SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models

Dongyang Liu [*] [1] [2] Renrui Zhang [*] [1] [2] Longtian Qiu [*] [2] Siyuan Huang [*] [2] Weifeng Lin [*] [2] Shitian Zhao [2]
Shijie Geng [3] Ziyi Lin [1] [2] Peng Jin [2] Kaipeng Zhang [2] Wenqi Shao [2] Chao Xu [2] Conghui He [2]
Junjun He [2] Hao Shao [1] Pan Lu [4] Yu Qiao [†] [2] Hongsheng Li [†] [1] [5] Peng Gao [†] [‡] [*] [2]

Abstract

We propose SPHINX-X, an extensive Multi-modality Large Language Model (MLLM) series developed upon SPHINX. To improve the architecture and training efficiency, we modify the SPHINX framework by removing redundant visual encoders, bypassing fully-padded sub-images with skip tokens, and simplifying multi-stage training into a one-stage all-in-one paradigm. To fully unleash the potential of MLLMs, we assemble a comprehensive multi-domain and multi-modal dataset covering publicly available resources in language, vision, and vision-language tasks. We further enrich this collection with our curated OCR intensive and Set-of-Mark datasets, extending the diversity and generality. By training over different base LLMs including TinyLlama-1.1B, InternLM2-7B, LLaMA2-13B, and Mixtral-8×7B, we obtain a spectrum of MLLMs that vary in parameter size and multilingual capabilities. Comprehensive benchmarking reveals a strong correlation between the multi-modal performance with the data and parameter scales. Code and models are released at `https://github.com/Alpha-VLLM/LLaMA2-Accessory`.

1. Introduction

Since the release of OpenAI's GPT-4 (V) (OpenAI, 2023) and Google's Gemini (Gemini Team, 2023), Multi-modal Large Language Models (MLLMs) have become an increasingly popular research area (Fu et al., 2023c; Yang et al., 2023e). By aligning multi-modal encoders with Large Language Models (LLMs), MLLMs demonstrate the potential to unlock myriad novel applications and further push the boundary of next-level artificial general intelligence, spanning from embodied intelligence (Geng et al., 2023), autonomous driving (Wen et al., 2023; Cao et al., 2023; Yang et al., 2023d) to graphical user interfaces (GUI) agents (He et al., 2024; Yang et al., 2023f).

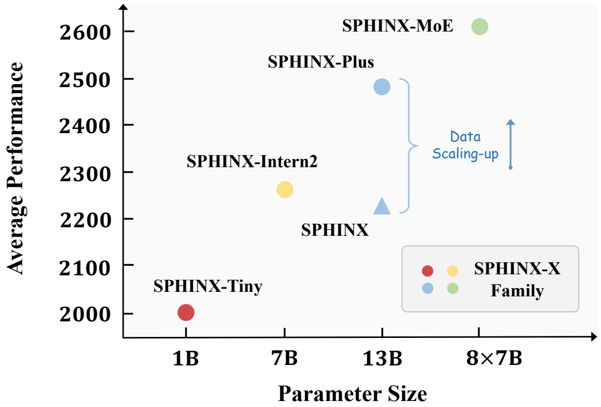


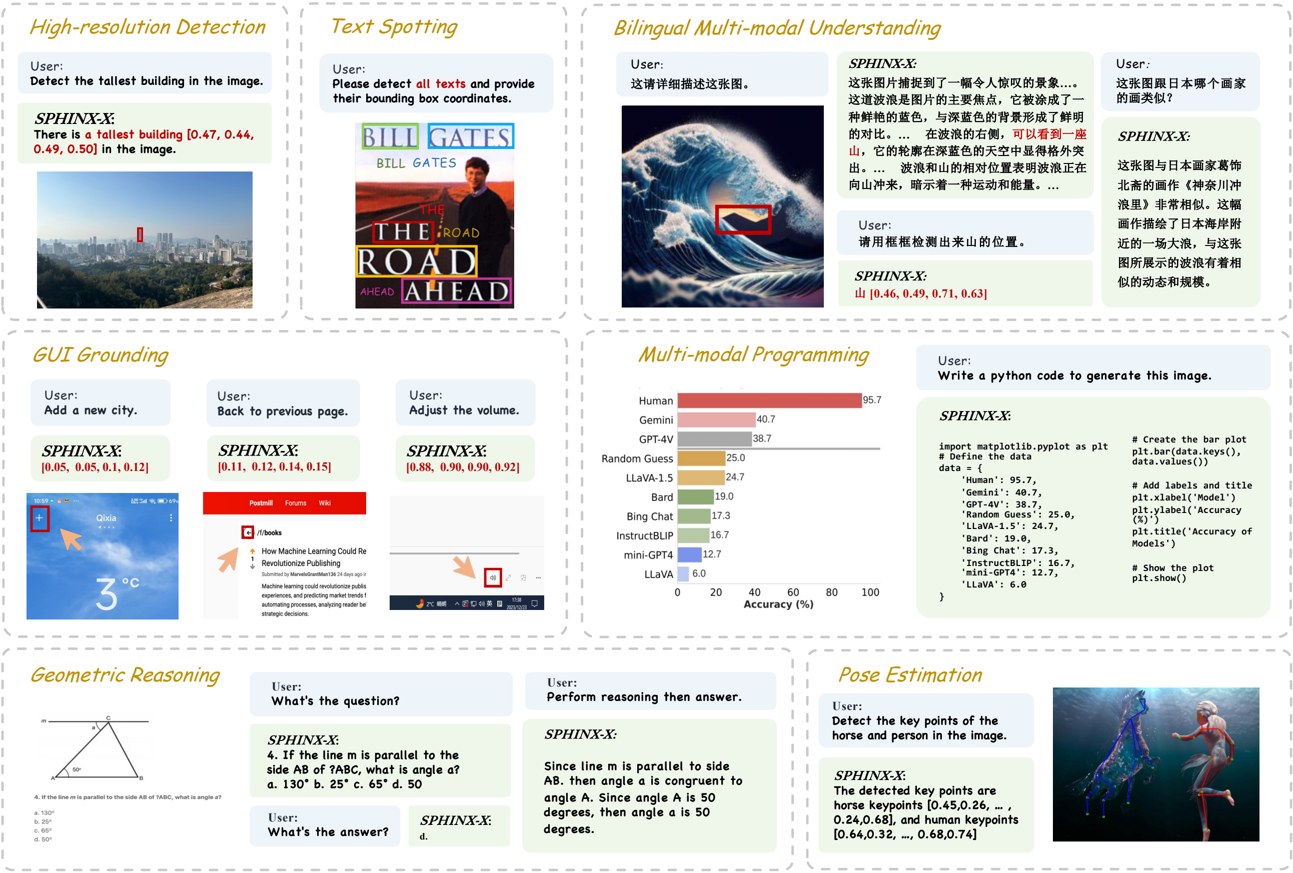
Figure 1. **Performance comparison with data and parameter scaling.** We introduce SPHINX-X, a general and well-performing MLLM family developed upon SPHINX (Lin et al., 2023).

Inspired by this, a wide array of open-source MLLMs have been developed within merely one year, including BLIP series (Li et al., 2023b; Dai et al., 2023), LLaMA-Adapter (Zhang et al., 2024c; Gao et al., 2023), LLaVA (Liu et al., 2023b;a; Li et al., 2024), MiniGPT-4 (Zhu et al., 2023a), mPLUG-Owl (Ye et al., 2023b;c), and SPHINX (Lin et al., 2023). Although these open-source MLLMs demonstrate promising multi-modal capabilities, their performance is still constrained by the training data from few task domains and limited choices of LLM parameters:

Limited Data Coverage for Tasks. Popular open-source MLLMs, such as BLIP-2, LLaVA, and LLaMA-Adapter, are typically trained on raw vision-language data from the natural image domain (e.g., LAION (Schuhmann et al., 2021; 2022), SBU (Ordonez et al., 2011), and Conceptual Cap-

[*] Equal Contribution [†] Corresponding Authors [‡] Project Lead [1] MMLab, CUHK [2] Shanghai AI Laboratory [3] Rutgers University [4] University of California, Los Angeles [5] Centre for Perceptual and Interactive Intelligence (CPII). Correspondence to: Peng Gao <gaopeng@pjlab.org.cn>, Hongsheng Li <hsli@ee.cuhk.edu.hk>, Yu Qiao <qiaoyu@pjlab.org.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



User:
Detect the tallest building in the image.

SPHINX-X:
There is a tallest building [0.47, 0.44, 0.49, 0.50] in the image.

User:
Please detect all texts and provide their bounding box coordinates.

User:
这请详细描述这张图。

SPHINX-X:
这张图片捕捉到了一幅令人惊叹的景象….。这道波浪是图片的主要焦点，它被涂成了一种鲜艳的蓝色，与深蓝色的背景形成了鲜明的对比。… 在波浪的右侧，可以看到一座山，它的轮廓在深蓝色的天空中显得格外突出。… 波浪和山的相对位置表明波浪正在向山冲来，暗示着一种运动和能量。…

User:
请用框框检测出来山的位置。

SPHINX-X:
山 [0.46, 0.49, 0.71, 0.63]

User:
这张图跟日本哪个画家的画类似？

SPHINX-X:
这张图与日本画家葛饰北斋的画作《神奈川冲浪里》非常相似。这幅画作描绘了日本海岸附近的一场大浪，与这张图所展示的波浪有着相似的动态和规模。

User:
Add a new city.

SPHINX-X:
[0.05, 0.05, 0.1, 0.12]

User:
Back to previous page.

SPHINX-X:
[0.11, 0.12, 0.14, 0.15]

User:
Adjust the volume.

SPHINX-X:
[0.88, 0.90, 0.90, 0.92]

Submitted by MarvelsGrantMan136 24 days ago

Machine learning could revolutionize publis experiences, and predicting market trends f automating processes, analyzing reader bel strategic decisions.

User:
Write a python code to generate this image.

SPHINX-X:

```
import matplotlib.pyplot as plt
# Define the data
data = {
    'Human': 95.7,
    'Gemini': 40.7,
    'GPT-4V': 38.7,
    'Random Guess': 25.0,
    'LLaVA-1.5': 24.7,
    'Bard': 19.0,
    'Bing Chat': 17.3,
    'InstructBLIP': 16.7,
    'mini-GPT4': 12.7,
    'LLaVA': 6.0
}

# Create the bar plot
plt.bar(data.keys(),
data.values())

# Add labels and title
plt.xlabel('Model')
plt.ylabel('Accuracy
(%)')
plt.title('Accuracy of
Models')

# Show the plot
plt.show()
```

4. If the line *m* is parallel to the side AB of ?ABC, what is angle *a*?

a. 130°
b. 25°
c. 65°
d. 50°

User:
What's the question?

SPHINX-X:
4. If the line m is parallel to the side AB of ?ABC, what is angle a?
a. 130° b. 25° c. 65° d. 50

User:
What's the answer?

