EVLM: An Efficient Vision-Language Model for Visual Understanding

Anonymous authors
Paper under double-blind review

Abstract

In the field of multi-modal language models, the majority of methods are built on an architecture similar to LLaVA. These models use a single-layer ViT feature as a visual prompt, directly feeding it into the language models alongside textual tokens. However, when dealing with long sequences of visual signals or inputs such as videos, the self-attention mechanism of language models can lead to significant computational overhead. Additionally, using single-layer ViT features makes it challenging for large language models to perceive visual signals fully. This paper proposes an efficient multi-modal language model to minimize computational costs while enabling the model to perceive visual signals as comprehensively as possible. Our method primarily includes: (1) employing cross-attention to image-text interaction similar to Flamingo. (2) utilize hierarchical ViT features. (3) introduce the Mixture of Experts (MoE) mechanism to enhance model effectiveness. Our model achieves competitive scores on public multi-modal benchmarks and performs well in tasks such as image captioning and video captioning.

1 Introduction

Recently, both academia and industry have seen the emergence of numerous outstanding large language models (Brown et al., 2020; Achiam et al., 2023; Anil et al., 2023; Gao et al., 2023; Team et al., 2023; Bai et al., 2023; Team, 2023; Zeng et al., 2022; Young et al., 2024), these models not only reduce the gap to the performance of GPT-4 but also excel across significant benchmarks. These powerful language models have fostered the development of vision-language models. Today's vision-language models can handle various visual tasks, including object recognition, object localization, OCR (optical character recognition), and document analysis. These advancements have significantly improved the model's ability to process complex visual information and generate accurate textual descriptions.

Researchers have explored many methods to enhance the perceptual capabilities of vision-language models for details. Some studies (Lin et al., 2023; Jain et al., 2024; Hong et al., 2024; Lee et al., 2024) focus on using multiple visual encoders to enhance the encoding of visual signals, while others (Liu et al., 2024b; Li et al., 2024b; Liu et al., 2024a; Chen et al., 2024; Dong et al., 2024) aim to improve the model's perception of small objects by increasing image resolution or slicing high-resolution images. These methods have notably boosted the performance of vision-language models in detail recognition, particularly in tasks such as OCR and document analysis. Once visual signals are encoded into features, the resulting features often require long token lengths. Attempts to use multiple encoders to extract visual features, increase input image resolution, or slice images, especially when dealing with video data or multiple image inputs, these operations significantly increase the length of visual tokens. In models like LLaVA (Liu et al., 2023a), directly feeding excessively long visual tokens into language models will result in substantial computational overhead. On the other hand, adopting approaches akin to Q-former Li et al. (2023a) for compressing visual features may lead to loss of visual information. Therefore, choosing appropriate strategies to balance computational efficiency and information richness is crucial in designing efficient vision-language models.

This paper proposes an efficient visual-language model that adopts a cross-attention mechanism similar to Flamingo (Alayrac et al., 2022) for interaction between visual and textual inputs. Adopting cross-attention ensures that even with long visual tokens, controlling the feature dimensions in cross-attention does not lead to excessive computational overhead. To feed sufficient visual features into the language model, hierarchical

[†]Corresponding author

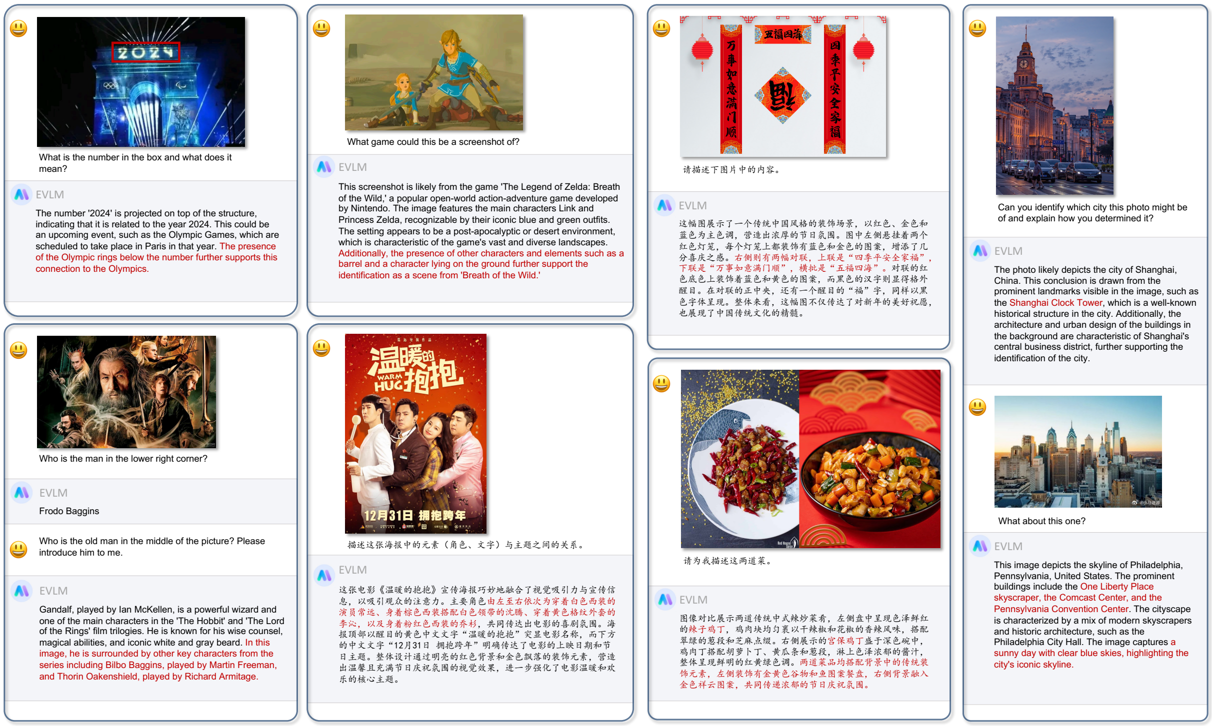

Figure 1: Some qualitative examples generated by our model.

ViT features are employed, enabling the large-scale language model to perceive visual signals at different levels, thus aiding in understanding tasks of varying granularity. Additionally, to enhance model performance, the Mixture of Experts (MoE) is applied on the Cross Attention to scale trainable model parameters. Extensive pre-training on a large-scale dataset of bilingual image-text pairs enables our visual-language model to acquire rich visual-linguistic knowledge. Leveraging our pre-trained model and refined visual feature input design, our model achieves competitive scores on public multimodal benchmarks and demonstrates exemplary performance in tasks such as image and video captioning. Fig. 1 shows some qualitative examples generated by our model.

2 Model Architecture

Our model architecture is based on Flamingo (Alayrac et al., 2022), primarily consisting of a visual encoder, a large language model, and a Gated Cross Attention Layer. To enable the multi-modal model to capture more fine-grained visual signals, we extracted hierarchical visual features from different layers of the visual encoder and increased the length of Flamingo's media tokens. Fig. 2 is our model framework diagram.

Visual Encoder: To enhance our multi-modal model's visual perception capability, we utilized the 4.4B EVA2-CLIP-E-Plus (Sun et al., 2023) model. In practice, we removed the norm and head layers after the last transformer block. To extract hierarchical visual features, we uniformly sampled 8 feature sequences from the last 40 layers of the transformer and sequentially fed these 8 feature sequences into different Gated Cross Attention layers of Flamingo.

