Comparison Visual Instruction Tuning

Wei Lin[1] Muhammad Jehanzeb Mirza[2] Sivan Doveh[3,4] Rogerio Feris[7] Raja Giryes[5]
Sepp Hochreiter[1,6] Leonid Karlinsky[7]
[1]ELLIS Unit, LIT AI Lab, Institute for Machine Learning, JKU Linz [2]MIT CSAIL [3]IBM Research
[4]Weizmann Institute of Science [5]Tel-Aviv University [6]NXAI GmbH [7]MIT-IBM Watson AI Lab

Project Page: https://wlin-at.github.io/cad_vi

Dataset Repo: https://huggingface.co/datasets/wlin21at/CaD-Inst

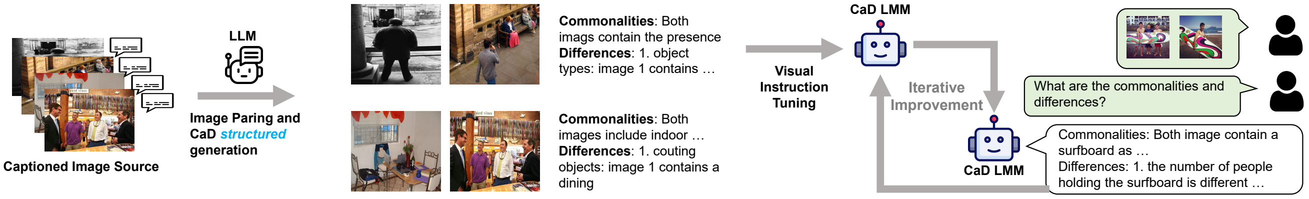

Captioned Image Source
LLM
Image Paring and CaD *structured* generation
Commonalities: Both imags contain the presence **Differences**: 1. object types: image 1 contains …
Commonalities: Both images include indoor … **Differences**: 1. couting objects: image 1 contains a dining
Visual Instruction Tuning
CaD LMM
Iterative Improvement
CaD LMM
What are the commonalities and differences?
Commonalities: Both image contain a surfboard as …
Differences: 1. the number of people holding the surfboard is different …

Figure 1. **CaD-VI concept**. We collect and pair densely captioned source images to form synthetic CaD instructions using an LLM. The resulting synthetic CaD Visual Instruction dataset is used to train our first CaD enabled LMM. It is then improved by iterative self-refinement where it is used to annotate new paired images from additional sources that are employed for its re-training. This results with both an improved CaD enabled LMM and a comprehensive CaD-Inst dataset (contributed in this work).

Abstract

Comparing two images in terms of Commonalities and Differences (CaD) is a fundamental human capability that forms the basis of advanced visual reasoning and interpretation. It is essential for the generation of detailed and contextually relevant descriptions, performing comparative analysis, novelty detection, and making informed decisions based on visual data. However, surprisingly, little attention has been given to these fundamental concepts in Large Multimodal Models (LMMs). We develop and contribute a new two-phase approach, CaD-VI, for collecting synthetic visual instructions, together with an instruction-following dataset, CaD-Inst, containing 349K image pairs with CaD instructions collected using CaD-VI. Our approach significantly improves the CaD spotting capabilities in LMMs, advancing the SOTA on a diverse set of related tasks by up to 17.5%. It is also complementary to existing difference-only instruction datasets, allowing automatic targeted refinement of those resources increasing their effectiveness for CaD tuning by up to 10%. Additionally, we propose an evaluation benchmark with 7.5K open-ended QAs to assess the CaD understanding abilities of LMMs.

1. Introduction

Understanding the Commonalities and Differences (CaD) between two signals (e.g., images) is a basic capability innate to humans [18]. Spotting changes and differences alerts us to interesting events happening in our surroundings, warns us of hazard, and drives us toward learning new concepts exposed after the change or relative movement. Understanding what is common helps structure visual information and allows differences to emerge by elimination. Together, these form powerful tools for human learning and acquiring world knowledge.