SPHINX-X:
d.

User:
Perform reasoning then answer.

SPHINX-X:

Since line m is parallel to side AB. then angle a is congruent to angle A. Since angle A is 50 degrees, then angle a is 50 degrees.

User:
Detect the key points of the horse and person in the image.

SPHINX-X:
The detected key points are horse keypoints [0.45,0.26, … , 0.24,0.68], and human keypoints [0.64,0.32, …, 0.68,0.74]

Figure 2. **Demonstrations of SPHINX-X.** With our proposed data and training strategies, SPHINX-X can achieve superior multi-modal understanding and reasoning capabilities in diverse domains, e.g., bilingual, serving as a multi-purpose vision generalist.

tions (Sharma et al., 2018)), and visual instruction-following data (Chen et al., 2023b; Wang et al., 2023a) generated by GPT-4 (V) (OpenAI, 2023). As a result, they normally exhibit a superior multi-modal understanding performance in natural images. However, they display limited or degraded results in out-of-domain scenarios, such as Optical Character Recognition (OCR), table, chart, and mathematics fields, where in-depth domain-specific knowledge is critical. In contrast, domain-specific MLLMs like Shikra (Chen et al., 2023a), mPLUG-DocOwl (Ye et al., 2023a), and Kosmos-2.5 (Lv et al., 2023) are tailored to excel in specific tasks, but at the expense of their general multi-modal capabilities.

Limited Choices of Model Parameters. Most open-source MLLMs are developed on top of dense LLMs, e.g., LLaMA (Touvron et al., 2023a;b), with 7B or 13B parameters. While such parameter counts are often prohibitively large for deployment on portable devices, the same number of parameters remains inadequate to fully explore the performance boundaries of MLLMs. Therefore, scaling down the model scale of MLLMs could facilitate the broader adoption of mobile devices. Meanwhile, scaling up the parameter count through the integration of sparsely-activated Mixture-of-Experts (MoE) architecture (Shazeer et al., 2017) could also unlock the full potential of MLLMs in addressing complex real-world multi-modal challenges.

To resolve the aforementioned limitations of existing MLLMs, we introduce a family of MLLMs termed **SPHINX-X** by extending the data coverage of tasks and parameter scales in SPHINX, as shown in Figure 1. The superior multi-modal generalization capacity of SPHINX-X for a diversity of tasks is exhibited in Figure 2. Importantly, we adjust the training process and model architecture of SPHINX to better accommodate the efficient and large-scale multi-modal training:

① ***Modifications over SPHINX.*** For the mixed four vision encoders in SPHINX, we only preserve two of them, i.e., CLIP-ConvNeXt (Liu et al., 2022) and DINOv2 (Oquab et al., 2023). Considering their distinct methodologies and architectures, the two encoders can provide the most complementary visual semantics, denoted as **M**ixture **o**f **V**isual experts (**MoV**). Then, for the sub-image division strategy of high-resolution images, if the input image has a large aspect ratio, we observe a frequent occurrence of fully-padded sub-images, where all pixels are zeros. To address this, we adopt a learnable skip token to represent them within LLMs, thereby shortening the sequence length for efficiency, while still preserving the relative positions of sub-images. Furthermore, given the increased training data volume, we condense the previous multi-stage training pipeline into a more straightforward single-stage paradigm. Instead of fine-tuning different parts of LLM parameters in two stages with different datasets, we directly train all the parameters of LLMs on all our collected datasets.

② ***Multi-Domain and Multi-Modal Datasets.*** To fully unleash the potential of MLLMs, we assemble an extensive collection of public datasets that span a wide array of tasks, and carefully extend two self-curated multi-modal datasets. In detail, we collect the public datasets from the realms of vision, language, and vision-language tasks, and reformulate them into a unified multi-turn conversational format. Moreover, to specifically enhance the targeted capacity of MLLMs, we further construct an OCR-intensive dataset and a Set-of-Mark (SoM) dataset. The expansion of OCR data processed from substantial PDFs can unlock the visual language understanding power of MLLMs, e.g., text spotting and document layout detection. The specialized SoM data also compensates for the SoM prompting (Yang et al., 2023c) potentials of SPHINX-X, for which we construct delicate SoM annotations in diverse domains by GPT-4.

③ ***LLM Parameter Scaling of SPHINX-X.*** With the aforementioned techniques and large-scale datasets, we marry SPHINX-X with various base LLMs of increasing parameter scales: TinyLlama-1.1B (Zhang et al., 2024b), InternLM2-7B (Team, 2023), LLaMA2-13B (Touvron et al., 2023b), and Mixtral-8×7B (Jiang et al., 2024a). Respectively, we develop a family of MLLMs that facilitate fast mobile deployment (SPHINX-Tiny), provide bilingual support (SPHINX-Intern2), possess moderate parameters with data scaling (SPHINX-Plus), and exhibit strong reasoning capabilities through Mixture-of-Expert architectures (SPHINX-MoE).

Extensive evaluations across a wide range of benchmarks reveal that SPHINX-Plus surpasses the original SPHINX, confirming that enriching dataset scales and diversity can benefit the performance. Furthermore, a comparison of base LLMs from 1.1B to 7×8B demonstrates that under the same training pipeline, scaling up the parameters can consistently boost the multi-modal understanding capabilities. Overall, we summarize the key contributions as follows:

- We release a family of well-performing MLLMs tailored from fast inference on mobile devices to complex reasoning tasks on high-end computers. A comprehensive range of experiments demonstrates that the scale of training data and the size of LLM parameters both play a critical role in the performance of MLLMs.
- We perform several modifications over SPHINX by eliminating redundant visual encoders, avoiding fully-padded sub-images with learnable skip tokens, as well as streamlining the complex multi-stage training pipeline into a single-stage all-in-one paradigm.
- We collected an extensive multi-modal dataset covering a broad spectrum of tasks and modalities. On top of that, we curated two new datasets for enhancing the OCR-intensive and Set-of-Marks prompting capabilities of MLLMs.

2. Related Work

Large Language Models (LLMs) Advancements in recent MLLM research are based on the breakthrough of LLMs constructed upon the Transformer architecture (Vaswani et al., 2017), where progress has stemmed from both an expansion of training data and a significant increase in model parameters. For instance, GPT-3 (Brown et al., 2020), boasting 175B parameters, excels at few-shot in-context learning, while GPT-2 (Radford et al., 2019) with 1.5B parameters falls short of reaching this level of performance. Inspired by GPT-3's success, several LLMs like PaLM (Chowdhery et al., 2022), OPT (Zhang et al., 2022b), BLOOM (Workshop et al., 2022), and LLaMA have emerged. Mistral (Jiang et al., 2023) further introduced window attention for enhanced long-context modeling, while Mixtral 8×7B leveraged sparse MoE layers (Fedus et al., 2022; Lepikhin et al., 2020; Shazeer et al., 2017) to upscale parameters efficiently, outperforming with fewer active parameters. Concurrently, models such as Qwen (Bai et al., 2023), Baichuan (Yang et al., 2023a), and InternLM (Team, 2023) have advanced bilingual LLM capabilities, whereas TinyLlama (Zhang et al., 2024b) and Phi-2 (Microsoft, 2023) focus on reducing parameters for edge deployment. Our SPHINX family extends LLMs to multimodal domains for visual understanding and reasoning. We select four LLMs with different pre-training and parameter scales, comparing their performance under multi-modal scenarios.

Multi-modal Large Language Models (MLLMs) Continual attempts are made to connect non-text encoders to LLMs for perception beyond natural languages, forming MLLMs. Efforts to extend LLMs to perceive beyond text have birthed MLLMs, with vision as the primary modality. Representative architectures include BLIP (Li et al., 2022), BLIP-2 (Li et al., 2023b), and MiniGPT-4 (Zhu et al., 2023a), which employ query Transformers to summarize visual features and align them to LLMs; Flamingo (Alayrac et al., 2022), which uses gated cross-attention for mixing visual representations and supports interleaved image-text inputs; The LLaMA-Adapter series (Zhang et al., 2024c; Gao et al., 2023) which introduce zero-initialized attention to minimize interference between visual and language tokens; and LLaVA (Liu et al., 2023b;a), which connects visual tokens to LLMs with a simple linear layer and directly fine-tunes LLM parameters to incorporate visual knowledge. There are also recent advances in fine-grained MLLMs that have demonstrated remarkable capabilities in understanding detailed information. For example, Shikra (Chen et al., 2023a) and VisionLLM (Wang et al., 2023b) excel in referring object detection, while ChartAssistant (Meng et al., 2024), mPLUG-DocOwl/PaperOwl (Ye et al., 2023a; Hu et al., 2023) focus on specialized domains such as tables, documents, and scientific diagrams analysis. Many efforts also extend LLMs into more modalities, such as ImageBind-LLM (Han et al., 2023a), Point-LLM (Guo et al., 2023), and others (Zhu et al., 2023b; Zhang et al., 2022a; 2023b). In this paper, we upgrade SPHINX (Lin et al., 2023) to an MLLM family for more general visual instruction following, achieving superior performance over various benchmarks.

3. Method

We first revisit the design principles of SPHINX in Section 3.1. We then respectively detail the three improvements made to SPHINX-X in Section 3.2 concerning the succinctness of visual encoders, learnable skip tokens for useless visual signals, and simplified one-stage training. Lastly, we illustrate the composition of our large-scale multi-modality dataset in Section 3.3, as well as introduce different base LLMs adopted by the SPHINX-X family in Section 3.4.

3.1. A Revisit of SPHINX

SPHINX (Lin et al., 2023) proposes three types of mixing strategies to develop a multi-purpose MLLM – mixing of model weights, tuning tasks, and visual embeddings. Following previous efforts (Gao et al., 2023; Liu et al., 2023b), SPHINX adopts a two-stage training pipeline, in which the first stage aligns pre-trained vision encoders with LLaMA2 (Touvron et al., 2023b), and the second stage integrates a variety of tasks for instruction tuning. For more robust visual representations, SPHINX incorporates the embeddings of four different vision encoders, including CLIP-ViT (Radford et al., 2021; Dosovitskiy et al., 2020), CLIP-ConvNeXt (Liu et al., 2022), DINOv2 (Oquab et al., 2023), and Q-former (Li et al., 2023c). SPHINX then introduces a multi-scale mixing strategy to tackle high-resolution images, which divides the high-resolution input into several sub-images along with a downsampled image for concurrent encoding. In addition, to further mix various domain semantics, SPHINX fuses the first-stage weights of LLMs that are tuned by different data domains. Despite its superior performance, SPHINX is still constrained by the cumbersome two-stage training process and mixed architectures, and it has yet to fully capitalize on the potential benefits of data and model scaling. Motivated by this, we develop SPHINX-X, an extensive series of MLLMs to explore a more general and comprehensive multi-modal learning paradigm.