Gated Cross-Attetion Layer: Similar to Flamingo, we use gated cross-attention to interact between vision and text. Unlike Flamingo, we replace the media token <image> with a set of learnable tokens of sequence length 16, hoping these learnable tokens can carry visual features similar to Qformer. Because not all text sequences are necessarily related to visual features, we pad a set of all-zero vectors on the visual feature sequence. The attention mask for learnable tokens, text sequences, and visual features is shown in Fig. 2, where each set of learnable tokens can only interact with the corresponding image, and text sequences can

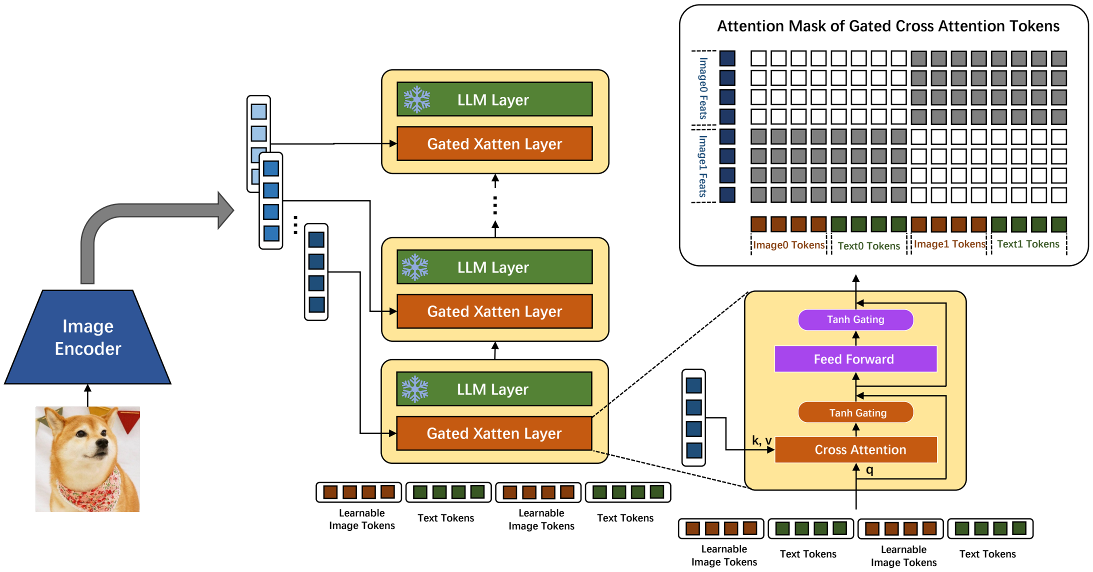

Figure 2: The framework diagram of our multi-modal model.

only interact with the previous image in the multi-modal sequence. The overall parameter size of gated cross-attetion layer is roughly 1.6 billion.

Large Language Model: We used the Qwen-14B-Chat 1.0 (Bai et al., 2023)version of the language model, showing remarkable performance in content understanding and logical reasoning. To condition the language model on visual inputs, we insert a gated cross-attention layer before every transformer layer of the language model.

Disscusion on Efficient Training: In this section, we analyze the computing budget estimation of our EVLM and compare it with the result of current mainstream LLaVA family architectures. In the FLOPs estimation, we evaluate the attention and FFN layers within a single Transformer layer in the LLM. As shown in Figure 3, h_{llm} denotes the hidden state size of the LLM, while d_{img} represents the dimension of visual representations. Moreover, in the Gated Cross-Attention Layer, the ratio of the attention layer to h_{llm} is denoted as r_{x_c}, and the ratio of the FFN layer to h_{llm} is also denoted as r_{x_f}. We distinguish between concatenation and cross-attention interaction modes, referred to as $\textbf{FLOPs}_{\text{full-attention}}$ and $\textbf{FLOPs}_{\text{cross-attention}}$, respectively. The total FLOPs can be estimated as follows:

$$\textbf{FLOPs}_{\text{full-attention}} = 24B(s_{\text{img}} + s_{\text{txt}})h_{\text{llm}}^2 + 4B(s_{\text{img}} + s_{\text{txt}})^2 h_{\text{llm}}, \tag{1}$$

$$\begin{aligned}\textbf{FLOPs}_{\text{cross-attention}} =& 4(6 + r_{x_c} + r_{x_f})B(16 + s_{\text{txt}})h_{\text{llm}}^2 + 4B(16 + s_{\text{txt}})^2 h_{\text{llm}} \\ &+ 4r_{x_c} B s_{\text{img}} d_{\text{img}} h_{\text{llm}} + 4r_{x_c} B(16 + s_{\text{txt}}) s_{\text{img}} h_{\text{llm}},\end{aligned} \tag{2}$$

where B denotes the batch size, s_{img} and s_{txt} denote the length of visual embeddings and text token sequences, respectively. The token sequence length is significantly reduced by employing a gated cross-attention layer, effectively lowering FLOPS and achieving efficient training. Specifically, with a length h of 5120 and d of 1792, and with r_{x_c} and r_{x_f} set to 0.2 and 0.5 respectively, we observed significant FLOPS reductions across various pre-training stages. FLOPS were reduced to S times the original, where $\mathbf{S} = \frac{\textbf{FLOPs}_{\text{cross-attention}}}{\textbf{FLOPs}_{\text{full-attention}}}$. For example, in multi-modal pre-training, s_{img} was 256 and s_{txt} was 64, yielding an $\mathbf{S}_P$ of 0.24. During continual pre-training, s_{img} was 1024 and s_{txt} was 64, resulting in an $\mathbf{S}_{CP}$ of 0.077. These results show a significant improvement in training efficiency.

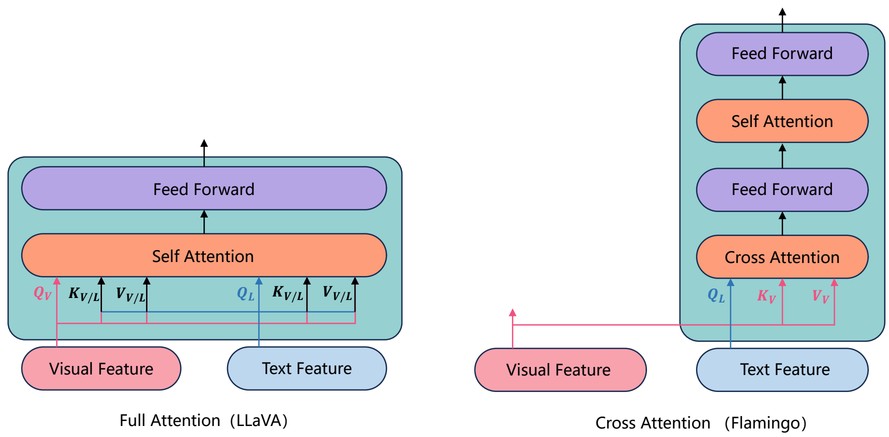

Figure 3: Full Attention and Cross Attention used in multi-modal model.

3 Training

Our training process consists of three stages: multi-modal pre-training, multi-task continual pre-training, and multi-modal instruction fine-tuning.

