
SPHINX: THE JOINT MIXING OF WEIGHTS, TASKS, AND VISUAL EMBEDDINGS FOR MULTI-MODAL LARGE LANGUAGE MODELS

Ziyi Lin^{1,2*}, Chris Liu^{1*}, Renrui Zhang^{1,2*}, Peng Gao^{1*†‡}, Longtian Qiu^{1,3*}
Han Xiao¹, Han Qiu¹, Chen Lin¹, Wenqi Shao¹, Keqin Chen¹, Jiaming Han^{1,2}
Siyan Huang¹, Yichi Zhang¹, Xuming He³, Hongsheng Li^{1,2†}, Yu Qiao^{1†}

¹Shanghai AI Laboratory, ²MMLab, CUHK, ³ShanghaiTech University

ABSTRACT

We present ***SPHINX***, a versatile multi-modal large language model (MLLM) with a joint mixing of model weights, tuning tasks, and visual embeddings. First, for stronger vision-language alignment, we unfreeze the large language model (LLM) during pre-training, and introduce a weight mix strategy between data from distinct domains. By directly mixing the weights of LLMs, the model can efficiently integrate diverse cross-modal semantics with favorable scalability. Then, to enable multi-purpose capabilities, we mix a variety of tasks for joint visual instruction tuning, including visual question answering, region-level understanding, caption grounding, document layout detection and human pose estimation, which contributes to mutual enhancement over different scenarios. Additionally, we propose to extract comprehensive visual embeddings from various network architectures, pre-training paradigms, and information granularity, providing language models with more robust image representations. Via our proposed three-fold mixing, ***SPHINX*** exhibits superior multi-modal understanding powers for a wide range of applications. On top of this, we further propose an efficient strategy aiming to better capture fine-grained appearances of high-resolution images. With a mixing of different scales and high-resolution sub-images, ***SPHINX*** attains exceptional visual parsing and reasoning performance on existing evaluation benchmarks. We hope our work may cast a light on exploring the significance of mixing different elements in future MLLM research. Code is released at <https://github.com/Alpha-VLLM/LLaMA2-Accessory>.

1 INTRODUCTION

Since the era of big data, large language models (LLMs) have attained tremendous strides (OpenAI, 2023a;b; Brown et al., 2020; Touvron et al., 2023a; Zhang et al., 2022), showcasing unprecedented levels of applications scenarios and generalization capabilities. To further expand their capacity ceiling, many efforts have been made to introduce additional visual input, developing powerful multi-modal large language models (MLLMs) (Zhang et al., 2023a; Li et al., 2023d; Liu et al., 2023d; Zhu et al., 2023; Zhao et al., 2023). These methods can not only generate well-organized language responses inherited from LLMs, but also unlock the multi-modal understanding capability for a wide range of applications, such as providing detailed image captions, answering visual questions, localizing different objects on the image, etc.

Existing MLLMs explore various strategies to endow LLMs with visual instruction-following capacities. **1)** Freezing the LLMs during pre-training, and only learning a projection network for vision-language alignment, e.g., a simple MLP layer of LLaMA-Adapter V2 (Gao et al., 2023b) and

* Equal contribution, † Equal advisory, ‡ Project leader

an attention-based visual abstractor of mPLUG-Owl (Ye et al., 2023). **2)** Constructing training data of new tasks to endow MLLMs with new visual understanding abilities, e.g., referential dialogues of Kosmos-2 (Peng et al., 2023b) and Shikra (Chen et al., 2023b) for region-level grounding and description. **3)** Employing advanced image encoders for extracting visual embeddings, e.g., a CLIP encoder (Radford et al., 2021) in LLaVA (Liu et al., 2023c) and a Q-Former (Li et al., 2023d) in MiniGPT-4 (Zhu et al., 2023).

In this paper, we propose a versatile MLLM, ***SPHINX***, with a mixing of three significant aspects: model weights, tuning tasks, and visual embeddings. The main characteristics and findings of our approach is illustrated as follows:

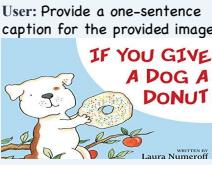
- **Unfreezing LLMs for pre-training.** Although the frozen LLM can effectively preserve its long-sentence generation capability, it constrains the potential of better vision-language alignment upon large-scale pre-training data. Therefore, we unfreeze the entire LLM, and supplement the vision-language datasets (Schuhmann et al., 2021) with RefinedWeb (Penedo et al., 2023) for language-specific tuning. This pre-training strategy not only enables LLMs to learn more cross-modal knowledge, but also alleviates the forgetting issue to generate detailed language responses.
- **Mixed model weights.** Vision-language data from particular domains might contain special semantics, e.g., synthetic captions (Christoph Schuhmann, 2022) compared to real-world ones (Schuhmann et al., 2021). Considering that directly mixing such data might confuse the MLLM, we introduce a weight mixing strategy to efficiently combine such domain-specific knowledge. Based on the MLLM pre-trained on real-world data, we fine-tune it on the synthetic data, and then aggregate the obtained LLM’s weights with the real-world ones in a training-free manner. In this way, our ***SPHINX*** can effectively integrate semantics of data from both synthetic and real-world domains with flexible scalability.
- **Mixed tuning tasks.** Different from existing task-specific MLLM models, we integrate a diverse array of visual instruction tasks to tune the pre-trained model, aiming to achieve a wide range of capabilities. Our mixing of tasks includes basic visual question answering (VQA), region-level referring expression comprehension/generation (REC/REG), multi-object detection and relation reasoning, text-oriented chart/document VQA, human pose estimation, etc. By such a comprehensive multi-task training paradigm, our ***SPHINX*** learns to be a well-performed generalist for visual instruction following.
- **Mixed visual embeddings.** To benefit from the advantages of different image encoding, we propose to mix the visual embeddings from various vision backbones (Oquab et al., 2023; Li et al., 2023d; Radford et al., 2021), including the ones with different network architectures (CNN vs. ViT), pre-training paradigms (supervised vs. self-supervised), and information granularity (global vs. local). By mixing different image tokens channel-wisely and sequence-wisely, ***SPHINX*** obtains stronger visual representations and leads to better vision-language alignment efficacy.

On top of this, we further investigate another challenging issue within existing MLLMs, i.e., the limited resolution of input images. As the pre-trained image encoders normally adopt a relatively low image resolution, e.g., 224×224 , it severely hinders fine-grained visual comprehension and reasoning for MLLMs. However, simply upsampling the images for encoders would harm the pre-trained positional prior, and, more importantly, lead to expensive computational overhead (the complexity increases quadratically to image size in self-attention mechanisms). Therefore, we propose to endow ***SPHINX*** with a longer sequence of visual embeddings by mixing different scales and high-resolution sub-images, as illustrated below.

- **Mixed scales and high-resolution sub-images.** we first spatially divide the input high-resolution image (448×448) into four sub-images (224×224), and also downsample it into a low-resolution one (224×224). Then, we feed the five images concurrently into the mixed visual encoders, and concatenate the extracted five token groups to represent the entire high-resolution visual features. By mixing visual embeddings of different scales and sub-images, our ***SPHINX*** can adaptively explore more fine-grained visual semantics from multi-scale image representations, while remaining encoding efficiency.

Captioning

User: Provide a one-sentence caption for the provided image.



SPHINX: A book cover for a children's book called If You Give a Dog a Donut.

Referring Expression Generation

User: How [0.05,0.32,0.54,0.83] feeling in the image?



SPHINX: Based on the image, it is difficult to determine the exact emotions of the man. However, he is sitting in a chair and appears to be engaged in a conversation with the group of people around him. The man is holding a book, which suggests that he might be discussing or teaching a subject related to the book. The presence of multiple people in the room indicates that the man is likely in a social or educational setting, where he is sharing his knowledge or opinions with others. The man's emotions could range from being calm and focused to being passionate and engaged in the discussion.

Visual Reasoning with Detection

User: Describe the image concisely. Include the bounding box for each mentioned object.



SPHINX: A deer[0.39,0.19,1.00,0.71] is eating a pumpkin[0.38,0.41,0.66,0.70].

User: Why this image funny?

SPHINX: The image is funny because a deer[0.39,0.19,1.00,0.71] is eating a pumpkin[0.38,0.41,0.66,0.70], which is a humorous sight because deer[0.39,0.19,1.00,0.71] typically eat leaves[0.00,0.19,1.00,0.81] and grass[0.00,0.19,1.00,0.81], not pumpkins[0.38,0.41,0.66,0.70].



Referring Expression Comprehension

User: Please provide the bounding box coordinate of the region this sentence describes: The cat.



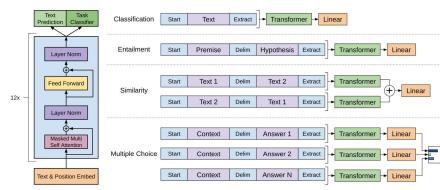
SPHINX: [0.37,0.32,0.90,0.58]

SPHINX:

The image is a satirical political cartoon featuring two men in military uniforms, one of whom is cutting into a globe with a sword. The cartoon is likely commenting on a political or social issue, such as the partitioning of a territory or the division of resources. The men's military uniforms and the act of cutting the globe suggest a sense of power and control, while the sword symbolizes the use of force or aggression. The cartoon's deep meaning is to convey a message about the consequences of such actions, possibly warning against the dangers of partitioning or division.