3.2. SPHINX-X

To better handle large-scale multi-task and multi-modal instruction-tuning, we perform the following improvements over SPHINX-X, enabling the training pipeline and model architecture to be concise. We present the upgraded SPHINX-X training pipeline in Figure 3.

Eliminating Redundant Visual Encoders. SPHINX employs four complementary vision encoders to capture diverse visual representations. Although the mixture of visual experts can improve the performance, it inevitably leads to a significant increase in computational costs, especially for a group of sub-images generated from a high-resolution input. To obtain better computational efficiency, we eliminate the CLIP-ViT and Q-former encoders, only preserving two visual encoders – DINOv2 and CLIP-ConvNeXt. As these two models are pre-trained by distinct learning approaches (self-supervised vs. weakly-supervised) and network architectures (ViT vs. CNN), they can already provide the most complementary and refined visual knowledge. We denote them as the **M**ixture **of** **V**isual experts (**MoV**).

Bypassing Fully-padded Sub-images with Skip Tokens. The superior performance of SPHINX can be attributed to its effective handling of high-resolution images with several local sub-images and one global downsampled image. During the training stages of SPHINX, all images by default are scaled and zero-padded to a high resolution 448×448, and then divided into four 224×224 sub-images. However, for images with large aspect ratios, say $2:1$, this operation will result in fully-padded sub-images filled entirely with zero-value pixels. Such fully-padded sub-images not only contain noisy and useless visual signals, but also produce spare visual tokens that waste computational resources within both MoV and LLM. To alleviate the issue, we propose a learnable skip token to replace the fully-padded sub-image, which provides explicit relative positional information for LLMs to identify the positions between useful sub-images. In this way, MoV can avoid encoding these zero-pixel sub-images, which allows for a reduction in the input sequence length for LLMs, achieving enhanced computational efficiency.

One-Stage All-in-One Training. The original training pipeline of SPHINX comprises two stages and utilizes a weight mixing strategy. However, it requires to manually assign various tunable parameters and dataset combinations to different training stages, which is a labor-intensive task. To simplify the overall paradigm, we design a single-stage all-in-one training pipeline, which treats all collected datasets equally and uniformly transforms them into multi-modal multi-turn dialog formats. During the one-stage training, we unfreeze all the parameters of SPHINX (i.e., LLM and intermediate projection layers) except for the two visual en-

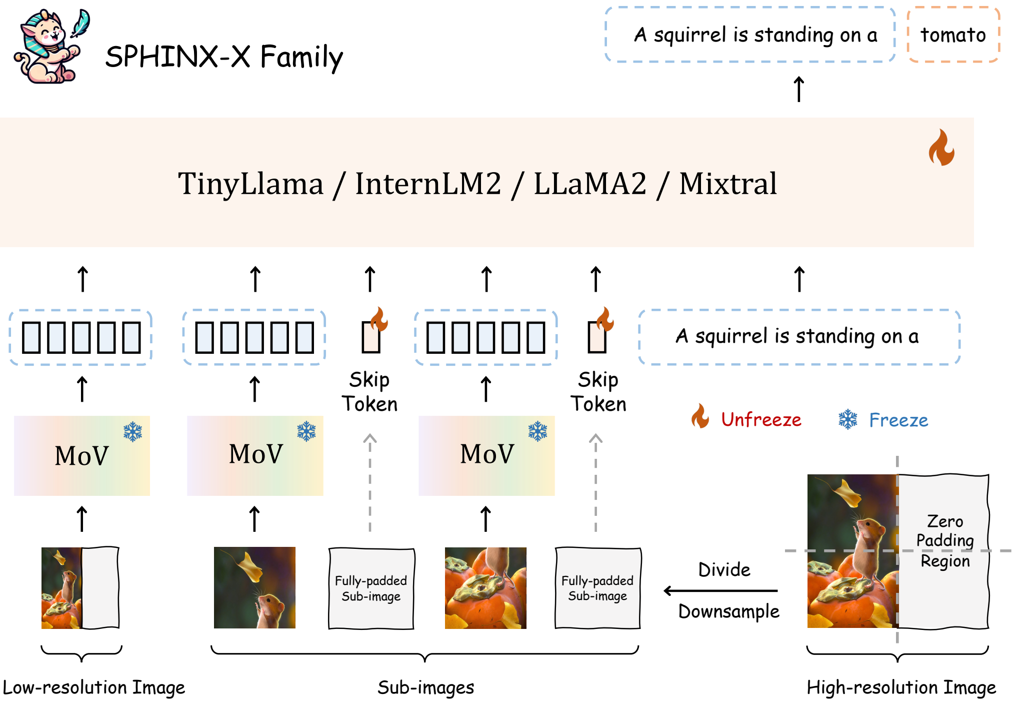


Figure 3. **Overall paradigm of SPHINX-X family.** On top of SPHINX (Lin et al., 2023), we adopt three modifications for a more general and concise architecture: removing redundant visual encoders in Mixture of Visual Experts (MoV), bypassing fully-padded sub-images with skip tokens, and simplifying multi-stage training into a one-stage all-in-one approach.

coders in MoV. Due to the large volume of training data and high reasoning capacity of LLMs, our one-stage all-in-one strategy can significantly streamline the training procedure for MLLMs while maintaining high performance.

3.3. Training Data of SPHINX-X

To obtain remarkable multi-modal capabilities, we widely convert three categories of public training data into instruction-following formats (language, vision, and vision-language), and carefully curate two targeted multi-modal datasets (OCR-intensive and Set-of-Mark) for SPHINX. All data is combined for the one-stage all-in-one training. Specifically, for natural language data, we utilized datasets that include multi-turn dialog, mathematical reasoning, and code generation. For vision data, we convert data from diverse computer vision tasks including image-level and object-level understanding into multi-turn conversation formats. For vision-language data, we collect various visual question-answering, visual instruct-tuning, and fine-grained image captioning datasets. On top of this, we generate an OCR dataset from large-scale PDF data, and a multi-domain Set-of-Marks dataset with fine-grained multi-modal knowledge.

For the three parts that are mainly composed of existing datasets (namely *language instruction-following, visual instruction-following and vision-language instruction-following*), we defer the details to the appendix (Sec. A.3). The data statistic information is also provided in the appendix (Table 9).

OCR-intensive Data. Most previous MLLMs can only leverage external tools and pre-extracted OCR tokens to obtain satisfactory OCR-related understanding. To enhance such capabilities for MLLMs, we compile an OCR-intensive dataset from extensive Internet PDF data. Different from previous synthetic OCR data (Yim et al., 2021; Kim et al., 2021) that are too simple and far from real-world application, our dataset is more challenging and larger-scale. Specifically, we first collect large-scale PDF datasets from Common Crawl [1] and arXiv websites. Then, we utilize PyMuPDF [2] to get the rendering results of each page in the PDF file and also save all the text annotations along with their bounding boxes. To ensure the OCR quality, we adopt multiple processing methods, including Unicode characteristic checking, text splits merge, etc. In this way, we constructed an in-house PaperText dataset with about 3M text-dense pages. Finally, we transform them into a unified question-answering format to strengthen the OCR documentation understanding ability.

[1] Common Crawl:https://commoncrawl.org/

[2] PyMuPDF: https://github.com/pymupdf/PyMuPDF

Multi-Domain Set-of-Mark Data. We notice that existing multi-modal datasets lack the fine-grained correspondence between images and texts. Thus, we construct a multi-domain dataset similar to Set-of-Marks techniques (Yang et al., 2023c) to endow MLLMs with dense multi-modal captioning knowledge. Initially, we collect diverse image datasets from various domains. Then, we utilize dataset annotations such as bounding boxes and object masks to place various marks like points, boxes, polygons, and identifiers, on the raw images. After that, we craft domain-specific instructions for each data type, and prompt GPT-4V with the masked images for multi-scale captioning, which generates captions of global image understanding, detailed region captioning, and object-relation analysis. Such SoM prompting for GPT-4V can motivate its power to produce higher-quality and fine-grained multi-modal data. During training, we do not utilize the marked images, but the raw images, and describe the marks by language within the multi-turn conversations for uniformity with other data domains.

3.4. SPHINX-X with Different LLMs

Built upon the aforementioned techniques and large-scale datasets, we provide four choices of base LLMs in SPHINX-X with increasing parameter scales: TinyLlama-1.1B (Zhang et al., 2024a), InternLM2-7B (Team, 2023), LLaMA2-13B (Touvron et al., 2023b), and Mixtral-8×7B (Jiang et al., 2024a). We introduce their features compared to the original SPHINX with LLaMA2-13B.

SPHINX-Tiny with TinyLlama-1.1B. TinyLlama can be regarded as a lightweight version of LLaMA. The compactness of 1.1B parameters allows TinyLlama to apply to a diversity of scenarios with limited computation resources. Therefore, we train SPHINX-Tiny to observe how the multi-modal performance varies given the smaller-scale LLM.

SPHINX-MoE with Mixtral-8×7B. As a sparse Mixture-of-Experts (MoE) LLM, Mixtral-8×7B utilizes 8 feed-forward networks at each transformer layer as experts, and relies on a router network to activate two experts each time. With this sparse mechanism, we expect to analyze the characteristics of different experts for multi-modal instruction following.

SPHINX-Plus with LLaMA2-13B. SPHINX-Plus utilizes the same scaled LLaMA, i.e., 13B parameters, with the original SPHINX, but is tuned by our constructed multi-modal dataset (more diverse and larger-scale) with one stage. This to some extent can illustrate the efficacy of data scaling-up. Note that, referring to SPHINX-2K (Lin et al., 2023), we also perform an improved version, termed **SPHINX-Plus-2K**, which increases the image resolution from 448×448 to 672×672 and splits the input to 3×3 sub-images for fine-grained visual understanding.

SPHINX-Intern2 with InternLM2-7B. InternLM (Team, 2023) is a strong bilingual LLM pre-trained on large Chinese and English corpus. The recently released InternLM2-7B shows stronger bilingual language understanding ability. We adopt InternLM2-7B as the base LLM to explore the performance pattern under the regular 7B-parameter LLM setup and the potential of bilingual multi-modal reasoning.

4. Experiment

4.1. Experimental Settings

All SPHINX-X models presented in the paper follow the one-stage all-in-one training strategy, and all modules except the visual encoders are optimized. The learning rate is set to 5e-6 for SPHINX-MoE, and 2e-5 for others. During training, the learning rate first linearly warmups to the target value within the first 0.01 epoch, and then gradually decays to 0 following the cosine schedule. We use the AdamW optimizer with weight decay = 0 and betas = $(0.9, 0.95)$. To accommodate the large model volume, a combination of ZeRO2-style (Rajbhandari et al., 2020) data parallel and Megatron-style (Shoeybi et al., 2019) model parallel is used. The model parallel size is set to 8 for SPHINX-MoE, 2 for SPHINX-Plus, and 1 for others. The effective batch size is 256. Note that SPHINX-Plus is initialized from the SPHINX model, while other SPHINX-X models use the original visual encoders/LLMs and randomly initialized linear projection layers. In Figure 1, we report the cumulative scores of SPHINX-X models over benchmarks mentioned in Tables 1, 4, and 5. However, due to the excessively high overall score, we excluded MME (Fu et al., 2023b) perception to balance the results.