3.1 Multi-modal Pre-training

Our multi-modal pre-training aims primarily at two objectives: 1) Cross-modal alignment of images and text, and 2) Modeling the intrinsic relationships within multi-modal data. We collected a large-scale dataset of image-text captions and web-type multi-modal data based on these objectives. For the image-text caption data, we implemented a data cleaning process to filter out anomalies such as images with unusual aspect ratios and text with repetitive words and to ensure relevance between images and text. We applied relevance filtering similar to MMC4 (Zhu et al., 2024) for web-type multi-modal data to retain highly correlated images. The detailed data processing procedures are documented in the appendix A.1. The Table 1 illustrates the distribution of our pre-training data. We obtained 2.5 billion image-text caption data and 50 million web-type multi-modal data. It is worth noting that 60% of this data consists of Chinese, including a significant amount of self-built Chinese caption data. This was done to enhance the fine-grained alignment capability of our multi-modal model, covering specific visual concepts such as celebrity, landmark building, and dish.

During model training, we concatenated the caption and multi-modal web-type data separately to ensure each sample had up to 64 images and a sequence length of 2048, resulting in a total of 60 million training samples. In the first 25% phase of training, only the parameters of the Gated Cross Attention Layer were trained. In the subsequent 75% phase, we unfrozen the parameters of the latter half of the Visual Encoder for training. The input image size was 224×224 during this phase. The training objective was to minimize the cross-entropy of the text tokens. We employed a cosine learning rate strategy with a maximum learning rate of $6.4e^{-4}$. We completed training on the entire set of 60 million training samples. The detailed training hyperparameter settings are documented in the appendix B.

3.2 Multi-task Continual Pre-training

We introduce the multi-task continual pre-training stage between the multi-modal pre-training and instruction fine-tuning. Compared with the pre-training stage, the continual pre-training stage pays more attention to MLM's high-level visual question-answering ability. Compared with the SFT stage, the continual pre-training stage is still about acquiring ability, not activating ability.

In the continual pre-training stage, our training data sources are categorized into five distinct parts: Visual Question Answering (VQA) data, Natural Language Processing (NLP) data, OCR data, detection data, and data which are sampled from the first pre-training stage to prevent catastrophic forgetting. The VQA data

Table 1: Details of Our pre-training data. LAION-en and LAION-zh are the English and Chinese language subset of LAION-5B (Schuhmann et al., 2022a). LAION-COCO (Schuhmann et al., 2022b) is a synthetic dataset generated from LAION-en. DataComp (Gadre et al., 2023) and Coyo (Byeon et al., 2022) are collections of image-text pairs. BLIP-cap is the bootstrapped pre-training datasets used by BLIP (Li et al., 2022). MMC4 (Zhu et al., 2024) and WanJuan (He et al., 2023) are the corpus of images interleaved with text.

Language	**Dataset**	**Type**	**Cleaned**
English	BLIP-cap	Caption	100M
	LAION-COCO	Caption	40M
	LAION-en	Caption	200M
	Coyo	Caption	160M
	DataComp	Caption	500M
	MMC4	Web	40M
Chinese	LAION-zh	Caption	100M
	In-house Data	Caption	1.4B
	WanJuan	Web	10M
	Total	Caption	2.5B
		Web	50M

mainly comes from open-source data. The OCR and detection datasets combine open-source data and data generated through our simulations. The detailed data processing procedures of OCR are documented in the appendix A.2. The NLP data is obtained from internal resources. Table 2 shows the specific data proportions and sources. Finally, We create interleaved image-text data by packing the same task data into sequences of length 2048 and increasing the image resolution from 224×224 to 448×448.

In this phase, we unfrozen the parameters of the latter half of the Visual Encoder and gated cross-attention layer for training. The training objective was to minimize the cross-entropy of the text tokens. We employed a cosine learning rate strategy with a maximum learning rate of $1e^{-4}$. The model obtained at this stage is called EVLM-Base. The detailed training hyperparameter settings are documented in the appendix B.

Table 2: Details of multi-task continual pre-training data.

Task	**# Samples**	**Dataset**
partially sampled stage1 data	30M	The data were clustered and then sampled according to their cluster IDs.
VQA	9M	GQA, VGQA, VQAv2, DVQA, OCR-VQA, DocVQA, TextVQA, ChartQA, AI2D, mmicl, Simulation data
Detection	17M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
OCR	26M	SynthDoG-en & zh, Common Crawl pdf & HTML, Simulation data
nlp data	10M	In-house Data
total	92M	

3.3 Supervised Fine-tuning

3.3.1 Dense Baseline Model

During this stage, we finetuned our EVLM-Base through instruction finetuning to activate its instruction-following abilities. We used a broad range of high-quality instruction tuning data, totaling 2.3 M samples. As illustrated in the Table 3, these include: 1) User Instruct Data: We incorporate the ShareGPT-4V and LLaVA-ZH datasets. 2) Multimodal Document/Chart Data: We used DocVQA and SynDog-EN to enhance the model's document comprehension capabilities. Following Qwen VL-7B Chat, we also used ChartQA, DVQA, and AI2D to understand charts and diagrams better. 3) Math Problems: We used MathInstruct, MathPlus, and geoqa+ data to improve the model's mathematical reasoning ability.

In this phase, we froze the LLM and tuned only the cross-attention layers and the last quarter ViT layers, achieving robust performance. The model obtained at this stage is called EVLM-Chat.

3.3.2 Scaling via Mixture-of-Experts

In order to achieve better performance, we get more training parameters by scaling the Gated Xaaten Layer. As depicted in Fig 4, we employ a fine-grained MoE architecture. Initially, we replicate the parameters of the FFN of EVLM-Base N times. Subsequently, each replicated FFN is segmented into M fine-grained experts, resulting in a total of NM fine-grained experts. We choose a routing layer that selects the appropriate set of k fine-grained experts to compute the output for the current token. We have set $n = 4$, $m = 4$, and $k = 4$ in our configuration.

Drawing from established practices(Dai et al., 2024), we introduce the world expert tasked with learning general knowledge. This expert is involved in the processing of every token. The output from the world expert is then combined with the outputs from the fine-grained experts to derive the final result.

We employ the same training data and configuration of the dense baseline model, and we freeze the LLM and tune only the cross-attention layers and the last quarter ViT layers. The model obtained at this stage is called EVLM-MoE.

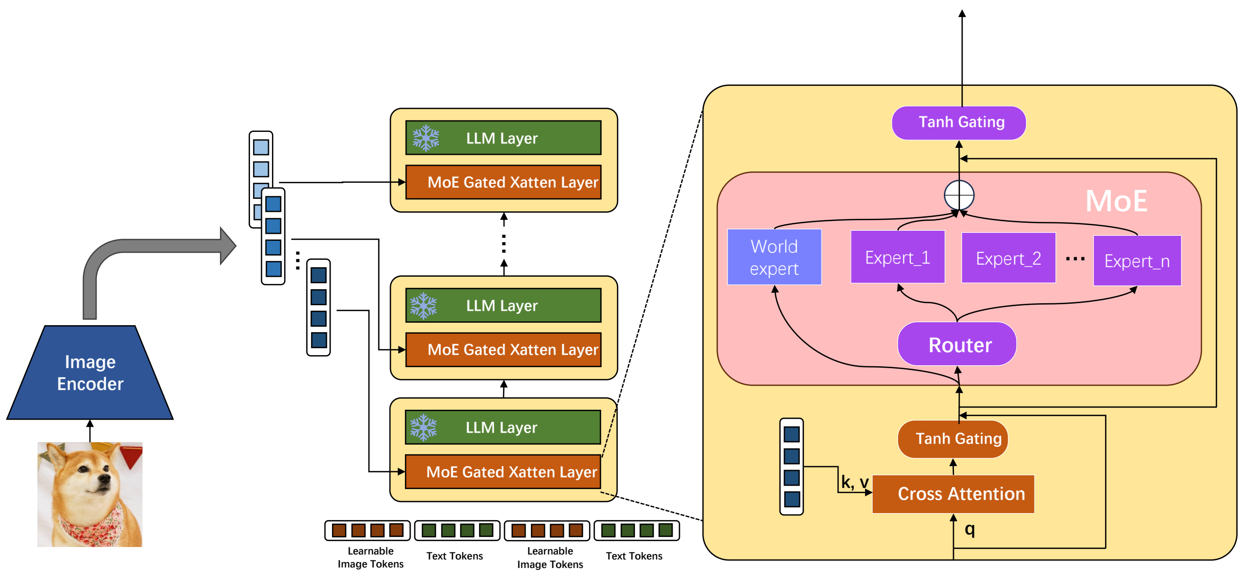

Figure 4: MoE structure.