The forefront of modern AI shifted with the recent emergence of foundation Large Language Models (LLMs) [4], where the top-performing ones [1, 2, 8, 9] closely align to human reasoning and world-knowledge capabilities. LLMs' great performance and wide applicability quickly led to their wide adoption into most of the current ML pipelines. In the computer vision community, this impacted the development of Large Multi-modal Models (LMMs) [7, 8, 16, 26, 31, 38, 42]. While multiple methods for adding multi-modal support to LLMs have been proposed, currently the more popular and better performing open LMMs largely rely on tuning using Visual Instructions (VI) [31, 52]. These methods align image tokens produced by visual encoders to be 'understandable' by an LLM decoder, allowing images to be seamlessly integrated into the LLM decoder input context stream together with the query text during inference.

In recent methods [7, 16, 26, 31], VI takes the form of a multi-turn conversation: with 'human turns' providing im-

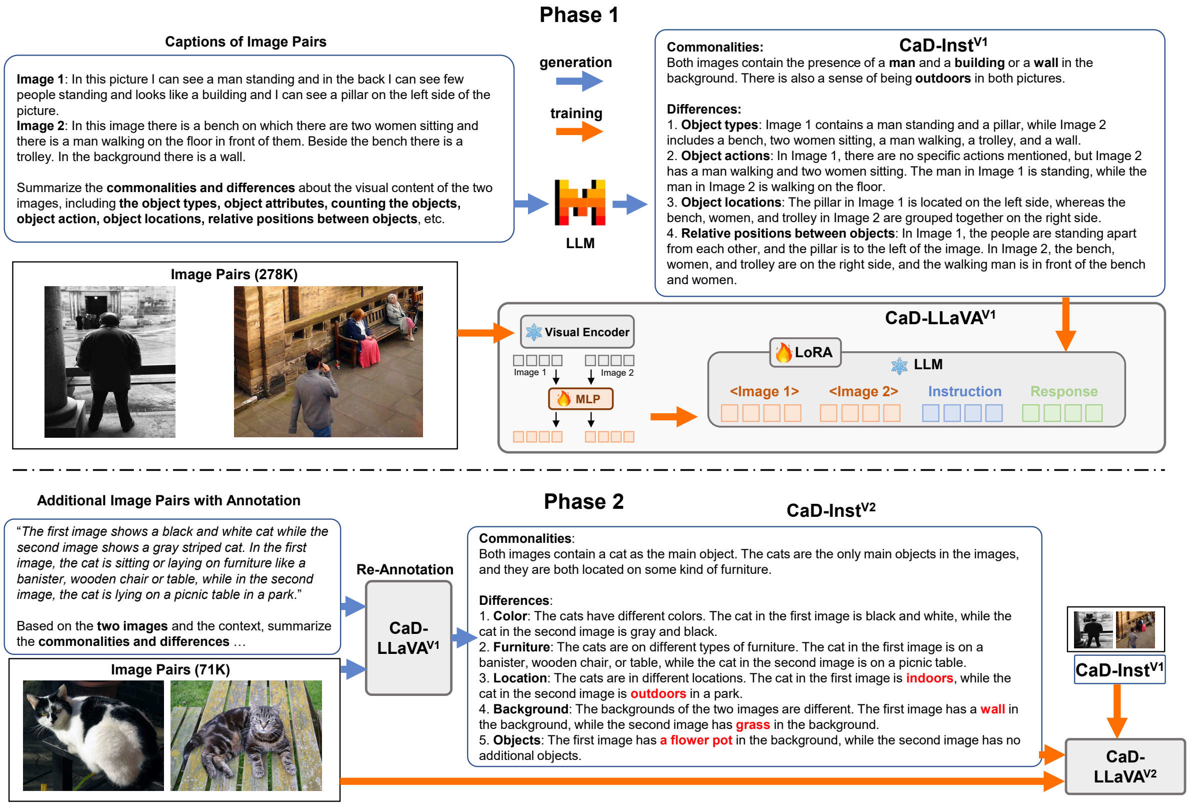


Image 1: In this picture I can see a man standing and in the back I can see few people standing and looks like a building and I can see a pillar on the left side of the picture.
Image 2: In this image there is a bench on which there are two women sitting and there is a man walking on the floor in front of them. Beside the bench there is a trolley. In the background there is a wall.