4.2. Performance Evaluation

In this section, we conduct a thorough assessment and present outcomes across various benchmarks, offering an extensive overview and evaluation of our SPHINX-X family.

MLLM Benchmarks. We evaluate SPHINX-X on recently introduced benchmarks, such as MME (Fu et al., 2023a), Seedbench (Li et al., 2023a), POPE (Li et al., 2023f), LLaVA-Bench (In-the-Wild) (Liu et al., 2023b), MM-Vet (Yu et al., 2023b), MathVista (Lu et al., 2023), MM-bench (Liu et al., 2023d), CCbench (Contributors, 2023), Tiny LVLM (Shao et al., 2023) and BenchLLM (Cai et al., 2023a), InfiMM-Eval (Han et al., 2023b), Qbench (Cai et al., 2023b) for multi-modal language models (MLLM) to provide a comprehensive assessment of its characteristics. The results, presented in Table 1, showcase SPHINX-X's state-of-the-art performance across various multi-modal tasks, including mathematical reasoning, complex scene understanding, low-level vision tasks, and visual quality assessment, as well as resilience when facing illusions.

Visual Question Answering. The evaluation on general

Table 1. **Performance comparison with state-of-the-art methods on popular MLLM benchmarks.**

Methods	POPE	MME^P	MME^C	MMB	SEED	LLaVA^W	MM-Vet	CCbench	MathVista	Tiny LVLM	BenchLMM	InfiMM-Eval	Qbench
BLIP-2	85.3	1293.8	-	-	46.4	38.1	22.4	-	-	284.7	-	-	-
InstructBLIP-7B	-	-	-	36.0	53.4	60.9	26.2	12.1	25.3	300.6	44.63	-	56.7
InstructBLIP-13B	78.9	1212.8	-	-	-	58.2	25.6	-	-	-	45.03	-	-
LLaMA-AdapterV2	-	1328.4	356.4	-	-	-	-	-	-	229.2	-	30.5	59.5
Qwen-VL-7B	-	-	-	38.2	56.3	-	-	5.5	-	-	-	-	59.4
Qwen-VL-7B-Chat	-	1487.6	360.7	60.6	58.2	-	-	**39.3**	-	316.8	-	37.4	-
LLaVA1.5-7B	85.9	1510.7	-	64.3	58.6	63.4	30.5	16.4	-	-	46.8	-	58.7
LLaVA1.5-13B	85.9	1531.3	295.4	67.7	61.6	70.7	35.4	26.5	-	307.2	55.5	32.62	62.1
SPHINX	**90.8**	**1560.2**	310.0	67.1	71.6	74.3	36.6	27.9	27.5	288.9	-	30.7	65.8
SPHINX-Tiny	82.2	1261.2	242.1	56.6	17.1	52.3	23.8	17.5	26.4	301.5	50.0	21.9	19.7
SPHINX-Intern2	86.9	1260.4	294.6	57.9	68.8	57.6	36.5	21.0	35.5	312.9	47.0	31.5	60.0
SPHINX-Plus	89.1	1457.7	283.6	71.0	**74.8**	71.7	**47.9**	25.6	36.8	282.1	**57.4**	**39.5**	**68.6**
SPHINX-MoE	89.6	1485.3	**367.1**	**71.3**	73.0	70.2	40.9	15.4	**42.7**	**335.3**	50.7	38.6	66.2

Table 2. **Performance on 7 academic VQA benchmarks.**

Method	OKVQA	VQAV2	VizWiz	GQA	SQA	IVQA
BLIP-2	45.9	-	19.6	41.0	-	40.6
InstructBLIP	-	-	33.4	49.5	-	44.8
LLaMA-AdapterV2	49.6	70.7	39.8	45.1	-	-
Shikra	47.2	77.4	-	-	-	-
Fuyu-8B	60.6	74.2	-	-	-	-
MiniGPT-v2	57.8	-	53.6	60.1	-	51.5
Qwen-VL-7B	58.6	79.5	35.2	59.3	67.1	-
Qwen-VL-7B-Chat	56.6	78.2	38.9	57.5	68.2	-
LLaVA1.5-7B	-	78.5	50.0	62.0	66.8	-
LLaVA1.5-13B	-	80.0	53.6	63.3	71.6	-
SPHINX	62.2	80.2	46.8	62.9	69.1	52.7
SPHINX-Tiny	53.6	74.7	49.2	58.0	21.5	40.7
SPHINX-Intern2	55.5	75.5	49.6	56.2	70.4	49.0
SPHINX-Plus	-	-	57.8	-	74.2	54.7
SPHINX-MoE	**62.7**	**81.1**	**61.9**	**63.8**	**74.5**	**57.3**

visual question answering (VQA) benchmarks such as VQAV2 (Agrawal et al., 2015), GQV (Hudson & Manning, 2019), OK-VQA (Marino et al., 2019), VizWiz (Gurari et al., 2018), ScienceQA (Lu et al., 2022), IconQA (Lu et al., 2021b) are presented in Table 2. SPHINX-X excels across diverse visual question-answering benchmarks, showcasing its state-of-the-art performance in general visual understanding, relational reasoning, scientific contexts, and symbolic visual reasoning. Moreover, we conduct experiments on text-oriented VQA benchmarks such as TextVQA (Singh et al., 2019), OCRVQA (Mishra et al., 2019), DocVQA (Mathew et al., 2021b), ChartQA (Masry et al., 2022), AI2D (Kembhavi et al., 2016), DeepForm (Svetlichnaya, 2020), InfoVQA (Mathew et al., 2021a), TabFact (Chen et al., 2019), VisualMRC (Tanaka et al., 2021). As shown in Table 4, SPHINX-X achieves competitive performance on text-related benchmarks with a limited portion of OCR data.

Visual grounding. To evaluate SPHINX-X's ability to precisely locate and comprehend referred objects or regions within images, we conduct experiments on Referring Expression Comprehension (REC) benchmarks, including RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Mao et al., 2015), and RefCOCOg (Mao et al., 2015). The results are presented in Table 5, SPHINX-X consistently outperforms the majority of state-of-the-art models, surpassing even specialist model G-DINO-L (Liu et al., 2023c) and other visual-language generalist models.

4.3. SPHINX-MoE on other MLLM Benchmarks

To investigate the ability of SPHINX-MoE more concretely and locate its ability level among many developed MLLMs, we evaluate SPHINX-MoE on some recently curated benchmarks, which are listed below:

- **MathVerse** (Zhang et al., 2024d): A mathematical benchmark in visual contexts to explore the multi-modal diagram interpretation and reasoning capabilities of MLLMs, which annotate math problems into different versions for fine-grained evaluation.
- **SciVerse** (Guo et al., 2024): A comprehensive scientific problem benchmark (physics, chemistry, and biology) in visual contexts to reveal the domain-specific knowledge comprehension proficiency of MLLMs.
- **MMVP** (Tong et al., 2024): A benchmark specially crafted to measure MLLMs' visual understanding capability.
- **HallusionBench** (Guan et al., 2023): A benchmark to agnostic MLLMs' language hallucination and visual illusion.
- **AesBench** (Huang et al., 2024): An expert benchmark aiming to comprehensively evaluate the aesthetic perception capacities of MLLMs.
- **MMMU** (Yue et al., 2023a) & **CMMMU** (Ge et al., 2024): An English and a Chinese benchmark, respectively, aiming to solve massive multi-discipline tasks, which need college-level subject knowledge and deliberate reasoning ability.
- **ScreenSpot** (Cheng et al., 2024): A benchmark across various GUI platforms and designed to assess MLLM's capability to localize elements based on the human's instructions.

The results for MathVerse and SciVerse are showcased in Table 3. Our SPHINX-MoE attains the best performance among open-source models, indicating superior math problem-solving and scientific understanding capabilities. The results of our model on other benchmarks are included in Table 6. As we can see SPHINX-MoE performs well on all benchmarks, so we can infer that (i) SPHINX-MoE

Table 3. **Evaluation results of SPHINX-MoE on MathVerse (Zhang et al., 2024d) and SciVerse (Guo et al., 2024).**

Model	MathVerse							SciVerse						
	All	Text Dominant	Text Lite	Text Only	Vision Intensive	Vision Dominant	Vision Only	All	Text Only	Knowledge Lite	Knowledge Rich	Knowledge Professional	Vision Dominant	Vision Only
Open Source MLLM														
LLaMA-Adapter V2	5.8	7.8	6.3	3.9	6.2	4.5	4.4	9.3	9.4	10.7	11.2	11.9	11.8	10.4
ImageBind-LLM	10.0	13.2	11.6	12.9	9.8	11.8	3.5	27.2	23.6	27.8	28.1	28.2	28.9	23.2
mPLUG-Owl2	10.3	11.6	11.4	13.8	11.1	9.4	8.0	-	-	-	-	-	-	-
MiniGPT-v2	10.9	13.2	12.7	15.3	11.1	11.3	6.4	30.0	28.5	31.4	29.6	30.5	31.6	26.9
LLaVA-NeXT	15.6	19.4	15.2	18.1	16.8	15.2	11.3	34.9	33.4	34.0	36.8	37.2	35.1	31.3
Closed Source MLLM														
Qwen-VL-Plus	11.8	15.7	11.1	14.5	9.0	13.0	10.0	-	-	-	-	-	-	-
Gemini-Pro	24.1	26.3	23.5	27.3	23.0	22.3	22.2	-	-	-	-	-	-	-
Qwen-VL-Max	25.9	30.7	26.1	28.9	24.1	24.1	21.4	-	-	-	-	-	-	-
GPT-4V	**41.0**	**54.7**	**41.4**	**48.7**	**34.9**	**34.4**	**31.6**	-	-	-	-	-	-	-
SPHINX-MoE	15.6	22.2	16.4	18.3	14.8	12.6	9.1	**37.3**	**41.1**	**38.9**	**38.8**	**41.3**	**36.3**	**31.4**

Table 4. **Performance on text-oriented VQA tasks.** '†' denotes to use ground-truth OCR tokens during inference and training.