4 Evaluation

In this section, we evaluate various multi-modal tasks to assess our models' visual understanding ability comprehensively.

4.1 Convergence of Multi-modal Pre-training Stage

We visualized the convergence of the model during the multi-modal pre-training phase. As shown in Fig. 5a, the loss steadily decreases as training progresses. To better monitor the model's alignment between images and text, we randomly sampled 10 examples from each class in the ImageNet-1K (Deng et al., 2009) validation set to assess the model's discriminative ability. During the evaluation, we input a prompt and computed the loss for each of the 1,000 candidate classes, selecting the class with the lowest loss as the model's predicted category to calculate accuracy. From Fig. 5b, it can be observed that as training progresses, the recognition

Table 3: Details of supervised fine-tuning data.

Range	Dataset	Type
User Instruct Data	ShareGPT-4V	665K
	LLaVA-ZH	150K
Multimodal Document/Chart Data	DocVQA	10K
	SynDog-EN	30K
	ChartQA	18K
	DVQA	200K
	AI2D	12K
Math Problems	MathInstruct	262K
	MathPlus	894K
	geoqa+	72K
Total		2.3M

accuracy on the ImageNet-1K validation set continues to improve, which serves as an effective monitoring mechanism.

From the evaluation set of ImageNet-1K, we observe rapid convergence of multi-modal large models. This is primarily attributed to the pre-trained parameters of ViT and LLM that we have initialized, enabling effective coarse-grained alignment of multi-modal data with relatively small amounts of image-text pairs. We constructed a finer-grained evaluation set to better monitor the information gain brought by large-scale multi-modal pre-training. This set comprises seven fine-grained categories, including POI, dish, game, and so on. We evaluated these seven categories using methods similar to those used for ImageNet-1K. Fig. 5c illustrates the average accuracy across these categories, indicating a steep increase in accuracy for fine-grained recognition as training progresses. This underscores the necessity of large-scale multi-modal pre-training; while coarse-grained alignment is achievable with limited image-text data, a comprehensive understanding of many fine-grained concepts necessitates extensive multi-modal knowledge. The appendix C.1 presents the variability in accuracy for each of the seven fine-grained categories. Despite extensive pre-training, the accuracy for the "Star" category remains relatively low, suggesting that our current multi-modal pre-training data may not sufficiently cover comprehensive multi-modal knowledge, necessitating further expansion of the dataset scale.

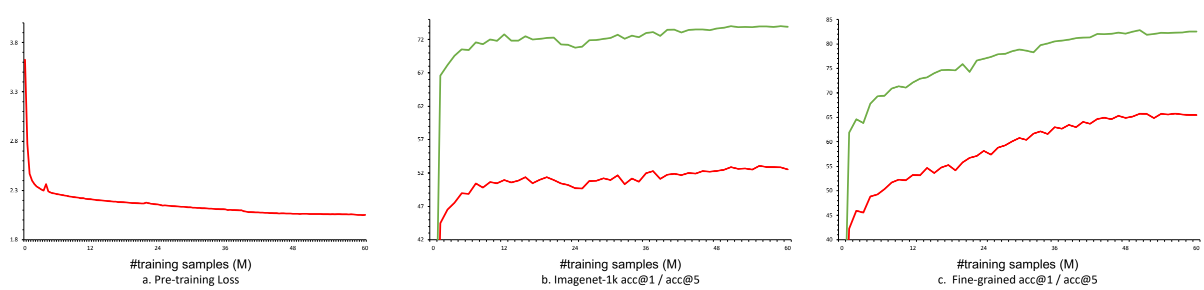

Figure 5: Visualization of the Convergence of the Pre-training Stage

4.2 Comparison with State-of-the-Art VLMs

In this section, we conduct extensive evaluations on a series of benchmarks to assess our model's multimodal understanding and reasoning capabilities. The benchmarks utilized in our study include general VQA, text-oriented VQA, and general Multimodal Benchmarks. As illustrated in Table 4, EVLM-Chat and EVLM-MoE demonstrate superior performance compared to its competitors across most of these benchmarks.

General VQA Benchmarks. We utilize four benchmarks: VQA^{v2}, GQA, ScienceQA (Image Set), and VizWiz. For VQA^{v2}, GQA, and VizWiz, we employ a greedy decoding strategy and report the Top-1 accuracy.

Table 4: **Comparison with SoTA models on 13 multimodal benchmarks.**General VQA benchmarks include: VQA^{v2} Antol et al. (2015), GQA Hudson & Manning (2019), SciQA-Img (Lu et al., 2022) and VizWiz (Gurari et al., 2018).Text-oriented VQA benchmarks include: TextVQA val (Sidorov et al., 2020), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022) and AI2D (Kembhavi et al., 2016).General multimodal benchmarks encompass: MME Fu et al. (2023), MMB (Liu et al., 2023b), MMB_{CN} Liu et al. (2023b) and POPE (Li et al., 2023b). '*' denotes specialist models obtained from separately fine-tuning on each task.

Method	LLM	Res.	General VQA				Text-oriented VQA				General Multimodal Benchmarks			
			VQA^{v2}	GQA	SciQA-Img	VizWiz	TextVQA	DocVQA	ChartQA	AI2D	MME	MMB	MMB_{CN}	POPE
Qwen-VL	Qwen-7B	448^2	79.5	59.3	67.1	35.2	63.8	65.1	65.7	62.3	–	38.2	–	–
Qwen-VL-Chat	Qwen-7B	448^2	78.2	57.5	68.2	38.9	61.5	62.6	66.3	57.7	1487.58/360.71	60.6	–	–
CogVLM*	Vicuna-7B	490^2	82.25	–	91.0	–	70.5	–	–	–	–	76.5	–	87.88
LLaVA-1.5	Vicuna-13B	336^2	80.0	63.3	71.6	53.6	61.3	–	–	–	1531/–	67.7	63.6	85.9
InternVL	Vicuna-13B	336^2	81.2	66.6	–	58.5	61.5	–	–	–	1586.4/–	–	–	87.6
VILA	LLaMA2-13B	336^2	80.8	63.3	73.7	60.6	66.6	–	–	–	1570.1	70.3	64.3	84.2
InfiMM-HD	Vicuna-13B	448^2-1344^2	82.0	63.5	83.6	–	70.7	55.1	–	–	1472.3/329.4	71.6	–	87.9
EVLM-Base	Qwen-14B-Chat 1.0	448^2	82.92	62.19	85.57	49.62	64.51	53.16	59.92	63.14	1579/345	78.1	71.47	94.56
EVLM-Chat	Qwen-14B-Chat 1.0	448^2	81.93	64.39	86.37	47.28	67.52	53.27	63.36	76.0	1593.56/402.5	76.89	76.89	89.65
EVLM-MoE	Qwen-14B-Chat 1.0	448^2	83.76	62.89	86.81	49.19	68.31	54.44	63.12	75.5	1607/351	78.09	76.55	93.3
LLava-Next-34B	Yi-34B	336^2*4	–	–	–	–	–	–	–	74.9	2030.4	79.3	79	–
InternVL1.2	Yi-34B	448^2	–	64.0	83.3	60.0	72.5	57.7	68.0	79.0	1687/489	82.2	81.2	88.0
CogVLM2	LLaMA3-Chinese		–	–	–	–	85.0	88.4	74.7	–	–	78.9	–	–
InternVL-1.5	InternLM2-20B	448^2*40	–	–	–	–	80.6	90.9	83.8	80.7	–	82.2	82.0	–

