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

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pp. 202–227. PMLR, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023b.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023a.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023b.

A Description of Evaluation Benchmarks

- **MM-Vet** (Yu et al., 2023) dataset is a benchmark designed to evaluate large vision-language models (LVLMs) across six core vision-language (VL) capabilities: recognition, knowledge, optical character recognition (OCR), spatial awareness, language generation, and mathematical reasoning. The dataset includes open-ended, real-world questions based on image-text pairs, requiring models to integrate multiple capabilities to solve complex tasks. MM-Vet benchmark consists of 200 images paired with 218 open-ended questions.
- **Q-Bench** (Wu et al., 2023) evaluates the capabilities of large vision-language models in three main areas related to low-level vision tasks. These tasks focus on evaluating how well LVLMs can perform basic low-level perception tasks that are traditionally associated with human visual perception. In the Q-Bench dataset, the questions are of three types: Yes-or-No, What, and How.
  - **Low-Level Visual Perception**: Assesses how accurately LVLMs can answer questions about low-level image attributes (e.g., clarity, color, distortion). LLVisionQA dataset includes 2,990 images, each with a corresponding question about low-level features.
  - **Low-Level Visual Description**: Evaluates the ability of LVLMs to describe images. LLDescribe dataset has 499 images with expert-labeled descriptions averaging 58 words each. LVLMs are compared against these to assess completeness, preciseness, and relevance.
  - **Visual Quality Assessment**: Evaluates LVLMs' ability to predict quantifiable quality scores for images by assessing how well they align with human-rated mean opinion scores (MOS) on low-level visual appearances, using 81,284 samples.
- **SQA-IMG** (Lu et al., 2022a) is a portion of the Science Question Answering (SQA) dataset that contains questions from a wide range of scientific domains, each paired with corresponding image contexts. The dataset includes 10,332 examples of multimodal multiple-choice questions, along with lectures and explanations that detail the reasoning behind the correct answers.

- **ChartQA** (Masry et al., 2022) dataset is a benchmark designed to test AI models on their ability to perform question-answering tasks over various types of charts. It focuses specifically on questions requiring complex reasoning, such as visual and logical interpretation, going beyond simpler template-based datasets. ChartQA includes 9,608 human-authored open-ended questions as well as 23,111 questions that are automatically generated from chart summaries.
- **SEED-IMG** (Li et al., 2023), a subset of SEED-Bench, focuses on evaluating spatial comprehension of images by testing models on various dimensions like scene understanding, object identification, and spatial relationships. In terms of scale, the dataset includes 19,000 multiple-choice questions that evaluate both image and video comprehension, covering 12 evaluation dimensions such as scene understanding, instance identity, spatial relations, and action recognition.
- **MME** (Fu et al., 2023) evaluates both perception and cognition abilities of LVLMs. It features 14 subtasks, including recognition tasks (such as object existence, count, position, color) and reasoning tasks (such as commonsense reasoning, numerical calculation, text translation, and code reasoning). MME uses manually created instruction-answer pairs, ensuring no overlap with public datasets. MME uses "yes/no" responses for quantitative evaluations.
- **MathVista** (Lu et al., 2023) is a benchmark designed to evaluate the mathematical reasoning capabilities of foundation models in visual contexts. It integrates challenges from diverse mathematical and visual tasks, with a focus on fine-grained, deep visual understanding and compositional reasoning. MathVista consists of 6,141 examples including 3,392 multiple-choice questions and 2,749 free-form questions derived from 28 existing multimodal datasets and 3 newly created datasets: IQTest, FunctionQA, and PaperQA.
- **LLaVA-W** (Liu et al., 2024c) is a challenging evaluation benchmark created to assess the generalization and instruction-following capabilities LVLMs in complex, real-world situations. It consists of 24 images and 60 questions, including diverse scenes like indoor environments, outdoor settings, memes, paintings, and sketches. Each image is associated with a highly detailed and manually curated description, and the questions focus on extracting intricate details and reasoning about the visual content. LLaVA-W involves a variety of tasks, including detailed descriptions, conversational answers, and complex reasoning.
- **MMStar** (Chen et al., 2024a) is a vision-dependent multimodal benchmark designed to evaluate the multimodal capabilities of LVLMs. It addresses two main issues identified in previous benchmarks: the reliance on textual information without visual input and data leakage during training. MMStar is composed of 1,500 samples carefully selected to ensure that visual content is necessary for solving each problem. MMStar evaluates six core capabilities across 18 detailed axes, which include tasks like image perception and logical reasoning. MMStar uses multiple-choice as the primary answer type.
- **MMVP** (Tong et al., 2024) evaluates the visual grounding capabilities of large vision-language models by identifying scenarios where they fail to distinguish simple visual patterns in images. These patterns include aspects like orientation, counting, viewpoint, and relational context. The benchmark is constructed using 150 pairs of images, resulting in 300 multiple-choice questions.

