-30 17:00, 风向: 西南风, 风力: 0-3级微风<5.4m/s, 温度: 28°C, 天气: 小雨
时间: 2024-07-30 18:00, 风向: 西南风, 风力: 0-3级微风<5.4m/s, 温度: 27°C, 天气: 中雨
时间: 2024-07-30 19:00, 风向: 南风, 风力: 0-3级微风<5.4m/s, 温度: 27°C, 天气: 中雨
时间: 2024-07-30 20:00, 风向: 南风, 风力: 0-3级微风<5.4m/s, 温度: 27°C, 天气: 中雨
时间: 2024-07-30 21:00, 风向: 南风, 风力: 0-3级微风<5.4m/s, 温度: 26°C, 天气: 小雨
时间: 2024-07-30 22:00, 风向: 南风, 风力: 0-3级微风<5.4m/s, 温度: 26°C, 天气: 小雨
时间: 2024-07-30 23:00, 风向: 南风, 风力: 0-3级微风<5.4m/s, 温度: 26°C, 天气: 小雨
(避免显示过长, 省略部分信息。)

Model Response

根据航班信息, 您乘坐的中国国航CA1723航班预计将于2024年7月30日20:30抵达北京首都机场T3航站楼。
在您到达时, 北京的天气预报显示会有中雨, 并且温度约为27°C。请注意携带雨具并做好防雨措施。祝您旅途愉快!
(Translation: According to flight information, your Air China flight CA1723 is expected to arrive at Beijing Capital Airport T3 Terminal at 20:30 on July 30, 2024. When you arrive, the weather forecast in Beijing shows moderate rain and the temperature is about 27°C. Please remember to bring rain gear and take rain protection measures. Have a pleasant journey!)

Figure 26: The model identified the destination and arrival time in the image, called the 24-hour weather query function, correctly input the destination, and extracted the weather at the arrival time from the query results, thus enabling it to answer the user's question.

Function Calling - Code Interpreter



Figure 29 Architecture of our speech interface for Llama 3.

Implement the flowchart, representing modules as functions without implementation details. Write the code and validate it using the code interpreter.

Function Input

```
# Placeholder functions for the modules
def speech_adapter(input_audio):
    # Placeholder for speech adapter implementation
    pass

# To avoid excessive length,
# the remaining functions without implementation details are omitted here.

# Example usage
input_audio = "talk about the weather in San Francisco."
speech_adapter_output = speech_adapter(input_audio)
speech_encoder_output = speech_encoder(speech_adapter_output)
system_prompt_output = system_prompt()
dialogue_history_output = dialogue_history()
llama_3_output = llama_3(speech_encoder_output, system_prompt_output, \
dialogue_history_output)
text_normalization_output = text_normalization(llama_3_output)
prosody_model_output = prosody_model(text_normalization_output)
output_audio = synthesizer(prosody_model_output)
```

Function Result

Finished execution.

Model Response

The code has been executed without any errors, which means that the structure of the code matches the flowchart and there are no syntax issues. However, since the functions are placeholders and do not perform any actual operations, we cannot validate the correctness of the implementation. The code simply represents the flowchart by defining each module as a function and demonstrating how the functions could be called in sequence to process input audio and produce output audio.

Figure 27: The model identified the steps in the flowchart, drafted the simulation code as required while omitting the details, and then successfully executed it through the code interpreter. Image source: ([Dubey et al., 2024](#))

Function Calling - Code Interpreter

Model	HumanEval	HumanEval+	MBPP	MBPP EvalPlus (base)
Llama 3 8B	72.6 ±6.8	67.1 ±7.2	60.8 ±4.3	72.8 ±4.5
Gemma 2 9B	54.3 ±7.6	48.8 ±7.7	59.2 ±4.3	71.7 ±4.5
Mistral 7B	40.2 ±7.5	32.3 ±7.2	42.6 ±4.3	49.5 ±5.0
Llama 3 70B	80.5 ±6.1	74.4 ±6.7	75.4 ±3.8	86.0 ±3.5
Mixtral 8×22B	75.6 ±6.6	68.3 ±7.1	66.2 ±4.1	78.6 ±4.1
GPT-3.5 Turbo	68.0 ±7.1	62.8 ±7.4	71.2 ±4.0	82.0 ±3.9
Llama 3 405B	89.0 ±4.8	82.3 ±5.8	78.8 ±3.6	88.6 ±3.2
GPT-4	86.6 ±5.2	77.4 ±6.4	80.2 ±3.5	83.6 ±3.7
GPT-4o	90.2 ±4.5	86.0 ±5.3	81.4 ±3.4	87.8 ±3.3
Claude 3.5 Sonnet	92.0 ±4.2	82.3 ±5.8	76.6 ±3.7	90.5 ±3.0
Nemotron 4 340B	73.2 ±6.8	64.0 ±7.3	75.4 ±3.8	72.8 ±4.5

Table 18 Pass@1 scores on code generation benchmarks. We report results on HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), as well as EvalPlus (Liu et al., 2024a) versions of these benchmarks.