Table 4 presents the overall performance on general VQA tasks. It is crucial to highlight that the evaluations in VQA^{v2}, GQA, and VizWiz are designed to test the models' visual perception abilities and their capacity to apply prior knowledge effectively. Additionally, ScienceQA, collected from elementary and high school science curricula, contains 21,208 multimodal multiple-choice science questions spanning a wide array of scientific topics, significantly broadening the benchmarking scope.

As shown in Table 4, our EVLM-Chat and EVLM-MoE achieve significantly better outcomes than previous generalist models. Specifically, on the ScienceQA task, EVLM-Chat and EVLM-MoE achieved 86.4% accuracy and 86.8 % accuracy, respectively. This result even surpasses that of previous generalist models with higher resolution, such as InfiMM-HD, which utilizes a dynamic resolution ranging from 448^2 to 1344^2. Moreover, our model demonstrates substantial performance improvements in VQA^{v2}, GQA, and VizWiz. These results underscore EVLM's superior capability to integrate multimodal information and utilize extensive prior knowledge for robust reasoning.

Text-oriented VQA Benchmarks. In addition to the general VQA evaluation, we further investigate our model's detailed visual perception capabilities by assessing its performance on text-oriented VQA datasets with broad real-world applications. These datasets include TextVQA (Sidorov et al., 2020), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), and AI2Diagram (Kembhavi et al., 2016).

The quantitative results, summarized in Table 4, demonstrate that our model outperforms previous general models and recent VLMs on most benchmarks. Notably, on the AI2Diagram dataset, which requires fine-grained visual perception for diagram understanding and associated question answering, EVLM-MoE and EVLM-Chat achieve accuracy of 75.5% and 76.0%, respectively. These findings underscore the effectiveness of our proposed deep vision-text fusion in comprehending complex text details within images.

General Multimodal Benchmarks. In addition to previous VQA evaluations, we further evaluate our model's visual understanding and reasoning abilities of real-world user behavior on general multimodal benchmarks, including MME, MMB, MMB_{CN}, and POPE. Compared to traditional VQA datasets, these benchmarks encompass a broader range of evaluation aspects, necessitating more complex reasoning capabilities.

As summarized in Table 4, EVLM-MoE and EVLM-Chat demonstrate commendable overall performance, highlighting its adaptability and capability across various disciplines. Specifically, our model possesses bilingual capabilities due to the large-scale interleaved data of captions, web pages, videos, images, and text. It outperforms previous generalist models on the MMBench and MMBench-Chinese benchmarks. Additionally, our model performs best on the POPE benchmark, showcasing its ability to reduce hallucinations. Our results demonstrate the benefits of vision-language pre-training on downstream tasks. These findings underscore our model's versatility and effectiveness in handling complex visual and textual information.

4.3 Image Caption

One of the key capabilities of the Multimodal Large Language Model (MLLM) is DenseCaption of images, which is its most direct application scenario. To enhance MLLM's performance in this area, we integrate high-quality description data into the Supervised Fine-Tuning (SFT) dataset based on existing pre-trained models. This enables the MLLM to generate fluent, detailed, accurate, and illusion-free image descriptions. Given the challenges of annotating DenseCaption, including the high cost and inefficiency of manual rewriting, we have designed a comprehensive process for generating high-quality, detailed image description data.

The process comprises several key steps:

1. **Multiple Descriptions Generation**: Multiple descriptions are generated to ensure comprehensive coverage of all image details.
2. **Authenticity Check**: These descriptions are split into short sentences for authenticity verification.
3. **Coherent Description Recombination**: The verified sentences are recombined into a coherent description.
4. **Stylization Using GPT-4**: Finally, the descriptions are stylized using GPT-4 to ensure they meet specifications and are expressive.

Using the data generated by this process, we effectively guided the MLLM's SFT training, significantly enhancing its image description capability to meet or even exceed human satisfaction. This not only improves MLLM's performance in practical applications but also provides a robust data foundation for future research and development.

Auto Caption Pipeline. To generate high-quality dense descriptions (DenseCaption), we employ various visual language models (VLMs), including self-developed models, internVL, GPT-4V, and GPT-4o. Initially, these models generate image descriptions which are then split into multiple phrases using the Llama2-70B model.

The split phrases are de-duplicated with the help of Llama2-70B, ensuring each phrase is unique. Subsequently, a powerful multimodal large language model (MLLM), such as GPT-4o, checks the authenticity of these phrases, retaining those that accurately match the image details. Following the authenticity check, GPT-4o integrates these phrases into coherent and fluent image descriptions. This step ensures the resulting descriptions are both accurate and low in illusions.

To further enhance the detail and completeness of the final image descriptions, we integrate multiple MLLM-generated descriptions. This integration improves the quality of the descriptions by making them more detailed and comprehensive. Finally, we use GPT-4 to refine the linguistic expressions, making the generated image descriptions more fluent, elegant, and consistent with human expression preferences.

Through this multi-step, multi-model synergistic approach, we can generate high-quality, detailed, and accurate image descriptions, significantly enhancing the application of multimodal macrolanguage modeling. This method provides robust data support and technical assurance for the development and practical application of multimodal macrolanguage modeling.

Using the image dense descriptions generated by aforementioned process, we fine-tuned our multi-modal large model. Leveraging our model's powerful visual feature perception capabilities, as shown in Fig. 6, we have achieved promising results on the image-dense captioning task. This has notably reduced hallucination phenomena in visual descriptions.

