LLaVA-OneVision: Easy Visual Task Transfer

Bo Li[2,♡] **Yuanhan Zhang**[2,♡] **Dong Guo**[1] **Renrui Zhang**[3,♡] **Feng Li**[4,♡] **Hao Zhang**[4,♡]
Kaichen Zhang[2] **Yanwei Li**[3,♡] **Ziwei Liu**[2] **Chunyuan Li**[1]

[1]ByteDance [2]S-Lab, NTU [3]CUHK [4]HKUST

https://llava-vl.github.io/blog/llava-onevision

Abstract

We present LLaVA-OneVision, a family of open large multimodal models (LMMs) developed by consolidating our insights into data, models, and visual representations in the LLaVA-NeXT blog series. Our experimental results demonstrate that LLaVA-OneVision is the first single model that can simultaneously push the performance boundaries of open LMMs in three important computer vision scenarios: single-image, multi-image, and video scenarios. Importantly, the design of LLaVA-OneVision allows strong transfer learning across different modalities/scenarios, yielding new emerging capabilities. In particular, strong video understanding and cross-scenario capabilities are demonstrated through task transfer from images to videos.

1 Introduction

It is a core aspiration in AI to build general-purpose assistants with Large Multimodal Models (LMM) [67]. LLaVA-OneVision is an open model, continuing to advance the line of research in building large vision-and-language assistant (LLaVA) [83] that can follow diverse instructions to complete a variety of computer vision tasks in the wild. As a cost-efficient recipe, it is typically developed by connecting vision encoders with large language models (LLM) using a simple connection module.

The first LLaVA model [83] demonstrates impressive multimodal chat abilities, sometimes exhibiting the behaviors similar to GPT-4V on previously unseen images and instructions for the first time. LLaVA-1.5 [81] significantly expands and improves the capabilities by incorporating more academic-related instruction data, achieving SoTA performance on a dozens of benchmarks with a data-efficient recipe. LLaVA-NeXT [82] inherits this property, further pushing performance boundaries through three key techniques: AnyRes for handling high-resolution images, expanding high-quality instruction data, and utilizing the best open LLM available at the time.

LLaVA-NeXT provides an extendable and scalable prototype, which facilitates several parallel explorations, reported in the LLaVA-NeXT blog series [82, 169, 65, 64, 68]:

https://llava-vl.github.io/blog/

- The *Video* blog [169] shows that the image-only-trained LLaVA-NeXT model is surprisingly strong on video tasks with zero-shot modality transfer, due to the design of AnyRes to digest any vision signals as a sequence of images.
- The *Stronger* blog [65] demonstrates the LLM model scaling succuss of this cost-efficient strategy. By simply scaling up the LLM, it achieves performance comparable to GPT-4V on selected benchmarks.

♡ Work collaborated with ByteDance

- The *Ablation* blog [64] summarizes our empirical exploration except the visual instruction data itself, including the choice of architectures (scaling of LLM & vision encoder), visual representations (resolution & #tokens), as well as training strategies (trainable modules & high-quality data) in the pursuit of data scaling success.
- The *Interleave* blog [68] describes the strategies to extend and improve the capability in new scenarios including multi-image, multi-frame (video) and multi-view (3D), while maintaining the single-image performance.

These explorations, conducted within a fixed compute budget, aimed to offer useful insights along the way as we navigate the project, rather than push performance limits. During the process, we have also been accumulating and curating a large collection of the high-quality datasets from January to June. By consolidating these insights and execute the experiments with "yolo run" on newly accumulated larger datasets, we introduce LLaVA-OneVision. We implement the new model with the available compute, without extensively de-risking individual components. This leaves room for further improvements in capabilities through additional data and model scaling following our recipe, Please see the detailed development timeline in Section A. In particular, our paper makes the following contributions:

- *Large multimodal models*. We develop LLaVA-OneVision, a family of open large multimodal models (LMMs) that improves the performance boundaries of open LMMs in three important vision settings, including single-image, multi-image, and video scenarios.
- *Emerging Capabilities with Task Transfer*. Our design in modeling and data representations allow task transfer across different scenarios, suggesting a simple approach to yield new emgerging capabilities. In particular, LLaVA-OneVision demonstrate strong video understanding through task transfer from images.
- *Open-source*. To pave the way towards building a general-purpose visual assistant, we release the following assets to the public: the generated multimodal instruction data, the codebase, the model checkpoints, and a visual chat demo.

2 Related Work

The SoTA proprietary LMMs, such as GPT-4V [109], GPT-4o [110], Gemini [131] and Claude-3.5 [3], exhibit excellent performance in versertile vision scenarios, including single-image, multi-image and video settings. In the open research community, existing works typically develop models tailored to each individual scenario separately. Specifically, most focus on pushing the performance limits in single-image scenarios [26, 83, 173, 73, 164, 35], only a few recent papers have begun to explore multi-image scenarios [70, 47]. While video LMMs excel in video understanding, they often do so at the expense of image performance [72, 76]. It is rare to have a single open model that reports excellent performance in all three scenarios. LLaVA-OneVision aims to fill this gap by demonstrating state-of-the-art performance across a broad range of tasks, and showcasing interesting emerging capabilities through cross-scenario task transfer and composition.

To the best of our knowledge, LLaVA-NeXT-Interleave [68] is the first attempt to report good performance in all three scenarios, LLaVA-OneVision inherits its training recipe and data for improved performance. Other versatial open LMMs with potentials to excel include VILA [77], InternLM-XComposer-2.5 [162]. Unfortunately, their results are not fully evaluated and reported; we compare with them in the experiments. In addition to building systems with versatial capabilities, LLaVA-OneVision is benefited from large-scale high-quality data training, including model-synthesized knowledge and the new collection of diverse instruction tuning data. For the former, we inherit all the knowledge learning data in [64]. For the latter, our are motivated by FLAN [136, 88, 145]. The data collection process is con-current with Idefics2 [63] and Cambrian-1 [133], but we focus on a smaller but more carefully curated collection of datasets. A similar conclusion is observed: a large amount of visual instruction tuning data can significantly improve performance. For comprehensive investigations on design choices of LMMs, we refer to several recent studies [51, 63, 64, 104, 133, 10].

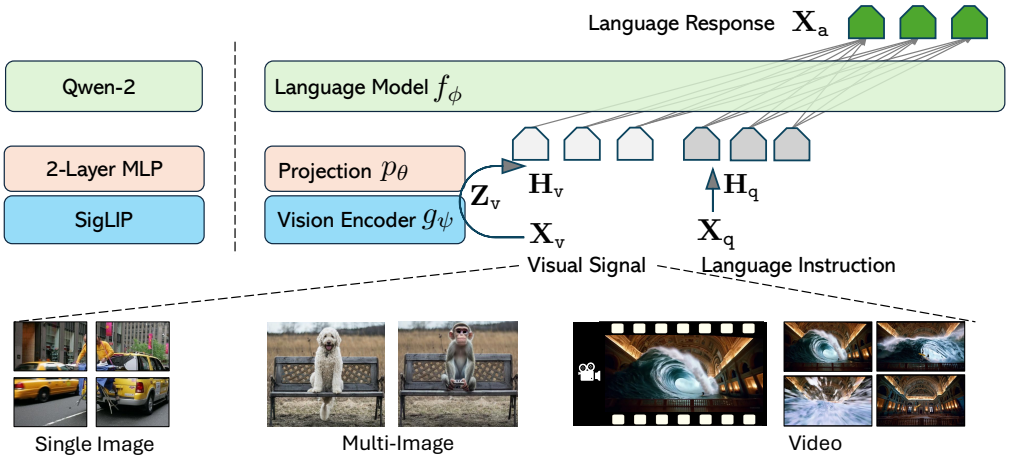


Figure 1: LLaVA-OneVision network architecture. Left: The current model instantiation; Right: the general form of LLaVA architecture in [83], but is extended to support more visual signals.

3 Modeling

3.1 Network Architecture

The model architecture inherits the minimalism design of LLaVA series, whose primary goals are (i) effectively leverage the pre-trained capabilities of both the LLM and visual model, as well as (ii) facilitate strong scaling behavior in terms of both data and model. The network archtecture is illustrated in Figure 1.

- *LLM*. We choose Qwen-2 [148] as our LLM $f_{\boldsymbol{\phi}}(\cdot)$ parameterized by $\boldsymbol{\phi}$, as it offers various model size and exhibits strong language capabilities to date among publicly available checkpoints.
- *Vision Encoder*. We consider the SigLIP [158] as the visual encoder $g_{\boldsymbol{\psi}}(\cdot)$ parameterized by $\boldsymbol{\psi}$, encoding an input image $\mathbf{X}_\text{v}$ into its visual feature $\mathbf{Z}_\text{v} = g(\mathbf{X}_\text{v})$. The grid features before and after the last Transformer layer are considered in our experiments.
- *Projector*. We consider a 2-layer MLP [81] $p_{\boldsymbol{\theta}}(\cdot)$ parameterized by $\boldsymbol{\theta}$, to project image features into the word embedding space, yielding a sequence of visual tokens $\mathbf{H}_\text{v} = p(\mathbf{Z}_\text{v})$.

The model choice is based on our empirical insights in [65, 64] that stronger LLM typically supercharge stronger multimodal capabilities in the wild, while SigLIP yields higher LMM performance among open vision encoders.

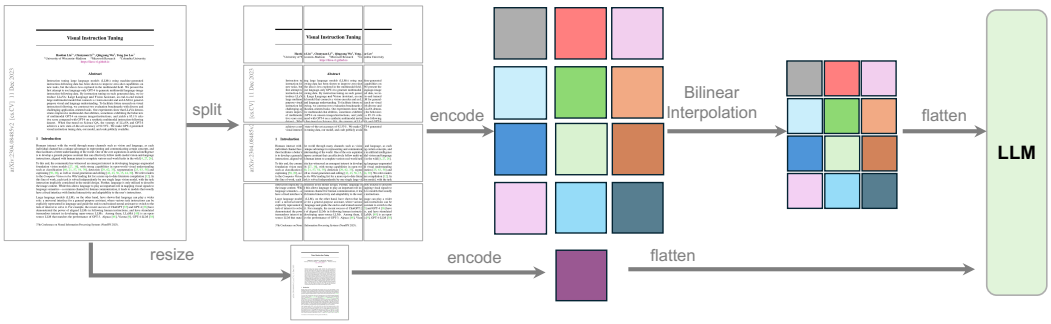
For a sequence of length L, we compute the probability of the target answers $\mathbf{X}_\text{a}$ by:

$$p(\mathbf{X}_\text{a}|\mathbf{X}_\text{v}, \mathbf{X}_\text{q}) = \prod_{i=1}^{L} p(\boldsymbol{x}_i|\mathbf{X}_\text{v}, \mathbf{X}_{\text{q},<i}, \mathbf{X}_{\text{a},<i}), \tag{1}$$

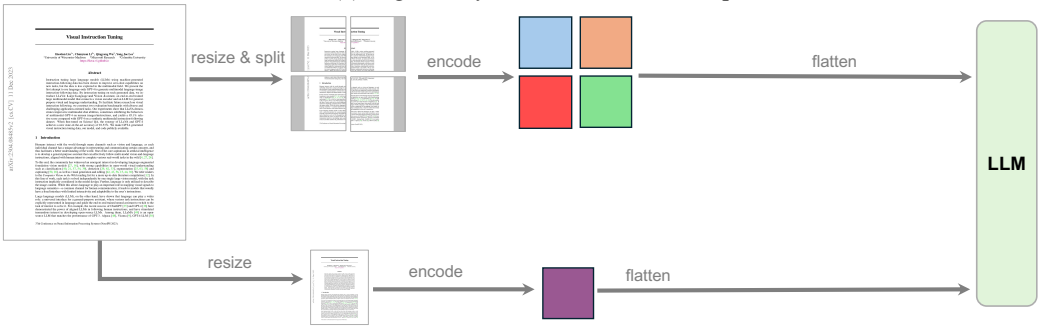
where $\mathbf{X}_{\text{q},<i}$ and $\mathbf{X}_{\text{a},<i}$ are the instruction and answer tokens in all turns before the current prediction token $\boldsymbol{x}_i$, respectively. For the conditionals in (1), we explicitly add $\mathbf{X}_\text{v}$ to emphasize the fact that the visual signal is grounded for all answers. As explained in Section 3.2, the form of visual signal $\mathbf{X}_\text{v}$ is general. The visual input fed into the vision encoder depends on the corresponding scenarios: the invidiual image crop in the single-image sequence, the invidiual image in a multi-image sequence and the invidiual frame in the video sequence, respectively.

3.2 Visual Representations

The representation of visual signals is key to the success of the visual encoding. It relates to two factors, *the resolution in the raw pixel space* and *the number of tokens in the feature space*, leading to the visual input representation configuration (resolution, #token). The scaling of both factors leads to improved performance, especially on tasks that require visual details. To strike a balance of performance and cost, we observe that the scaling of resolution is more effective than that of token numbers, and recommend an AnyRes strategy with pooling. The comparison is illustrated in Figure 2.



(a) Higher AnyRes with Bilinear Interpolation



(b) The original AnyRes

Figure 2: The visual representations. Top: The new Higher AnyRes scheme with Bilinear Interpolation to deal with images of higher resolution; Bottom: the original AnyRes in [82].

Figure 3: The visual representation strategy to allocate tokens for each scenario in LLaVA-OneVision. The maximum number of visual tokens across different scenarios is designed to be similar, ensuring balanced visual representations to accommodate cross-scenario capability transfer. Note that 729 is the #tokens for SigLIP to encode a visual input of resolustion 384×384.