Document Layout Detection

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



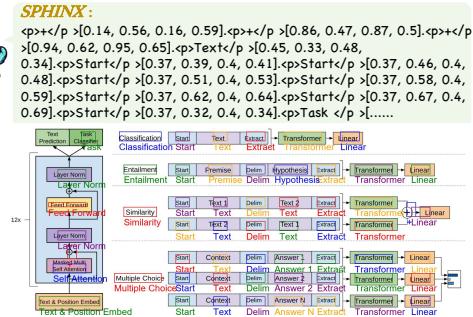
SPHINX:

```

<p></p>[0.14, 0.56, 0.16, 0.59]<p></p>[0.86, 0.47, 0.87, 0.5]<p></p>[0.94, 0.62, 0.95, 0.65]<p>Text</p>[0.45, 0.33, 0.48, 0.34]<p>Start</p>[0.37, 0.39, 0.4, 0.41]<p>Start</p>[0.37, 0.46, 0.4, 0.48]<p>Start</p>[0.37, 0.51, 0.4, 0.53]<p>Start</p>[0.37, 0.58, 0.4, 0.59]<p>Start</p>[0.37, 0.62, 0.4, 0.64]<p>Start</p>[0.37, 0.67, 0.4, 0.69]<p>Start</p>[0.37, 0.32, 0.4, 0.34]<p>Task </p>[...]

```

SPHINX:



Visual Question Answering

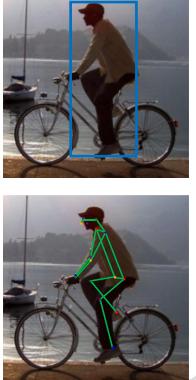
User: What is the deep meaning of this picture?



SPHINX: The image is a satirical political cartoon featuring two men in military uniforms, one of whom is cutting into a globe with a sword. The cartoon is likely commenting on a political or social issue, such as the partitioning of a territory or the division of resources. The men's military uniforms and the act of cutting the globe suggest a sense of power and control, while the sword symbolizes the use of force or aggression. The cartoon's deep meaning is to convey a message about the consequences of such actions, possibly warning against the dangers of partitioning or division.

Human Pose Estimation

User: Detect the key points of the person in the region [0.34,0.06,0.73,0.89].



SPHINX: The detected key points are nose[0.42,0.15], left eye[0.43,0.13], left ear[0.50,0.13], left shoulder[0.55,0.20], right shoulder[0.50,0.18], left elbow[0.47,0.35], right elbow[0.47,0.33], left wrist[0.39,0.44], right wrist[0.39,0.43], left hip[0.65,0.45], right hip[0.60,0.44], left knee[0.52,0.56], right knee[0.47,0.46], left ankle[0.59,0.82], and right ankle[0.61,0.63].

Figure 1: Examples of multi-task visual reasoning by our proposed **SPHINX**, which excels in diverse visual understanding and perception tasks, such as object detection, caption grounding, and region-level description.

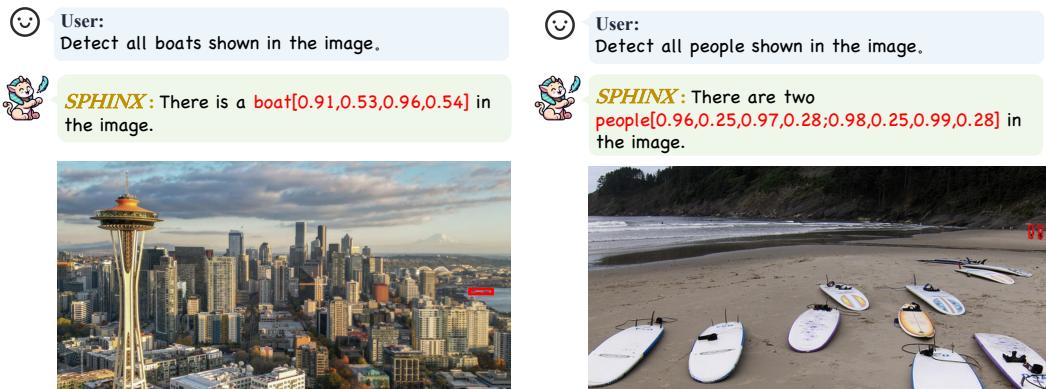


Figure 2: **Examples of SPHINX for fine-grained visual perception.** With a longer sequence of visual tokens, our model exhibits superior fine-grained understanding capacity.

Note that, as the different sub-images of high-resolution images do not interact with each other in the visual encoder, they are forced to interchange information within the attention layers of LLMs, which motivates LLMs to process visual conditions more thoroughly and deeply. By the proposed three-fold mixer along with a longer visual token sequence, **SPHINX** fine-tunes LLMs, e.g., LLaMA-2 (Touvron et al., 2023b), to be a powerful MLLM with superior visual instruction-following capacity. As shown by the examples in Figure 1, our model excels in a variety of vision tasks, e.g., detecting different objects with remarkable precision and parsing their relations, or accurately interpret the content within complicated figures. Importantly, as shown in Figure 2, **SPHINX** can achieve impressive fine-grained visual perception for high-resolution images, which exhibits *state-of-the-art* performance on extensive evaluation benchmarks, e.g., MMBench (Liu et al., 2023f), MME (Fu et al., 2023a), and POPE (Li et al., 2023e).

2 RELATED WORK

Large language models (LLMs). The field of Natural Language Processing (NLP) has witnessed significant progress over the years, particularly with the advent of LLMs. With Transformer (Vaswani et al., 2017) as the fundamental architecture, LLMs (OpenAI, 2023a; Radford et al., 2019; OpenAI, 2023b) have demonstrated unprecedented performance in modeling intricate language patterns over extensive contexts. Therein, BERT (Devlin et al., 2018) showcases the benefits of pre-training on vast text corpora and fine-tuning on specific tasks, setting new standards on various benchmarks. OpenAI’s GPT series (Radford & Narasimhan, 2018; Radford et al., 2019; OpenAI, 2023a;b), especially GPT-3 (Brown et al., 2020), harness the power of massive model scaling, with billions and even trillions of parameters. To obtain better instruction following ability, InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2023a) are presented to exhibit exceptional fluency and versatility in open-domain conversation tasks, ranging from text generation to question answering. Recently, the instruction tuning based on LLaMA (Touvron et al., 2023a) and LLaMA-2 (Touvron et al., 2023b) has gained great popularity as open-source LLMs to the community. Therein, Alpaca (Taori et al., 2023) and LLaMA-Adapter (Zhang et al., 2023a) respectively adopt full and parameter-efficient fine-tuning to acquire favorable instruction-following LLMs. Vicuna (Chiang et al., 2023) and GPT-4-LLM (Peng et al., 2023a) further showcase the improvement brought by higher-quality instruction datasets. Other efforts also extend LLMs for match problem solving (Wang et al., 2023a; Zhou et al., 2023), visual model system (Wu et al., 2023; Yang et al., 2023), and open-world recognition (Zhang et al., 2023b; Zhu et al., 2022). In this paper, we develop our **SPHINX** based on the superior language understanding of LLaMA-2 (Touvron et al., 2023b) and instruction tuning experience of LLaMA-Adapter series (Zhang et al., 2023a; Gao et al., 2023b), which introduce a three-fold mixer to extend the capability ceiling of instruction-following LLMs for multi-modal input.

Multi-modal large language models (MLLMs). In addition to language instruction following, many efforts have been made to inject multi-modal conditions into LLMs for wider application scenarios. As prior attempts, VisualGPT (Chen et al., 2022) and BLIP series (Li et al., 2023d; 2022;

Dai et al., 2023) indicate the potentials of aligning LLMs with visual input for image captioning and question answering. Flamingo (Alayrac et al., 2022) and Kosmos-1 (Huang et al., 2023) further exhibit promising multi-modal understanding performance for image-text interleaved contexts. With large-scale pre-training and model sizes, GPT-4 (OpenAI, 2023b) and Bard (Google, 2023) both showcase remarkable proficiency in vision-language understanding and reasoning over diverse multi-modal tasks. In parallel, a bunch of works have been proposed to align LLaMA with vision modality for advanced visual instruction-following capabilities. LLaVA (Liu et al., 2023d) and MiniGPT-4 (Zhu et al., 2023) utilize a simple projection layer to connect vision encoders (Li et al., 2023d; Radford et al., 2021) with LLMs. LLaMA-Adapter V2 (Gao et al., 2023a) introduces zero-initialized attention mechanisms for efficient visual instruction tuning, and mPLUG-Owl (Ye et al., 2023) adopts delicately designed intermediate networks for cross-modal alignment. For more modality input, ImageBind-LLM (Han et al., 2023) and PandaGPT (Su et al., 2023) further incorporate audio and video conditions guided by ImageBind (Girdhar et al., 2023). Besides, recent MLLMs are also extended to region-level parsing (Chen et al., 2023b; Peng et al., 2023b), in-context learning (Li et al., 2023a;b), arbitrary image resolutions (Bavishi et al., 2023), text-to-image generation (Wen et al., 2023; Dong et al., 2023), and 3D question answering (Xu et al., 2023; Guo et al., 2023; Hong et al., 2023). Different from previous works, our **SPHINX** aims for image-conditioned MLLM, and proposes a three-fold mixer, i.e., model weights, tuning tasks, and visual embeddings, attaining superior generalization capacity for multi-modal learning.

3 SPHINX

In this section, we introduce a versatile MLLM, **SPHINX**. In Section 3.1, we first present the details of our proposed three-fold mixing for MLLM, i.e., mixed model weights, tuning tasks, and visual embeddings. Then, in Section 3.2, we illustrate how to tackle high-resolution images with longer image tokens. Finally, in Section 3.3, we introduce several extended applications of **SPHINX**.