Count the average value of each row in this chart and draw a bar chart to show it. Write the code and verify it using the code interpreter.

Function Input

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Extracting the numerical values from the table
data = {
    "Model": ["Llama 3 8B", "Gemma 2 9B", "Mistral 7B", "Llama 3 70B", \
              "Mixtral 8x22B", "GPT-3.5 Turbo", "Llama 3 405B", "GPT-4", \
              "GPT-4o", "Claude 3.5 Sonnet", "Nemotron 4 340B"], 
    "HumanEval": [72.6, 54.3, 40.2, 80.5, 75.6, 68.0, 89.0, 86.6, 90.2, \
                  92.0, 73.2], 
    "HumanEval+": [67.1, 48.8, 32.3, 74.4, 68.3, 62.8, 82.3, 77.4, 86.0, \
                   82.3, 64.0], 
    "MBPP": [60.8, 59.2, 42.6, 75.4, 66.2, 71.2, 78.8, 80.2, 81.4, 76.6, \
              75.4], 
    "MBPP EvalPlus (base)": [72.8, 71.7, 49.5, 86.0, 78.6, 82.0, 88.6, \
                             83.6, 87.8, 90.5, 72.8]
}

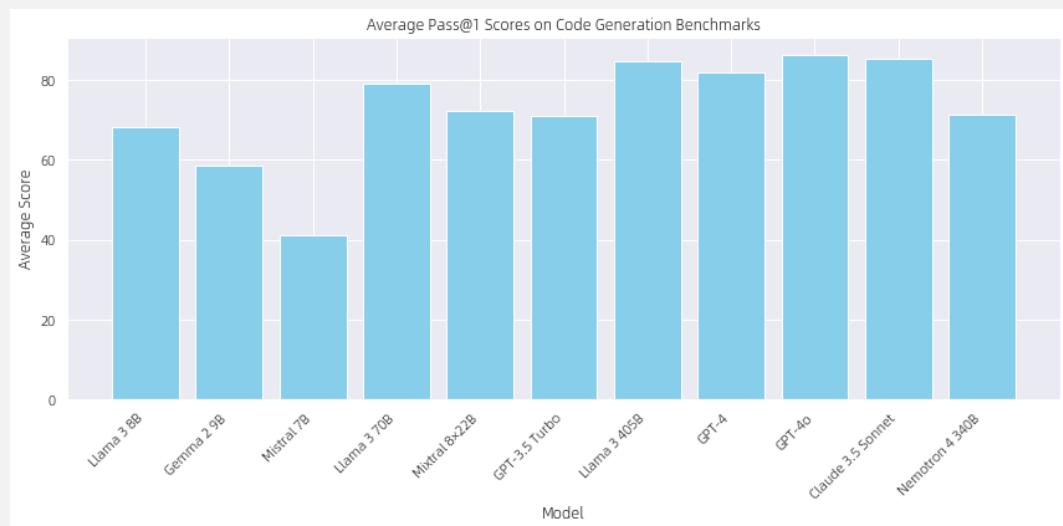
# Create a DataFrame
df = pd.DataFrame(data)

# Calculate the average for each row
df["Average"] = df.mean(axis=1)

# Plotting the bar chart
plt.figure(figsize=(12, 6))
plt.bar(df["Model"], df["Average"], color="skyblue")
plt.xlabel("Model")
plt.ylabel("Average Score")
plt.title("Average Pass@1 Scores on Code Generation Benchmarks")
plt.xticks(rotation=45, ha="right")
plt.tight_layout()
plt.show()

```

Function Result



Model Response

The average Pass@1 scores on code generation benchmarks for each model have been calculated and are displayed in the table above. The bar chart shows the average scores for each model, providing a visual comparison of their performance across the different benchmarks.

Figure 28: The model analyzed the table, and wrote the code to calculate the mean of each row in the table and to plot the results as a bar chart, which was successfully executed and visualized by the code interpreter. Image source: ([Dubey et al., 2024](#))

Function Calling - Code Interpreter

2.2.2 QWEN2 MIXTURE-OF-EXPERTS MODEL

The architecture of Qwen2 MoE models closely mirrors that of Qwen1.5-MoE-A2.7B (Qwen Team, 2024c). As a substitute for the original FFN, the MoE FFN consists of n individual FFNs, each serving as an expert. Each token is directed to a specific expert E_i for computation based on probabilities assigned by a gated network G :

$$\mathbf{p} = \text{softmax}(G(\mathbf{x})), \quad (1)$$

$$\mathbf{y} = \sum_{i \in \text{top}_k(\mathbf{p})} \mathbf{p}_i E_i(\mathbf{x}). \quad (2)$$

In the following, we present critical design considerations of Qwen2 MoE.

Implement the formula, write the code and verify it using the code interpreter.

Function Input