For AnyRes with a configuration of width a, height b, it divides the image into $a \times b$ crops, each with the shape (a, b). Each crop has the same resolution suitable for the vision encoder. Assuming there are T tokens per crop, the total number of visual tokens is $L = (a \times b + 1) \times T$, where the base image is resized before being fed into the vision encoder. We consider a threshold τ, and reduce the #token per crop, using bilinear interpolation if needed:

$$T_{\text{new}} = \begin{cases} \frac{\tau}{(a \times b + 1)} & \text{if } L > \tau \\ T & \text{if } L \leq \tau \end{cases} \tag{2}$$

A set of spatial configurations (a, b) is defined to specify various methods for cropping images, thereby accommodating images of different resolutions and aspect ratios. Among them, the configuration

that requires a minimum number of crops is selected. Please see our detailed ablations of visual representation in [64].

The proposed Higher AnyRes strategy can serve as a flexible visual representation framework, adaptable for multi-image and video representation. The optimal configuration for performance and cost can be adjusted accordingly. We illustratie the configuration in Figure 3, describe the detailed in Section C.1 and provide high-level encoding strategies as below:

- *Single-image*. We consider a large maximum spatial configuration (a, b) for single-image representation to maintain the original image resolution without resizing. Additionally, we purposefully allocate a large number of visual tokens per image, resulting in a long sequence to effectively represent the visual signal. This is based on the observation that there is a larger number of high-quality training samples with diverse instructions for images compared to videos. By representing an image with a long sequence that mimics video representation, we facilitate a smoother capability transfer from image to video understanding [169, 64].
- *Multi-image*. Only the base image resolution is considered and fed into the vision encoder to obtain feature maps, eliminating the need for multi-crop of high resolution image and thus saving computational resources [68].
- *Video*. Each frame of the video is resized to the base image resolution and processed by the vision encoder to generate feature maps. Bilinear interpolation is employed to reduce the number of tokens, allowing the consideration of a larger number of frames by reducing tokens per frame. Empirical evidence suggests this provides a better trade-off between performance and computational cost [169].

These representation configurations are designed for capability transfer with a fixed compute budget in our experiments. With increased computational resources, the number of tokens per image or frame can be increased during both training and inference stages to boost performance.

4 Data

In the realm of multimodal training from LLM, the axiom "quality over quantity" is especially true. This principle is paramount due to the extensive knowledge stored within pre-trained LLMs and Vision Transformers (ViTs). While it is essential to accumulate balanced, diverse, and high-quality instruction data by the end of the LMM's training lifecycle, an often-overlooked aspect is the continuous exposure of the model to new, high-quality data for further knowledge acquisition whenever it is available. In this section, we discuss the data sources and strategies for high-quality knowledge learning and visual instruction tuning.

4.1 High-Quality Knowledge

The web-scale public image-text data is often of low-quality, rendering the data scaling of multimodal pre-training less efficient. Instead, we recommend to focus on high-quality knowledge learning, given a limited compute budget. This approach acknowledges that the pre-trained LLMs and ViTs already possess a substantial knowledge base, and the goal is to refine and enhance this knowledge with carefully curated data. By prioritizing the quality of data, we can maximize compute efficiency.

We consider data from three major categories for high-quality knowledge learning:

- *Re-Captioned Detailed Description Data*. LLaVA-NeXT-34B [82] is known for its strong detailed caption ability among open-source LMMs. We used the model to generate new captions for the images from the following datasets: COCO118K, BLIP558K, and CC3M. We combined them to form the Re-Captioned Detailed Description Data, totaling 3.5M samples. This can be viewed as an simple attempt of self-improvement AI, where the training data is generated by an early version of the model itself.
- *Document / OCR Data*. We utilized the Text Reading subset from the UReader dataset, totaling 100K, which is easily accessible through PDF rendering. We used this text reading data along with the SynDOG EN/CN, to form the Document / OCR Data, totaling 1.1M samples.
- *Chinese and Language Data*. We used the original ShareGPT4V [20] images and utilized GPT-4V provided by the Azure API to generate 92K detailed Chinese caption data, aiming to improve the model's capability in Chinese. Since we used a large portion of detailed caption

data, we also aim to balance the model's language understanding ability. We collected 143K samples from the Evo-Instruct dataset [16].

It is interesting to note that almost all (accounting for 99.8%) of the high-quality knowledge data is synthetic. This is due to the high cost and copyright constraints associated with collecting large-scale, high-quality data in the wild. In contrast, synthetic data can be easily scaled. We believe that learning from large-scale synthetic data is becoming a trend as AI models continue to grow more powerful.

4.2 Visual Instruction Tuning Data

Visual instruction tuning [83] refers to the capability of an LMM to understand and act upon visual instructions. These instructions can be in the form of language, combined with visual media such as images and videos, which the LMM processes and follows to perform a task or provide a response. This involves integrating visual understanding with natural language processing to interpret the instructions and execute the required responses.

Data Collection and Curation. As demosntrated in previous works [81, 133, 63], visual instruction tuning data is crutial for LMM capaiblity. Therefore, maintaining a high-quality dataset collection is crucial and beneficial to the community. We started to collect a large pool of instruction tuning datasets from various original sources, with an unbalanced data ratio among categories. Additionally, we utilize a few new subsets from the Cauldron [63] and Cambrian [133] dataset collections.

We categorize the data based on a three-level hierachy: vision, instruction, and response.

- *Vision Input*. Three vision scenarios are considered, depding which visual input is considered in the multimodal sequence, including single-image, multi-image, video.
- *Language Instruction*. The instructions, which often appears as questions, define the tasks to perform to deal with the visual input. We classify the data into five major categories: *General QA*, *General OCR*, *Doc/Chart/Screen*, *Math Reasoning*, and *Language*. These instructions define the skill sets that a trained LMM could cover. We use task categorization to help maintain and balance the skill distribution.
- *Language Response*. The answer not only responds the user request, but also specifies the model behavior. It can be broadly categorized into free-form and fixed-form.

Free-form data is typically annotated by advanced models like GPT-4V/o and Gemini, while fixed-form data is derived from academic datasets, e.g. VQAv2, GQA, Visual Genome. For free-form data, we keep the original answers. However, for fixed-form data, we manually review the content and make necessary corrections to the question and answer formats. We adhere to the LLaVA-1.5 prompting strategy for multiple-choice data, short answer data, and specific task data (e.g., OCR). This step is crucial for guiding the model's behavior to correctly balance QA performance, conversational ability, and reasoning skills in more complicated tasks, as well as preventing potential conflicts from different data sources. We list the full details about each dataset in our collection, and their categorization and formatting prompt in Appendix E.3.

We divide the instruction data into two separate groups: one for single-image scenario and the other for all vision scenarios. This division is based on insights from our earlier studies [68, 169], which highlight the relationship between image and video models: a stronger image model can better transfer to multi-image and video tasks. Additionally, the quantity and quality of training datasets available for single images are significantly higher than those for videos and multi-image tasks.

Single-Image Data. Since single-image data is crucial for multimodal capabilities, we explicitly compile a large single-image data collection for model learning. We select from collected data sources to form a balanced collection, resulting in a total of 3.2 million samples. The overall distribution of single-image data is shown in Figure 4, with detailed information and the roadmap of data collection presented in Appendix E.1.

OneVision Data. In addition to the single-image stage training, we further fine-tune the model using a mixture of video, image, and multi-image data. We introduce a total of 1.6 million mixed data samples, comprising 560K multi-image data from [68], 350K videos collected in this project, and 800K single-image samples. Notably, in this stage, we do not introduce new single-image data but instead sample high-quality and balanced portions from the previous single-image data, as described

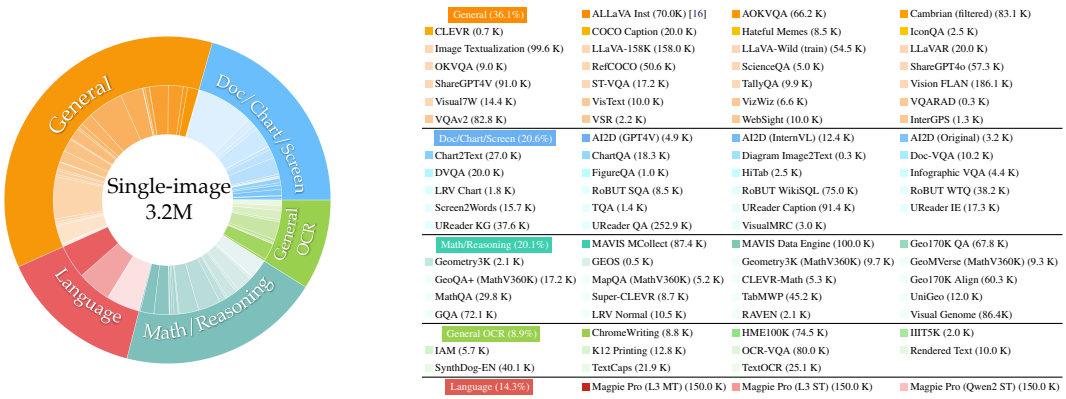


Figure 4: **Single-Image 3.2M.** A High-Quality Single-Image Dataset Collection. Left: Data Distribution within Each Category. The outer circle shows the distribution of all data categories and the inner circle shows the distribution of data subsets. Right: The detailed quantities of datasets.

Figure 5: **OneVision 1.6M.** A high-quality single-image, multi-image and video dataset collection. Left: Data Distribution within each category. The outer circle shows the distribution of all data categories and the inner circle shows the distribution of data subsets. Right: The detailed quantities of datasets. "MI" means it is the multi-image version dataset proposed by DEMON [69].

in [68]. The data distribution and details are presented in Figure 5, with additional information available in Appendix E.2.

5 Training Strategies

To enable LLM for multimodal capabilities, we identify three critical functionalities, and systematically divide them into three distinct learning stages for the purpose of ablation studies. As with most existing research, prior LLaVA models mainly explore the single-image instruction tuning. However, other parts are less frequently investigated and therefore constitute the primary focus of this section.

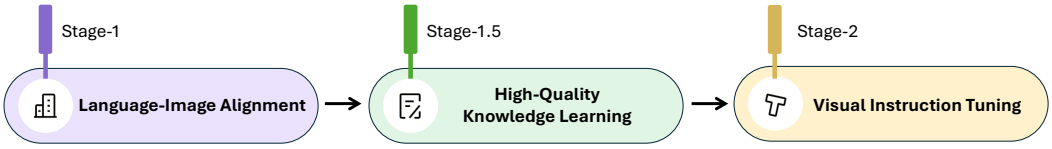
We train the model via a curriculum learning principle, where training objectives and examples of increasing difficulty are observed in a stage-wise manner. With a fixed compute budget, this strategy helps decompose the training process and produces immediate checkpoints that can be re-used in more experiment trails.

- *Stage-1: Language-Image Alignment*. The goal is to well align the visual features into the word embedding space of LLMs.
- *Stage-1.5: High-Quality Knowledge Learning*. To strike a balance between compute-efficiency and injecting new knowledge into LMMs, we recommend to consider the high-quality knowledge for LMM learning. The training configuration mirrors the settings used in Stage-2, ensuring consistency and allowing the model to integrate new information seamlessly.

- *Stage-2: Visual Instruction Tuning*. To teach LMM to solve a diverse set of visual task with preferred responces, we organize the instruction data into different groups, described in Section 4.2. The model is scheduled to train on these groups in order.

Specifically, the visual instruction tuning process consists of two phases: (i) *Single-Image Training*: The model is first trained on 3.2 million single-image instructions, resulting in a model with strong performance in following a diverse set of instructions to complete visual tasks using a single image. (ii) *OneVision Training*: The model is then trained on a mixture of video, single-image, and multi-image data. In this phase, the model expands its capabilities from single-image scenarios to diverse scenarios. It learns to follow instructions to complete tasks in each new scenario and transfer the learned knowledge across different scenarios, resulting in new emergent capabilities. Note that the proposed OneVision training in the post-training stage is probably the simplest and most cost-efficient way to empower the LMMs with the multi-image and video understanding capabilities.

The training strategy is summarized in Table 1. We progressively train the model to deal with long sequence training. The maximum image resolution and the number of visual tokens gradually increase as training progresses. In Stage-1, the base image representation is considered with 729 tokens. In Stages 1.5 and 2, AnyRes is considered with up to 5 times and 10 times more visual tokens, respectively. Regarding trainable modules, Stage-1 updates only the projector, while the subsequent stages update the full model. It is also noted that the learning rate for the vision encoder is 5 times smaller than that for the LLM.



		Stage-1	Stage-1.5	Stage-2	
				Single-Image	OneVision
Vision	**Resolution**	384	384×{2×2, 1×{2,3}, {2,3}×1}	384×{{1×1}, ⋯, {6×6}}	384×{{1×1}, ⋯, {6×6}}
	#Tokens	729	Max 729×5	Max 729×10	Max 729×10 (See Fig. 3)
Data	**Dataset**	LCS	Image (Sec. 4.1)	Image (Sec. 4.2)	(Multi)-Image & Video (Sec. 4.2)
	#Samples	558K	4M	3.2M	1.6M
Model	**Trainable**	Projector	Full Model	Full Model	Full Model
	0.5B LLM	1.8M	0.8B	0.8B	0.8B
	7.6B LLM	20.0M	8.0B	8.0B	8.0B
	72.7B LLM	72.0M	73.2B	73.2B	73.2B
Training	**Batch Size**	512	256/512	256/512	256/512
	LR: ψ_{vision}	1×10^{-3}	2×10^{-6}	2×10^{-6}	2×10^{-6}
	LR: $\{\boldsymbol{\theta}_{\text{proj}}, \boldsymbol{\phi}_{\text{LLM}}\}$	1×10^{-3}	1×10^{-5}	1×10^{-5}	1×10^{-5}
	Epoch	1	1	1	1

Table 1: Detailed configuration for each training stage of the LLaVA-OneVision model. The table outlines the progression of vision parameters, dataset characteristics, model specifications, and training hyperparameters across different stages of the curriculum learning process. We use a global batch size of 512 for the 0.5B model, and 256 for the 7B and 72B models.