Method	Text VQA	OCR VQA	Doc VQA	Chart QA	AI 2D	Deep Form	Info VQA	KLC	WTQ	Tab Fact	Visual MRC
Specialist models											
Donut	43.5	-	67.5	41.8	-	**61.6**	11.6	30.0	18.8	54.6	93.9
UReader	57.6	-	65.4	59.3	-	49.5	**42.2**	**32.8**	29.4	**67.6**	**221.7**
Generalist models											
BLIP-2	42.5†	40.6	-	-	-	-	-	-	-	-	-
InstructBLIP	50.7†	44.8	-	-	-	-	-	-	-	-	-
LLaMA-AdapterV2	37.4	-	-	-	-	-	-	-	-	-	-
Qwen-VL-7B	63.8	**75.7**	65.1	**65.7**	62.3	-	-	-	-	-	-
Qwen-VL-7B-Chat	61.5	70.5	62.6	62.6	62.6	-	-	-	-	-	-
LLaVA1.5-7B	58.2	-	-	-	-	-	-	-	-	-	-
LLaVA1.5-13B	61.3	-	-	-	-	-	-	-	-	-	-
SPHINX	58.8	70.0	35.8	22.5	38.1	0	24.0	0	13.8	52.9	95.3
SPHINX-Tiny	57.8	60.3	53.0	34.1	24.6	11.8	26.3	22.2	15.3	51.1	147.5
SPHINX-Intern2	58.1	53.0	56.3	39.7	**63.0**	6.5	31.6	10.5	21.1	51.4	149.3
SPHINX-Plus	65.7	70.1	61.2	53.4	46.0	9.2	34.7	23.9	27.1	52.8	171.0
SPHINX-Plus-2K	**70.6**	68.9	**71.6**	55.1	47.4	23.2	39.1	31.1	**31.1**	54.0	178.4
SPHINX-MoE	68.0	64.8	68.4	55.0	55.6	20.7	41.8	25.5	29.9	52.7	184.4

Table 5. **Performance (Top-1 Accuracy@0.5) on Referring Expression Comprehension (REC) tasks.**

Method	RefCOCO+			RefCOCO			RefCOCOg	
	val	test-A	test-B	val	test-A	test-B	val-u	test-u
Specialist models								
UNINEXT	85.24	89.63	79.79	92.64	94.33	91.46	88.73	89.37
G-DINO-L	82.75	88.95	75.92	90.56	93.19	88.24	86.13	87.02
Generalist models								
OFA-L	68.29	76.00	61.75	79.96	83.67	76.39	67.57	67.58
Shikra 13B	82.89	87.79	74.41	87.83	91.11	81.81	82.64	83.16
MiniGPT-v2-7B	79.97	85.12	74.45	88.69	91.65	85.33	84.44	84.66
MiniGPT-v2-7B -Chat	79.58	85.52	73.32	88.06	91.29	84.30	84.19	84.31
Qwen-VL-7B	83.12	88.25	77.21	89.36	92.26	85.34	85.58	85.48
Qwen-VL-7B -Chat	82.82	88.59	76.79	88.55	92.27	84.51	85.96	86.32
SPHINX	86.64	91.08	80.35	91.05	92.65	86.56	88.19	88.35
SPHINX-Tiny	71.34	78.49	63.71	82.89	86.89	77.91	78.50	78.86
SPHINX-Intern2	76.80	84.86	69.01	86.08	89.70	81.78	83.99	83.40
SPHINX-Plus	**87.59**	**92.08**	**82.96**	**92.44**	**94.22**	**90.06**	**90.11**	**90.56**
SPHINX-MoE	85.50	90.48	79.88	90.64	93.74	86.85	88.26	88.51

has a better visual understanding ability and less language hallucination than other competitors. (ii) SPHINX-MoE can deal with the web and mobile domain data well. It should be noted that on some tasks or metrics, SPHINX-MoE performs even better than GPT-4V, *e.g.*, MMVP and AesP, AesE in AesBench. However, it is hard for SPHINX-MoE to solve the multi-discipline tasks, *i.e.*, the MMMU and CMMMU benchmark. And we think this is due to the lack of multi-modal multi-disciplinary data during the training stage. Thus we would consider to involving more multi-disciplinary data in SPHINX-MoE's training.

4.4. Performance of SPHINX-Plus on Video Analysis

To further assess the visual comprehension capabilities of our method, we conduct additional experiments on challenging video tasks. Since SPHINX-Plus is an image-based MLLM and is not trained on any video data, we need to do additional processing on video inputs. To be specific, we evenly sampled videos and selected the middle frame as the representative frame for input into the model. We conduct extensive experiments on Video-Bench (Ning et al., 2023) which evaluates the performance of models across three distinct capability levels: (i) video-exclusive understanding, (ii) prior knowledge-based question-answering, and (iii) comprehension and decision-making.

As shown in Table 7, SPHINX-Plus, despite being an image-based model, significantly outperforms existing models (Jin et al., 2023; Su et al., 2023; Zhang et al., 2023a) specifically tailored for video tasks. Especially in the aspects of video-exclusive understanding and prior knowledge-based question-answering, SPHINX-Plus showcases outstanding performance, signifying its proficiency in visual perception and knowledge extraction capabilities. In challenging datasets such as MOT, SPHINX-Plus demonstrates slightly

Table 6. **Evaluation results of SPHINX-MoE on other MLLM benchmarks.**

Methods	MMVP	HallusionBench					AesBench				MMMU-val	CMMMU		ScreenSpot		
		qAcc	*fAcc*	Easy *aAcc*	Hard *aAcc*	*aAcc*	AesP	AesE	AesA	AesI		val	test	Mobile	Desktop	Web
Open Source MLLM																
LLaVA	6.0	-	-	-	-	-	62.43	64.68	45.96	1.125	-	-	-	-	-	-
MiniGPT-4	12.7	8.79	10.12	31.87	27.67	35.78	41.93	39.35	38.57	0.999	26.8	-	-	6.4	3.7	2.8
InstructBLIP	16.7	9.45	10.11	35.60	**45.12**	45.26	54.29	53.89	46.54	1.126	32.9	-	-	-	-	-
LLaVA-v1.5	24.7	10.55	24.86	49.67	29.77	46.94	66.32	68.32	45.46	1.157	36.4	-	-	-	-	-
CogAgent-Chat	-	-	-	-	-	-	-	-	-	-	-	24.6	23.6	46.6	46.5	**49.7**
Qwen-VL-7B-Chat	-	5.93	6.65	31.43	24.88	39.15	63.21	64.18	46.25	1.192	35.9	30.7	31.3	6.9	5.9	0
Closed Source MLLM																
Gemini-Pro	40.7	-	-	-	-	-	71.99	71.37	49.38	1.222	47.9	-	-	-	-	-
GPT-4V	38.7	**28.79**	**39.88**	**75.60**	37.67	**65.28**	72.08	70.16	**50.86**	**1.301**	**56.8**	**42.5**	**43.7**	-	-	-
SPHINX-MoE	**49.3**	16.48	23.12	55.16	37.91	52.08	**72.93**	**73.32**	49.93	1.267	31.1	29.3	29.6	**55.1**	**50.5**	37.3

Table 7. **Comparison with state-of-the-art methods on Video-Bench.** '*' denotes the QA-pairs are re-constructed or annotated by Video-Bench. 'V' denotes a video training version of the model used.

Methods	Avg.	Video-Exclusive Understanding							Prior Knowledge-based QA			Comprehension and Decision-Making		
		Activitynet-QA	MSVD-QA*	MSRVTT-QA*	TGIF-QA	YouCook2*	UCF-Cirme*	MOT*	TV-QA*	MV-QA*	NBA-QA*	License Exam*	Decision-Making*	SQA3D*
Video-based MLLM														
Video-LLaMA	31.8	39.9	41.2	34.1	31.3	28.9	27.6	16.7	24.8	32.4	26.2	30.6	49.1	31.2
mPLUG-Owl^V	32.7	41.5	42.5	36.3	31.7	27.1	22.8	**27.8**	24.0	30.2	25.1	33.3	51.0	32.0
VideoChat	34.6	44.6	42.2	37.4	33.7	27.7	22.4	**27.8**	26.2	34.1	28.6	38.9	55.4	31.4
Chat-UniVi	35.2	49.0	48.6	41.7	41.3	29.0	**28.3**	16.7	23.1	33.6	25.7	38.9	53.1	29.1
PandaGPT	36.7	45.0	50.4	44.6	29.7	33.0	33.0	16.7	27.9	37.1	31.1	41.7	56.0	30.8
Otter^V	37.1	44.3	55.0	47.0	34.3	32.7	22.4	16.7	27.7	37.1	34.3	**52.8**	48.7	29.7
Video-ChatGPT	38.3	46.6	57.5	46.3	35.6	34.8	24.1	**27.8**	28.8	36.5	22.5	41.7	**58.2**	37.2
Image-based MLLM														
SPHINX	39.0	50.1	56.7	45.4	42.8	37.0	25.2	5.6	29.8	33.3	30.9	50.0	52.8	47.7
SPHINX-Plus	**45.1**	**53.1**	**68.5**	**54.0**	**53.4**	**42.0**	27.6	11.1	**36.5**	**44.0**	**45.0**	47.2	55.6	**48.8**

lower performance compared to existing state-of-the-art methods (Maaz et al., 2023; Li et al., 2023d). We attribute this to the need to model timing relationships in videos. SPHINX-Plus has not been fine-tuned by any video data, so its performance marginally underperforms others.

4.5. Demonstrations of SPHINX-X

In Figure 2, the demonstrates of SPHINX-X indicate that our models can 1) conduct fine-grained object detection in high-resolution images by the proposed sub-image division strategy; 2) conduct text spotting with accurate content and positions; 3) engage in bilingual image-based conversations, and generate coherent, accurate, and detailed Chinese descriptions for synthetic images; 4) generate accurate code for visual programming based on the precise understanding of the given plot; 5) analyze App screenshots based on the functional description and output the corresponding bounding box; 6) accurately interpret geometric questions from images, thanks to the math and extensive OCR datasets included in our training corpus; and 7) estimate the correct pose with rigorous body key points.

In the appendix (Figure 7), we respectively show the Set-of-marks (SoM) prompting and OCR understanding capabilities of SPHINX-X. With our curated SoM dataset, SPHINX-X can well understand the marks given in the prompt, i.e., a cat and bear bottle, and analyze the appearance and relations of designated objects. By the training of OCR-intensive data, our model can conduct accurate document layout detection and character recognition.

5. Conclusion

In this paper, we introduce SPHINX-X, a series of MLLMs for multi-purpose multi-modal instruction tuning with LLM parameters ranging from 1B to 8×7B. On top of the original SPHINX, we propose three aspects of improvements, i.e., removing redundant visual encoders, bypassing fully-padded sub-images with skip tokens, and simplifying multi-stage training into a one-stage all-in-one paradigm. We also curate a large-scale multi-domain dataset for MLLM training, which contains a wide range of public datasets and our constructed targeted data. Extensive benchmarks and evaluations demonstrate the superior performance and generalization capacity of SPHINX-X. We hope our work may cast a light on future MLLM research.

Acknowledgements

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd. under the Innovation and Technology Commission (ITC)'s InnoHK, by General Research Fund of Hong Kong RGC Project 14204021, and by National Natural Science Foundation of China (Grant No. 62206272). Hongsheng Li is a PI of CPII under the InnoHK.