```
import numpy as np

def softmax(x):
    """Compute softmax values for each sets of scores in x."""
    e_x = np.exp(x - np.max(x))
    return e_x / e_x.sum(axis=0)

def top_k(p, k):
    """Return the indices of the top k elements in the probability vector p."""
    return np.argsort(p)[-k:][::-1]

# Example input
x = np.array([1, 2, 3, 4, 5]) # Example input vector
G = lambda x: x # Simple gating network (identity function for demonstration)
E = lambda x, i: x + i # Simple expert functions (add the expert index to the input)
# Compute probabilities
p = softmax(G(x))
# Select top k experts
k = 3
top_experts = top_k(p, k)
# Compute output
y = np.sum([p[i] * E(x, i) for i in top_experts])
p, top_experts, y
```

Function Result

```
(array([0.01165623, 0.03168492, 0.08612854, 0.23412166, 0.63640865]), array([4, 3, 2]), 31.451165958047895)
```

Model Response

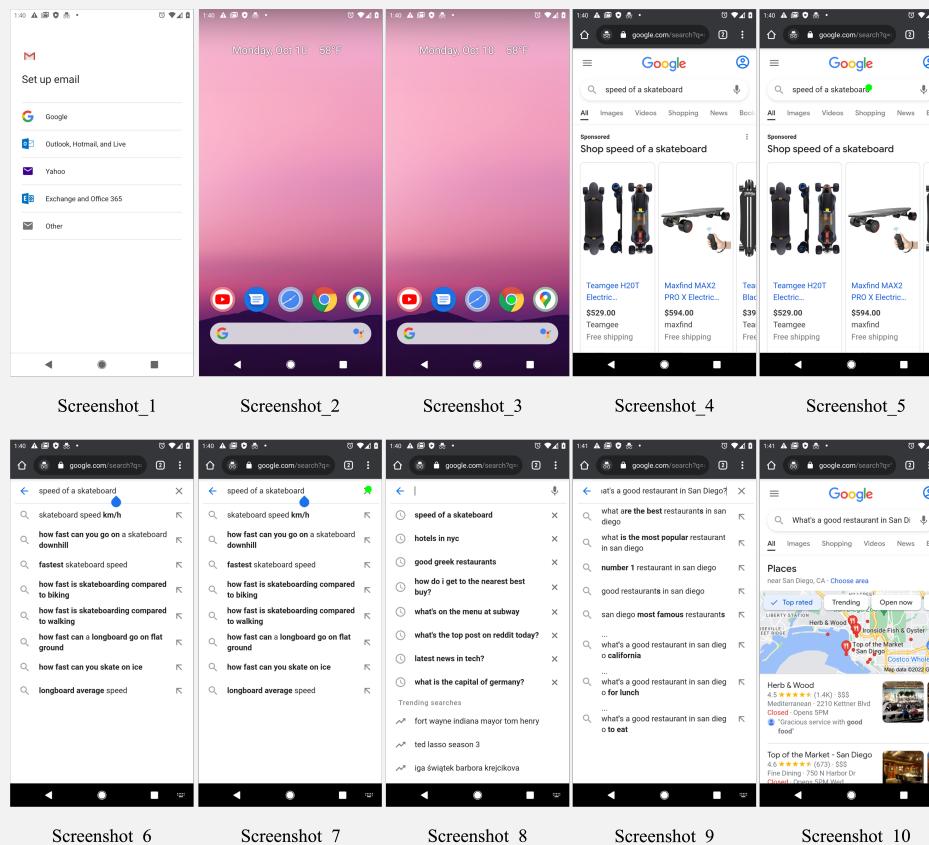
The probabilities assigned to each expert based on the input vector (\mathbf{x}) are: ($\mathbf{p} = [0.01165623, 0.03168492, 0.08612854, 0.23412166, 0.63640865]$).

The top ($k = 3$) experts selected based on these probabilities are the experts with indices 4, 3, and 2 (in descending order of probability). The output (\mathbf{y}) computed as a weighted sum of the outputs of these top experts is approximately 31.4512.

This example demonstrates a simple implementation of the MoE model's operation. In a real-world scenario, the gating network (G) and the experts (E_i) would be more complex, involving neural network layers and potentially non-linear transformations.

Figure 29: The model understood the formula, implemented the code as required, and successfully executed it in the code interpreter to obtain the results. Image source: [\(Dubey et al., 2024\)](#)

VL Agent - UI Operations



System Message

You are a helpful assistant.

Actions

You have the following actions.

Click

Click: A quick, light fingertip press that commands, selects, or navigates through a phone's user interface. Parameters: [{"name": "point", "description": "The specific point of interest on the screen, marked by the coordinate (x, y).", "required": true, "schema": "string"}]

Type

Type: Engaging with a smartphone's interface by entering text for various purposes like messaging, searching, or command execution. Parameters: [{"name": "text", "description": "The text to be typed on a smartphone.", "required": true, "schema": "string"}]

Swipe

Swipe: Touch the screen and move your finger to a new location for new content. Parameters: [{"name": "direction", "description": "Determine the swipe route from the cardinal directions: up, down, left, or right.", "required": true, "schema": "string", "choices": ["up", "down", "left", "right"]}]

Back

Back: Go in reverse to the prior interface or exiting the current context Parameters: []

Home

Home: Return to the device's home screen from any application or task. Parameters: []

Enter

Enter: Execute a command to confirm inputs or selections. Parameters: []

Done

Done: Indicate the completion of a task. Parameters: []

VL Agent - UI Operations