6 Experimental Results

We conduct standardized and reproducible evaluations for LLaVA-OneVision models on all benchmarks using LMMs-Eval [161]. For fair comparison with other leading LMMs, we primarily report results from original papers. When results are unavailable, we onboard the models in LMMs-Eval and evaluate them using consistent settings. All our results are reported with greedy decoding and 0-shot settings unless otherwise specified.

To reveal the generality and effectiveness of the designed paradigm, we comprehensively evaluate our LLaVA-OneVision models across different modalities in Table 2, including single-image, multi-image, and video benchmarks. Detailed results for each modality are presented in Table 3, Table 4, and Table 5, respectively. We denote the the model checkpoint trained after the single-image stage and one-vision stage as *LLaVA-OV (SI)* or *LLaVA-OV*, respectively

Three model sizes are provided (0.5B, 7B and 72B), to accomodate applications with different performance-throughput trade-off, ranging from edge device to cloud serving. The GPT-4V and GPT-4o results are presented as references. Our largest model LLaVA-OneVision-72B yields superior performance between GPT-4V and GPT-4o on most benchmarks. It suggests that the proposed recipe is effecitve, revealing a promising path for further scaling. However, a relatively larger gap remains in complex tasks such as visual chat scenarios, we leave it as future research in stronger LLMs, larger training data and better preference learning.

6.1 Single-Image Benchmarks

To validate the performance for single-image tasks in real-world scenories, we consider a comprehensive set of image benchmarks in Table 3. It can be categorized into three classes:

(1) Chart, Diagram, and Document Understanding. As the main visual formats for structured OCR data, we evaluate the results on AI2D [54], ChartQA [101], DocVQA [103], and InfoVQA [102] benchmarks. Though current open-source models such as InternVL [22] and Cambrian [133] achieve performance comparable to commercial models, LLaVA-OneVision goes a step further, surpassing GPT-4V [109] and approaching the performance level of GPT-4o [110].

(2) Perception and Multi-discipline Reasoning. Including visual perception scenarios, we reveal the potentials of our model for more complex and challenging reasoning tasks. Specifically, we adopt the perception benchmarks including MME [151], MMBench [86], and MMVet [154], and reasoning benchmarks such as MathVerse [165], MathVista [90], and MMMU [157]. The results of LLaVA-OneVision significantly outperforms GPT-4V on various benchmarks, and comparable to GPT-4o on MathVista. This further confirms the superiority of our framework in visual perception and reasoning tasks.

(3) Real-world Understanding and Visual Chat. We consider the evaluation of LMMs as general-purpose assistant in the wild as the most important metrics, beyond the lab environments. To validate the capabilities in real-world scenarios, we utilize several widely-adopted benchmarks, including RealworldQA [141], Vibe-Eval [111], MM-LiveBench [161], and LLaVA-Bench-Wilder [65]. While our model still has room for improvement compared to GPT-4V and GPT-4o, it achieves competitive performance with open-source models of similar parameter size. Notably, our model performs well on MM-LiveBench [161], a benchmark for real-world internet content with constantly updated content, demonstrating the model's broad world knowledge and strong generalization abilities.

6.2 Multi-Image Benchmarks

We further evaluate LLaVA-OneVision in multi-image interleaved settings, where users may ask questions between multiples images. In particular, we perform comprehensive assessment on the diverse subtasks of LLaVA-Interleave Bench [68], such as Spot the Difference [45], Image Edit Instruction (IEI) [68], Visual Storytelling (VST) [40], Text-rich VQA (TR-VQA) [85], Multi-image VQA (MI-VQA) [117], Raven Puzzle [24], Q-Bench (QB) [139], and NLVR2 [125]). We also utilize several multi-view benchmarks for evaluation, which depict 3D environments with multiple viewpoints, including 3D Dialogue (3D-Chat) and Task Decomposition (3D-TD) from 3D-LLM [38], ScanQA [5], ALFRED [122], and nuScenes VQA [9]. We refer to these datasets as in-domain evaluations, since our training data includes the training split of them.

Moreover, we conduct evaluations on different out-domain tasks, which reveals the generalization capability of our approach. They include the multi-image split of math QA benchmark MathVerse [165] and science QA benchmark SciVerse [34], multi-image perception benchmark BLINK [31], MMMU-(multi-image) [157] that contains all multi-image QA in MMMU, and MuirBench [135] spanning 12 diverse multi-image tasks.

As shown in Table 4, LLaVA-OneVision (SI) consistently outperforms existing multi-image LMMs in all benchmarks. After additional tuning on multi-image and video data, LLaVA-OneVision shows a marked improvement over GPT-4V in specific areas, with significant margins. This highlights its strong performance in complex tasks such as multi-image reasoning, identifying differences, and understanding 3D environments. In addition, we observe a consistent performance enhancement on after the one-vision training stage, which is more evident on multi-view benchmarks that are absent

Capability	Benchmark	LLaVA OneVision-0.5B	LLaVA OneVision-7B	LLaVA OneVision-72B	GPT-4V (V-Preview)	GPT-4o
Single-Image	†AI2D [53] Science Diagrams	57.1%	81.4%	85.6%	78.2%	94.2%
	†ChartQA [101] Chart Understanding	61.4%	80.0%	83.7%	78.5%	85.7%
	†DocVQA [103] (test) Document Understanding	70.0%	87.5%	91.3%	88.4%	92.8%
	†InfoVQA [102] (test) Infographic Understanding	41.8%	68.8%	74.9%	-	-
	MathVerse [165] (vision-mini) Professional Math Reasoning	17.9%	26.2%	39.1%	32.8%	50.2%
	MathVista [90] (testmini) General Math Understanding	34.8%	63.2%	67.5%	49.9%	63.8%
	MMBench [86] (en-dev) Multi-discip	52.1%	80.8%	85.9%	75.0%	-
	MME [28] (cog./perp.) Multi-discip	240/1238	418/1580	579/1682	517/1409	-
	MMStar [19] Multi-discip	37.5%	61.7%	66.1%	57.1%	-
	MMMU [157] (val) College-level Multi-disp	31.4%	48.8%	56.8%	56.8%	69.1%
	MMVet [153] Multi-discip	29.1%	57.5%	63.7%	49.9%	76.2%
	SeedBench [66] (image) Multi-discip; Large-scale	65.5%	75.4%	78.0%	49.9%	76.2%
	†ScienceQA [93] High-school Science	67.2%	96.0%	90.3%	75.7%	-
	ImageDC [65] Image Detail Description	83.3%	88.2%	91.2%	91.5%	-
	RealworldQA [141] Realwold QA	55.6%	66.3%	71.9%	61.4%	-
	Vibe-Eval [112] Chanllenging Cases	33.8%	51.7%	50.7%	57.9%	63.1%
	MM-LiveBench [161] (2406) Internet Content Understanding	49.9%	77.1%	81.5%	-	92.4%
	LLaVA-Wilder [65] (small) Realworld Chat	55.0%	67.8%	72.0%	81.0%	85.9%
Multi-Image	LLaVA-Interleave [68] Out-domain	33.3%	64.2%	79.9%	60.3%	-
	MuirBench [135] Comprehensive Multi-image	25.5%	41.8%	54.8%	62.3%	-
	Mantis [47] Multi-image in the Wild	39.6%	64.2%	77.6%	62.7%	-
	BLINK [31] Unusual Visual Scenarios	52.1%	48.2%	55.4%	51.1%	-
	†Text-rich VQA [84] OCR, Webpage, Ducument	65.0%	80.1%	83.7%	54.5%	-
Video	ActivityNetQA [155] Spatio-Temporal Reasoning	50.5%	56.6%	62.3%	57.0%	-
	EgoSchema [98] Egocentric Video Understanding	26.8%	60.1%	62.0%	-	-
	PerceptionTest [115] Perception and Reasoning	49.2%	57.1%	66.9%	-	-
	SeedBench [66] (video) Multi-discip; Video	44.2%	56.9%	62.1%	60.5%	-
	LongVideoBench [138] (val) Long Video	45.8%	56.3%	63.2%	60.7%	66.7%
	MLVU [170] Long Video Understanding	50.3%	64.7%	68.0%	49.2%	64.6%
	MVBench [71] Multi-discip	45.5%	56.7%	59.4%	43.5%	-
	VideoChatGPT [97] Video Conversation	3.12	3.49	3.62	4.06	-
	VideoMME [29] Multi-discip	44.0%	58.2%	66.2%	59.9%	71.9%

Table 2: Performance comparison to state-of-the-art commercial models with our LLaVA-OneVision models (0.5B to 72B parameters) across diverse evaluation benchmarks spanning multiple modalities. † indicates that the training set has been observed in our data mixture.

Model	AI2D	ChartQA	DocVQA	InfoVQA	MathVerse	MathVista	MMBench	MME	MMMU
	test	test	val/test	val/test	mini-vision	testmini	en-dev	test	val
Qwen-VL-Max [8]	79.3	79.8	-/93.1	-	23.0	51.0	77.6	2281	51.4
Gemini-1.5-Pro [130]	94.4	87.2	-/93.1	-/81.0	-	63.9	-	-	62.2
Claude 3.5 Sonnet [3]	94.7	90.8	-/95.2	49.7	-	67.7	-	-	68.3
GPT-4V [109]	78.2	78.5*	-/88.4	-	32.8	49.9	75.0	517/1409	56.8
GPT-4o [110]	94.2	85.7	-/92.8	-	50.2	63.8	-	-	69.1
Cambrian-34B [133]	79.7	73.8	-/75.5	-	-	53.2	81.4	-	49.7
VILA-34B [77]	-	-	-	-	-	-	82.4	1762	51.9
IXC-2.5-7B [162]	81.5	82.2	-/90.9	-/70.0	20.0	59.6	82.2	2229	42.9
InternVL-2-8B [22]	83.8	83.3	-/91.6	-/74.8	27.5	58.3	81.7	2210	49.3
InternVL-2-26B [22]	84.5	84.9	-/92.9	-/75.9	31.3	59.4	83.4	2260	48.3
LLaVA-OV-0.5B (SI)	*54.2*	*61.0*	*75.0/71.2*	*44.8/41.3*	*17.3*	*34.6*	*43.8*	*272/1217*	*31.2*
LLaVA-OV-0.5B	57.1	61.4	73.7/70.0	46.3/41.8	17.9	34.8	52.1	240/1238	31.4
LLaVA-OV-7B (SI)	*81.6*	*78.8*	*89.3/86.9*	*69.9/65.3*	*26.9*	*56.1*	*81.7*	*483/1626*	*47.3*
LLaVA-OV-7B	81.4	80.0	90.2/87.5	70.7/68.8	26.2	63.2	80.8	418/1580	48.8
LLaVA-OV-72B (SI)	*85.1*	*84.9*	*93.5/91.8*	*77.7/74.6*	*37.7*	*66.5*	*86.6*	*563/1706*	*57.4*
LLaVA-OV-72B	85.6	83.7	93.1/91.3	79.2/74.9	39.1	67.5	85.9	579/1682	56.8

Model	MMVet	MMStar	S-Bench	S-QA	ImageDC	MMLBench	RealWorldQA	Vibe-Eval	LLaVA-W	L-Wilder
	test	test	image	test	test	2024-06	test	test	test	small
Qwen-VL-Max [8]	-	-	-	-	-	-	-	-	-	-
Gemini-1.5-Pro [130]	-	-	-	-	-	85.9	70.4	60.4	-	-
Claude 3.5 Sonnet [3]	75.4	-	-	-	-	92.3	59.9	66.2	102.9	83.1
GPT-4V [109]	49.9	57.1	49.9	75.7	91.5	-	61.4	57.9	98.0	81.0
GPT-4o [110]	76.2	-	76.2	-	92.5	92.4	58.6	63.1	106.1	85.9
Cambrian-34B [133]	-	-	-	85.6	-	-	67.8	-	-	-
VILA-34B [77]	53.0	-	75.8	-	-	-	-	81.3	-	-
IXC-2.5-7B [162]	51.7	59.9	75.4	-	87.5	-	67.8	45.2	78.1	61.4
InternVL-2-8B [22]	60.0	59.4	76.0	97.0	87.1	73.4	64.4	46.7	84.5	62.5
InternVL-2-26B [22]	65.4	60.4	76.8	97.5	91.0	77.2	66.8	51.5	99.6	70.2
LLaVA-OV-0.5B (SI)	*26.9*	*36.3*	*63.4*	*67.8*	*83.0*	*43.2*	*53.7*	*34.9*	*71.2*	*51.5*
LLaVA-OV-0.5B	29.1	37.5	65.5	67.2	83.3	49.9	55.6	33.8	74.2	55.0
LLaVA-OV-7B (SI)	*58.8*	*60.9*	*74.8*	*96.6*	*85.7*	*75.8*	*65.5*	*47.2*	*86.9*	*69.1*
LLaVA-OV-7B	57.5	61.7	75.4	96.0	88.9	77.1	66.3	51.7	90.7	67.8
LLaVA-OV-72B (SI)	*60.0*	*65.2*	*77.6*	*91.3*	*91.5*	*84.4*	*73.8*	*46.7*	*93.7*	*72.9*
LLaVA-OV-72B	63.7	66.1	78.0	90.3	91.2	81.5	71.9	50.7	93.5	72.0

Table 3: LLaVA-OneVision performance on single-image benchmarks. *GPT-4V reports 4-shot results on ChartQA. All results are reported as 0-shot accuracy.

in single-image data. This demonstrates the significance of our one-vision paradigm for empowering LMMs with comprehensive visual capbalities.