Summarize the **commonalities and differences** about the visual content of the two images, including **the object types, object attributes, counting the objects, object action, object locations, relative positions between objects**, etc.

Commonalities:
Both images contain the presence of a **man** and a **building** or a **wall** in the background. There is also a sense of being **outdoors** in both pictures.

Differences:
1. **Object types**: Image 1 contains a man standing and a pillar, while Image 2 includes a bench, two women sitting, a man walking, a trolley, and a wall.
2. **Object actions**: In Image 1, there are no specific actions mentioned, but Image 2 has a man walking and two women sitting. The man in Image 1 is standing, while the man in Image 2 is walking on the floor.
3. **Object locations**: The pillar in Image 1 is located on the left side, whereas the bench, women, and trolley in Image 2 are grouped together on the right side.
4. **Relative positions between objects**: In Image 1, the people are standing apart from each other, and the pillar is to the left of the image. In Image 2, the bench, women, and trolley are on the right side, and the walking man is in front of the bench and women.

“*The first image shows a black and white cat while the second image shows a gray striped cat. In the first image, the cat is sitting or laying on furniture like a banister, wooden chair or table, while in the second image, the cat is lying on a picnic table in a park.*”

Based on the **two images** and the context, summarize the **commonalities and differences** …

Commonalities:
Both images contain a cat as the main object. The cats are the only main objects in the images, and they are both located on some kind of furniture.

Differences:
1. **Color**: The cats have different colors. The cat in the first image is black and white, while the cat in the second image is gray and black.
2. **Furniture**: The cats are on different types of furniture. The cat in the first image is on a banister, wooden chair, or table, while the cat in the second image is on a picnic table.
3. **Location**: The cats are in different locations. The cat in the first image is **indoors**, while the cat in the second image is **outdoors** in a park.
4. **Background**: The backgrounds of the two images are different. The first image has a **wall** in the background, while the second image has **grass** in the background.
5. **Objects**: The first image has **a flower pot** in the background, while the second image has no additional objects.

Figure 2. Pipeline of our two-phase CaD-VI: In Phase-1, we leverage captions for image pairs and an LLM to generate CaD VI data - CaD-Inst^{V1} (278K). We perform visual instruction tuning on it to arrive at the Phase-1 model: CaD-LLaVA^{V1}. In Phase-2, we leverage CaD-LLaVA^{V1} to generate CaD VI data on additional image pairs and collect CaD-Inst^{V2} (71K). Visual instruction tuning with CaD-Inst^{V1} and CaD-Inst^{V2} leads to our final model CaD-LLaVA^{V2}.

age context and asking the questions, and ‘LMM turns’ answering them [31]. Yet, the majority of VI data focused on providing merely a single image in the VI conversations [31]. Only few works included multi-image VI samples [3, 38], and surprisingly, very few used some form of CaD VI data [16, 25, 26] to add CaD support in the LMM.

Due to the fundamental importance of endowing LMMs with CaD capabilities, thus getting them closer to human capabilities, we propose CaD-VI- a multi-phase CaD generation approach, for progressive dense and structured CaD VI data collection (concept shown in Fig. 1), which we employ to build CaD-Inst training curriculum and associated CaD-QA benchmark comprised of CaD-related open-ended questions, both contributed in this work. In essence, the final CaD-Inst curriculum associates diverse and large-scale (349K) image pair collection with highly detailed and structured CaD summaries. CaD summaries computed for an additional set of 7.6K image pairs, are used for extracting open CaD-related QA resulting in CaD-QA.