3.1 THREE-FOLD MIXING

The overall mixing paradigm of **SPHINX** is shown in Figure 3. We adopt two steps for training: the first pre-training stage for vision-language alignment, and the second fine-tuning stage for visual instruction-following learning. During the two stages, we apply the proposed mixing of model weights and tuning tasks, respectively. The model is composed of an LLM, e.g., LLaMA-2 (Touvron et al., 2023b), a mixing of vision encoders, and two linear projection layers. In this subsection, we illustrate the unfrozen LLM during pre-training, and then introduce our mixing strategy.

Unfreezing LLM for stage-1 pre-training. Existing MLLMs (Zhu et al., 2023; Li et al., 2023d; Dai et al., 2023) actually freeze the entire LLM during pre-training, and only train intermediate projection layers for vision-language alignment. This strategy aims to prevent LLMs from overfitting to generating only short sentences, since the pre-training caption data only contains concise descriptions of images. However, the frozen weights largely constrain the cross-modal learning potential of LLMs with large-scale pre-training data. Therefore, we propose to unfreeze the entire LLM, along with learnable linear projection layers, for more sufficient vision-language adaption, while still keeping the vision encoders frozen for high-quality image representations. To particularly preserve the long-sentence generation ability, we supplement the existing image-caption data with additional text corpora data (RefinedWeb (Penedo et al., 2023)) for language-only tuning. To be more specific, for every iteration, we sample one text and image-caption data respectively from language and vision-language datasets. We denote their sequence length as N_l and N_{vl} , and represent the parameters of our MLLM as θ . Then, the cross-entropy loss function with a balance factor α is formulated as

$$\mathcal{L} = - \left(\alpha \cdot \sum_{i=1}^{N_l} \log p(w_i | w_{<i}; \theta) + (1 - \alpha) \cdot \sum_{i=1}^{N_{vl}} \log p(w_i | I; w_{<i}; \theta) \right), \quad (1)$$

where w_i denotes the i -th word, and I denotes the image corresponding to the caption data. In this way, we not only optimize the LLM’s weights with extensive visual captioning data, but also maintain its pre-trained generative capacity with curated corpora.

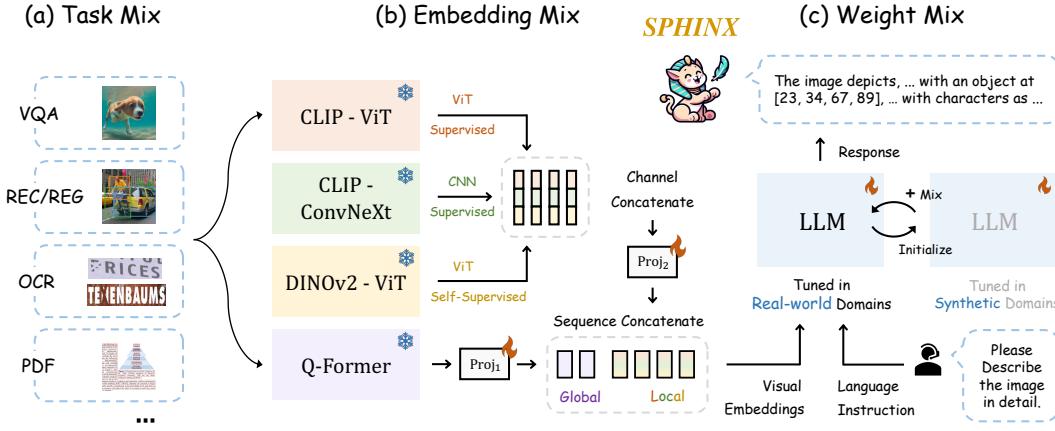


Figure 3: **Mixing paradigm of *SPHINX*.** We propose a three-fold mixing for developing versatile MLLMs: mixed tuning tasks (a), mixed visual embeddings (b), and mixed model weights (c).

Mixed model weights of different domains. Some vision-language data from particular domains contains distinct semantic knowledge, such as the synthetic captions of LAION-COCO (Christoph Schuhmann, 2022) compared to real-world descriptions of LAION-400M (Schuhmann et al., 2021). Considering that our unfrozen LLM allows for pre-training the model with different data domains, we propose a weight mixing strategy of domain-specifically tuned LLMs to integrate respective knowledge from real-world and synthetic data. We first utilize the most common domain data (LAION-400M (Schuhmann et al., 2021)) for pre-training, which endows the MLLM with fundamental visual understanding capabilities. Then, we regard such a pre-trained model as an initial checkpoint to further fine-tune LLM on synthetic domains, e.g., LAION-COCO (Christoph Schuhmann, 2022). Finally, to take advantages of the best of both worlds, we directly conduct a weighted mixing of two LLMs’ weights for semantic aggregation. In detail, we denote the parameters of the fundamental LLM as θ_{real} , and the fine-tuned parameters by synthetic data as θ_{syn} . The mixing process is formulated as

$$\theta_{mix} = \beta \cdot \theta_{real} + (1 - \beta) \cdot \theta_{syn}, \quad (2)$$

where β denotes the mixing coefficient, and θ_{mix} represents the mixed LLM weights with aggregated semantics. Compared to fusing different domain data for joint pre-training, our weight mix strategy can encourage every MLLM to better learn domain-unique knowledge, and exhibits flexible scalability for any new data domains.

Mixed tuning tasks for stage-2 fine-tuning. After pre-training and model weight mixing, the MLLM has acquired a satisfactory alignment between vision and language modalities. To further enhance the instruction-following capacity, we collect instruction data from a wide range of multi-modal tasks, and jointly fine-tune the model to learn a vision generalist, instead of a specialist for specific scenarios. Previous MLLMs can only perform simple visual question answer (VQA) and single large object referring. In contrast, *SPHINX* is jointly fine-tuned with general VQA, region-level referring expression comprehension/generation (REC/REG), multi-object detection and relation reasoning, text-oriented chart/document VQA, and human pose estimation. Thanks to the superior reasoning capacity of LLM and proper designs of task prompts, *SPHINX* can derive multi-purpose capabilities for visual understanding and perception, excelling in various application scenarios.

Mixed embeddings for visual encoding. To obtain a robust visual representation, we propose to ensemble a variety of vision backbones for image encoding. The pre-trained backbones are chosen according to three aspects to have effective complementary characteristics. **1)** Different network architectures. As CNN (He et al., 2016a) and ViT (Dosovitskiy et al., 2020) mainly aggregate different types of visual appearances, i.e., neighboring dependencies and long-range interactions, we adopt CLIP (Radford et al., 2021) models respectively with ConvNeXt (Woo et al., 2023) and ViT image encoders. **2)** Different pre-training paradigms. Supervised training can impose explicit semantic information from textual captions or category labels, while self-supervised learning enforces

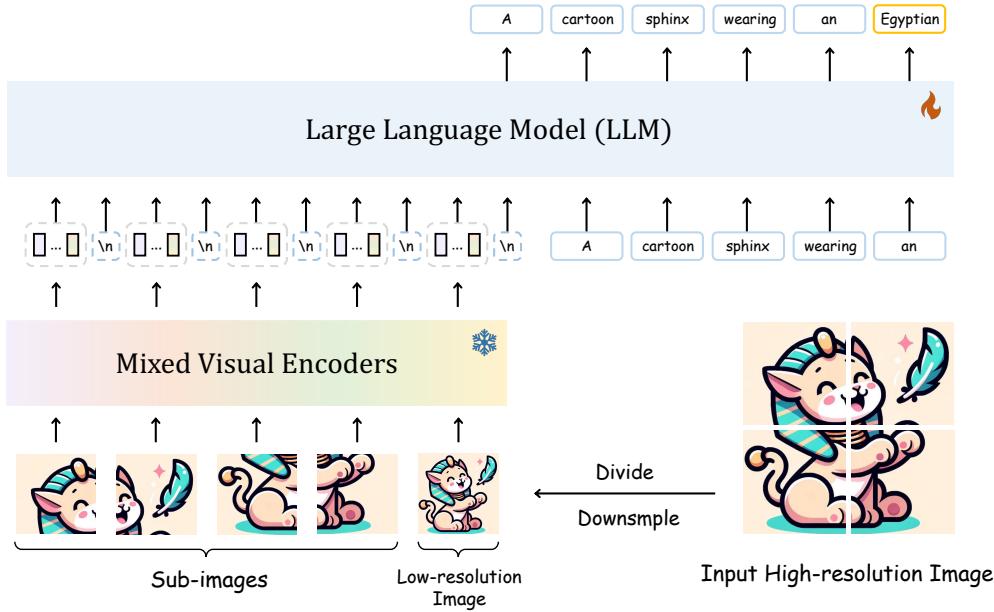


Figure 4: **Pipeline of SPHINX for high-resolution images.** We propose to further mix different scales and sub-images for better capture fine-grained semantics on high-resolution images.

the model to explore implicit pretext task signals. Thus, we further employ ViT self-supervised by DINOv2 (Oquab et al., 2023), compared to the text-supervised CLIP. 3) Different information granularity. The aforementioned visual encoding all produce image patch features in local semantic level. To better capture global features, we also adopt Q-Former (Li et al., 2023d) to summarize visual embeddings by independent queries from the global context. After all the aforementioned encoding, we first channel-wisely concatenate the three image features from different network architectures and pre-training paradigms. Then, using two projection layers for dimension alignment, we spatial-wisely concatenate the representations between Q-Former and the other three vision backbones. The obtained image tokens are directly placed in front of language instructions, which provides visual conditions for the subsequent instructions.