6.3 Video Benchmarks

Video is also a common modality to build world model, capturing the dynamic nature of the real world over time. We conduct experiments on several open-ended and multi-choice video benchmarks. These include ActivityNet-QA [155] that contains human-annotated action-related QA pairs derived from ActivityNet dataset, EgoSchema [98] and MLVU [170] focusing on long video understanding, PerceptionTest [115] designed to evaluate the perception skills, VideoMME [29] and NeXTQA [142] containing diverse video domains and durations (from minutes to hours), VideoDetailCaption [87] and Video-ChatGPT [96] for video detailed description and visua chat, respectively.

As shown in Table 5, LLaVA-OneVision achieves comparable or better results than previous open source models with much larger LLMs. The superiority of LLaVA-OneVision is particularly evident in complex benchmarks such as EgoSchema and VideoMME. Even compared to the advanced commercial model GPT-4V, LLaVA-OneVision performs competitively on the ActivityNet-QA, MLVU, and VideoMME benchmarks.

Model	IEI	MI-VQA	NLVR2	Puzzle	Q-Bench	Spot-Diff	TR-VQA	VST	3D-Chat	3D-TD	ScanQA	ALFRED	nuScenes	BLINK	Mantis	MathVerse	MuirBench	SciVerse
	in-domain multi-image								in-domain multi-view					out-domain				
GPT-4V [109]	11.0	52.0	88.8	17.1	76.5	12.5	54.5	10.9	31.2	35.4	32.6	10.3	63.7	51.1	62.7	60.3	62.3	66.9
LLaVA-N-Image-7B† [82]	13.2	39.4	68.0	9.0	51.0	12.9	59.6	10.1	-	-	-	-	-	41.8	46.1	13.5	-	12.2
VPG-C-7B [70]	15.2	46.8	73.2	2.4	57.6	27.8	38.9	21.5	-	-	-	-	-	43.1	52.4	24.3	-	23.1
Mantis-7B [47]	11.2	52.5	87.4	25.7	69.9	17.6	45.2	12.5	2.60	14.7	16.1	14.0	46.2	46.4	59.5	27.2	36.1	29.3
LLaVA-N-Inter-7B [68]	24.3	87.5	88.8	48.7	74.2	37.1	76.1	33.1	-	-	-	-	-	52.6	62.7	32.8	38.9	31.6
LLaVA-N-Inter-14B [68]	24.5	95.0	91.1	59.9	76.7	40.5	78.6	33.3	70.6	52.2	34.5	62.0	76.7	52.1	66.4	33.4	40.7	32.7
LLaVA-OV-0.5B (SI)	*15.6*	*44.8*	*56.1*	*30.0*	*45.8*	*8.5*	*36.7*	*7.6*	*22.1*	*22.1*	*16.9*	*25.5*	*8.2*	*37.9*	*38.2*	*20.9*	*22.7*	*26.7*
LLaVA-OV-0.5B	17.1	48.7	63.4	35.4	48.8	36.4	65.0	29.8	60.0	48.0	29.4	62.2	70.5	52.1	39.6	60.0	25.5	29.1
LLaVA-OV-7B (SI)	*20.5*	*60.3*	*75.9*	*24.6*	*56.0*	*7.9*	*52.8*	*8.4*	*24.5*	*29.9*	*22.1*	*32.0*	*70.8*	*45.6*	*54.2*	*26.3*	*32.7*	*30.0*
LLaVA-OV-7B	22.2	90.2	89.4	53.3	74.5	39.2	80.1	31.7	62.8	52.6	30.1	61.0	79.8	48.2	64.2	67.6	41.8	79.1
LLaVA-OV-72B (SI)	*22.1*	*61.2*	*78.9*	*44.2*	*61.5*	*15.6*	*67.9*	*12.1*	*30.8*	*25.4*	*21.9*	*43.5*	*75.5*	*46.0*	*56.8*	*58.6*	*33.2*	*65.8*
LLaVA-OV-72B	22.5	95.3	93.8	63.4	83.2	43.3	83.7	34.5	63.2	53.3	35.8	66.3	78.8	55.4	77.6	91.6	54.8	94.9

Table 4: LLaVA-OneVision performance on multi-image benchmarks with all results reported in accuracy. † denotes the LLaVA-NeXT-Vicuna-7B (2024-01). We use IEI for Image Edit Instruction, MI-VQA for Multi-image VQA, NLVR2 for Natural Language for Visual Reasoning, SDiff for Spot the Difference, VST for Visual Story Telling, TR-VQA for Text-rich VQA. For MathVerse and SciVerse, we report the accuracy on their multi-image splits.

Model	ActNet-QA	EgoSchema	MLVU	MVBench	NextQA	PercepTest	SeedBench	VideoChatGPT	VideoDC	VideoMME	L-VideoBench
	test	test	m-avg	test	mc	val	video	test	test	wo/w-subs	val
GPT-4V [109]	57.0	-	49.2	43.5	-	-	60.5	4.06	4.00	59.9/63.3	61.3
GPT-4o [110]	-	-	64.6	-	-	-	-	-	-	71.9/77.2	66.7
Gemini-1.5-Flash [131]	55.3	65.7	-	-	-	-	-	-	-	70.3/75.0	61.6
Gemini-1.5-Pro [131]	57.5	72.2	-	-	-	-	-	-	-	75.0/81.3	64.0
VILA-40B [77]	58.0	58.0	-	-	67.9	54.0	-	3.36	3.37	60.1/61.1	-
PLLaVA-34B [143]	60.9	-	-	58.1	-	-	-	3.48	-	-	-
LLaVA-N-Video-34B [169]	58.8	49.3	-	-	70.2	51.6	-	3.34	3.48	52.0/54.9	50.5
LongVA-7B [163]	50.0	-	56.3	-	68.3	-	-	3.20	3.14	52.6/54.3	-
IXC-2.5-7B [162]	52.8	-	37.3	69.1	71.0	34.4	-	3.46	3.73	55.8/58.8	-
LLaVA-N-Video-32B [169]	54.3	60.9	65.5	-	77.3	59.4	-	3.59	3.84	60.2/63.0	-
LLaVA-OV-0.5B (SI)	*49.0*	*33.1*	*47.9*	*43.3*	*53.6*	*48.6*	*43.4*	*3.08*	*3.51*	*41.7/40.4*	*41.9*
LLaVA-OV-0.5B	50.5	26.8	50.3	45.5	57.2	49.2	44.2	3.12	3.55	44.0/43.5	45.8
LLaVA-OV-7B (SI)	*55.1*	*52.9*	*60.2*	*51.2*	*61.6*	*54.9*	*51.1*	*3.54*	*3.51*	*55.0/59.1*	*54.3*
LLaVA-OV-7B	56.6	60.1	64.7	56.7	79.4	57.1	56.9	3.51	3.75	58.2/61.5	56.4
LLaVA-OV-72B (SI)	*62.1*	*58.6*	*60.9*	*57.1*	*67.2*	*62.3*	*60.9*	*3.55*	*3.66*	*64.8/66.9*	*58.3*
LLaVA-OV-72B	62.3	62.0	68.0	59.4	80.2	66.9	62.1	3.62	3.60	66.2/69.5	61.3

Table 5: LLaVA-OneVision performance on video benchmarks. We report the score out of 5 for VideoDC, VideoChatGPT while other results are reported in accuracy. All results are reported as 0-shot accuracy.

Within the LLaVA-OV split, the smallest performance difference occurs in PerceptionTest, with a minimal improvement of 0.5 points when scaling the LLM from 0.5B to 7B. This contrasts with at least a 5-point improvement in other datasets. The modest gain at PerceptionTest suggests that LLaVA-OV's perception capabilities may mainly depend on its vision module, supporting findings from recent studies such as those by Qiao et al. [116], which separate the roles of the image encoder and the LLM in perception and reasoning tasks. Notably, for datasets like EgoSchema that demand significant reasoning, a larger LLM substantially enhances performance.

Moreover, in comparing LLaVA-OV-7B (SI) with LLaVA-OV-7B, the smallest improvement is seen with ActivityNet-QA. This suggests that LLaVA-OV-7B (SI), which is trained only on images, can

already perform well on this dataset. Delving into ActivityNet-QA, it becomes apparent that many questions can be answered by observing just a single frame from the video. For instance, the question "What's the color of the ball?" can be answered throughout the video as the ball is visible from start to finish. This scenario does not require the model to understand the video sequence, allowing LLaVA-OV-7B (SI) to perform well.

7 Emerging Capabilities with Task Transfer

In addition to reporting the LLaVA-OneVision's capabilities across various benchmarks, we also observe the emerging behaviors of the proposed model with task transfer and composition, paving a promising way to generalize to tackle real-world computer vision tasks in the wild. We illustrate several emerging capabilities using examples as below.

S1: Joint understanding of diagram and chart (Transfer from single-image to multi-image) The capability to understand tables and charts are seperately learned from single image diagram and single-image chart understanding data, and the joint understanding task of table and chart do not appear in multi-image data. As shown in Table 6, LLaVA-OneVision is capable of understanding and reasoning over the joint of diagram and chart.

S2: GUI for multi-modal agent (Transfer from single-image and multi-image). Understanding GUIs and applying multimodal models to agentic tasks is of great value. In Table 7, LLaVA-OneVision recognizes the graphical user interface (GUI) screenshots of an iPhone and provides operational instructions to search for and open the TikTok app. This task requires strong OCR capabilities learned from single-image scenarios and relational reasoning skills developed from multi-image scenarios. The example highlights LLaVA-OneVision's proficiency in GUI understanding and task execution.

S3: Set-of-mark Prompting (Transfer from single-image task composition). Different from existing open LLMs, LLaVA-OneVision demonstrates excellent set-of-marks (SoM) reasoning [149], an emerging capability shown in Table 8. To the best of our knowledge, this is the first time that open LMMs report good emerged SoM ability, as we observe that LLaVA-OneVision is able to produce SoM reasoning for many examples in [149]. This task is not explicitly included in our training data, it is hypothsized that the ability is composed by visual referring and OCR.

S4: Image-to-Video Editing Instruction (Transfer from single-image and video). LLaVA-OneVision could generate detailed video creation prompts based on a static image in Table 9. Given an image and a target video, the model constructs a coherent and vivid narrative for the video, detailing elements such as characters, actions, background settings, and scene specifics. This task leverages both single-image analysis and video comprehension. It is hypothesized that this ability is generalized from the composition of single-image editing instruction task and video detailed description task.

S5: Video-to-Video Difference (Transfer from multi-image and video). Understanding differences in images is a common ability in recent large multimodal models (LMMs), but our models extend this capability to videos. Table 10 showcases LLaVA-OneVision's ability to analyze differences between two video sequences with the same beginning frame but different endings. The model provides a detailed comparison, describing characters, actions, and scene changes. In Table 11, LLaVA-OneVision's describe the differences one by one between videos with a similar background but different main object in the foreground. This task leverages spot the difference in the multi-image analysis to generalize to video scenarios.

S6: Multi-camera Video Understanding in Self-driving (Transfer from single-image and multi-image to video). Understanding videos in a normal aspect ratio is straightforward, what about the videos with multi-views? In Table 12, we observe that LLaVA-OneVision could analyze and interprets multi-camera video footage from self-driving cars. Given video showing four camera views, the model describes each view in detail and plans the ego car's next move. This task combines multi-panel comprehension, video detailed description, and spatial-temporal reasoning.

S7: Composed Sub-video Understanding (Transfer from multi-image to video). Besides multi-view video, we see our model generalize to vertical videos with two sub-scenes. Table 13 demonstrates LLaVA-OneVision's ability to understand and describe the content and layout of a composed sub-video. Given a vertical video with a series of frames featuring a consistent background and a person in the foreground, the model provides a detailed analysis of visual elements, their arrangement, and the narrative context. This task requires single-image analysis, multi-image sequence comprehension, and contextual reasoning.

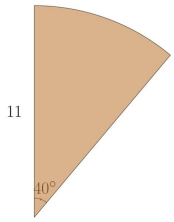
S8: Visual prompting in video (Task transfer from single-image to video). In Table 14, LLaVA-OneVision is able to understand the highlighed area with a semi-transparent circle in the video, and clearly see the number "10" on the back of the player. The capability of understanding visual prompts and OCR is a capablity of single-image LMMs. Our model displays the capablity of understanding visual prompts in videos, without training on video data with visual prompts.

S9: Visual Referring in Image in Video Understanding. The ability to refer to image query when answering questions about a video as shown in Table 15. This capbility is not seen in LLaVA-NeXT or LLaVA-Interleave, this is proabably because strong base single-image training is required for such capabilty to appear.