As shown in Fig. 2, the Phase-1 of CaD-VI is a ‘cold start’ where, in the absence of LMMs with substantial CaD capabilities, we leverage image captions and an LLM to hallucinate (coarse) CaD VI data - CaD-Inst^{V1} (278K), where we collect *structured* and *detailed* CaD summaries for our paired images sourced from a dense & large-scale image collection [34]. Training on the first phase CaD-Inst^{V1} data we arrive at CaD-LLaVA^{V1}- an LMM that has strong CaD capabilities compared to a large variety of leading LMMs including the very few trained with some CaD data (see Sec. 5). Next, leveraging our CaD-LLaVA^{V1} model to produce non-hallucinated, image-informed CaD data, we generate additional CaD instructions into the collection CaD-Inst^{V2} (71K). Combining CaD-Inst^{V1} and CaD-Inst^{V2} we form CaD-Inst and train our final CaD-LLaVA^{V2} 7B and 13B LMMs to achieve (a) significant (up to 17.5%) absolute improvement over a large variety of recent SOTA LMMs employing a variety of 5 CaD-related existing closed-QA evaluation benchmarks (namely BISON[15], SVO Probes[13], NLVR2[37], EQBEN[41], and COLA[36]), and (b) strong (up to over 20%) relative improvements on our contributed open-QA CaD benchmark - CaD-QA. Additionally, as CaD-Inst can be safely mixed with the LLaVA VI data [30], we show in Tab. 4 that our CaD-LLaVA^{V2} models effectively avoid forgetting the general capabilities of the corresponding LLaVA LMMs.

Our contributions are as follows: (i) we contribute CaD-Inst- a large-scale visual instruction tuning dataset for enhancing CaD reasoning capabilities of LMMs; (ii) we contribute CaD-QA- an open QA evaluation benchmark for assessing CaD capabilities; (iii) we contribute and open source a CaD-VI methodology for collecting CaD instruction tuning data and re-purposing datasets with existing difference annotations; (iv) we demonstrate significant (up to 17.5%) improvements in CaD reasoning for LMMs trained using CaD-Inst as well as potential to scale CaD-Inst via self-improvement by CaD-Inst-trained models.

2. Related Work

Large Multimodal Models. LMMs have shown significant advancements in integrating visual and textual data, enhancing the ability of deep neural networks to understand and generate multimodal content. BLIP-2 employs a bootstrapping approach that leverages frozen image encoders and large language models through a querying transformer, achieving remarkable results on various vision-language tasks with fewer parameters compared to previous models [27]. Similarly, MiniGPT-4 [51] and LLaMA-Adapters [45] utilize pretrained visual and language models, with adapters aligning image tokens to language tokens, improving the efficiency and performance of multimodal understanding and generation. In addition to these early models, the LLaVA series [31], including LLaVA 1.5 [30] and LLaVA 1.6 [32], have enhanced visual instruction tuning, enabling better handling of single-image inputs and more accurate multimodal outputs. The InternLM XComposer 2.0 VL [44], EMU2 [39], Otter [26], SparklesChat [16], and MMICL [48] extend these capabilities by incorporating multiple images as input, thereby enriching the models' understanding and generation of text based on complex visual scenes. These models showcase the evolution from single-image to multi-image inputs, highlighting the progress in multimodal learning architectures and applications.

Visual Instruction Tuning Datasets. The success of LMMs builds on the collection of high-quality visual instruction tuning data, either constructed from existing VQA datasets [6, 11, 12, 17, 28], curated image-text pairs [51] and LLM-generated instruction-following data with input of rich human annotations [25, 30, 31, 46, 47]. However, the collection of multimodal data for learning commonalities and differences between two images is still under-explored.

Image Commonalities and Differences. Only a few datasets contain difference-only related annotation [19, 25]. Spot-the-diff [20] collects human-annotated short change descriptions for surveillance video frames. Our CaD-Inst^{V1} data collection is partially inspired by the differences-only data collection done by [25] as a small part of their VI strategy. However, different from [25] we: (i) collect both differences *and commonalities* (compared to only differences in [25]); (ii) we leverage a significantly more *dense* caption-source of [34] compared to [5] used in [25]; (iii) we are *structuring* our differences in CaD according to 6 axes (whichever applicable on case basis) - object types, attributes, counting, actions, locations, and relative positioning, also explicitly asking the LLM to extract (from the dense captions) information along these axes, while [25] produced unstructured difference description text; (iv) unlike [25] we are not relying on the existence of manually collected object bounding boxes; (v) the scale of our data is approx. 4 times larger than of [25]. Due to these differences, as evident from the direct comparison in Tab. 5, training the same model on CaD-Inst^{V1} has significant performance advantages over training on CaD instructions of [25]. To summarize, our work focuses on CaD understanding, largely neglected by the visual instruction tuning community. We propose a new CaD-VI approach for collecting synthetic visual instructions and enhancing the CaD analysis capabilities in LMMs. CaD-VI not only advances the state-of-the-art in related tasks by significant margins but also complements existing datasets [19, 25] by enabling their automatic targeted refinement, thereby improving their effectiveness for CaD tuning.