3.2 HIGH-RESOLUTION IMAGES

With above three-fold mixing strategy, **SPHINX** already showcases superior performance for diverse visual perception and reasoning tasks. However, one key problem of existing MLLMs still remains, i.e., the limited resolution of input images. To this end, we further propose to utilize longer sequence of visual tokens by mixing high-resolution sub-images, as shown in Figure 4.

Low-resolution constraints of MLLMs. Previous MLLM works adopt frozen image encoders during all training stages, in order to preserve the pre-trained visual semantics. Therefore, the image resolution of MLLMs is normally limited to 224×224 , severely hindering their efficacy for fine-grained visual perception, especially region-level grounding and description. However, directly processing the upsampled image is not optimal for two reasons. First, to align the image size, the pre-trained positional encoding vectors in ViT are also required to be upsampled correspondingly, which would harm the prior spatial cues. Second, the computation complexity of ViT increases quadratically to the input image size. Thus, naively upsampling the image leads to extensive inference time and GPU memory consumption.

Mixed scales and high-resolution sub-images. In our **SPHINX**, we extend the mixing of visual embeddings to more scales and sub-images, allowing for efficient high-resolution image encoding. For an input high-resolution image, e.g., 448×448 , we construct five corresponding images of 224×224 , and feed them as independent images into the mixed vision encoders. Specifically, we first downsample the input image into 224×224 resolution as an abstract representation, and also divide



Figure 5: **Examples of language-referred segmentation** by integrating **SPHINX** and Segment Anything Model (SAM) (Kirillov et al., 2023).

it into four sub-images of 224×224 from the four corners of the original image, which preserve its detailed visual information. In this way, we enable MLLMs to not only capture fine-grained visual appearances with 224×224 positional encodings, but also achieve favorable computation efficiency. Afterwards, the five groups of image tokens are encoded and concatenated as a long sequence for feeding into LLM, where the first one group encodes global semantics, and the other four record more fine-grained local features. Importantly, as the image tokens of different patches do not have interaction through the vision encoders, they are forced to interact within the attention layers of LLMs for obtaining complete visual information. Such a strategy, in turn, motivates LLMs to parse the relations within visual conditions for better cross-modal learning. From this perspective, our **SPHINX** can be regarded as a new paradigm for similar to ViT (Dosovitskiy et al., 2020), where the mixed vision encoders serve as a patch embedding layer, and the LLM plays the role for patch interaction as a vision decoder. On high-resolution visual understanding tasks, **SPHINX** achieves significant improvement with the mixed visual representations of scales and high-resolution sub-images.

3.3 EXTENSION

In this section, we respectively introduce some extended applications derived from **SPHINX**.

3.3.1 INTEGRATION WITH SAM AND STABLE DIFFUSION

In addition to multi-purpose visual instruction-following capabilities, we can also collaborate **SPHINX** with other visual foundation models to tackle more challenging vision tasks. Figure 5 and 6 respectively show two applications for language-referred segmentation and image editing.

Language-referred segmentation. Given that our MLLM is able to output accurate detection results indicated by a user-provided description or semantic category, we can cascade the Segment Anything Model (SAM) (Kirillov et al., 2023) for language-referred instance or semantic segmentation. In detail, we regard the predicted bounding boxes from **SPHINX** as box prompts, and feed them into SAM for segmenting a specific instance or all objects within a semantic category. In this way, we effectively incorporate the semantic reasoning power of LLMs and class-agnostic segmentation abilities of SAM, achieving a wider range of functionalities.

Image inpainting and editing. Based on the segmentation results from SAM, we refer to Inpaint Anything (Yu et al., 2023a) to integrate image inpainting models (LaMa (Suvorov et al., 2021)) and text-to-image generative models (Stable Diffusion (Rombach et al., 2021)) for high-quality image inpainting and editing. Specifically, we first detect and segment the user-indicated objects via **SPHINX** and SAM as illustrated in the previous paragraph. Then, we feed the segmentation

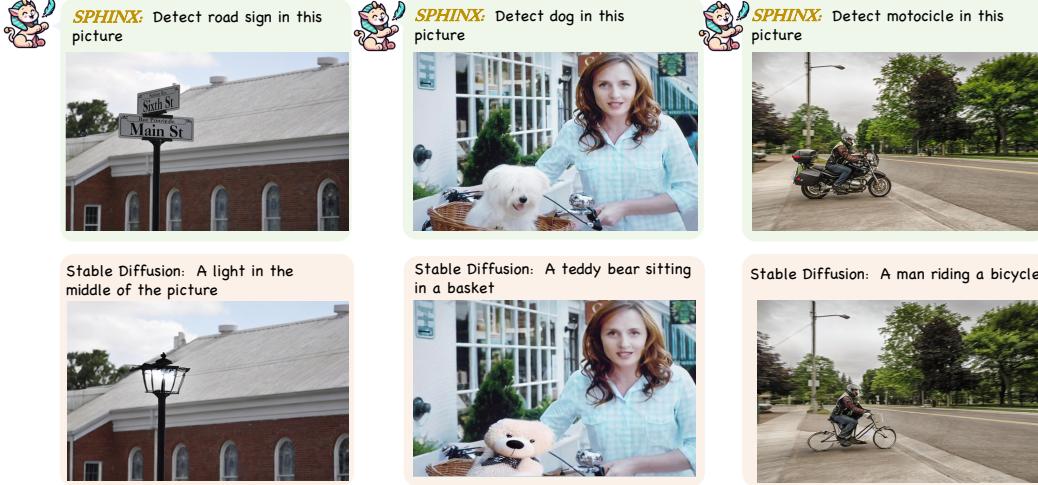


Figure 6: **Examples of image inpainting and editing** by integrating **SPHINX** and Stable Diffusion (Rombach et al., 2021).

mask into LaMa (Suvorov et al., 2021) for removing the corresponding objects with contextual data. After this, the user can prompt Stable Diffusion to further generate new visual concepts to replace the original ones. This allows for flexible and language-driven image inpainting and editing.

3.3.2 FINE-TUNING SPHINX FOR VISUAL RECOGNITION

Empowered by the joint mixing of weights, tasks and visual embeddings, our **SPHINX** has comprehended robust and diverse visual category semantics. Considering this, we propose to regard **SPHINX** as a universal initialization for traditional visual recognition tasks. For instance, given a classification task, e.g., ImageNet-1K (Russakovsky et al., 2015), we transform it into a single-turn conversation data format, e.g., “Classify the image.” as the instruction and “This is a [CLASS]” as the response. By performing supervised fine-tuning on the transformed dataset, we observe significantly fast training convergence on ImageNet-1K. Surprisingly, with only one epoch, **SPHINX** can achieve 70.8% classification accuracy without any data augmentation. This convergence speed is much faster than traditional approaches, such as ResNet (He et al., 2016b) and ViT (Dosovitskiy et al., 2020) that normally cost around 300 epochs and require strong data augmentation techniques.

4 EXPERIMENTS

4.1 TRAINING DETAILS

As mentioned in Section 3.1, our training pipeline consists of two stages. In stage 1, or the Pre-training stage, we start from a text-only LLM, and build the multi-modal capabilities from scratch with large-scale noisy datasets. In stage 2, or the fine-tuning stage, we extract the strong capabilities learned in stage 1 on practical tasks by further training with diverse and high-quality instruct-following datasets. The construct of the datasets and the training configuration for both stages are detailed as follows.

Pre-training datasets. We use two image captioning datasets LAION-400M (Schuhmann et al., 2021) and LAION-COCO (Christoph Schuhmann, 2022) for multi-modal alignment. As we full-fine-tune the language model backbone for long steps, we also jointly train with a text-only dataset RefinedWeb (Penedo et al., 2023) to avoid harming its text reasoning capability due to catastrophic forgetting.

Pre-training configuration. We fine-tune the weight of large language model and the visual projections in the pre-training stage, among which the weight of large language model is initialized from off-the-shelf open-source weights such as LLaMA-2 (Touvron et al., 2023b) and the visual

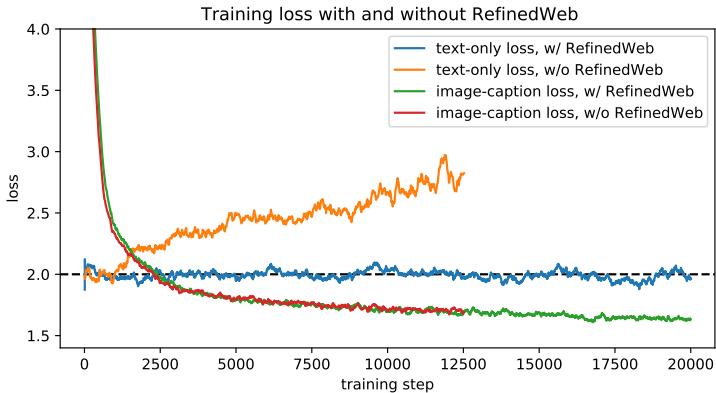


Figure 7: Loss curve in the pre-training stage with and without optimizing on RefinedWeb. The text-only loss is on RefinedWeb and the image-caption loss is on LAION-400M. Without jointly training on RefinedWeb, the image-caption loss descends similarly but the text-only loss grows significantly even in less than 1/10 of the complete training schedule. We early-stop the *without-RefinedWeb* experiments after the forgetting phenomenon is clear.

projections are initialized randomly. The visual encoders themselves are kept frozen at their original weights throughout the training. We use the AdamW optimizer (Kingma & Ba, 2014) with $(\beta_1, \beta_2) = (0.9, 0.95)$, a cosine annealing learning rate schedule for 180,000 steps from 5×10^{-5} to 5×10^{-6} with the first 2,000 steps being a linear warm-up from 0 to 5×10^{-5} , and a constant weight decay of 0.1. For the joint training on both images and texts, we form each batch with 640 image-text pairs from LAION-400M or LAION-COCO and 65,536 text tokens from RefinedWeb, and add the losses from the two parts for back-propagation. Since captions in LAION-400M and LAION-COCO are based on web-crawled data and generally do not contain much fine-grained information, we use only one global view of each image for faster training. We do not apply any form of prompts during pre-training. The pre-training time is around 125 hours on 32 A100 GPUs with a 7B language model and about twice with a 13B language model.