8 Conclusions

LLaVA-OneVision is a new, open LMM that shines when transferred to a broad range of tasks in the scenarios of single-image, multi-image and videos. The model is developed by consolidating the insights in the LLaVA-NeXT blog series, and is trained by scaling the recipe with a larger dataset and stronger LLMs. Our design allows new capabilities to emerge, through training multiple scenarios together and task transfer, eg, strong visual understanding ability from image to video. Our results demonstrate that LMMs trained with this open recipe and resources achieve state-of-the-art performance across various benchmarks. We also hope that LLaVA-OneVision serves as a valuable starting point for the community to build specific applications, and develop stronger LMMs for diverse vision scenarios through further scaling.

S1: Joint Understanding of Diagram and Chart from Multi-Image

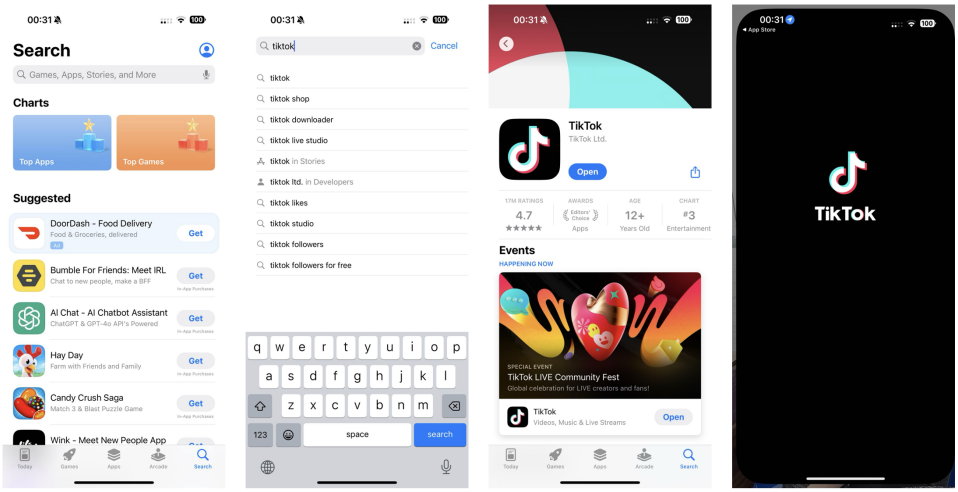


Insurance Company	Price
State Farm	68
Allstate	63
Liberty Mutual	59
USAA	90

User	Ross owns a house similar to the brown sector in the following image . They want to get the insurance from Allstate. The price per unit area is given in the following figure . What would be the cost of insuring the whole house? All the computations are rounded to two places of decimal.
LLaVA-OV	To find the cost of insuring the whole house, we need to calculate the area of the sector and then multiply it by the price per unit area. The area of the sector can be calculated using the formula $A = (\theta/360) * \pi * r^2$, where θ is the central angle and r is the radius. In this case, $\theta = 40°$ and $r = 11$. Plugging these values into the formula, we get $A = (40/360) * \pi * 11^2 \sim 38.01$. The price per unit area is given in the table as \$63. Multiplying the area by the price per unit area, we get the cost of insuring the whole house as $38.01 * \$63 \sim \$2,386.03$.

Table 6: LLaVA-OneVision transfers its ability to *understand diagram and table to multi-image scenarios*, interpreting multiple images in a coherent manner.

S2: Multi-image GUI Understanding for Agents



User	There are four sequential mobile UI screens, and are transformed by three steps of tap operations. Please respectively describe the three tap operations.
LLaVA-OV	The image shows a sequence of four mobile application interface screens. Here's a breakdown of the three tap operations based on the visible changes: 1. First Tap Operation: - The user starts by tapping on the search bar at the top of the screen. - The keyboard is shown at the bottom of the screen, indicating that the user can now type a search query. - The user types "TikTok" in the search bar and taps the search icon on the keyboard to initiate a search. 2. Second Tap Operation: - After tapping the search icon or pressing 'Enter', the screen displays a list of search results. - The user taps on the first result, which is the TikTok application. - This action opens the TikTok's app store page, which includes the app icon, developer information ... 3. Third Tap Operation: - The user taps the "Open" button on the app store page. - This action opens the TikTok application, and the screen transitions to the app's main interface.

Table 7: LLaVA-OneVision *plays the role of agent*. It recognizes multiple screenshots on the iPhone and take action to interact with the iPhone, providing operation instructions for automating tasks.

S3: Set-of-mark Prompting

User	Here's an image with objects labeled in numbers. Please describe the objects that are marked as 4, 5 and 7.
LLaVA-OV	The image shows a room with various items marked with numbers. Here's the description for the items marked as 4, 5, and 7: The item marked as 4 is a framed picture hanging on the wall. It features a black and white illustration of a dandelion with seeds blown away, creating a sense of movement. The item marked as 5 is a white bookshelf with multiple shelves. It appears to be a storage unit that is part of the room's furniture. It is located between the two desks. The item marked as 7 is a black chair with a modern design. It has a metal frame and a woven seat, suggesting a contemporary style. This chair is positioned in front of one of the desks.

Table 8: LLaVA-OneVision exhibits *excellent set-of-mark prompting capabilities*, ie, referring to marks when answering questions. This example demonstrates that describing specific objects based on numerical labels within an image highlights its comprehension skills in handling fine-grained visual content.

S4: Image-to-Video Editing Instruction

User	Please provide the video generation instruction that can change the source image to the target video.
LLaVA-OV	To generate a video from the provided image, you would need to create a sequence of images that depict the progression of the mushrooms being transformed into penguins. Here's a step-by-step guide on how to create this video: 1. Start with the initial image of the mushrooms on a plate. 2. In the next image, begin to change the appearance of the mushrooms to resemble penguins. Start with the smallest mushrooms and gradually move to the larger ones. 3. Continue to transform the mushrooms into penguins, adding details such as eyes, beaks, and flippers. 4. As the video progresses, more mushrooms should be transformed, and the penguins should appear to interact with each other and the mushrooms. 5. The final images should show a group of penguins interacting with the mushrooms, creating a playful and whimsical scene. Remember to maintain a consistent style and color palette throughout the video to ensure a cohesive look. The transformation should be smooth and realistic, making it seem as though the mushrooms are naturally turning into penguins.

Table 9: LLaVA-OneVision learns to *generate detailed video creation prompts based on a static image.* This capability is generalized to videos from the image-to-image language editing generation.

S5: Video-to-Video Difference

User	Here are two videos, please provide the analysis on their differences.
LLaVA-OV	The first video shows a sequence of images where a man in a blue plaid shirt approaches a woman in a red dress, engages in a conversation, and eventually gives her a kiss on the cheek. The woman in the red dress is looking at the man with a slight smile, and another woman in a light blue top is observing the scene. The setting appears to be a street with buildings in the background. The second video shows the same man in the blue plaid shirt walking away from the woman in the red dress, who is now looking in the opposite direction. The woman in the light blue top is no longer visible in this sequence. The man is walking down the street, passing by the buildings and other pedestrians. The overall mood of the video changes from the intimate and personal interaction in the first video to a more solitary and neutral scene in the second video.

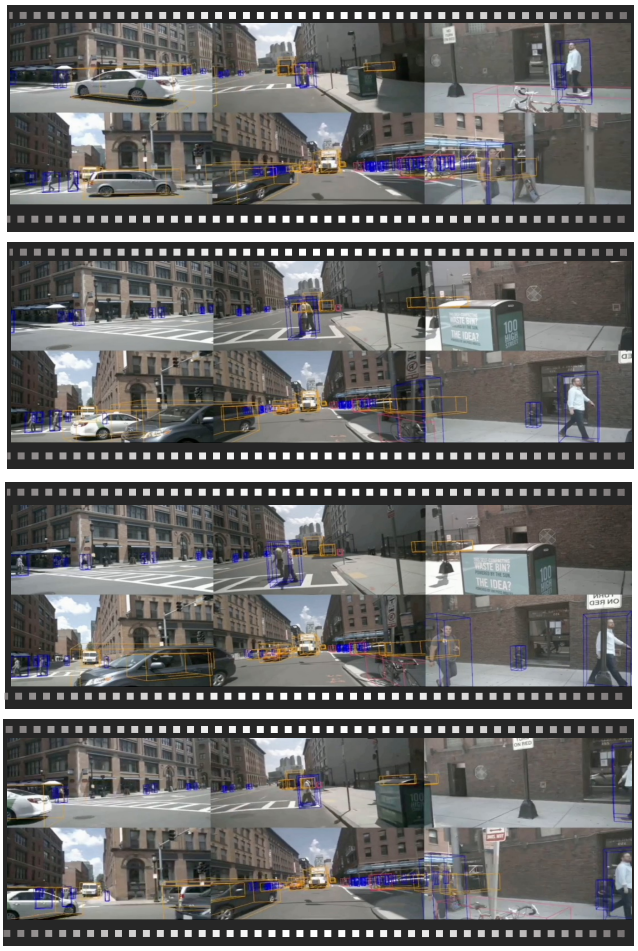
Table 10: LLaVA-OneVision learns to *analyze differences between videos* with the same starting frame but different endings.

S5: Video-to-Video Difference

User	What are the differences between the two videos?
LLaVA-OV	The first video features a caterpillar with green and black stripes on its body, resting on a leaf. The caterpillar has white, bristly hair-like structures along its back. The background is blurred, but it appears to be a garden or a natural setting with green foliage and some sunlight filtering through. The second video shows a butterfly with blue and black wings, resting on a leaf. The butterfly has white spots on its wings and an orange patch near the base of the wings. The background is also blurred but suggests a similar garden or natural setting with greenery. The differences between the two videos are: 1. The first video shows a caterpillar, while the second video shows a butterfly. 2. The caterpillar has green and black stripes, while the butterfly has blue and black wings with white spots. 3. The caterpillar has white, bristly hair-like structures along its back, which are not present on the butterfly. 4. The butterfly has a more delicate and transparent appearance compared to the fuzzy texture of the caterpillar. 5. The wings of the butterfly are spread out, while the caterpillar is in a curled position. 6. The butterfly has long antennae, which are not visible in the caterpillar's frame.

Table 11: LLaVA-OneVision learns to *analyze differences between videos* with similar backgrounds but different foreground objects.

S6: Multi-camera Video Understanding in Self-driving



User	Here is a video showing four camera views of a driving car. The top two show front views, and the bottom two show back views. First describe the video of each view in detail, and then plan the next move of the ego car.
LLaVA-OV	The video consists of six different camera views of a driving car, arranged in two columns with three rows. Each view captures the car from a different angle, showing both front and back perspectives. In the front views: 1. The first image shows the car approaching an intersection with a pedestrian crossing. There are cars ahead, and the pedestrian is visible on the sidewalk. 2. The second image shows the car at the intersection, with the pedestrian crossing and cars ahead. The pedestrian is now closer to the car. 3. The third image shows a pedestrian visible on the sidewalk. In the back views: 1. The first image shows the car approaching the intersection with pedestrians crossing and a black car aside. 2. The second image shows the car at the intersection, with several cars behind. 3. The third image shows the car passing through the intersection, with the pedestrian visible on the sidewalk. The next move of the ego car would be to continue driving forward, following the road and traffic rules, and ensuring the pedestrian's safety as they cross the street.The next move of the ego car would be to continue driving forward, following the road and traffic rules, and ensuring the pedestrian's safety as they cross the street.

Table 12: LLaVA-OneVision *analyzes and interprets multi-camera video footage in self-driving contexts.*

S7: Composed Sub-video Understanding

User	Please describe the video's content and layout in detail.
LLaVA-OV	The video showcases a series of frames featuring a central, obscured foreground view of a room, presumably a game room or office setup. The room is adorned with several posters on the wall, a model of a humanoid figure wearing armor, and various other items that suggest a space dedicated to hobbies or gaming. In the foreground, there is a person, likely a woman, who is wearing headphones and has a neutral to slightly surprised facial expression, indicating that she may be reacting to something off-screen. The backdrop throughout the video is a dynamic, digital illustration from the television show "Game of Thrones," specifically from the series "House of the Dragon." This series is a prequel to "Game of Thrones" and centers around the House Targaryen, a prominent family within the show's fictional world. The images in question feature characters from the series, including members of House Targaryen. Some characters are seen in battle attire, riding dragons, and engaged in combat scenarios. The scenes are detailed and vivid, depicting intense moments from the show, such as the White Walkers and dragon battles. The video starts with the person in the foreground wearing headphones and a black shirt. Behind her, the consistent backdrop from "House of the Dragon" unfolds, focusing on characters from the series. It is evident that the series elements are meant to evoke a sense of drama and tension that is central to the narrative of "House of the Dragon." As the video progresses, the scenes transition through various moments from the series. The characters are shown in different scenarios, such as riding dragons, engaging in combat, and experiencing intense moments. The visuals are rich with color and detail, highlighting the epic scale and dramatic tone of the series. The consistent presence of the game room setup in the foreground provides a sense of continuity throughout the video. The video captures the essence of "House of the Dragon" by using dramatic and action-packed scenes from the series. The person in the foreground remains a constant, adding an element of personal reaction to the intense moments presented in the background. In summary, the video juxtaposes a calm, possibly surprised individual in a game room setup with dynamic and intense scenes from "House of the Dragon," creating a contrast between the peaceful foreground and the dramatic background

Table 13: LLaVA-OneVision learns to *understand and describe composed sub-videos in detail.*

S8: Visual Prompting in Video

User	Describe the player highlighted in the video.
LLaVA-OV	The player highlighted in the video is wearing a white kit, and the number "10" is visible on their shirt. This player is positioned in the center of the pitch and is actively involved in the play, demonstrating control and dribbling skills as they navigate around the field. The player is surrounded by opponents in red kits, indicating that they are likely a key player for their team, possibly a playmaker or attacking midfielder. Throughout the video, the player maintains possession of the ball and creates opportunities for their team, showcasing their importance in the game's dynamics.