3. Two-Phase Visual Instruction Tuning

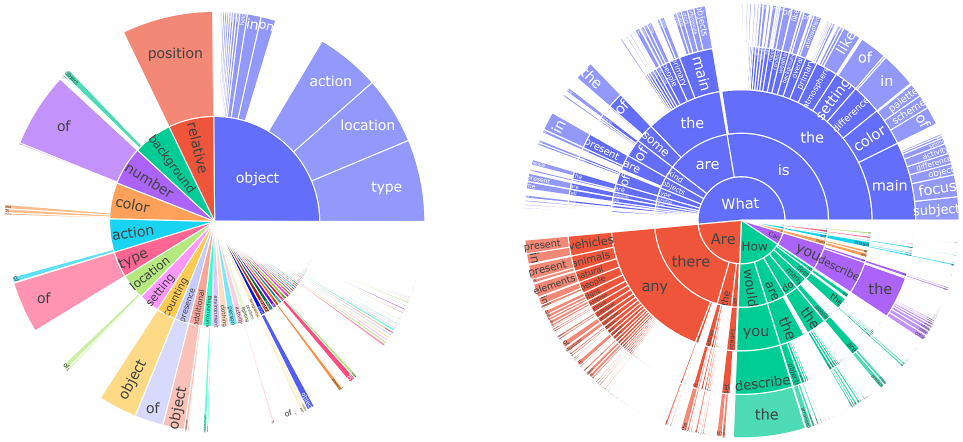
As illustrated in Fig. 2, our CaD-VI consists of two phases: in Phase-1, we employ an LLM to generate summary of CaD for image pairs (Sec. 3.1) and perform visual instruction tuning on the collected data (Sec. 3.2); in Phase-2, we leverage the Phase-1 model to generate CaD on additional image pairs and perform training with combined instruction data from both phases (Sec. 3.3).

3.1. Phase-1a: LLM Instruction Data Collection

In our first phase, we leverage an LLM to generate a summary of commonalities and differences for a pair of two images, as shown in Fig. 2 (top row). Specifically, we construct image pairs and prompt an LLM, supplying it with two image captions (one per image) and an instruction prompt asking it to summarize all the commonalities and differences according to the provided captions, contributing to our first phase CaD instruction data collection denoted as CaD-Inst^{V1}.

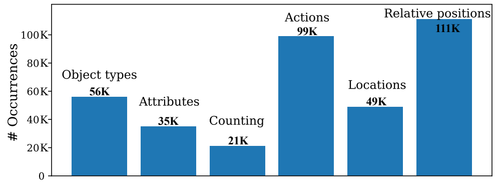
Image Source. We select the Localized Narratives dataset [34] which consists of 873K image-caption pairs with diverse samples sourced from COCO [5, 29], Flickr30K [43], ADE20K [50] and Open Images [23]. The captions are generated by transcription from spoken descriptions of the image content, which are dense, detailed, and descriptive with an average length of 36.5 words. To cover comprehensive visual contents and increase the diversity in terms of commonalities and differences, we collect 278K image pairs with different levels of similarity between their captions. We compute similarity by counting the number of overlapping nouns in the corresponding captions.