Impact Statement

The SPHINX-X Multi-modality Large Language Model series has the potential to impact society in several ways:

Enhanced Multimodal AI Applications: SPHINX-X could lead to the development of more sophisticated AI systems capable of understanding and interacting with both text and visual input. This can improve services like automated translations, image recognition, and assistive technologies for individuals with disabilities.

Inclusivity and Language Diversity: By training on a diverse, multi-domain, and multi-modal dataset, the model may offer broader language support, which can bridge communication gaps and foster inclusivity.

Ethical and Bias Considerations: The expansive dataset used for training must be carefully curated to avoid perpetuating biases, ensuring that the model's responses are fair and ethical.

Misuse Risks: The misuse of MLLMs for generating deceptive content is a risk, underscoring the need for robust governance and ethical usage frameworks.

The responsible deployment of SPHINX-X requires careful consideration of these potential impacts to maximize benefits and minimize negative consequences.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31, 2015.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in neural information processing systems*, pp. 1877–1901, 2020.

Cai, R., Song, Z., Guan, D., Chen, Z., Luo, X., Yi, C., and Kot, A. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. *ArXiv*, abs/2312.02896, 2023a.

Cai, R., Song, Z., Guan, D., Chen, Z., Luo, X., Yi, C., and Kot, A. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*, 2023b.

Cao, Y., Xu, X., Sun, C., Huang, X., and Shen, W. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023a.

Chen, L., Li, J., wen Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., and Lin, D. Sharegpt4v: Improving large multi-modal models with better captions. *ArXiv*, abs/2311.12793, 2023b. URL https://api.semanticscholar.org/CorpusID:265308687.

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., LI, S., Zhou, X., and Wang, W. Y. Tabfact: A large-scale dataset for table-based fact verification. *ArXiv*, abs/1909.02164, 2019.

Cheng, K., Sun, Q., Chu, Y., Xu, F., Li, Y., Zhang, J., and Wu, Z. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/, March 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Contributors, O. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., Liu, Z., Sun, M., and Zhou, B. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023a.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., and Ji, R. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023b.

Fu, C., Zhang, R., Lin, H., Wang, Z., Gao, T., Luo, Y., Huang, Y., Zhang, Z., Qiu, L., Ye, G., et al. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*, 2023c.

Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., and Qiao, Y. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Ge, Z., Xinrun, D., Bei, C., Yiming, L., Tongxu, L., Tianyu, Z., Kang, Z., Yuyang, C., Chunpu, X., Shuyue, G., Haoran, Z., Xingwei, Q., Junjie, W., Ruibin, Y., Yizhi, L., Zekun, W., Yudong, L., Yu-Hsuan, T., Fengji, Z., Chenghua, L., Wenhao, H., Wenhu, C., and Jie, F. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.20847*, 2024.

Gemini Team, G. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Geng, H., Wei, S., Deng, C., Shen, B., Wang, H., and Guibas, L. Sage: Bridging semantic and actionable parts for generalizable articulated-object manipulation under language instructions. *arXiv preprint arXiv:2312.01307*, 2023.

Ghosal, D., Chia, Y. K., Majumder, N., and Poria, S. Flacuna: Unleashing the problem solving power of vicuna using flan fine-tuning, 2023.

Guan, T., Liu, F., Li, X. W. R. X. Z., Wang, X. L. X., Yacoob, L. C. F. H. Y., and Zhou, D. M. T. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pp. arXiv–2310, 2023.

Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., 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.

Guo, Z., Zhang, R., Chen, H., Gao, J., Gao, P., Li, H., and Heng, P.-A. Sciverse. *arXiv preprint*, 2024. URL https://sciverse-cuhk.github.io/.

Gupta, A., Dollár, P., and Girshick, R. B. Lvis: A dataset for large vocabulary instance segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359, 2019.

Gurari, D., Li, Q., Stangl, A., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023a.

Han, X., You, Q., Liu, Y., Chen, W., Zheng, H., Mrini, K., Lin, X., Wang, Y., Zhai, B., Yuan, J., Wang, H., and Yang, H. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models. 2023b.

He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., and Yu, D. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024.

Hu, A., Shi, Y., Xu, H., Ye, J., Ye, Q., Yan, M., Li, C., Qian, Q., Zhang, J., and Huang, F. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. *arXiv preprint arXiv:2311.18248*, 2023.

Huang, Y., Yuan, Q., Sheng, X., Yang, Z., Wu, H., Chen, P., Yang, Y., Li, L., and Lin, W. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024a.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de Las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts. *Arxiv 2401.04088*, 2024b.

Jin, P., Takanobu, R., Zhang, C., Cao, X., and Yuan, L. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2016.

Kafle, K., Cohen, S. D., Price, B. L., and Kanan, C. Dvqa: Understanding data visualizations via question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, 2018. URL https://api.semanticscholar.org/CorpusID:4445015.

Kazemzadeh, S., Ordonez, V., andre Matten, M., and Berg, T. L. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. URL https://api.semanticscholar.org/CorpusID:2682274.

Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7:15, 2021.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Kuznetsova, A., Rom, H., Alldrin, N. G., Uijlings, J. R. R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. The open images dataset v4. *International Journal of Computer Vision*, 128:1956 – 1981, 2018.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023a.

Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., and Li, C. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023c. URL https://api.semanticscholar.org/CorpusID:256390509.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023d.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023e.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023f.

Lian, W., Goodson, B., Pentland, E., Cook, A., Vong, C., and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/Open-Orca/OpenOrca, 2023.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023b.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., yue Li, C., Yang, J., Su, H., Zhu, J.-J., and Zhang, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023c.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023d.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., and Zhu, S.-C. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2021a. URL https://api.semanticscholar.org/CorpusID:234337054.

Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., and Zhu, S.-C. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *ArXiv*, abs/2110.13214, 2021b.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513, 2022.

Lu, P., Bansal, H., Xia, T., Liu, J., yue Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023.

Luo, Z., Xu, C., Zhao, P., Sun, Q., Geng, X., Hu, W., Tao, C., Ma, J., Lin, Q., and Jiang, D. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.

Lv, T., Huang, Y., Chen, J., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., Dong, L., Luo, W., et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023.

Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Mao, J., Huang, J., Toshev, A., Camburu, O.-M., Yuille, A. L., and Murphy, K. P. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2015.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199, 2019.

Masry, A., Long, D., Tan, J. Q., Joty, S., and Hoque, E. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Mathew, M., Bagal, V., Tito, R. P., Karatzas, D., Valveny, E., and Jawahar, C. Infographicvqa. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2021a.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021b.

Meng, F., Shao, W., Lu, Q., Gao, P., Zhang, K., Qiao, Y., and Luo, P. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*, 2024.

Microsoft. Phi-2, 2023. URL https://huggingface.co/microsoft/phi-2.

Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. Ocr-vqa: Visual question answering by reading text in images. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019.

Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., and Yuan, L. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

OpenAI. GPT-4V(ision) system card, 2023. URL https://openai.com/research/gpt-4v-system-card.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ordonez, V., Kulkarni, G., and Berg, T. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Pasupat, P. and Liang, P. Compositional semantic parsing on semi-structured tables. In *Annual Meeting of the Association for Computational Linguistics*, 2015. URL https://api.semanticscholar.org/CorpusID:9027681.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. Objects365: A large-scale, high-quality dataset for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438, 2019.

Shao, W., Hu, Y., Gao, P., Lei, M., Zhang, K., Meng, F., Xu, P., Huang, S., Li, H., Qiao, Y., et al. Tiny lvlm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*, 2023.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318, 2019.

Stanislawek, T., Grali'nski, F., Wr'oblewska, A., Lipi'nski, D., Kaliska, A., Rosalska, P., Topolski, B., and Biecek, P. Kleister: Key information extraction datasets involving long documents with complex layouts. In *IEEE International Conference on Document Analysis and Recognition*, 2021.

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., and Cai, D. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Svetlichnaya, S. Deepform, 2020. URL https://wandb.ai/stacey/deepform_v1/reports/DeepForm-Understand-/Structured-Documents-at-Scale.

Tanaka, R., Nishida, K., and Yoshida, S. Visualmrc: Machine reading comprehension on document images. *ArXiv*, abs/2101.11272, 2021.

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

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Arxiv 2401.06209*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, 2017.

Wang, J., Meng, L., Weng, Z., He, B., Wu, Z., and Jiang, Y.-G. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *ArXiv*, abs/2311.07574, 2023a. URL https://api.semanticscholar.org/CorpusID:265150580.

Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023b.

Wen, L., Fu, D., Li, X., Cai, X., Ma, T., Cai, P., Dou, M., Shi, B., He, L., and Qiao, Y. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023a.

Yang, J., Zeng, A., Zhang, R., and Zhang, L. Unipose: Detecting any keypoints. *ArXiv*, abs/2310.08530, 2023b.

Yang, J., Zhang, H., Li, F., Zou, X., Li, C., and Gao, J. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023c.

Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Guo, Y., and Zhang, S. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023d.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., and Wang, L. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9 (1):1, 2023e.

Yang, Z., Liu, J., Han, Y., Chen, X., Huang, Z., Fu, B., and Yu, G. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023f.

Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023a.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Jiang, C., Li, C., Xu, Y., Chen, H., Tian, J., Qian, Q., Zhang, J., and Huang, F. mplug-owl: Modularization empowers large language models with multimodality, 2023b.

Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023c.

Yim, M., Kim, Y., Cho, H.-C., and Park, S. Synthtiger: Synthetic text image generator towards better text recognition models. In *International Conference on Document Analysis and Recognition*, pp. 109–124. Springer, 2021.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023a.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490, 2023b.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023a.

Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023b.

Zhang, H., Li, X., and Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Zhang, P., Zeng, G., Wang, T., and Lu, W. Tinyllama: An open-source small language model. *ArXiv*, abs/2401.02385, 2024a. URL https://api.semanticscholar.org/CorpusID:266755802.

Zhang, P., Zeng, G., Wang, T., and Lu, W. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024b.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., and Li, H. Pointclip: Point cloud understanding by clip. In *CVPR 2022*, 2022a.

Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Li, H., Qiao, Y., and Gao, P. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *CVPR 2023*, 2023b.

Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., and Qiao, Y. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=d4UiXAHN2W.

Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.-W., Gao, P., et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024d.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022b.

Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *ArXiv*, abs/2306.17107, 2023c. URL https://api.semanticscholar.org/CorpusID:259287523.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023a.

Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., and Gao, P. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. *ICCV 2023*, 2023b.

A. Appendix

A.1. Analysis of Routing Mechanisms in SPHINX-MoE

A.1.1. Inference with different numbers of activating experts

For SPHINX-MoE, the LLM backbone is based on Mixtral-8×7B (Jiang et al., 2024b), which is a mixture-of-experts-based large language model. Thus, during the inference time, only some of the experts will be activated when dealing with each token. In the training stage of SPHINX-MoE, only two of the eight experts will be activated, so we set the default number of activating experts to 2 when inference. To investigate how the activating experts' amount will affect the inference performance, we change it from one to eight, and the results are shown in Figure 5.