Table 14: LLaVA-OneVision learns to *provide detailed descriptions of highlighted subjects in video content.*

S9: Referring Image in Video Understanding

	 
User	There are 2 images. Does the person in image 1 appear in image 2? If yes, what is he doing ?
LLaVA-OV	Yes, the person from image 1 appears in image 2. He is playing soccer with two other individuals in a residential outdoor area.
User	Who is the man in image 1?
LLaVA-OV	The man is Lionel Messi.
	 
User	Here is an image and a video. Does the man shown in the image also appear in the video? If yes, describe his behavior in detail.
LLaVA-OV	Yes, the man shown in the image appears in the video. He is seen engaging in a soccer activity in a grassy outdoor area. He is dressed in a white shirt, dark pants, and white shoes. Throughout the video, he is seen standing near a yellow and black soccer ball, and at various points, he is either preparing to kick the ball or has just kicked it. His actions suggest he is actively participating in the game, and he appears to be in motion, indicating movement and activity.
	 
User	Here is a video and an image. Does the man shown in the image also appear in the video? Explain it.
LLaVA-OV	The man shown in the image does not appear in the video. The image shows a person in a red sports jersey with the number 7, while the video features individuals playing with a soccer ball in a grassy outdoor area. The clothing, background, and activity are different between the image and the video, indicating that they are separate and not related.
User	Who is the man in the image?
LLaVA-OV	The man is Cristiano Ronaldo.

Table 15: LLaVA-OneVision's capability in *referring image and video understanding*. It accurately identifies the same individual in two images in the first instance. It identifies the same individual in both the image and the video in the second instance and correctly concludes the absence of the individual in the third instance, indicating its understanding capability to relate visual query in both image and video understanding.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, 2019. 39

[2] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms, 2019. 39

[3] Anthropic. Claude-3.5. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. 2, 11

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 39

[5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 9

[6] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 40

[7] Haoping Bai, Shancong Mou, Tatiana Likhomanenko, Ramazan Gokberk Cinbis, Oncel Tuzel, Ping Huang, Jiulong Shan, Jianjun Shi, and Meng Cao. Vision datasets: A benchmark for vision-based industrial inspection, 2023. 40

[8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Technical Report*, 2023. 11, 37

[9] Ankan Bansal, Yuting Zhang, and Rama Chellappa. Visual question answering on image sets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 51–67. Springer, 2020. 9

[10] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2

[11] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, 2019. 39

[12] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. 40

[13] Jimmy Carter. Textocr-gpt4v. https://huggingface.co/datasets/jimmycarter/textocr-gpt4v, 2024. 39

[14] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps, 2022. 39

[15] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. *arXiv preprint arXiv:2109.00590*, 2021. 40

[16] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 6, 7, 39

[17] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression, 2022. 39

[18] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning, 2022. 39

[19] 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 preprint arXiv:2403.20330*, 2024. 10

[20] 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 preprint arXiv:2311.12793*, 2023. 5

[21] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 38, 40

[22] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 9, 11, 37, 39

[23] Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. Hitab: A hierarchical table dataset for question answering and natural language generation. In *ACL*, 2022. 39

[24] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024. 9

[25] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 40

[26] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2024. 2

[27] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons, 2019. 40

[28] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 10, 36, 38

[29] 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 preprint arXiv:2405.21075*, 2024. 10, 11

[30] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. 40

[31] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 9, 10

[32] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model, 2023. 39

[33] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. 38, 40

[34] Ziyu Guo, Renrui Zhang, Hao Chen, Jialin Gao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Sciverse. *https://sciverse-cuhk.github.io*, 2024. 9

[35] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 2

[36] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos, 2018. 40

[37] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. 39, 40

[38] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 9

[39] Mehrdad Hosseinzadeh and Yang Wang. Image change captioning by learning from an auxiliary task. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2724–2733, 2021. 40

[40] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016. 9

[41] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 39

[42] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing, 2024. 40

[43] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 40

[44] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III au2, and Larry Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives, 2017. 40

[45] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018. 9

[46] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images, 2018. 40

[47] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 2, 10, 12, 40

[48] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 39

[49] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018. 37, 39

[50] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018. 39

[51] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. *Technical Report*, 2024. 2

[52] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023. 39

[53] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 10, 37, 39

[54] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 9, 36, 38

[55] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017. 39

[56] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017. 40

[57] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020. 39

[58] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 37, 39

[59] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 39

[60] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online, May 2022. Association for Computational Linguistics. 40

[61] Shanghai AI Laboratory. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2023. 38

[62] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 39

[63] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Technical Report*, 2024. 2, 6, 37

[64] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, May 2024. 1, 2, 3, 5, 34, 35

[65] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. 1, 3, 9, 10, 34, 36, 38

[66] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 10

[67] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 2024. 1

[68] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next: Tackling multi-image, video, and 3d in large multimodal models, June 2024. 1, 2, 5, 6, 7, 9, 10, 12, 34, 35, 36, 38

[69] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions, 2024. 7, 40

[70] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions. *arXiv preprint arXiv:2308.04152*, 2023. 2, 12

[71] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2023. 10

[72] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, 2024. 2

[73] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *Technical Report*, 2024. 2

[74] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization, 2019. 40

[75] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning, 2023. 39

[76] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

[77] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 2, 11, 12

[78] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 37, 39

[79] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 39

[80] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 39

[81] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1, 3, 6, 37

[82] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 1, 4, 5, 12, 34, 37

[83] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2, 3, 6, 39

[84] Xuejing Liu, Wei Tang, Xinzhe Ni, Jinghui Lu, Rui Zhao, Zechao Li, and Fei Tan. What large language models bring to text-rich vqa?, 2023. 10

[85] Xuejing Liu, Wei Tang, Xinzhe Ni, Jinghui Lu, Rui Zhao, Zechao Li, and Fei Tan. What large language models bring to text-rich vqa? *arXiv preprint arXiv:2311.07306*, 2023. 9

[86] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, and Ziwei Liu. Mmbench: Is your multi-modal model an all-around player? *Technical Report*, 2023. 9, 10, 36

[87] LMMs-Lab. Video detail caption, 2024. 11

[88] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023. 2

[89] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, and Yaofeng Sun. Deepseek-vl: towards real-world vision-language understanding. *Technical Report*, 2024. 37

[90] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv preprint arXiv:2310.02255*, 2023. 9, 10, 36, 38

[91] 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, 2021. 39

[92] 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. 39

[93] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 10, 39

[94] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023. 39

[95] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS*, 2021. 39, 40

[96] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 11

[97] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 10

[98] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 10, 11

[99] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 39

[100] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46, 2002. 39

[101] 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. In *ACL*, 2022. 9, 10, 36, 37, 39

[102] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 9, 10, 36, 39

[103] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 9, 10, 36, 37, 39, 40

[104] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 2

[105] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012. 39

[106] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 39

[107] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952, 2019. 40

[108] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model, 2020. 39

[109] OpenAI. Gpt-4v. https://openai.com/index/gpt-4v-system-card/, 2023. 2, 9, 11, 12

[110] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 2, 9, 11, 12

[111] Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, et al. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*, 2024. 9

[112] Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, Ethan Yeo, Eugenie Lamprecht, Qi Liu, Yuqi Wang, Eric Chen, Deyu Fu, Lei Li, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Mikel Artetxe, and Yi Tay. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models, 2024. 10, 36, 38

[113] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning, 2019. 40

[114] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions, 2024. 39

[115] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. 10, 11

[116] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms, 2024. 12

[117] Harsh Raj, Janhavi Dadhania, Akhilesh Bhardwaj, and Prabuchandran KJ. Multi-image visual question answering. *arXiv preprint arXiv:2112.13706*, 2021. 9

[118] Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. Aesop: Abstract encoding of stories, objects, and pictures. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2032–2043, 2021. 40

[119] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022. 39

[120] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015. 39

[121] ShareGPT. https://sharegpt.com/, 2023. 37, 39

[122] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 9, 40

[123] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. 39

[124] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ivan Laptev, Ali Farhadi, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv e-prints*, 2016. 38, 40

[125] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 9

[126] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019. 40

[127] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language, 2019. 40

[128] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, 2021. 39

[129] Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning, 2023. 39

[130] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 11

[131] 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 preprint arXiv:2312.11805*, 2023. 2, 12

[132] Ting-Hao, Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling, 2016. 40

[133] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 6, 9, 11, 39

[134] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning, 2021. 36, 39

[135] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 9, 10

[136] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2

[137] Chris Wendler. wendlerc/renderedtext, 2023. 39

[138] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 10

[139] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 9

[140] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison, 2024. 40

[141] x.ai. Grok-1.5 vision preview. 9, 10

[142] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, June 2021. 11, 38, 40

[143] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 12

[144] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *ArXiv*, abs/2406.08464, 2024. 36, 37, 39

[145] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690*, 2024. 2

[146] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024. 37, 39

[147] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes, 2018. 40

[148] 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 preprint arXiv:2407.10671*, 2024. 3, 35

[149] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 13

[150] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model, 2023. 37, 39

[151] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 9

[152] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 39

[153] 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, 2023. 10

[154] 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. *arXiv preprint arXiv:2308.02490*, 2023. 9

[155] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 10, 11, 38, 40

[156] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. *arXiv preprint arXiv:2203.01601*, 2022. 39

[157] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, and Yuxuan Sun. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 9, 10, 36, 38

[158] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3, 35

[159] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, 2019. 39, 40

[160] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. 40

[161] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 8, 9, 10, 36

[162] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 2, 11, 12

[163] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 12

[164] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2

[165] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. 9, 10

[166] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, and Hongsheng Li. Mavis: Mathematical visual instruction tuning, 2024. 39

[167] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 38

[168] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 39

[169] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. 1, 5, 6, 12, 34, 35, 36

[170] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 10, 11

[171] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding, 2024. 36

[172] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos, 2017. 38, 40

[173] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

[174] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 39

A Development Roadmap from LLaVA-NeXT to LLaVA-OneVision

LLaVA-OneVision is built upon techniques developed in the LLaVA-NeXT blog series [82, 169, 65, 64, 68] from January to June 2024. The initial LLaVA-NeXT provided an extendable and scalable prototype, which facilitated several parallel explorations. These explorations, conducted within a fixed compute budget, aimed to offer useful insights along the way, rather than push performance limits. LLaVA-OneVision consolidates these insights and execute with "yolo run" – implements the new model with the available compute, without extensively de-risking individual components.

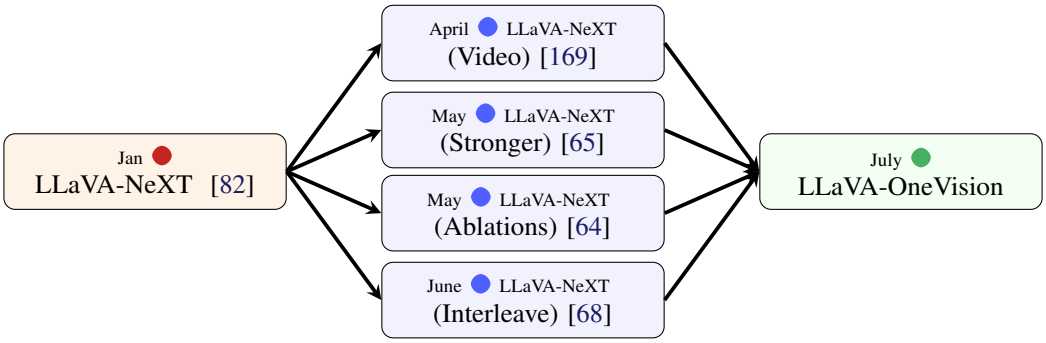


Figure 6: The development timeline from LLaVA-NeXT to LLaVA-OneVision.

1. **LLaVA-NeXT:**
   Improved reasoning, OCR, and world knowledge [82]
   - Blog: https://llava-vl.github.io/blog/2024-01-30-llava-next/
   - A cost-efficient training recipe for LMMs with strong performance

2. **LLaVA-NeXT (Video):**
   A Strong Zero-shot Video Understanding Model [169]
   - Blog: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
   - Thanks to the design of AnyRes to digest vision signal, the image-only-trained LLaVA-NeXT model is surprisingly strong on video tasks with zero-shot modality transfer. DPO training with AI feedback on videos can further yield significant improvement.

3. **LLaVA-NeXT (Stronger):**
   Stronger LLMs Supercharge Multimodal Capabilities in the Wild [65]
   - Blog: https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/
   - The same cost-efficient recipe, supporting LLaMA3 (8B) and Qwen (72B &110B). Simply scaling up LLM catches up with GPT-4V on selected benchmarks. Developed an evaluation benchmark for daily-life visual chat, LLaVA-Bench (Wilder).

4. **LLaVA-NeXT (Ablation):**
   What Else Influences Visual Instruction Tuning Beyond Data? [64]
   - Blog: https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/
   - Ablating the choice of Architectures (Scaling LLM & Vision Encoder), Visual Representations (Resolution & #Tokens), and Training Strategies (Trainable modules & High-quality data).

5. **LLaVA-NeXT (Interleave):**
   Tackling Multi-image, Video, 3D in Large Multimodal Models [68]
   - Blog: https://llava-vl.github.io/blog/2024-06-16-llava-next-interleave/
   - Extending the capability to new scenarios including multi-image, multi-frame (video) and multi-view (3D), with new training data (M4-Instruct) and benchmark (LLaVA-Interleave Bench).