LLM Data Generation. In this work, we focus on employing open-source foundation models for data collection. The



(a) Characteristics in CaD summaries

(b) Questions in CaD-QA



(c) Axis counts in CaD summaries

User: Image 1: **<image>** Image 2: **<image>** Summarize the commonalities and differences about the visual content of the two images, including the object types, object attributes, counting the objects, object actions, object locations, relative positions between objects, etc.
Assistant: <summary>
User: Which image suits the caption **<chosen_caption>** better? A. Image 1 B. Image 2 Answer with the option's letter from the given choices directly.
Assistant: <choice>

(d) Instruction template

Figure 3. (a) Distribution of characteristics (first two words) in the CaD summary collected in CaD-Inst^{V1}; (b) Distribution of question types (first five words) in the evaluation benchmark CaD-QA; (c) Axis counts in CaD summaries; (d) Two-turn conversation template.

current open-source LMMs do not have strong capabilities of visual reasoning and instruction following when processing multiple input images. In this case, using caption as a symbolic representation of each image and employing an LLM with strong text instruction-following ability for generation of comparison summary of multiple input images is a more robust way of data collection than using open-source LMMs. The practice of this data collection pipeline with LLMs and dense captions is verified in the original LLaVA [31] and many following works [16, 25, 46].

We leverage the Mixtral 8×7B LLM [21] for generating detailed and structured summaries of commonalities and differences for pairs of images. As the LLM can only accept text as input, in Phase 1 we use image captions to represent visual content of images. This is a rather crude approximation, which is alleviated in Phase 2 of our CaD-VI approach. To encourage the diverse and creative generation of commonalities and differences, we do not provide in-context examples of expected output in the prompt to the LLM. Furthermore, we specifically prompt the LLM *to structure* the commonalities and differences summaries according to the following 6 visual aspects: (i) object types; (ii) attributes; (iii) counts; (iv) actions; (v) locations; and (vi) relative positions; as illustrated in Fig. 2. We provide detailed prompts in the supplementary. Importantly, LLM is not forced to produce all 6 aspects in every summary; they are generated adaptively according to the available content.

Generated Data Statistics. In CaD-Inst^{V1} we collected structured summaries of CaD for 278K image pairs, with average length of 157 words (40 for commonalities and 117 for differences). The summaries are structured according to 6 axes, appearing unevenly on a case-to-case basis based on the LLM decision. We illustrate the distribution of data characteristics in Fig. 3(a), and the total observed axis counts in Fig. 3(c). More statistics and details are provided in the supplementary.

CaD visual instructions data. We construct a two-turn conversation for each image pair. In the first turn, we define the task of summarizing CaD by providing the encoded visual tokens of the two images and instructing the model to summarize the CaD, where the response part of the turn is the LLM-generated structured summary collected above. In this instruction, we do not provide the image captions, forcing the model to rely only on image tokens to complete the task. In the second turn, we reinforce the image-text alignment by employing a simple task of text-to-image retrieval to avoid forgetting the model's general capabilities. We randomly sample one of the two captions and request the model to select the image (from the current pair) to which the caption belongs. Through ablation study in Tab. 7, we show that while this task itself does not lead to satisfying results, combining it with the task of summarizing commonalities and differences results in significant improvement. The template for the two-turn conversation is illustrated in Fig. 3(d).

3.2. Phase-1b: CaD Visual Instruction Tuning

Architecture. As illustrated in Fig. 2, we use our collected CaD-Inst^{V1} data to perform visual instruction tuning using the open-sourced code of LLaVA-1.5 [30] LMM. The LLaVA-1.5 model consists of $\phi_L(\cdot;\theta_L)$ - a pretrained Vicuna 1.5 [49] LLM (finetuned from LLama 2 [40]); $\phi_V(\cdot;\theta_V)$ - a pretrained visual encoder CLIP ViT-L/14@336px [35]; and $\phi_M(\cdot;\theta_M)$ - a two-layer MLP projector converting the visual encoder tokens to post-embedding layer LLM tokens.

Given a pair of two images x_{V_1}, x_{V_2} and the instruction x_I, the MLP projects the visual features computed by the visual encoder into embedded language tokens, *i.e.* $v_k = \phi_M(\phi_V(x_{V_k};\theta_V);\theta_M), k \in \{1,2\}$. Then the projected visual features and instruction text tokens are concatenated and fed into the LLM, where the response text tokens are generated in an autoregressive manner, *i.e.*

$$\hat{x}^i_R = \phi_L([v_1, v_2, x_I, \hat{x}^{<i}_R];\theta_L), \tag{1}$$

where $\hat{x}^i_R$ denotes the i-th token in the generated response.