Fine-tuning datasets. In the multi-task fine-tuning phase, our objective is to equip the Multimodal Language Model (MLLM) with the versatility needed for diverse downstream tasks. Building upon insights from prior research (Liu et al., 2023d; Dai et al., 2023; Chen et al., 2023b; Zhu et al., 2023; Liu et al., 2023b), we include instruction following data such as LLaVA (Liu et al., 2023d) and ShareGPT (ShareGPT, 2023), exposing the model to tasks involving explicit directives. For general Vision Question Answering (VQA), we leverage datasets like VQAV2 (Agrawal et al., 2015) and GQA (Hudson & Manning, 2019). Expanding the scope to outside knowledge, we integrate datasets like OKVQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022), providing the model with information beyond the training data. Optical Character Recognition (OCR) datasets, such as OCRVQA (Mishra et al., 2019) and TextCaps (Sidorov et al., 2020). We introduce abundant general object detection datasets, such as COCO (Lin et al., 2014) to inspire the model’s capabilities of localization and classification. To address grounding tasks, we incorporate RefCOCO (Kazemzadeh et al., 2014) and VG (Krishna et al., 2017) datasets, exposing the model to more challenges related to object localization. Additionally, Grounding Caption datasets, such as those from Flickr30k (Plummer et al., 2015), further refine the understanding of descriptions in the context of image regions. Despite the diversity of data sources, we streamline the training by converting all datasets into a multi-turn conversation format. This not only reduces training costs but also enhances overall efficiency.

Fine-tuning configuration. The trained and frozen parts are identical as the pre-training stage. The optimizer settings are similar to the pre-training stage, except that we use a batch size of 128, a maximum learning rate of 2×10^{-5} , a minimum learning rate of 0 and a linear warmup for 0.03 epoch during fine-tuning. Training data are sampled from the mixture of datasets uniformly without down- or up-sampling, *i.e.*, the chance of a dataset being sampled from is proportional to its original size. We follow the image preprocessing of (Chen et al., 2023b; Liu et al., 2023b), *i.e.*, padding the

| Prompts | Benchamrks |
|--|--------------------------------------|
| - | LLaVA-Bench, MM-Vet, MathVista |
| Answer the question using a single word or phrase. | VQAV2, GQA, OKVQA, VSR, MME, OCR-VQA |
| Answer with the option’s letter from the given choices directly. | SeedBench, ScienceQA, IconVQA |
| Please provide the bounding box coordinate of the region this sentence describes: {description}. | RefCOCO, RefCOCO+, RefCOCOg |
| Reference OCR token: {OCR} | TextVQA |
| Answer the question using a single word or phrase. | |
| When the provided information is insufficient, respond with ‘Unanswerable’. Answer the question using a single word or phrase. | VizWiz |
| There are several options: {options} | CCBench, MMBench |
| Detect all objects shown in the image. detect all {category name} shown in the image. | Object Detection |
| Detect all people shown in the image. | Human Pose Detection |
| Detect the key points of the person in the region {coordinate}. | Document Layout |
| Please detect all texts and provide their bounding box coordinate. | Grounded Caption |
| Describe the image concisely. Include the bounding box for each mentioned object. | Relation Detection |
| What is the relationship between the object in {coordinate} and the object in {coordinate}? | Referring Relationship |
| Please provide the bounding box coordinate of the region this sentence describes: {description} | |

Table 1: **Prompts used in evaluation.**

image along the shorter edge to make it a square before resize, for better handling of images with extreme aspect ratios. The fine-tuning takes about 38 hours with 16 A100 GPUs with a 13B language model. The maximum training sequence length is set to 3072.

4.2 QUANTITATIVE EVALUATION

In this section, we provide a comprehensive evaluation of **SPHINX** and showcase results across multiple benchmarks. Our evaluation encompasses both quantitative metrics and qualitative assessments, providing a holistic understanding of our VLM model’s performance.

Image-text joint pre-training. We show in Figure 7 the effects of introducing a text-only task (*i.e.*, RefinedWeb) to jointly train with image captioning in the pre-training stage. For the *without-RefinedWeb* experiments, we set the balance factor α to 0 which means only image-text pairs is optimized. It is a clear tendency that the text-only loss grows if the model is *not* trained with RefinedWeb, showing that our joint-training scheme is effective in preserving the text-modelling capability while adapting for cross-modal understanding.

Evaluation prompt design. In our model evaluation, we prioritize aligning with each benchmark’s desired output format. To achieve this, we employ distinct prompts tailored to benchmarks that necessitate long answers, short answers, and multiple-choice responses. This approach ensures that our model is adept at satisfying the specific requirements of diverse evaluation scenarios. For tasks involving visual grounding, we directly utilize the prompts during training to enhance the model’s performance on these particular challenges. The detailed information is provided in Table 1.

Benchmarks on multi-modal large language models. We test our model on recently proposed MLLM benchmarks to comprehensive evaluation of the model’s characteristic such as MME (Fu et al., 2023b), Seedbench (Li et al., 2023c), POPE (Li et al., 2023e), LLaVA-Bench (In-the-Wild) (Liu et al., 2023d), MM-Vet (Yu et al., 2023b), MathVista (Lu et al., 2023), MMBench (Liu et al., 2023g), CCBench (Contributors, 2023). We show the result in Table 2. We observe the **SPHINX** surpasses previous state-of-the-art MLLM result on 5 out of 9 benchmarks. We compare our model with strong

| Method | POPE | MME ^P | MME ^C | MMB | MMB ^{CN} | SEED | LLava ^W | MM-Vet | CCbench | MathVista |
|-------------------------------------|-------------|------------------|------------------|-------------|-------------------|-------------|--------------------|-------------|-------------|-------------|
| BLIP-2 (Li et al., 2023d) | 85.3 | 1293.8 | - | - | - | 46.4 | 38.1 | 22.4 | - | - |
| InstructBLIP-7B (Dai et al., 2023) | - | - | - | 36 | 23.7 | 53.4 | 60.9 | 26.2 | 12.1 | 25.3 |
| InstructBLIP-13B (Dai et al., 2023) | 78.9 | 1212.8 | - | - | - | - | 58.2 | 25.6 | - | - |
| Shikra (Chen et al., 2023b) | - | - | - | 58.8 | - | - | - | - | - | - |
| LLaMA-AdapterV2 (Gao et al., 2023a) | - | 1328.40 | 356.43 | - | - | - | - | - | - | - |
| Qwen-VL-7B (Bai et al., 2023) | - | - | - | 38.2 | 7.4 | 56.3 | - | - | 5.5 | - |
| Qwen-VL-7B-Chat (Bai et al., 2023) | - | 1487.58 | 360.71 | 60.6 | 56.7 | 58.2 | - | - | 39.3 | - |
| LLaVA1.5-7B (Liu et al., 2023b) | 85.9 | 1510.7 | - | 64.3 | 58.3 | 58.6 | 63.4 | 30.5 | 16.4 | - |
| LLaVA1.5-13B (Liu et al., 2023b) | 85.9 | 1531.3 | 295.36 | 67.7 | 63.6 | 61.6 | 70.7 | 35.4 | 26.5 | - |
| SPHINX | 80.7 | 1476.1 | 322.2 | 66.9 | 56.2 | 69.14 | 73.5 | 36.0 | 25.6 | 27.0 |
| SPHINX-1k | 90.8 | 1560.2 | 310.0 | 67.1 | 59.5 | 71.6 | 74.3 | 36.6 | 27.9 | 27.5 |
| SPHINX-2k | 87.2 | 1470.6 | 326.8 | 65.9 | 57.9 | 71.6 | 76.9 | 40.2 | 27.4 | 27.8 |

Table 2: Comparison with SoTA methods on 9 MLLM benchmarks.