B Description of Evaluation LVLMs

- **LLaVA-1.5** (Liu et al., 2024a) incorporates academic task-oriented datasets to enhance performance in VQA tasks and features an MLP vision-language connector, which improves upon the original linear layer utilized in LLaVA (Liu et al., 2024c). It uses CLIP ViT-L/14 (Radford et al., 2021) with a 336px resolution as its vision encoder, resulting in a total of $(336/14)^2$ = 576 visual tokens. LLaVA-1.5 is built on Vicuna with either 7B or 13B parameters. The training dataset includes 558K samples for pre-training and 665K for fine-tuning, totaling 1.2M image-text pairs from publicly available datasets
- **LLaVA-NeXT** (Liu et al., 2024b) (also known as LLaVA-1.6) enhances visual reasoning, OCR, and world knowledge, offering four times higher image resolution (up to 1344x336)

and improved performance in visual conversations. Its architecture includes a CLIP ViT-L/14 as a vision encoder, paired with Vicuna models ranging from 7B to 34B as a backbone language model. It utilizes 1.3M visual instruction tuning data samples for training, maintaining efficiency with approximately one day of training on 32 A100 GPUs. The architecture's high resolution and dynamic grid scheme improve detailed image processing capabilities.

- **LLaVA-OneVision** (Li et al., 2024b) is a LVLM designed for task transfer across single-image, multi-image, and video scenarios, with strong capabilities in video understanding through image-to-video task transfer. Its architecture consists of a Qwen2 language model (Yang et al., 2024) with 8B to 72B parameters, and the SigLIP vision encoder (Zhai et al., 2023), which processes images at a base resolution of 384x384, producing 729 visual tokens. The model employs a 2-layer MLP projector. The training utilized 3.2M single-image data samples and 1.6M multi-modal data samples, focusing on high-quality visual instruction tuning data to enhance its multimodal capabilities.
- **Meteor** (Lee et al., 2024c) is a large vision-language model that uniquely embeds multi-faceted rationales using a Mamba-based architecture (Gu & Dao, 2023), enabling efficient processing of lengthy rationales to enhance its vision-language understanding. This approach allows Meteor to achieve superior performance without scaling up model size or employing additional vision encoders. Its architecture includes a CLIP-L/14 vision encoder with an image resolution of 490x490, comprising 428M parameters, and InternLM2-7B (Cai et al., 2024) as a foundational LLM. Meteor was trained on 2.1M question-answer pairs, with 1.1M curated triples.
- **TroL** (Lee et al., 2024b) uses a unique characteristic called layer traversing, which reuses layers in a token-wise manner, allowing it to simulate retracing the answering process without physically adding more layers, making it efficient despite smaller model sizes. TroL uses CLIP-L and InternViT as vision encoders, containing 428M and 300M parameters, respectively, and supports 24 layers. The image resolution is adjusted using MLPs in the vision projector. For its foundational LLM, TroL utilizes Phi-3-mini with 3.8B parameters and InternLM2 with 1.8B and 7B parameters. The training dataset comprises 2.3M visual instruction tuning samples.