As we can see, on most datasets, *i.e.*, ScienceQA (Lu et al., 2022), TextQA (Singh et al., 2019), RefCOCO (Lin et al., 2014) and Mathvista (Lu et al., 2023), when activating two experts, keeping it the same with training setting, SPHINX-MoE performs the best. However, for MME (Fu et al., 2023b), when setting the number of activating experts to four, SPHINX-MoE works the best. Two activating experts actually make the second low-performance. This inconsistency with the training setting is interesting.

A.1.2. Experts' usage distribution on different domains and different modalities

In some previous works, each expert in the mixture-of-experts model is a specialist for a specific domain or modality, *e.g.*, VLMO (Bao et al., 2022). So we explore that, in SPHINX-MoE, how each expert in each layer deals with data from different domains and different modalities. So we pick the artwork, celebrity and OCR subtasks from the MME (Fu et al., 2023b) benchmark, and infer SPHINX-MoE on these subtasks with two activating experts, recording the expert's usage distribution of each layer, as shown in Figure 4. Subfigure (a), (b) and (c) show the results on vision modality, language modality and vision&language modalities separately. From the distribution record, we don't see an obvious pattern that experts are specialists for different domains or modalities. (i) For different domains, the experts' usage is similar for the three different domain data: artwork, celebrity and OCR. (ii) For different modalities, there are no specific experts that mainly deal with one specific modality. But there is an interesting scenario that the experts' usage distribution of the layers at both ends of the model is more flat than that of the the middle layers.

A.1.3. Prune some of the experts when inference

Different from the dense model, for the sparse model only part of the parameters are activated during the inference time. So if we prune some experts, the ability of the model could be partly saved. To investigate how the number of pruned experts will affect SPHINX-MoE's ability, we prune different numbers of experts of SPHINX-MoE's each layer, and the results are shown in Figure 6. In Figure 6, the x-axis means the number of retained experts after pruning. For each value on the x-axis, termed as "n", we randomly choose $8 - n$ experts in each layer to be pruned in SPHINX-MoE. For each n, we run it three times for average performance.

We find that some experts are "important", *i.e.*, there is a huge performance variance in the three runs. If we prune some specific experts, SPHINX-MoE will lose most ability, while in another run, the most ability of the model will be retained even if we prune the same number of experts in each layer. Thus, if we keep these "important" experts and prune other less "important" experts in SPHINX-MoE, most ability could be saved, as the upper tendency in Figure 6.

A.2. Video Analysis on MVBench

To further evaluate the video understanding capacity, we evaluate SPHINX-X on MVBench (Li et al., 2023e), which breaks down video understanding into 20 sub-aspects, allowing for a more detailed comparison of model performance at a finer granularity. As shown in Table 7, SPHINX-Plus, despite being an image-based model, significantly outperforms existing models (Jin et al., 2023; Su et al., 2023; Zhang et al., 2023a) specifically tailored for video tasks. Especially in the aspects of video-exclusive understanding and prior knowledge-based question-answering, SPHINX-Plus showcases outstanding performance, signifying its proficiency in visual perception and knowledge extraction capabilities. In challenging datasets such as MOT, SPHINX-Plus demonstrates slightly lower performance compared to existing state-of-the-art methods (Maaz et al., 2023; Li et al., 2023d). We attribute this to the need to model timing relationships in videos. SPHINX-Plus has not been fine-tuned by any video data, so its performance marginally underperforms others.

A.3. Additional details on the training dataset

Language Instruction-following Data. Unlike previous works (Zhu et al., 2023a; Liu et al., 2023b;a) that utilize instruction-tuned LLMs such as Vicuna (Chiang et al., 2023), SPHINX-X is directly trained on top of the basic pre-trained LLM, i.e., LLaMA2 (Touvron et al., 2023b). This is to investigate the training characteristics of multi-modal models from LLMs more clearly. Therefore, we are required to collect a high-quality dataset combination for language instruction-following. The dataset includes multi-turn dialog, question-answering, code generation, and math word problems. In detail, UltraChat (Ding et al., 2023) and OpenOrca (Lian et al., 2023) are utilized for basic multi-turn conversation abilities. MetaMath (Yu et al., 2023a) and MathInstruct (Yue et al., 2023b) are high-quality mathemat-

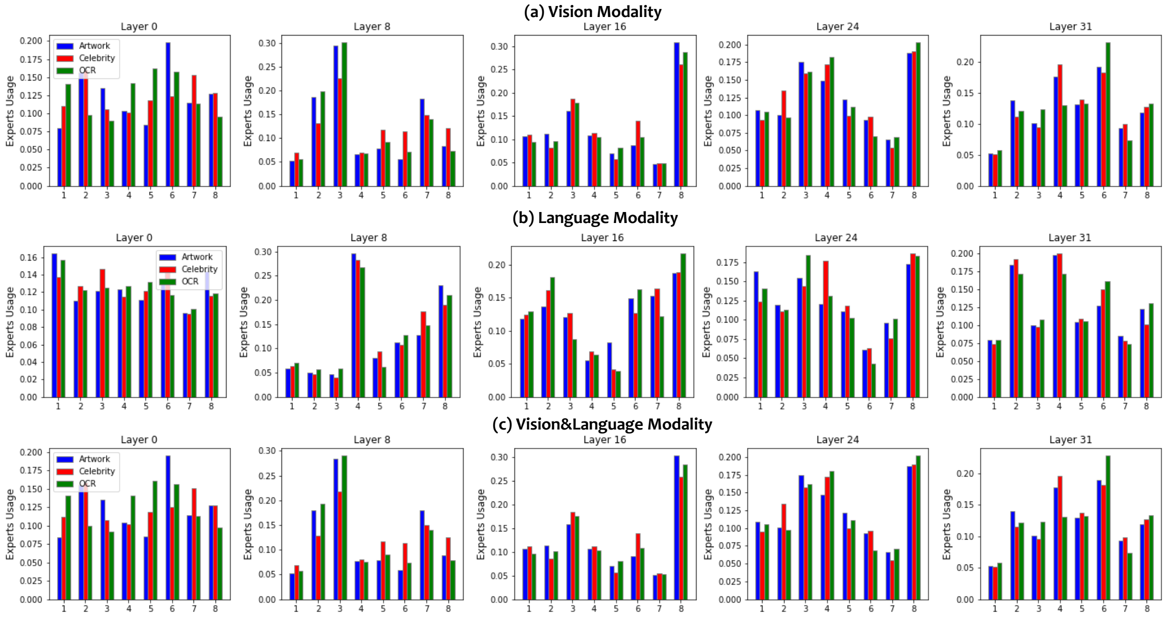


Figure 4. **Experts' usage distribution on different domains and different modalities.**

Table 8. **Comparison with state-of-the-art methods on MVBench.** '^{V}' denotes a video training version of the model used.

Methods	Avg.	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI
										Video-based MLLM											
Otter^{V}	26.8	23.0	23.0	27.5	27.0	29.5	53.0	28.0	33.0	24.5	23.5	27.5	26.0	28.5	38.5	22.0	18.0	22.0	23.5	19.0	19.5
mPLUG-Owl^{V}	29.7	22.0	28.0	34.0	29.0	29.0	40.5	27.0	31.5	27.0	23.0	29.0	31.5	27.0	44.0	24.0	40.0	31.0	26.0	20.5	29.5
Video-ChatGPT	32.7	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	48.5	29.0	39.5	33.0	29.5	26.0	35.5
Video-LLaMA	34.1	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	45.5	32.5	32.5	40.0	30.0	21.0	37.0
VideoChat	35.5	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	**27.0**	48.5	35.0	20.5	46.0	26.5	42.5	**41.0**	23.5	23.5	36.0
VideoChat2	**51.1**	**66.0**	**47.5**	**83.5**	**49.5**	**60.0**	**58.0**	**71.5**	**42.5**	23.0	23.0	**88.5**	**39.0**	**42.0**	44.0	**49.0**	**58.5**	36.5	**35.0**	**40.5**	**65.5**
										Image-based MLLM											
SPHINX	37.5	32.5	31.5	65.0	38.5	43.5	54.0	37.5	28.5	22.5	26.5	45.5	**39.0**	41.0	47.5	40.0	23.5	37.5	31.0	35.0	30.5
SPHINX-Plus	39.7	47.5	32.0	58.0	42.5	43.5	45.0	44.0	35.5	**29.0**	**27.0**	52.0	38.0	41.0	**59.5**	37.5	23.0	**41.0**	29.0	40.0	29.5

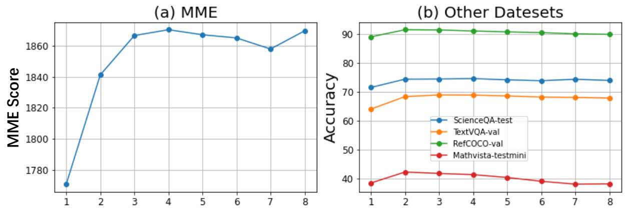


Figure 5. **Performance with different numbers of activating experts when inference.** We respectively report the performance on MME and other benchmarks.

ical datasets with reasoning process. WizardCoder (Luo et al., 2023) is adopted for increasing the coding ability of LLMs. Flan-mini (Ghosal et al., 2023) is a subset of FLAN datasets and is included for question-answering capabilities.