| Method | General VQA | | | | | | | Text-Oriented VQA | |
|-------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------------|-------------|
| | OKVQA | VQAV2 | VizWiz | GQA | VSR | ScienceQA | IconVQA | TextVQA | OCR-VQA |
| BLIP-2 (Li et al., 2023d) | 45.9 | - | 19.6 | 41.0 | 50.9 | - | 40.6 | - | 40.6 |
| InstructBLIP (Dai et al., 2023) | - | - | 33.4 | 49.5 | 52.1 | - | 44.8 | - | 44.8 |
| LLaMA-AdapterV2 (Gao et al., 2023a) | 49.6 | 70.7 | 39.8 | 45.1 | - | - | - | 37.4 | - |
| Shikra (Chen et al., 2023b) | 47.2 | 77.4 | - | - | - | - | - | - | - |
| Fuyu-8B (Bavishi et al., 2023) | 60.6 | 74.2 | - | - | - | - | - | - | - |
| MiniGPT-v2 (Chen et al., 2023a) | 57.8 | - | 53.6 | 60.1 | 62.9 | - | 51.5 | - | - |
| Qwen-VL-7B (Bai et al., 2023) | 58.6 | 79.5 | 35.2 | 59.3 | 63.8 | 67.1 | - | 63.8 | 75.7 |
| Qwen-VL-7B-Chat (Bai et al., 2023) | 56.6 | 78.2 | 38.9 | 57.5 | 61.5 | 68.2 | - | 61.5 | 70.5 |
| LLaVA1.5-7B (Liu et al., 2023b) | - | 78.5 | 50.0 | 62.0 | - | 66.8 | - | 58.2 | - |
| LLaVA1.5-13B (Liu et al., 2023b) | - | 80.0 | 53.6 | 63.3 | - | 71.6 | - | 61.3 | - |
| SPHINX | 62.1 | 78.1 | 39.9 | 62.6 | 58.5 | 69.3 | 50.4 | 51.63 | 66.0 |
| SPHINX-1k | 62.2 | 80.2 | 46.8 | 62.9 | 65.4 | 69.1 | 52.7 | 58.78 | 70.0 |
| SPHINX-2k | 62.6 | 80.7 | 44.9 | 63.1 | 57.1 | 70.6 | 50.5 | 61.19 | 67.8 |

Table 3: Performance comparison on 10 academic task-oriented benchmarks.

baselines including BLIP-2 (Li et al., 2023d), InstructBLIP (Dai et al., 2023), Shikra (Chen et al., 2023b), Qwen (Bai et al., 2023), Fuyu (Bavishi et al., 2023) and LLaVA1.5 (Liu et al., 2023b). The gap between **SPHINX** and **SPHINX-1k** on POPE suggests that the introduction of high resolution can significantly improve visual hallucination problems.

Visual question answering. Furthermore, we evaluate general VQA benchmarks, such as VQAV2 (Agrawal et al., 2015), OKVQA (Marino et al., 2019), GQA (Hudson & Manning, 2019), vizwiz (Gurari et al., 2018), ScienceQA (Lu et al., 2022), visual spatial reasoning (VSR) (Liu et al., 2023a), IconQA (Lu et al., 2021). Additionally, we conduct experiments on Text-oriented VQA such as TextVQA (Singh et al., 2019), OCR-VQA (Mishra et al., 2019). We provide the results in Table 3. **SPHINX** achieves comparative results across all benchmarks. We observe that **SPHINX-1k** and **SPHINX-2k** outperform **SPHINX** in VQA datasets that demand fine-grained visual information, showcasing the effectiveness of our visual mixed-up approach for achieving high resolution without relying on a visual encoder trained specifically on high-resolution images.

Visual grounding. Table 4 presents a comparison between the **SPHINX** model and baseline models on REC benchmarks. Our **SPHINX** exhibits robust performance in visual grounding tasks such as RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Mao et al., 2015), and RefCOCOg (Mao et al., 2015), surpassing other vision-language generalist models. Notably, **SPHINX** outperforms specialist models G-DINO-L (Liu et al., 2023e) by more than 1.54% in accuracy across all tasks within RefCOCO/RefCOCO+/RefCOCOg. When compared to the formidable baseline, Qwen-VL (7B) (Bai et al., 2023), which is trained on 20.9M grounding data. **SPHINX-1k** achieves 88.14% accuracy compared to 86.45% on average with less than 1M data. These findings offer direct evidence of the competitive visual grounding capabilities of **SPHINX**.

4.3 DEMONSTRATIONS

In this section, we present the qualitative outcomes of **SPHINX**, showcasing its remarkable capabilities in SAM-assisted segmentation, general object detection, human pose estimation, document layout detection, anomaly detection and etc. Surprisingly, **SPHINX** also exhibits improved performance by chain of thoughts and obtains emergent cross-task abilities.

| Methods | RefCOCO+ | | | RefCOCO | | | RefCOCOg | | Avg |
|---|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | val | test-A | test-B | val | test-A | test-B | val-u | test-u | |
| <i>Specialist models</i> | | | | | | | | | |
| UNINEXT (Yan et al., 2023) | 85.24 | 89.63 | 79.79 | 92.64 | 94.33 | 91.46 | 88.73 | 89.37 | 88.90 |
| G-DINO-L (Liu et al., 2023e) | 82.75 | 88.95 | 75.92 | 90.56 | 93.19 | 88.24 | 86.13 | 87.02 | 86.60 |
| <i>Generalist models</i> | | | | | | | | | |
| VisionLLM-H (Wang et al., 2023b) | - | - | - | - | 86.70 | - | - | - | - |
| OFA-L (Wang et al., 2022) | 68.29 | 76.00 | 61.75 | 79.96 | 83.67 | 76.39 | 67.57 | 67.58 | 72.65 |
| Shikra 7B (Chen et al., 2023b) | 81.60 | 87.36 | 72.12 | 87.01 | 90.61 | 80.24 | 82.27 | 82.19 | 82.93 |
| Shikra 13B (Chen et al., 2023b) | 82.89 | 87.79 | 74.41 | 87.83 | 91.11 | 81.81 | 82.64 | 83.16 | 83.96 |
| MiniGPT-v2 7B (Chen et al., 2023a) | 79.97 | 85.12 | 74.45 | 88.69 | 91.65 | 85.33 | 84.44 | 84.66 | 84.29 |
| MiniGPT-v2 7B-chat (Chen et al., 2023a) | 79.58 | 85.52 | 73.32 | 88.06 | 91.29 | 84.30 | 84.19 | 84.31 | 83.70 |
| Qwen-VL-7B (Bai et al., 2023) | 83.12 | 88.25 | 77.21 | 89.36 | 92.26 | 85.34 | 85.58 | 85.48 | 86.45 |
| Qwen-VL-7B-Chat (Bai et al., 2023) | 82.82 | 88.59 | 76.79 | 88.55 | 92.27 | 84.51 | 85.96 | 86.32 | 85.74 |
| <i>SPHINX</i> | 82.77 | 87.29 | 76.85 | 89.15 | 91.37 | 85.13 | 84.87 | 83.65 | 84.12 |
| <i>SPHINX-1k</i> | 86.64 | 91.08 | 80.35 | 91.05 | 92.65 | 86.56 | 88.19 | 88.35 | 88.14 |
| <i>SPHINX-2k</i> | 85.51 | 90.62 | 80.45 | 91.10 | 92.88 | 87.07 | 88.07 | 88.65 | 88.04 |

Table 4: Performance comparison in 8 REC benchmarks.

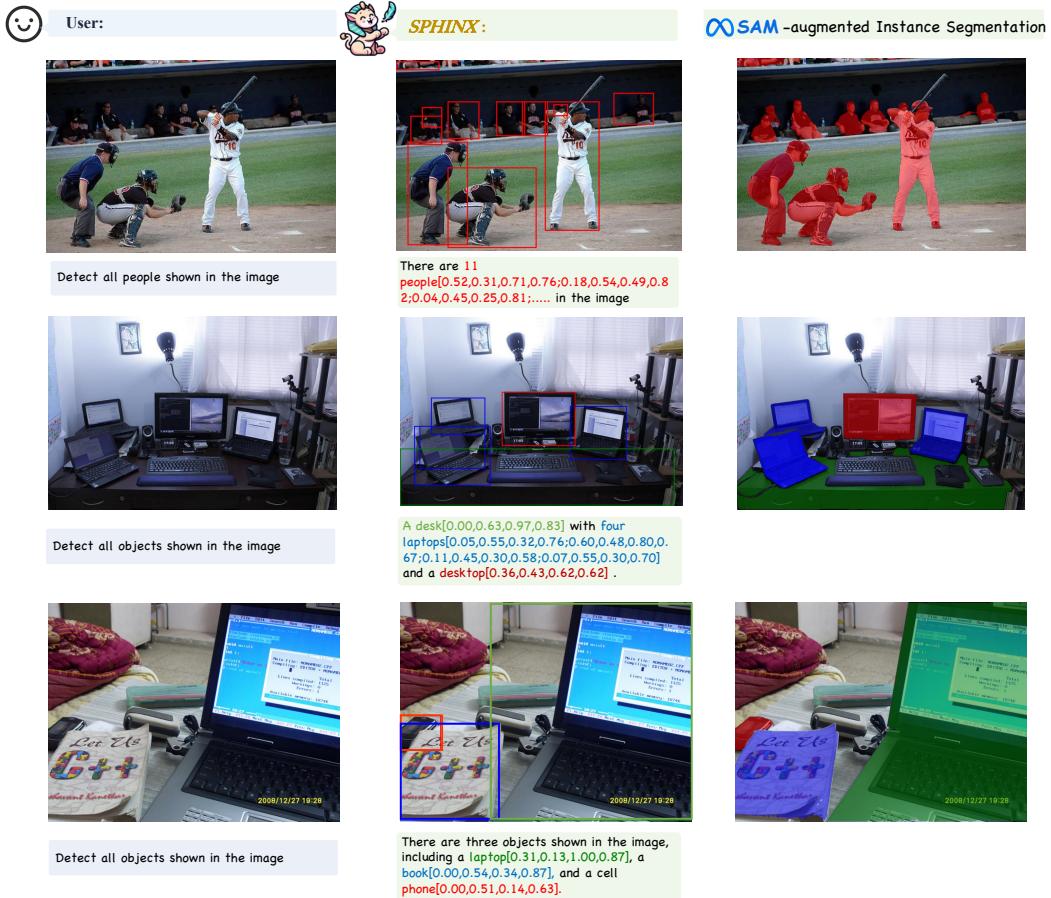


Figure 8: Examples of **SPHINX** integrating with Segment Anything Model (SAM) (Kirillov et al., 2023) for language-referred segmentation.