4.4 Video Caption

4.4.1 Attention Mask

We can also use our EVLM model to understand video. In order to better extract sequence information in the video, such as the action changes of characters in the sequence, the position changes of objects, OCR information in the image, etc., it is necessary to extract information from each image separately when inputting the image sequence into the model to avoid mutual interference between the information of each image. Therefore, it is necessary to design the attention mask in the model during the SFT stage. As shown in Fig. 7, in order to enable the model to acquire all visual information about the video, we ensure that each textual

这幅图描绘了电影《泰坦尼克号》的经典场景，以深蓝色和橙色渐变的天空为背景，营造出一种沉静而浪漫的氛围。图中，杰克·道森（由莱昂纳多·迪卡普里奥饰演）和罗丝·德维特·布卢埃特（由凯特·温丝莱特饰演）身着1912年风格的服装，杰克身着黑色上衣和白色衬衫，而罗丝则身着白色蕾丝上衣和黑色外套，两人紧靠在一起，杰克的手臂环绕在罗丝的腰间，罗丝的头靠在杰克的肩上。在他们下方，一艘巨大的豪华邮轮“泰坦尼克号”正在波涛汹涌的海面上航行，船体上清晰地标有“TITANIC”字样。这艘船的轮廓在夕阳的映照下显得格外壮观，而船身的灯光在海面上投射出明亮的光芒。整个画面以深蓝和橙色为主色调，营造出一种既神秘又浪漫的氛围。

这幅图呈现了西藏著名的布达拉宫，一座宏伟的宫殿建筑群，其壮观的白色外墙和红色屋顶在蓝天的映衬下显得格外引人注目。布达拉宫的建筑风格融合了藏族传统与佛教元素，其错落有致的布局和精致的装饰细节彰显了其悠久的历史和文化价值。宫殿前的广场上，几只鸟儿在空中飞翔，为宁静的场景增添了一抹生动的色彩。广场的倒影在清澈的水面上清晰可见，形成了一幅对称的美丽画面。整个场景在柔和的阳光照耀下，散发出一种宁静而庄严的氛围，让人仿佛能感受到布达拉宫所承载的深厚历史和文化内涵。

In this scene, a family of three stands hand in hand on the beach, facing the sea with their backs to the camera. The father is dressed in a white short-sleeved shirt and dark shorts, the mother wears a blue dress, and the young girl is dressed in a blue top paired with dark shorts. The sand beneath their feet is fine and warm, gently kissed by the sea waves, creating white foam. Above, the sky is azure with a few pristine white clouds drifting by. Sunlight filters through the clouds, casting a warm and tranquil atmosphere over them.

In this image, we see five hot air balloons floating in the sky during sunrise or sunset. The most prominent balloon occupies the central position of the composition, its silhouette clearly visible under the sunlight. Surrounding it are several smaller balloons, each varying in size and color , some displaying gradients of orange and yellow, while others are monochrome. These balloons are dispersed across the scene, creating a serene and dreamy atmosphere.
Below the balloons, there is a wide expanse of grassland scattered with trees and bushes, their silhouettes contrasting with the outlines of the balloons. On the distant horizon, the silhouette of a hill is faintly visible, adding a touch of mystery to the scene. The entire setting is bathed in warm sunlight, with the sky displaying gradients of light yellow and orange, contrasting sharply with the silhouettes of the balloons and the ground.
Overall, this image depicts a beautiful scene where hot air balloons float peacefully over a tranquil countryside during early morning or evening hours.

In this image, we see two people dressed casually sitting on outdoor folding chairs, enjoying a relaxed picnic time. In front of them is a small table with orange and yellow beverage bottles and an orange kettle. Beside them is a yellow jeep adorned with black hubcaps and tires. The background features a serene riverbank scattered with stones of various sizes, while in the distance, rolling mountains are sharply outlined against the sky. The entire scene is bathed in soft light, creating a peaceful and pleasant atmosphere. Clear in the bottom left corner of the image is a watermark from “视觉中国”.

Figure 6: Image dense caption.

token accesses all visual features related to the video. However, media tokens still only access visual features corresponding to their respective frames.

4.4.2 Evaluation Benchmarks

We use the video-dense caption task to verify the performance of our EVLM model. When constructing the video dense caption, we used five publicly available data sources: YouTube1B (Zellers et al., 2022), ActivityNet (Caba Heilbron et al., 2015), Ego4D (Grauman et al., 2022), with a total of 3596 videos, including 39 categories: Online Courses, Beauty & Skincare, Computer, etc. As shown in Fig. 8 to meet the diversity of the evaluation benchmark and better test the model's performance. All scores are scored using gpt4o (Achiam et al., 2023) as the referee, abandoning the original cider, bleu, and other evaluations, which are more authoritative. The statistical data of each category is shown in the image.

The evaluation results are shown in Table 5.

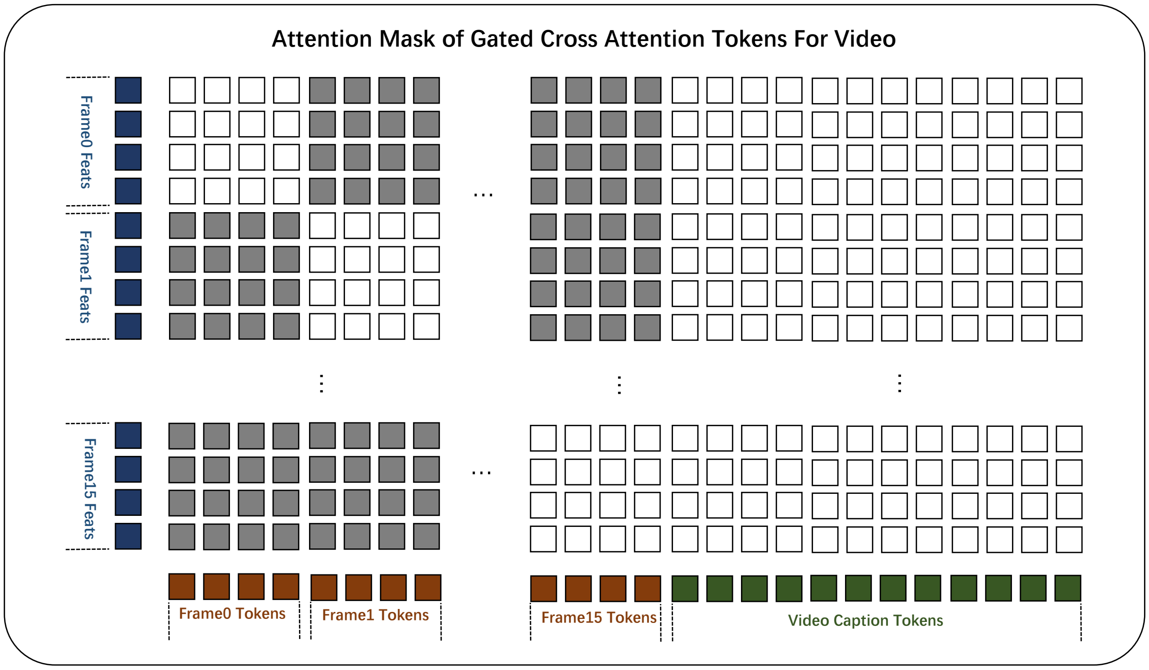

Figure 7: Attention Mask For Video Captioning.