```
## Continuously take action until the task is completed.  
*FUNCTION*: The action to take, should be one of {Click,Type,Swipe,Back,Home,Enter,Done}  
*ARGS*: The input of the action.  
*RESULT*: Action results.  
*RETURN*: Reply based on action results.
```

User

Your task is: What's a good restaurant in San Diego?

Assistant

The screen shows the email setup page, which is not relevant to finding a restaurant. Possible actions are to exit the email setup and navigate to a search engine or map application where I can look up restaurants in San Diego.

FUNCTION: Home *ARGS*: {}

RESULT:

RETURN: The screen shows the home screen with various apps, including Google Chrome and Maps which could be used to search for a restaurant in San Diego. Possible actions are to tap on the Google Chrome app to open it and then use it to search for a good restaurant in San Diego.

FUNCTION: Click *ARGS*: {"point": "(687, 792)"}
RESULT:

RETURN: The current screen shows search results for skateboards, which is not related to the query about finding a good restaurant in San Diego. Possible actions are to click on the Google search bar at the top of the screen and enter the correct search term "good restaurant in San Diego" to find relevant information.

FUNCTION: Click *ARGS*: {"point": "(596, 213)"}
RESULT:

RETURN: The current screen shows search suggestions related to skateboard speeds, which is not relevant to the query about finding a good restaurant in San Diego. Possible actions are to click on the "X" button to correct the query for finding a good restaurant in San Diego.

FUNCTION: Click *ARGS*: {"point": "(927, 135)"}
RESULT:

RETURN: The screen shows the Google search bar with a list of previous searches, none of which are related to finding a restaurant in San Diego. Possible actions are to type "good restaurant in San Diego" into the search bar and then press enter to perform the search.

FUNCTION: Type *ARGS*: {"text": "What's a good restaurant in San Diego?"}

RESULT:

RETURN: The screen shows a list of search suggestions related to finding good restaurants in San Diego, indicating that the user has typed in a query but not yet executed it. Possible actions are to select one of the suggested searches or press enter on the keyboard to execute the current search and view the results.

FUNCTION: Enter *ARGS*: {}
RESULT:

RETURN: The screen shows a list of top-rated restaurants in San Diego with details such as ratings, price range, and opening hours. Possible actions are to stop and set the query as completed.

FUNCTION: Done *ARGS*: {}

Figure 30: Qwen2-VL as an agent understands the query with respect to UI operation, utilizes the pre-defined actions in system message, and fulfill the task step-by-step.

VL Agent - Card Game



Screenshot_1



Screenshot_2



Screenshot_3

System Message

You are a helpful assistant.

Actions

You have the following actions.

Hit

Hit: Get a new card. Parameters: []

Stand

Stand: Stop to get a new card. Parameters: []

Continuously take action until the task is completed.

FUNCTION: The action to take, should be one of {Hit,Stand}

ARGS: The input of the action.

RESULT: Action results.

RETURN: Reply based on action results.

User

You are a blackjack player. You are observing the current game state in the image. With the given functions, your goal is to dealer by getting as close to 21 points as possible without exceeding it.

Assistant

I have 8 points, and the dealer has 10 points. I should hit.

FUNCTION: Hit *ARGS*: {}

RESULT:

RETURN: I have 15 points, and the dealer has 10 points. I should hit.

FUNCTION: Hit *ARGS*: {}

RESULT:

RETURN: I have 17 points, and the dealer has 10 points. I should stand.

FUNCTION: Stand *ARGS*: {}

RESULT: Dealer Card: Q, 5, Q. The total point of dealer is 25 > 21.

Player Card: 4, 4, 7, 2. The total point of player is 17.

Player win!

Figure 31: Qwen2-VL recognizes these cards and utilizes Hit and Stand to play the blackjack.