B Author Contributions

- Bo Li contributes to maintaining the LLaVA-OneVision codebase, conducting the large-scale training of the LLaVA-OneVision models of all stages (including the stage with single-image, multi-image, and video data), based on upon our previous LLaVA-NeXT series. He contributes significantly to the single-image development such as LLaVA-NeXT-Ablations [64], high-quality recpationing, as well as collection and curation of the single-image data mixture.

- Yuanhan Zhang contributes to a series of works in LLaVA-NeXT-Video [169], including video training and inference codebase, an effective pipeline for high-quality video data generation, and all the video training data.

- Dong Guo contributes to collection and curation of the single-image data mixture and consistently provides technical support throughout the project.

- Feng Li, Renrui Zhang, and Hao Zhang contribute to LLaVA-NeXT-Interleave [68], including the multi-image instruction data mixture, the multi-image evaluation benchmarks, and the early prototype of LLaVA-OneVision, i.e., a joint training stage with single-image, multi-image, and videos. They also contribute to the collection and curation of the single-image data mixture.

- Kaichen Zhang maintains the training codebase and contributes to the integration of LLaVA-OneVision model into LMMs-Eval's evaluation pipeline.

- Yanwei Li contributes to revising the paper.

- Ziwei Liu makes valuable suggestions throughout the projects.

- Chunyuan Li initiates and leads the series of projects, designs the roadmap and milestones, drives the excution, as well as leads the the paper writing.

C Implmenetation Details

C.1 Token Strategy for Mixed-Modality Data

We provide a detailed explanation of our token strategy for handling mixed-modality data within LLaVA-OneVision's architecture, which is illustrated in Figure 3.

For single-image data, we employ the *AnyResMax-9* strategy, as previously outlined in blog [64]. Using SO400M [158] as the Vision Encoder, each input image (or grid) is processed into 729 visual tokens. Consequently, the maximum number of visual tokens for a single image is $729 \times (1 + 9)$, where 1×729 represents the base tokens and 9×729 accounts for the grid tokens.

For multi-image data, we utilize a simple padding strategy. Each image is first resized to fit within a 384x384 frame by zero-padding, as required by SO400M, while maintaining the aspect ratio. After processing through the vision encoder, the zero-padding is removed from the tokens. Our training data includes up to 12 images per instance, resulting in a maximum of 12×729 multi-image tokens.

For video data, we adopt a strategy similar to LLaVA-NeXT-Video [169]. Each frame is processed through the vision encoder and then subjected to 2×2 bilinear interpolation, resulting in 196 tokens per frame. We sample up to 32 frames per video, leading to a maximum of 32×196 video tokens.

As shown in Figure 3, the maximum number of tokens across different modalities is approximately equal. This design strategy aims to balance the data from various modalities, ensuring more equitable representation that is transferable from the perspective of the language model. For instance, a high-resolution image can be interpreted as a composition of multiple images, and multiple images can be understood as a shorter video.

C.2 Language Templates and Special Tokens

We utilize the Qwen-2 series [148] language models with the template as OpenAI's ChatML[1]. During training, we adopt `<image>` as the marker for image tokens, following previous LLaVA models. This image special token is represented as -200 in the input index after tokenization. For multi-image

[1] OpenAI Release v0.28.0/chatml.md

scenarios, we use multiple <image> interleaved with text to denote the positions of the images. For video scenarios, we place a single <image> at the beginning to indicate the inclusion of a video.

One more aspect related to the handling of image tokens is ensuring that there are no extra <image> in the data. For instance, in some code writing tasks, there could be <image>...</image> related to HTML code. To avoid potential misunderstandings, we manually removed around 10 such samples from the Magpie [144] and Screen2Words [134] datasets.

D Evaluation Steers Development

D.1 Post-Evaluation as a Development Tool

With the help of our comprehensive evaluation toolkit, LMMs-Eval [161], we conduct post-evaluations on a selected set of benchmarks after each training experiment concludes.

Our preference for selecting benchmarks is based on whether the targeted scenarios are sufficiently important and specific. These evaluations should not be too resource-intensive, meaning the benchmarks should not contain too many items, take too long to evaluate, or consume a large number of GPT-4V tokens (when using it as the judge model).

In our development, we evaluate on AI2D [54], ChartQA [101], DocVQA [103], and InfoVQA [102] to examine the model's fine-grained understanding of tables, charts, and diagrams, as well as MME [28] for formatting control, since it requires only Yes or No answers. We also include MMBench-Dev [86] and MMMU-Val [157] for multi-discipline evaluation. Quickly obtaining evaluation results on these benchmarks will guide our next steps in model development and data curation.

D.2 Improving Model Performance on Key Scenarios

During our development process, we gradually recognized the significance of using static evaluation benchmarks as perfprmance indicators. Our primary goal at this stage is not to overfit the model to certain datasets to achieve exceptionally high performance. Instead, we benchmark our models against GPT-4V's performance to set our target thresholds (e.g., initially 80%, gradually increasing to 95%-100%). Once the model meets the score requirements in static evaluations, it indicates that the model has sufficient capabilities in the selected scenarios. Furthermore, we cannot blindly pursue results on benchmarks, as even the test data for AI2D may have certain issues [2].

Ultimately, our focus is on optimizing the model's visual chat and reasoning capabilities. In this stage, we monitored the model's performance on benchmarks such as MathVista [90], LLaVA-Wilder [65], MM-LiveBench [171], and Vibe-Eval [112]. These benchmarks require the model to engage in visual dialogue with challenging questions, and demand a diverse skill set with extensive world knowledge. This helps us create a model with strong generalization capabilities in real-world scenarios.

D.3 Evaluation Task Information

In this section, we provide information on all the tasks used during the evaluation. Specifically, we use the default `post_prompt` and `pre_prompt` from the `LMMs-Eval` framework. These prompts are consistent with the evaluation of our previous LLaVA-NeXT [65, 169, 68]. The table below details the specific tasks used in `LMMs-Eval` and their corresponding task names.

[2] Discussion on AI2D Evaluation

Tasks Information

- **Single-image:**
  - ai2d, chartqa, docvqa_val, infovqa_val, mme, realworldqa, mathvista_testmini, llava_in_the_wild, mmvet, mmbench_en_dev, ocrbench, mmmu, llava_wilder_small, vibe_eval, wildvision_0617, live_bench_2406, mathverse_testmini_vision, seedbench, scienceqa_img, mmstar, dc100_en
- **Videos:**
  - activitynetqa, videochatgpt, nextqa_mc_test, egoschema, video_dc499, videmme, videomme_w_subtitle, perceptiontest_val_mc, mlvu, mvbench
- **Multi-image:**
  - llava_interleave_bench, muirbench

By referring to the task names listed here, the audience can directly retrieve the generation arguments and specific prompt information. For instance, the details for `tasks=ai2d` are available at lmms-eval/ai2d. By following these settings, researchers can easily reproduce our results.

E Data Curation Roadmap of LLaVA-NeXT Series

In this section, we provide the in-depth experience and roadmap of data curation in the LLaVA-NeXT series. To achieve strong multimodal performance, we need to collect and curate high-quality data from various sources, which is crucial for the model's generalization capabilities.

E.1 Single-Image Data Curation

As the primary data source, our principle for single-image data has always been that quality outweighs quantity. Given limited resources, we strive to use high-quality data to maximize the performance.

The first version of the LLaVA-NeXT models (LLaVA-NeXT-Vicuna-7B/13B, Mistral-7B, Hermes-Yi-34B), comprising 760K data samples [82], includes 665K samples from LLaVA-1.5 [81], 3,247 samples from AI2D [53], 18,317 samples from ChartQA [101], 10,194 samples from DocVQA [103], 20,000 samples from DVQA [49], 40,093 samples from SynthDOG-EN [58], and 15,131 samples from user requests on LLaVA's demo, re-annotated with GPT-4V. In the subsequent iteration, we added 20,000 samples from COCO Caption [78], forming a new 790K version. This 790K dataset supported the second release of LLaVA-NeXT models (LLaVA-NeXT-LLaMA3-8B, LLaVA-NeXT-Qwen-72B, LLaVA-NeXT-Qwen-110B).

In subsequent collections, we accumulated open-sourced datasets from the Internet and referred to the dataset collection processes of other advanced LMMs, such as Qwen-VL [8], DeepSeek-VL [89], Intern-VL [22], Vision-Flan [146], UReader [150], Idefics-2 (Cauldron) [63], and Cambrian. During the data iteration process, we strictly adhered to the initial LLaVA-1.5 strategy. For each dataset, we manually inspected and ensured its quality and QA format. We also designed specific formatting prompts to make data from different sources compatible with each other, thus avoiding conflicts.

Some data sources, such as AI2D and ChartQA, appear in different dataset collections and may be duplicated. Since Cauldron includes special formatting prompts, its data is not straightforward to re-format. Therefore, we prioritize using data from other collections that are closer to the raw format. For the Cambrian dataset, we only selected a subset of the GPT-4o re-annotated data. We also collected math-related data from the MathV and MAVIS datasets.

For the pure language data, we replaced the ShareGPT [121] text data that LLaVA has been using since version 1.5. Given that our largest Qwen2-72B model has achieved performance levels close to latest GPT-4 model in language tasks, we need to use higher quality language data to maintain or further enhance its language capabilities. To achieve this, we sourced the highest quality language SFT data available, the Magpie-Pro dataset [144].

After undergoing the aforementioned process, we have obtained approximately 4 million raw SFT data samples, ensuring their quality and accuracy. Additionally, we utilized Azure's OpenAI GPT-4V and GPT-4o services to re-annotate our data, focusing on scenarios that were not adequately covered by the original data but are crucial. These scenarios include:

(1) Detailed Descriptions on Charts and Diagrams: For this scenario, we used images from the AI2D and InfoVQA training sets and employed GPT-4V to provide detailed descriptions of the images, resulting in 4,874 detailed descriptions for AI2D and 1,992 samples for InfoVQA.

(2) Chinese Language: We used images from the LLaVA-158K dataset and employed GPT-4o to provide detailed descriptions in Chinese, resulting in a total of 91,466 samples.

(3) Multi-turn Dialogue: Also with the LLaVA-158K dataset, we employed GPT-4o to create long dialogues with an average of more than 3 turns per conversation, obtaining a total of 26,048 samples.

When resources permit, we recommend a data validation process we used in early stage data sourcing. We extract approximately 100K samples from each newly added data source or collection (if the selected data source can form a collection) and add them to the 790K version of the dataset. We validate newly added data under the SO400M-Qwen-1.5-0.5B experimental setting. If the addition of new data results in a performance decline compared to the baseline, we conduct further manual inspections of the data and adjust the formatting prompt accordingly. This step requires abundant resources and must be carried out by highly professional researchers, as it cannot be substituted with average human annotators.

During the collection process, we manually labeled the datasets with two tags: {General, Language, Math/Reasoning, General OCR, Doc/Chart/Screen} and {Fixed-form, Free-form}. Based on these tags, we formed the final distribution of 3.2 million single-image data samples.

Starting with the initial distribution, we gradually increased the amount of free-form (most of them are GPT-4V/o annotated) data and observed the model's performance on various benchmarks and try to balance among them. These benchmarks include academic datasets, such as AI2D [54], MME [28], MMMU [157], MathVista [90], and visual chat datasets, such as LLaVA-Wilder [65], and Vibe-Eval [112]. Ultimately, we gradually established an optimal data distribution for single-image tasks under the 7B setting.

E.2 OneVision Data Curation

In addition to single-image data, we incorporate multi-image and video datasets to support a wider scope of visual scenarios. We aim to balance the capability among different data modalities, and achieve an overall superior performance with one framework as LLaVA-OneVision.

For multi-image data, we adopt the diverse interleaved multimodal tasks within M4-Instruct dataset from LLaVA-NeXT-Interleave [68]. This dataset mainly comprises general multi-image tasks, such as spotting the difference, visual story telling, image editing instruction generation, interleaved multi-image dialogue, multi-image puzzle, low-level multi-image assessment, etc. Besides, we also utilize the multi-view datasets in M4-Instruct to indicate spatial information in the 3D world, including embodied VQA (dialogue and planning) and 3D scene VQA (captioning and grounding).

For video data, we first integrate the multi-frame data from M4-Instruct, including NExT-QA [142] and ShareGPT4Video [21]. Then, to enable more detailed temporal cues, we select several datasets commonly used in recent academic research for re-annotation, including Charades [124], ActivityNet [155], YouCook2 [172], and Ego4D [33]. Initially, we annotated captions. Following ShareGPT-4o [61], we sampled video frames at 1 frame per second (FPS) and used the pre-defined instructions to prompt GPT-4o for generating video captions. Additionally, following LLaVA-Hound [167], we developed open-ended question-answering pairs and their corresponding multiple-choice versions using the captions created by GPT-4o. We also employed GPT-4o to generate question-answer pairs, obtaining high-quality video data for OneVision training.

E.3 Detailed Dataset Statistics

We primarily use tables to present the statistical information of all datasets utilized in both the Single-Image and OneVision stages. The information includes the dataset category, dataset name, number of samples, and prompt type. The dataset statistics are summarized in Table 16.