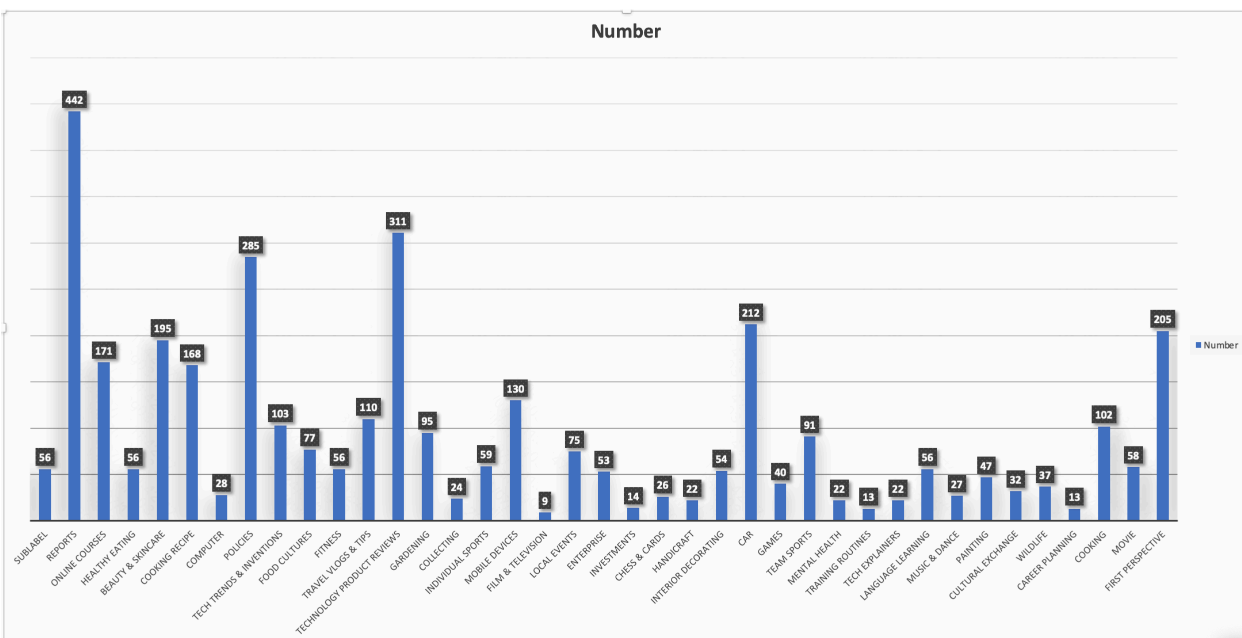

Figure 8: Video caption statistics

4.4.3 Caption Analysis

As shown in figure 9, our EVLM large model can generate dense captions in Chinese and English for videos and can well depict the actions and environment of the characters in the video, action categories, and other information.

Table 5: Details of experiments.

Model	Verbosity	Accurate description
Video-llava	4.06	7.0
Video-llava2	5.69	7.18
Video-llama2	5.32	7.1
Ours	**5.73**	**7.22**

5 Related Work

Recently, multimodal large models have garnered increasing attention, with a multitude of notable works (Alayrac et al., 2022; Li et al., 2023a; Achiam et al., 2023; Liu et al., 2023a; Zhu et al., 2023; Dai et al., 2023; Bai et al., 2023; Chen et al., 2024; Wang et al., 2023; Peng et al., 2023) emerging in the field. Most of these studies focus on exploring how to more effectively integrate Large Language Models (LLMs) with other modalities to accomplish multimodal tasks.

MLLM Input Project Most studies employ visual encoders to extract visual features mapped into Large Language Models (LLMs). Some approaches Liu et al. (2023a); Chen et al. (2024) directly feed the output of visual features through a multilayer perceptron (MLP) and concatenate it with the input of the LLM. Another method (Zhu et al., 2023; Li et al., 2023a) adopts a transformer-based structure, commonly referred to as a "q-former," which uses a fixed number of learnable tokens to represent the visual features. Additionally, there are studies Alayrac et al. (2022); Wang et al. (2023) that integrate visual feature outputs into each layer of the LLM, facilitating a deep fusion of modalities.

MLLM Vision Encoder In the field of multimodal large models, vision models such as CLIP (Radford et al., 2021; Ilharco et al., 2021), EvaCLIP (Sun et al., 2023), and SigLip (Zhai et al., 2023) are commonly used as visual encoders. However, to mitigate the potential for information loss inherent in the features extracted by CLIP, some studies (Tong et al., 2024) opt to employ an additional vision encoder, such as DINOv2 (Oquab et al., 2023), which is designed to enhance the feature representation. Furthermore, to capture features at varying resolutions and to accommodate the need for efficient computation, some works integrate a lightweight Convolutional Neural Network (CNN) (He et al., 2016) model.

MoE The structure of MoE (Mixture of Experts)Jacobs et al. (1991), characterized by sparse activation, can significantly expand the scale of models or datasets under the same computational resources, thereby enhancing model performance. MoE is widely utilized in LLM and MLLM Fedus et al. (2022); Lin et al. (2024); Dai et al. (2024); Jiang et al. (2023). UpcyclingKomatsuzaki et al. (2022) proposes to train moe from dense models to reduce training costs. DeepSeek-MoEDai et al. (2024) enhances the specialization of experts through fine-grained MoE. Additionally, there are studiesMcKinzie et al. (2024); Li et al. (2024a) applying MoE in MLLMs to further improve model performance.

6 Conclusion and Future Work

We propose an efficient multimodal visual-language model. We can efficiently handle large-scale image-text pre-training by leveraging our refined approach to visual inputs. Our model achieves competitive results on public benchmarks, particularly in image and video-dense captioning. Looking forward, several directions can further enhance model performance:

- Employing more powerful, larger-scale language models.
- Exploring the capability of video understanding under extremely long sequences using cross-attention mechanisms.

References

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

用中文详细描述该视频

阳光明媚的一天，一位女士身穿浅黄色衬衫，套着深色马甲，蓝色牛仔裤，正在修剪草坪。她背着一个黑色的婴儿背带，背带里坐着一个头戴卡其色遮阳帽，身穿蓝色短袖和短裤的小宝宝。女士双手握着割草机的把手，沿着草坪慢慢地推动，一边割草一边低头温柔地注视着怀中的宝宝。宝宝好奇地四处张望，似乎对周围的一切都充满兴趣。画面温馨，充满了生活气息。

Describe the video details

On a sunny day, a woman wearing a light yellow shirt, a dark vest, and blue jeans is mowing the lawn. She is wearing a black baby carrier, in which sits a baby wearing a khaki sun hat, a blue short-sleeved shirt and shorts. The woman holds the handle of the lawn mower with both hands and pushes it slowly along the lawn. While mowing the lawn, she lowers her head and looks at the baby in her arms tenderly. The baby looks around curiously, and seems to be interested in everything around him. The picture is warm and full of life.

用中文详细描述该视频

在一个阳光明媚的日子里，一位身穿黄色T恤和黑色运动裤的运动员，正在一个圆形投掷区内练习链球投掷。他双手紧握链球手柄，双脚分开，站在圆圈中央。 运动员开始旋转身体，链球在他头顶画出一个越来越大的圆。随着旋转速度的加快，链球的离心力也越来越大。 最终，运动员松开手，链球在空中划出一道抛物线，飞向远方。运动员保持着投掷后的姿势，目送链球远去。整个过程充满了力量和动感。

Describe the video details

On a sunny day, an athlete in a yellow T-shirt and black sweatpants was practicing hammer throwing in a circular throwing area. He held the hammer handle tightly with both hands and stood in the center of the circle with his feet apart. The athlete began to rotate his body, and the hammer drew a larger and larger circle above his head. As the speed of rotation increased, the centrifugal force of the hammer became stronger and stronger. Finally, the athlete let go, and the hammer drew a parabola in the air and flew into the distance. The athlete maintained the posture after throwing and watched the hammer go away. The whole process was full of power and movement.

Figure 9: Video caption.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pp. 2425–2433, 2015.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

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

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

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

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120):1–39, 2022.

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

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

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

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pp. 18995–19012, 2022.

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

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*, 2023.

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

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *CVPR*, pp. 14281–14290, 2024.

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

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27992–28002, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2023.

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

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*, 2022.

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Collavo: Crayon large language and vision model. *arXiv preprint arXiv:2402.11248*, 2024.

Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *arXiv preprint arXiv:2405.05949*, 2024a.