SAM-augmented instance segmentation. We integrate **SPHINX** with SAM to enhance segmentation capabilities. This integration involves detecting bounding boxes for the target objects and subsequently providing the bounding box coordinates to SAM for the generation of segmentation masks. The results, depicted in Figure 8, showcase a notable performance improvement achieved

Human Pose Estimation

User: Detect the key points of the five people in the region [0.38,0.39,0.57,0.64];[0.63,0.37,0.81,0.65];[0.82,0.36,0.94,0.65];[0.23,0.35,0.34,0.64];[0.04,0.38,0.16,0.64].



SPHINX: The detected key points are nose[0.46,0.44], left eye[0.47,0.43], right eye[0.46,0.43], left ear[0.48,0.43], right ear[0.45,0.43], left shoulder[0.50,0.45], right shoulder[0.44,0.45], left elbow[0.53,0.43], right elbow[0.41,0.43], left wrist[0.55,0.41], right wrist[0.39,0.40], left hip[0.50,0.53], right hip[0.47,0.53], left knee[0.54,0.59], right knee[0.48,0.60], left ankle[0.50,0.62], and right ankle[0.42,0.61].....

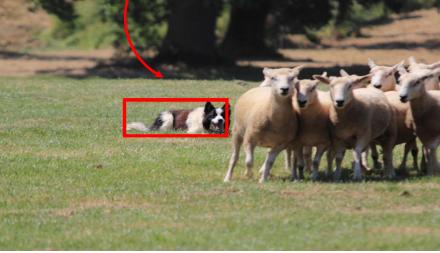


Relationship Detection

User: what is the [0.10,0.44,0.59,0.76] trying to do?



User: what is [0.28,0.44,0.51,0.53] intending to do?



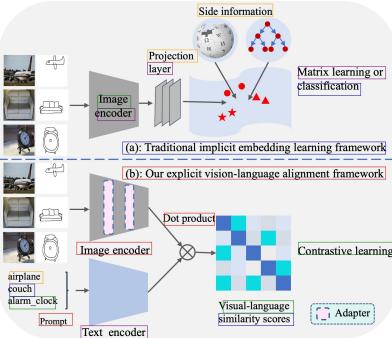
SPHINX: The bird is trying to eat the food, specifically the sandwich, that the baby is holding.

SPHINX: The dog, which is a black and white sheepdog, is likely intending to herd the sheep in the field. Sheepdogs are trained to work with sheep and are responsible for guiding, protecting, and controlling the flock. The dog's position in the field, along with the presence of the sheep, suggests that it is fulfilling its herding duties.

Accurate Diagram Understanding

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

SPHINX: layer[0.37, 0.24, 0.43, 0.27].Image [0.25, 0.32, 0.33, 0.35].couch[0.04, 0.77, 0.1, 0.8].Prompt[0.11, 0.85, 0.19, 0.88].airplane[0.04, 0.74, 0.12, 0.77].encoder[0.25, 0.35, 0.34, 0.38].alarm_clock[0.04, 0.8, 0.16, 0.83].Dot product[0.41, 0.6, 0.53, 0.63].classification[0.73, 0.29, 0.88, 0.32].Projection[0.37, 0.21, 0.49, 0.24].Text encoder[0.21, 0.88, 0.37, 0.91].Visual-language [0.54, 0.81, 0.72, 0.84].Image encoder[0.2, 0.68, 0.38, 0.71].similarity scores[0.54, 0.84, 0.72, 0.87].Side information[0.49, 0.1, 0.67, 0.13].Matrix learning or [0.73, 0.25, 0.94, 0.28].Contrastive learning[0.73, 0.68, 0.96, 0.71].a: Traditional implicit embedding learning framework and b: Our explicit vision-language alignment framework.



User: Generate a detailed description about the image.

SPHINX: The image depicts a diagram that illustrates a traditional implicit embedding learning framework and an explicit vision-language alignment framework. The traditional framework consists of a series of steps, including image encoding, projection, and matrix learning or classification. The explicit framework, on the other hand, consists of a text encoder, a dot product, and contrastive learning. The diagram also includes various images, such as a clock, a couch, and a chair, as well as a plane and a car.

Figure 9: Examples of **SPHINX** for human pose estimation (Top), relationship reasoning (Middle), and diagram understanding (Bottom).

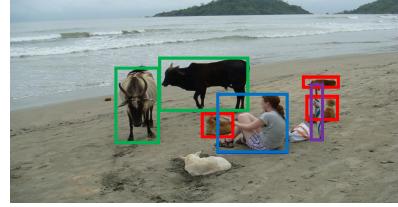
Improve Object Detection through Hints



User:
Detect all objects shown in the image.



SPHINX: There are seven objects shown in the image, including a umbrella[0.77,0.46,0.80,0.63], two cows[0.38,0.38,0.61,0.54;0.27,0.41,0.38,0.64], a person[0.53,0.49,0.71,0.67], and three dogs[0.76,0.50,0.84,0.57;0.49,0.55,0.57,0.62;0.75,0.44,0.84,0.47].



User:
There are four dogs shown in the image. Detect all dogs shown in the image.



SPHINX: There are four dogs[0.76,0.51,0.85,0.57;0.76,0.44,0.84,0.47;0.43,0.66,0.57,0.74;0.50,0.55,0.57,0.62] in the image.



Better Visual Reasoning with Object Detection



User: Describe the image concisely. Include the bounding box for each mentioned object.



SPHINX: A car[0.04,0.34,0.99,0.71] with wooden parts[0.26,0.41,0.94,0.68] is parked on the side of the road[0.00,0.50,1.00,0.71].



User: Why this picture funny?



SPHINX: The picture is funny because it is unusual to see a car[0.04,0.34,0.99,0.71] with wooden parts[0.26,0.41,0.94,0.68]. The wooden parts give the car a unique and quirky appearance, making it stand out and evoke laughter from viewers.



User: Describe the image concisely. Include the bounding box for each mentioned object.



SPHINX: A man[0.43,0.25,0.65,0.65] in a yellow shirt[0.43,0.29,0.63,0.46] is ironing clothes[0.34,0.44,0.65,0.59] on the back of a car[0.46,0.37,1.00,0.75].



User: Why this picture funny?



SPHINX: This picture is funny because it is an unusual and unconventional sight to see someone ironing clothes[0.34,0.44,0.65,0.59] on the back of a car[0.46,0.37,1.00,0.75]. Ironing is typically done indoors, in a dedicated space like a laundry room or a bedroom. The fact that the man is ironing clothes[0.34,0.44,0.65,0.59] on the back of a car[0.46,0.37,1.00,0.75] adds a humorous and quirky element to the scene.

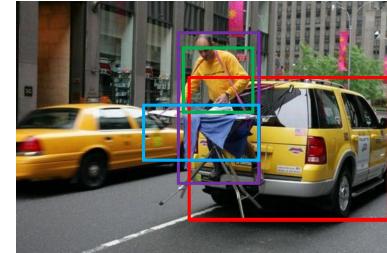


Figure 10: Examples of **SPHINX** for improved object detection through hints and better visual reasoning with object detection.

through the collaboration of **SPHINX** and SAM. Surprisingly, We observe that the predicted masks for small objects are extremely accurate such as the cell phone in the last row. The synergistic application of **SPHINX** and SAM underscores the considerable potential inherent in our methodology.

Region-level understanding. In Figure 9, the performance of **SPHINX**’s detection capabilities is showcased. The upper row displays the synchronized jumping of five teenagers, each assuming distinct poses. Notably, **SPHINX** accurately predicts the pose with key points for each individual, leaving no participant overlooked. The middle row illustrates the **SPHINX**’s reasoning ability focus on specified region. We observe that **SPHINX** successfully recognize the desired objects and detailed answer the question. The bottom row indicates **SPHINX**’s superior diagram understanding ability, which produces accurate layout detection and content comprehension.