Dataset	# Samples	Prompt ID	Dataset	# Samples	Prompt ID
General (1.14M, 36.1%)					
AOKVQA [119]	66160	1	Cambrian (filtered) [133]	83131	-
CLEVR [48]	700	1	COCO Caption [78]	20000	9
Hateful Memes [57]	8500	1	IconQA [95]	2494	5
Image Textualization [114]	99583	11	LLaVA-158K [83]	158000	-
LLaVA-Wild (train) [83]	54517	-	LLaVAR [168]	20000	-
OKVQA [99]	8998	1	RefCOCO [152]	50586	7,8
ScienceQA [93]	4976	5	ShareGPT4O [121]	57289	11
ShareGPT4V [121]	92025	11	ST-VQA [11]	17247	1
TallyQA [1]	9868	1	Vision FLAN [146]	186070	-
Visual7W [174]	14366	5	VisText [129]	9969	15
VizWiz [37]	6614	2	VQARAD [62]	313	1
VQAv2 [4]	82783	1	VSR [79]	2157	3
WebSight	10000	18	InterGPS [91]	1280	5
ALLaVA Instruct [16]	70000	-			
Doc/Chart/Screen (20.6%, 647K)					
AI2D (GPT4V Detailed Caption)	4874	12	AI2D (InternVL [22])	12413	4
AI2D (Original) [53]	3247	5	Chart2Text [108]	26961	13
ChartQA [101]	18317	1	Diagram Image2Text	300	17
DocVQA [103]	10194	1	DVQA [49]	20000	1
FigureQA [50]	1000	3	HiTab [23]	2500	1
Infographic VQA [102]	4404	1	LRV Chart [80]	1787	-
RoBUT SQA	8514	-	RoBUT WikiSQL	74989	-
RoBUT WTQ	38246	1	Screen2Words [134]	15730	10
TQA [55]	1365	5	UReader Caption [150]	91439	9
UReader IE [150]	17327	1	UReader KG [150]	37550	14
UReader QA [150]	252954	1	VisualMRC[128]	3027	-
Math/Reasoning (20.1%,632K)					
MAVIS Manual Collection [166]	87358	19	MAVIS Data Engine [166]	100000	19
CLEVR-Math [48]	5290	2	Geo170K Align [32]	60252	-
Geo170K QA [32]	67833	19	Geometry3K [91]	2101	6
GEOS [120]	508	6	Geometry3K (MathV360K) [92]	9734	6
GeoMVerse (MathV360K) [52]	9303	20	GeoQA+ (MathV360K) [18]	17172	6
MapQA (MathV360K) [14]	5235	1	MathQA [2]	29837	19
Super-CLEVR [75]	8652	2	TabMWP [94]	45184	2
UniGeo [17]	11959	6	GQA [41]	72140	1
LRV Normal [80]	10500	-	RAVEN [159]	2100	3
Visual Genome [59]	86417	7,8			
General OCR (8.9%,281K)					
ChromeWriting [137]	8835	21	HME100K [156]	74502	21
IIIT5K [105]	2000	22	IAM [100]	5663	22
K12 Printing	12832	22	OCR-VQA [106]	80000	1
Rendered Text [137]	10000	22	SynthDog-EN [58]	40093	16
TextCaps [123]	21952	9	TextOCR-GPT4V [13]	25114	11
Pure Language (450K) (14.3%, 647K)					
Magpie Pro [144] (L3 MT)	149999	-	Magpie Pro (L3 ST)	150000	-
Magpie Pro (Qwen2 ST)	149996	-			

Table 16: The detailed statistics of Single-Image datasets used in LLaVA-OneVision. Prompt ID denotes the ID of Formatting Prompt which is corresponding to the ID in Table 18. - denotes no fromatting prompt is used.

Dataset	# Samples	Prompt ID	Dataset	# Samples	Prompt ID
Multi-image Scenarios					
Spot-the-Diff [46]	10.8K	20	Birds-to-Words [27]	14.3K	21
CLEVR-Change [113, 39]	3.9K	22	HQ-Edit-Diff [42]	7.0K	3
MagicBrush-Diff [160]	6.7K	4	IEdit [127]	3.5K	19
AESOP [118]	6.9K	23	FlintstonesSV [36]	22.3K	24
PororoSV [74]	12.3K	25	VIST [132]	26K	4
WebQA [15]	9.3K	8	TQA (MI) [56]	8.2K	9
OCR-VQA (MI) [107]	1.9K	17	DocVQA (MI) [103]	1.9K	18
RAVEN [159]	35K	5	MIT-StateCoherence [43]	1.9K	11
MIT-PropertyCoherence [43]	1.9K	12	RecipeQA ImageCoherence [147]	8.7K	14
VISION [7]	9.9K	13	Multi-VQA [69]	5K	-
IconQA [95]	34.6K	-	Co-Instruct [140]	50.0K	-
DreamSim [30]	15.9K	-	ImageCoDe [60]	16.6K	-
nuScenes [12]	9.8K	10	ScanQA [6]	25.6K	7
ALFRED [122]	22.6K	16	ContrastCaption [47]	25.2K	-
VizWiz (MI) [37]	4.9K	6	ScanNet [25]	49.9K	7
COMICS Dialogue [44]	5.9K	15	NLVR2 [126]	86K	26
Multi-frame (Video) Scenarios					
NExT-QA [142]	9.5K	2	ActivityNet [155]	6.5k	1
Ego-4D [33]	0.8K	2	Charades [124]	23.6K	1
YouCook2 [172]	41.9K	2	ShareGPT4Video [21]	255K	-

Table 17: The detailed statistics of Multi-Image and Video datasets used in LLaVA-OneVision. Prompt ID denotes the ID of Formatting Prompt corresponding to the ID in Table 19. - denotes no fromatting prompt is used. “MI" means it is the multi-image version dataset from DEMON [69].

ID	Type	Postion	Prompt
1	VQA	Tail	Answer the question with a single word (or phrase).
2	VQA	Head	Hint: Please answer the question and provide the final answer at the end.
3	VQA (Yes/No)	Tail	Answer the question with Yes or No./Yes or No?/...
4	Choice	Tail	Answer with the given letter directly
5	Choice (Option Letter)	Tail	Answer with the option letter from the given choices directly. / Please respond with only the letter of the correct answer.
6	Choice (Option Letter)	Head	Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
7	Region Caption	All	Provide a short description for this region.
8	Grounding	All	Provide the bounding box coordinate of the region this sentence describes.
9	Breif Caption	All	Provide a one-sentence caption for the provided image./Create a compact narrative representing the image presented./...
10	Screen Summarization	All	Summarize the main components in this picture./Provide a detailed account of this screenshot./...
11	Detailed Caption	All	Describe this image in detail./Explain the visual content of the image in great detail./...
12	Science Books	All	Here is a diagram figure extracted from some Grade 1 - 6 science books.\nPlease first describe the content of this figure in detail, including how the knowledge visually displayed in the diagram.\nThen start with a section title \"related knowledge:\", briefly and concisely highlight the related domain knowledge and theories that underly this diagram. Note that you do not need to provide much detail. Simply cover the most important concepts.
13	Information Extraction	Head	Provide the requested information directly.
14	Graph Sumarization	All	Please clarify the meaning conveyed by this graph./Explain what this graph is communicating./...
15	Photo Sumarization	All	Highlight a few significant elements in this photo./Mention a couple of crucial points in this snapshot./...
16	Chart Sumarization	All	What insights can be drawn from this chart?/Explain the trends shown in this chart./...
17	OCR	Head	OCR this image section by section, from top to bottom, and left to right. Do not insert line breaks in the output text. If a word is split due to a line break in the image, use a space instead
18	Diagram Linkage	All	Dissect the diagram, highlighting the interaction between elements./Interpret the system depicted in the diagram, detailing component functions./...
19	Code Generation	All	Compose the HTML code to achieve the same design as this screenshot.
20	Choice (with Reasoning)	Head	First perform reasoning, then finally select the question from the choices in the following format: Answer: xxx.
21	Math Computing	Tail	Round computations to 2 decimal places.
22	LaTeX OCR	All	Please write out the expression of the formula in the image using LaTeX format.
23	Text Reading	All	What is written in the image? Answer this question using the text in the image directly./Read and list the text in this image.
24	Choice (Full Option)	Tail	Please provide your answer by stating the letter followed by the full option.

Table 18: The information of formatting prompts for Single-Image data. The “Position" means the position of the formatting prompt in the prompt where “All" means the formatting prompt is the prompt. Sometimes, there are multiple prompts of the same meaning. In this case, the prompt column is fomatted as “Prompt1/Prompt2/...".

ID	Type	Postion	Prompt
			Video
1	Choice (Option Letter)	Tail	Answer with the option letter from the given choices directly. / Please respond with only the letter of the correct answer.
2	Choice (Full Option)	Tail	Please provide your answer by stating the letter followed by the full option.
			Multi-Image
3	Open-Ended	Head	What's the difference between 2 images?
4	Open-Ended	Head	Given the stories paired with the first several images, can you finish the story based on the last image?/With the narratives paired with the initial images, how would you conclude the story using the last picture?/...
5	Multi-Choice	Head	Here is a Raven's Progressive Matrice in a three-by-three form. You are provided with the first eight elements in eight images, please select the last one from four choices following the structural and analogical relations.
6	Multi-Choice	All	There are ten possible explanations for the ten different answers to a VQA: ... I will give you two sets of pictures, questions, and answers to determine if they belong to the same 'Question-Answer Differences'. You must choose your answer from the Choice List.
7	Open-Ended	Head	This is a 3D scenario.
8	Open-Ended	Head	I will give you several images and a question, your job is to seek information in the slide and answer the question correctly./Based on the images, please answer the following question./...
9	Multi-Choice	Head	Provided with a series of diagrams from a textbook, your responsibility is to correctly answer the following question. You must choose your answer from the Choice List./Using a selection of textbook diagrams, your task is to provide an accurate response to the subsequent query. You must choose your answer from the Choice List./...
10	Open-Ended	Head	Given six images taken from different cameras on a street view car, your task is to answer questions about the depicted scene. You must choose your answer from the Choice List. /Upon receiving six photographs captured from various cameras on a street-view car, your responsibility is to provide accurate responses to questions about the scene. You must choose your answer from the Choice List. /...
11	Multi-Choice	Head	I will provide you with two sets of pictures, each of which shows an object in the opposite state. Can you tell me if the states of these two sets of pictures are the same? You must choose your answer from the Choice List. /I have two sets of pictures that show an object in opposite states. Can you tell me if the states of these two sets of pictures are the same? You must choose your answer from the Choice List. /...
12	Multi-Choice	Head	Are the following four images of the same class? You must choose your answer from the Choice List. /Do the following four images belong to the same category? You must choose your answer from the Choice List. /...
13	Multi-Choice	Head	Are these two workpieces the same type?/Are these two workpieces of the same kind?/...
14	Multi-Choice	Head	Presented with a textual recipe tutorial, your task is to scrutinize it carefully and select the image that is incoherent in the provided sequence of images. You must choose your answer from the Choice List. /Given a text-based recipe guide, your responsibility is to meticulously review it and identify the image that doesn't fit in the following sequence of images. You must choose your answer from the Choice List. /...
15	Multi-Choice	Head	I will give you a series of comic panels. The dialogue box of the last panel is masked. Can you choose the most relevant one from the candidates? You must choose your answer from the Choice List. /Given previous full panels and one masked panel, your job is to select the most appropriate dialogue among four candidates. You must choose your answer from the Choice List. /...
16	Open-Ended	Head	Give you a main goal, your job is to figure out what to do now by looking at current envirments. Your past views as well as decisions are also provided./Given a primary objective and your current surroundings, use your previous decisions and perspectives to determine your next move./...
17	Multi-Choice	Head	I will give you two pictures of the book cover. Please look at the pictures and answer a question You must choose your answer from the Choice List. /I will provide you with two images of the book cover. Please examine the images and answer a question. You must choose your answer from the Choice List. /...
18	Multi-Choice	Head	I will give you some pictures, and each group of pictures will correspond to a question. Please answer it briefly. You must choose your answer from the Choice List. /For each group of pictures, there is a question. Please give a short answer to it. You must choose your answer from the Choice List. /...
19	Open-Ended	Head	Please give a editing Request to describe the transformation from the source image to the target image./What is the correct image edit instruction that can transfrom the source image to target image?/...
20	Open-Ended	Head	What's the difference between 2 images? /Identify the alterations between these two images. /...
21	Open-Ended	Head	What's the difference between 2 birds? /Identify the alterations between these two birds. /...
22	Open-Ended	Head	What's the difference between 2 images? /Identify the alterations between these two images. /...
23	Open-Ended	Head	Given the stories paired with the first several images, can you finish the story based on the last image?/With the narratives paired with the initial images, how would you conclude the story using the last picture?/...
24	Open-Ended	Head	Given the stories paired with the first several images, can you finish the story based on the last image?/With the narratives paired with the initial images, how would you conclude the story using the last picture?/...
25	Open-Ended	Head	Given the stories paired with the first several images, can you finish the story based on the last image?/With the narratives paired with the initial images, how would you conclude the story using the last picture?/...
26	Multi-Choice	All	Answer the following multiple-choice question: Here is a statement describing 2 images: ... Is it true or false?

Table 19: The information of formatting prompts for One-Vision data. The "Position" means the position of the formatting prompt in the prompt where "All" means the formatting prompt is the prompt. Sometimes, there are multiple prompts of the same meaning. In this case, the prompt column is fomatted as "Prompt1/Prompt2/...".

E.4 Policy Information and Reproducibility

We will open-source most of the public datasets we used. These images and data are already publicly available for academic research; we incorporated them and converted the format for our use. However, a small portion of our data sources related to user data and those obtained using the Azure OpenAI Service cannot be directly released due to company policy. We will provide the exact data YAML files used in the final reproduction scripts and will offer reproducible experimental scripts, training logs, and final version checkpoints using fully public data as our compute resources allow.