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

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

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. *arXiv preprint arXiv:2403.18814*, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Haogeng Liu, Quanzeng You, Xiaotian Han, Yiqi Wang, Bohan Zhai, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Infimm-hd: A leap forward in high-resolution multimodal understanding. *arXiv preprint arXiv:2403.01487*, 2024a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.

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

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

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

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv:2210.08402*, 2022a.

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

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

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

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.

InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pp. 9568–9578, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*, 2022.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023.

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

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *NeurIPS*, 2024.

A Dataset details

A.1 Caption data and Web-type data

For the caption dataset, we conducted the following data-cleaning processes:

1. Removed data containing damaged images and solid color images.
2. Removed data with abnormal aspect ratios.
3. Removed data containing extremely low-resolution images.
4. Removed data with text consisting solely of numbers or symbols.
5. Removed data with text containing long sequences of digits.
6. Removed data where the text contained duplicate words.
7. Removed data containing specific terms such as "HTTP", ".com" and ".png" in the text.
8. Removed data with an excessively short text.
9. Removed data containing date-related text.
10. Converted traditional Chinese characters to simplified Chinese characters.
11. Utilized the CLIP model to calculate image-text relevance and removed data with low relevance scores.

For the web-type dataset, we performed the following straightforward processing steps:

1. Removed data containing damaged images and solid color images.
2. Removed data with abnormal aspect ratios.
3. Removed data containing extremely low-resolution images.
4. Removed data with an excessive number of images in web data.
5. Removed data where the text length exceeded 2048 characters.
6. Applied relevance filtering similar to MMC4 (Zhu et al., 2024) to retain highly correlated images.

A.2 OCR

To enhance the OCR capabilities of our model, we have meticulously curated an OCR dataset sourced from a combination of real-world data and synthetic data. The real-world data we gathered includes content from various sources such as videos uploaded to Kuaishou, the Wukong dataset, Common Crawl 2021, street view data, and a plethora of ebooks presented as images to represent authentic scenarios. To ensure the validity of the image-text pairs at the model's designated resolution, we have implemented the following key steps for processing real-world data:

1. Employing expert models to extract texts, coordinate boxes, and recognition confidence prob from the images.

2. Employing image inpainting techniques to eliminate text with characters that are excessively small or possess low recognition confidence.
3. Filtering out images with inadequate text content.
4. Eliminating images that contain redundant text across the entire dataset.

For the synthetic data component, we have harnessed the power of SynthDog to create a diverse OCR dataset, incorporating the use of LaTeX to produce OCR data with dense text. Throughout the data generation process, we begin by selecting text-free images from Kuaishou videos as backgrounds to simulate real-world scenarios. We then explore a wide array of Chinese and English fonts, encompassing both handwritten and standard styles, to generate text in various formats. Furthermore, we introduce uncommon characters, artistic fonts, and diverse data types to enrich the dataset. To bolster the model's ability to recognize dense text, we employ LaTeX to generate PDF data containing a higher volume of characters, subsequently converting them into image-text pairs.

B Hyperparameters

We report the detailed training hyperparameter settings in Table 6.

Table 6: Training hyperparameters

Configuration	Multi-modal Pre-training	Continual Pre-training	Supervised Fine-tuning
ViT init.	EVA2-CLIP-E-PLUS	1st-stage	2nd-stage
LLM init.	Qwen-14B-Chat 1.0	Qwen-14B-Chat 1.0	Qwen-14B-Chat 1.0
Gated Cross Attention init.	random	1st-stage	2nd-stage
Image resolution	224^2	448^2	448^2
ViT sequence length	257 * 8	1025 * 8	1025 * 8
LLM sequence length	2048	2048	2048
Optimizer		AdamW	
Optimizer hyperparameter		$\beta_1 = 0.9, \beta_2 = 0.999, eps = 1e^{-8}$	
Peak learning rate	$6e^{-4}$	$1e^{-4}$	$5e^{-5}$
Minimum learning rate	$3e^{-5}$	$5e^{-5}$	$1e^{-6}$
ViT Drop path rate		0	
Learning rate schedule		cosine decay	
Weight decay		0.05	
Gradient clip		10.0	
Training steps	125k	50k	12k
Warm-up steps	2000	2000	500
Global batch size	480	160	16
Gradient Acc.	1	1	1
Numerical precision		`bfloat16`	
Data parallel mode		FSDP SHARD_GRAD_OP	
Activation checkpointing		✓	

During the multi-modal pre-training phase, the model was trained using the AdamW optimizer with parameters set as $\beta_1 = 0.9, \beta_2 = 0.999, eps = 1e^{-8}$. A cosine learning rate schedule was employed, with a maximum learning rate of $6e^{-4}$ and a minimum of $3e^{-5}$, incorporating a linear warm-up over 2000 steps. We applied a weight decay of $5e^{-2}$ and gradient clipping set to 10.0. Initially, during the first 25% of training, only the parameters of the Gated Cross Attention Layer were trained. In the subsequent 75% phase, the parameters of the latter half of the Visual Encoder were unfrozen for training. The input image size was maintained at 224×224 pixels throughout this phase. Training encompassed the entire dataset comprising 60 million training samples.

During the continual multi-task training stage, we augmented the input resolution of the visual encoder from 224×224 to 448×448, thereby mitigating information loss associated with image down-sampling. We utilized a cosine learning rate schedule with a maximum learning rate of $1e^{-4}$ and a minimum of $5e^{-5}$, including a linear warm-up over 2000 steps.

C Additional experimental details

C.1 Convergence of Multi-modal Pre-training Stage

Figure 10 illustrates the evolution of accuracy across seven fine-grained categories throughout the training process.

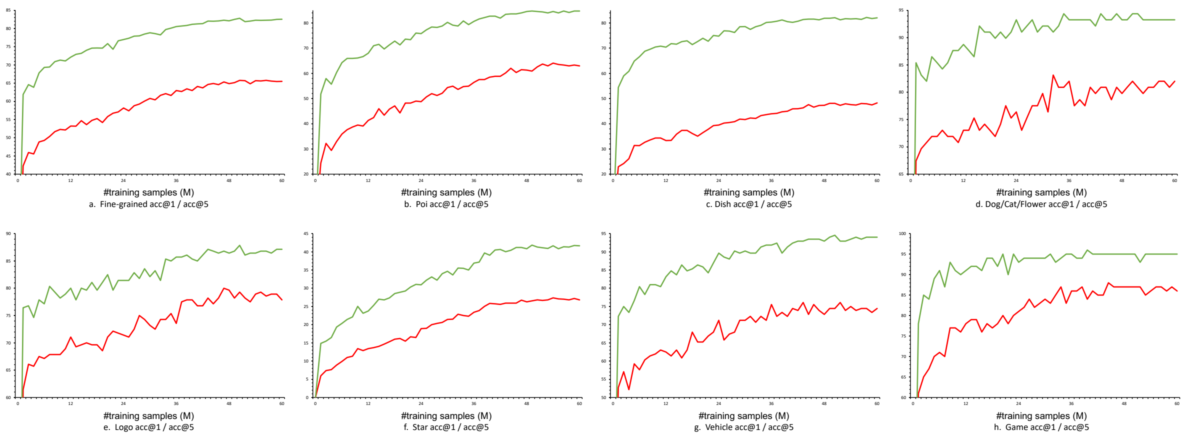

Figure 10: Visualization of the Convergence of the Pre-training Stage