Visual Instruction-following Data. For comprehensive visual understanding, we expand the data scale of SPHINX to incorporate a variety of vision tasks and transform their annotations into a unified question-answering format. The tasks include image classification (Russakovsky et al., 2014), object detection such as COCO (Lin et al., 2014),Open-Images (Kuznetsova et al., 2018),Object365 (Shao et al., 2019),Lvis (Gupta et al., 2019), human pose estimation such as UniPose (Yang et al., 2023b), COCO-Pose (Lin et al., 2014), and visual grounding. We utilize a task-specific prompt as the question, and regard the ground-truth labels as the answer by textualizing them in language space. For generality, we do not utilize any special tokens for different tasks, and treat them all as pure language problems. This

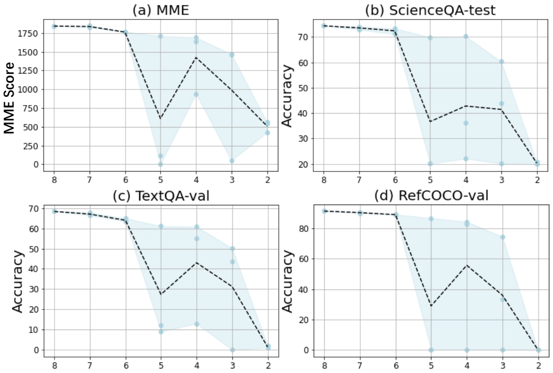


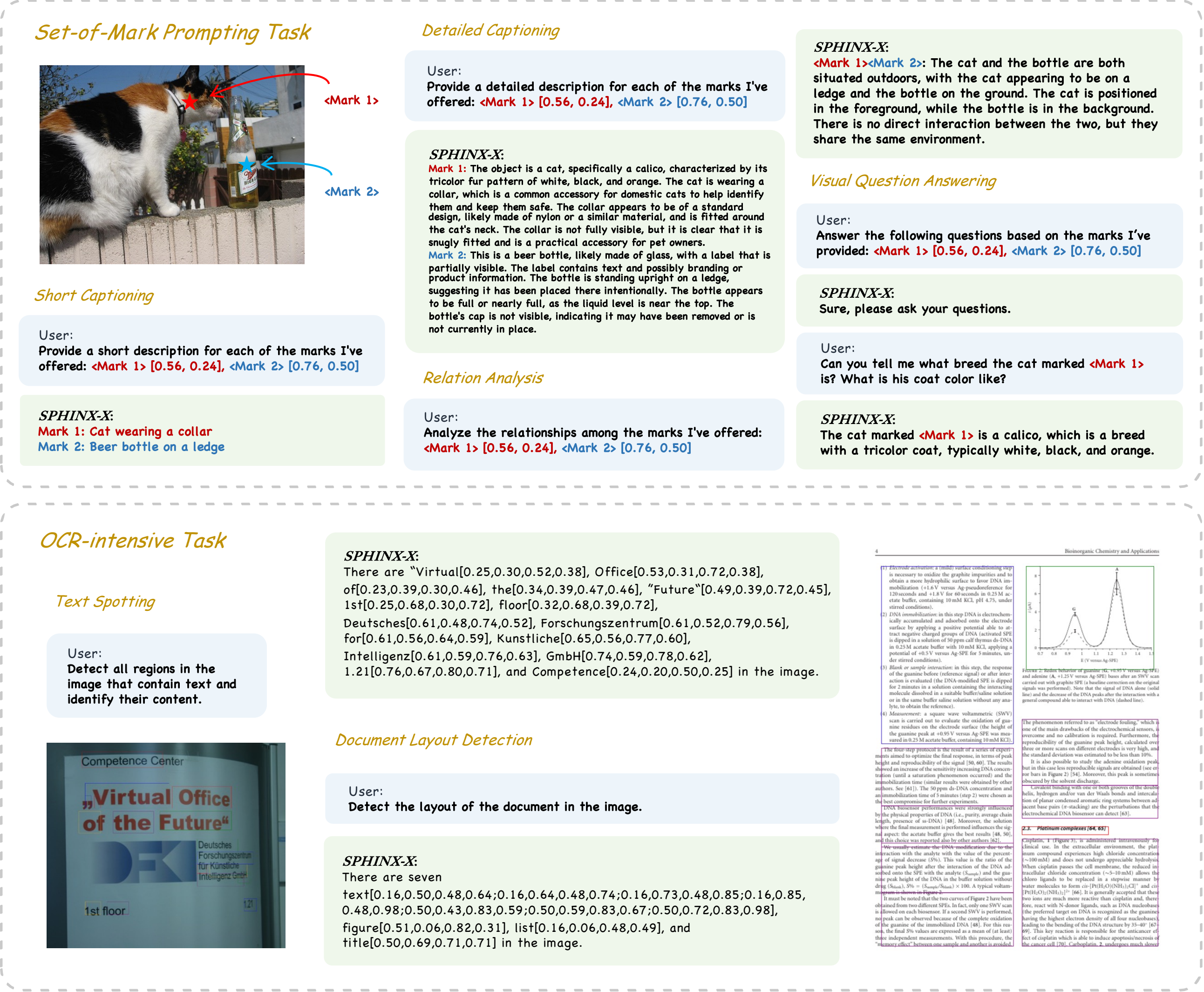
Figure 6. **Performance change when pruning different numbers of experts in each layer.** The black dotted line is the average of the three runs of random pruning.

visual supervised fine-tuning enhances SPHINX-X with the performance of image parsing, object localization, and relation reasoning, empowering MLLMs with in-built capacity to be a universal visual generalist.

Vision-language Instruction-following Data. To align MoV with LLMs and enable visual instruction following, we gather large-scale datasets from established visual question-answering sources such as VQAV2 (Agrawal et al., 2015), GQA (Hudson & Manning, 2019), OK-VQA (Marino et al., 2019), Visual Genome (Krishna et al., 2017), and CLEVR (Johnson et al., 2016). To specifically boost SPHINX-X's text-oriented VQA capabilities, we incorporate datasets including TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021b), ChartQA (Masry et al., 2022), AI2D (Kembhavi et al., 2016), Deepform (Svetlichnaya, 2020), DVQA (Kafle et al., 2018), InfographicsVQ (Mathew et al., 2021a), KleisterCharity (Stanislawek et al., 2021), TabFact (Chen et al., 2019), VisualMRC (Tanaka et al., 2021), and WikiTableQuestions (Pasupat & Liang, 2015). Leveraging the rich knowledge embedded in large foundation models, we also encompass high-quality MLLM-generated data, e.g., dense captioning data of ShareGPT4V (Chen et al., 2023b) and visual instruction data from LLaVA (Liu et al., 2023b), LVIS-INSTRUCT4V (Wang et al., 2023a), and LLaVAR (Zhang et al., 2023c). Additionally, we employ Geometry3K (Lu et al., 2021a) to enhance the model's geometry problem-solving abilities.

Table 9. **One-stage training data summary of SPHINX-X.**

Tasks	#Samples	Datasets
Language Instruction-following Data		
Multi-turn Dialog	1.8M	UltraChat,Flan-mini,OpenOrca
Math	0.6M	MetaMathQA,MathInstruct
Coding	80k	WizardCoder
Visual Instruction-following Data		
Detection	4.9M	V3Det,OpenImages,Lvis,COCO,Object365
Human Pose	0.3M	Unipose,COCO-Pose
Classification	1M	ImageNet1K
Grounding	1M	Visual Genome, RefCOCO, RefCOCO+,RefCOCOg,Flickr30k
Vision-language Instruction-following Data		
VQA	0.7M	VQAV2,OKVQA,GQA,Visual Genome CLEVR,ChartQA,DeepForm,DocVQA DVQA, InfographicsVQA,KleisterCharity VisualMRC,WikiTableQuestions TextVQA,TabFact
Caption	0.5M	MSCOCO,ShareGPT4V,LaionGPV4V
Visual Instruction	0.4M	LLaVA,LVIS-INSTRUCT4V,LLaVAR
OCR-intensive Data		
OCR	3M	PaperText: Arxiv, Common Crawl
Text Layout & Spotting	1.0 M	DocBank, M6Doc, Publaynet, DocLayNet, ICDAR, CTW1500
Set-of-Marks Instruction-following Data		
Natural Images	5k	COCO, LVIS, Visual Genome
Website/Mobile/Desktop agent	1k	SeeClick
OCR-related	2k	TotalText,CTW1500,IC13,IC15
Document Images	1k	M6Doc, DoclayNet, PublayNet
Multipanel Images	1k	In-house dataset


<Mark 1>
<Mark 2>
Competence Center
„Virtual Office of the Future“
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
1st floor
121

Set-of-Mark Prompting Task

Short Captioning

User:
Provide a short description for each of the marks I've offered: <Mark 1> [0.56, 0.24], <Mark 2> [0.76, 0.50]

SPHINX-X:
Mark 1: Cat wearing a collar
Mark 2: Beer bottle on a ledge

Detailed Captioning

User:
Provide a detailed description for each of the marks I've offered: <Mark 1> [0.56, 0.24], <Mark 2> [0.76, 0.50]

SPHINX-X:
Mark 1: The object is a cat, specifically a calico, characterized by its tricolor fur pattern of white, black, and orange. The cat is wearing a collar, which is a common accessory for domestic cats to help identify them and keep them safe. The collar appears to be of a standard design, likely made of nylon or a similar material, and is fitted around the cat's neck. The collar is not fully visible, but it is clear that it is snugly fitted and is a practical accessory for pet owners.
Mark 2: This is a beer bottle, likely made of glass, with a label that is partially visible. The label contains text and possibly branding or product information. The bottle is standing upright on a ledge, suggesting it has been placed there intentionally. The bottle appears to be full or nearly full, as the liquid level is near the top. The bottle's cap is not visible, indicating it may have been removed or is not currently in place.

Relation Analysis

User:
Analyze the relationships among the marks I've offered: <Mark 1> [0.56, 0.24], <Mark 2> [0.76, 0.50]

SPHINX-X:
<Mark 1><Mark 2>: The cat and the bottle are both situated outdoors, with the cat appearing to be on a ledge and the bottle on the ground. The cat is positioned in the foreground, while the bottle is in the background. There is no direct interaction between the two, but they share the same environment.

Visual Question Answering

User:
Answer the following questions based on the marks I've provided: <Mark 1> [0.56, 0.24], <Mark 2> [0.76, 0.50]

SPHINX-X:
Sure, please ask your questions.

User:
Can you tell me what breed the cat marked <Mark 1> is? What is his coat color like?

SPHINX-X:
The cat marked <Mark 1> is a calico, which is a breed with a tricolor coat, typically white, black, and orange.

OCR-intensive Task

Text Spotting

User:
Detect all regions in the image that contain text and identify their content.

SPHINX-X:
There are “Virtual[0.25,0.30,0.52,0.38], Office[0.53,0.31,0.72,0.38], of[0.23,0.39,0.30,0.46], the[0.34,0.39,0.47,0.46], “Future“[0.49,0.39,0.72,0.45], 1st[0.25,0.68,0.30,0.72], floor[0.32,0.68,0.39,0.72], Deutsches[0.61,0.48,0.74,0.52], Forschungszentrum[0.61,0.52,0.79,0.56], for[0.61,0.56,0.64,0.59], Kunstliche[0.65,0.56,0.77,0.60], Intelligenz[0.61,0.59,0.76,0.63], GmbH[0.74,0.59,0.78,0.62], 1.21[0.76,0.67,0.80,0.71], and Competence[0.24,0.20,0.50,0.25] in the image.

Document Layout Detection

User:
Detect the layout of the document in the image.

SPHINX-X:
There are seven text[0.16,0.50,0.48,0.64;0.16,0.64,0.48,0.74;0.16,0.73,0.48,0.85;0.16,0.85, 0.48,0.98;0.50,0.43,0.83,0.59;0.50,0.59,0.83,0.67;0.50,0.72,0.83,0.98], figure[0.51,0.06,0.82,0.31], list[0.16,0.06,0.48,0.49], and title[0.50,0.69,0.71,0.71] in the image.

Figure 7. **Set-of-Marks (SoM) prompting and OCR-intensive capabilities of SPHINX-X.** With our constructed two datasets, SPHINX-X exhibits outstanding visual performance on SoM prompting and OCR-related tasks. Note that the SoM marks are only utilized in the textual prompt, without rendering on input images.