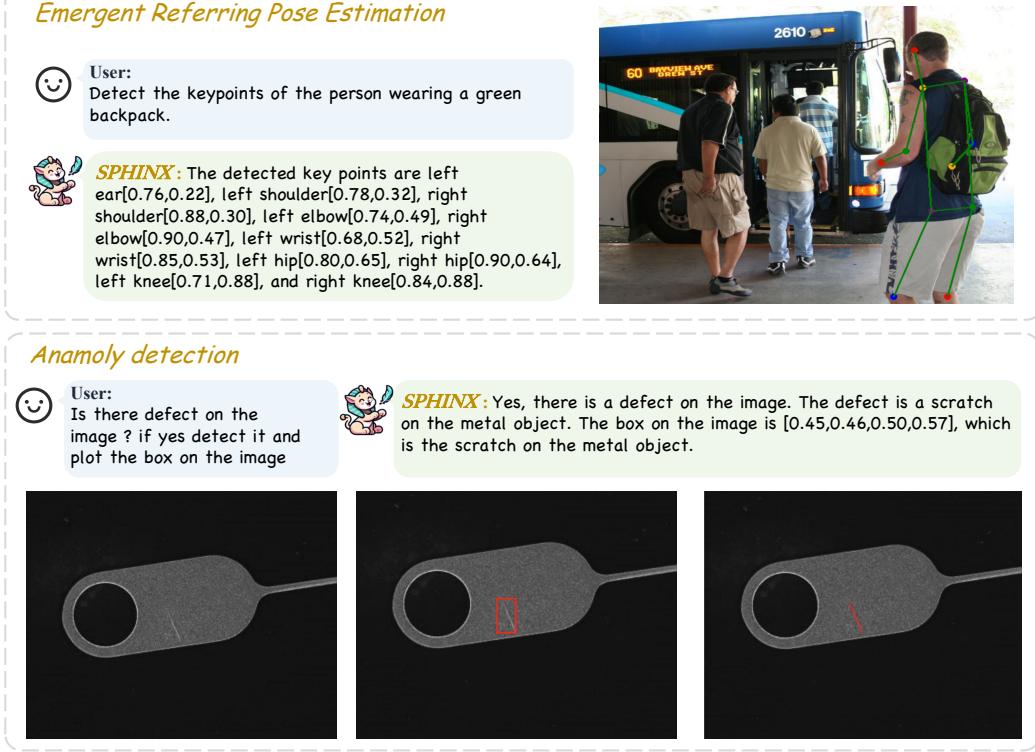


Figure 11: Examples of **SPHINX** for emergent referring pose estimation (Top) and the challenging anomaly detection (Bottom).

Better visual reasoning with object detection. The enhanced visual reasoning capabilities of our model with object detection are showcased in Figure 10. Notably, **SPHINX** leverages the object detection feedback by initially instructing **SPHINX** to generate object detection results and then requesting it to answer questions based on localization outcomes. The model will prioritize selecting the most relevant objects for coordinate feedback based on the query content, rather than all detected objects. This underscores the idea that in multitask training, the synergy between different tasks can significantly enhance overall performance. Furthermore, the model exhibits commendable Contextual Understanding (COT) by effectively integrating information from diverse elements in the image, resulting in more powerful reasoning ability.

Improved object detection through hints. We highlight **SPHINX**'s proficiency in understanding user hints. As depicted in Figure 10, initially requesting the model to predict all dogs in the image leads to misidentification of other objects. However, upon offering additional hints about the desired object, **SPHINX** demonstrates an improved comprehension of instructions and accurately predicts all dogs in the image.

Emergent referring pose estimation. The original referring object comprehension and pose estimation are two different tasks, where the former detects object bounding boxes according to textual descriptions, and the latter outputs human keypoints from given bounding boxes. Interestingly, as shown in Figure 11 (Top), by our mixing of the two tuning tasks, our **SPHINX** acquires the emergent capacity for referring pose estimation, i.e., generating human keypoints directly from textual descriptions. Such an observation indicates that our **SPHINX** fully comprehend the semantics across different vision-language tasks, and implicitly connect them via superior reasoning power.

Anomaly detection. It is important for industrial monitoring and healthcare to detect rare events or outliers that may indicate abnormal or suspicious behavior. As shown in Figure 11 (Bottom), our **SPHINX** also excels in anomaly detection. Although we do not explicitly involve related training

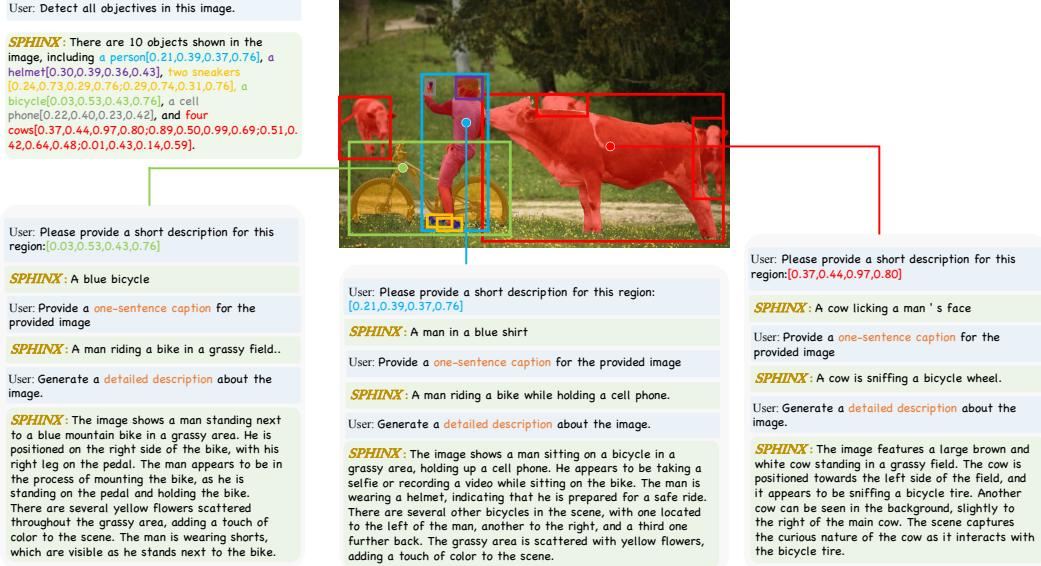


Figure 12: Examples of multi-level dense captioning by **SPHINX**.

data, our MLLM still demonstrates superior localization accuracy for unsharp defects. This indicates wide potentials of **SPHINX** in real-world applications.

Multi-level dense captioning. Endowed with diverse multi-task pre-training, **SPHINX** can perform multi-level dense captioning by iterative promoting itself. Given an input image, prompting **SPHINX** with “Detect all objects shown in the image” can localize the position of all objects. Then, we iteratively prompt each detected region with “Please provide a short description for this region : [x1, y1, x2, y2]” to extract a simple property on the localized region. To get a deeper understandings on the detected regions, we crop all images based on the detection results. Each cropped view is feed independently into **SPHINX** with two prompts, namely, “Provide a one-sentence caption for the provided image.” and “Generate a detailed description about the image.”. By doing so, we can detect all objects shown in the image and densely label all boxes with property, simple caption and detailed caption. The multi-level dense captioning results is illustrated on Figure 12.

5 CONCLUSION

In this paper, we propose **SPHINX**, a versatile multi-modal large language model (MLLM) with multi-purpose visual instruction-following capabilities. In our MLLM, we introduce a joint mixing of three different aspects: model weights of pre-trained LLMs by real-world and synthetic data, tuning tasks for diverse visual perception and reasoning tasks, and visual embeddings from different types of vision backbones. On top of this, we further devise to endow **SPHINX** with the capacity to process high-resolution images by mixing different visual scales and sub-images, which exhibits superior fine-grained visual understanding performance. Via our proposed three-fold mixing strategy, **SPHINX** achieves impressive performance over a wide range of multi-modality evaluation benchmarks, and can serve as a strong vision generalist to tackle object detection, region-level captioning, and human pose estimation, etc. Our MLLM can also be integrated with other visual foundation models for wider functionalities, e.g., SAM (Kirillov et al., 2023) for language-referred segmentation and Stable Diffusion (Rombach et al., 2021) for image editing. Our future work will focus on incorporating a wider range of vision-language tasks into **SPHINX** for all-purpose capabilities.

REFERENCES

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer*

Vision, 123:4 – 31, 2015.

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. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL <https://www.adept.ai/blog/fuyu-8b>.

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. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18030–18040, 2022.

Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023a.

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

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

Richard Vencu Theo Coombes Romain Beaumont Christoph Schuhmann, Andreas Köpf. Laion-coco. <https://laion.ai/blog/laion-coco/>, 2022.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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 preprint arXiv:2306.13394*, 2023a.

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 preprint arXiv:2306.13394*, 2023b.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023a.

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 preprint arXiv:2304.15010*, 2023b.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Google. Bard. <https://bard.google.com/>, 2023.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Ke Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *ArXiv*, abs/2309.00615, 2023.

Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016b.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *arXiv preprint arXiv:2307.12981*, 2023.

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

Drew A. Hudson and Christopher D. Manning. 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.

Sahar Kazemzadeh, Vicente Ordonez, Marc Andre Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

-
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, C. Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *ArXiv*, abs/2306.05425, 2023a.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726, 2023b.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023c.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 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 preprint arXiv:2301.12597*, 2023d.
- 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*, 2023e.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.
- Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023a.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023b.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023c.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023d.
- Siyi Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023e.
- 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*, 2023f.
- 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*, 2023g.
- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

-
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana-Maria Camбуру, Alan Loddon Yuille, and Kevin P. Murphy. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2015.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 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.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019.
- OpenAI. Chatgpt. <https://chat.openai.com>, 2023a.
- OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023b.
- 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.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023a.
- 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 preprint arXiv:2306.14824*, 2023b.
- Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74 – 93, 2015.
- Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022.

ShareGPT. Sharegpt. <https://sharegpt.com/>, 2023.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. *ArXiv*, abs/2003.12462, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318, 2019.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *ArXiv*, abs/2305.16355, 2023.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*, 2023a.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023b.

Song Wen, Guian Fang, Renrui Zhang, Peng Gao, Hao Dong, and Dimitris Metaxas. Improving compositional text-to-image generation with large vision-language models. *arXiv preprint arXiv:2310.06311*, 2023.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

-
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointilm: Empowering large language models to understand point clouds. *ArXiv*, abs/2308.16911, 2023.
- B. Yan, Yi Jiang, Jiannan Wu, D. Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15325–15336, 2023.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023a.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490, 2023b.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023a.
- Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *CVPR 2023*, 2023b.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023.
- Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*, 2023.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022.