Training. We finetune the LLaVA-1.5 model using the LLaVA [31] pipeline. Specifically, following LLaVA pretraining, we finetune only the pretrained projection MLP and the (frozen) LLM with LoRA adapters [14]. We minimize the CLM loss of the next token prediction in the responses:

$$\mathcal{L}_{CLM} = \sum_i -\log p(\hat{x}^i_R | V_1, V_2, x_I, x^{<i}_R) \tag{2}$$

To preserve the general VL capabilities of the LMM, we merge CaD-Inst^{V1} with the finetuning data of LLaVA-1.5 (665K samples). In Tab. 4 we show that CaD-VI indeed preserves the general LMM capabilities compared to LLaVA-1.5 as evaluated on the popular SEED benchmark [24]. The Phase-1 CaD visual instruction tuning results in our cold-start model CaD-LLaVA^{V1} which is an LMM that can be leveraged for annotating visual commonalities and differences.

3.3. Phase-2: Data Collection and Training

Phase-2a: LMM-based CaD Instruction Collection. While in Phase 1 we used an LLM to extract a CaD summary based on human-generated captions, for Phase 2 data collection we leverage our Phase 1 model CaD-LLaVA^{V1} and additional image pairs to extract the CaD summaries informed by the images directly. Here we select the Scene-Difference [25] collection as an additional image source. It contains 71K pairs of similar images from COCO [29] and provides annotation of unstructured difference-only summaries (see Fig. 2 bottom left for an example). We feed both the image pairs and the original annotations into our CaD-LLaVA^{V1} model, and generate a *structured summary* of *both* commonalities and differences. The exact prompt is provided in the supplementary. This leads to our phase-2 CaD instruction data - CaD-Inst^{V2}. As shown in Tab. 5, our collected CaD instructions significantly improve over the utility of the original [25] annotations. As part of our analysis in Tab. 5 and 6, and additional experiments provided in supplementary, we also show that similarly out-of-distribution image pair collections or even unlabeled image pair collections can be effectively leveraged for Phase-2.

In Phase-2, we generate CaD data leveraging both captions and the CaD image analysis capabilities of our Phase-1 model. This significantly reduces hallucinations and improves the quality of the Phase-2 stage CaD dataset as evident by the significant performance improvement obtained by Phase-2 model over Phase-1 model (Tab. 5 E and F). In the ablation in Sec. 6 (Tab. 6) we also show that image captions can be included in Phase-2 data collection.

In Phase-1, we have image pairs of different similarity levels while in Phase-2 we have highly similar image pairs which lead to more fine-grained difference summaries. We combine data of both phases.

Phase-2b CaD Visual Instruction Tuning We follow the Phase-1b introduced in Sec. 3.2 for CaD visual instruction tuning. Here we finetune on a combination of LLaVA 1.5 [30] finetune data (665K), CaD-Inst^{V1} data (278K) and CaD-Inst^{V2} data (71K). This phase of CaD visual instruction tuning leads to the Phase 2 model, denoted as CaD-LLaVA^{V2}.

4. Benchmark of Open-Ended CaD QA

In order to evaluate the capability of LMMs on answering open-ended questions regarding commonalities and differences of a pair of two images, we construct and contribute the CaD-QA benchmark.

Data Collection. Similar to the data collection pipeline introduced in Sec. 3.1, we employ Visual Genome [22] and the detailed image captions from SVIT [47] as image & caption source. We collect 7.5K image pairs with 8 or more overlapping nouns in their captions. For each pair, we employ the Mixtral 8×7B LLM to produce the structured CaD summaries from the captions. Next, we prompt Mixtral with both the image captions and the CaD summary, instructing it to generate a multi-turn conversation with several rounds of Q&A, providing some in-context examples of the desired layout (see supplementary for the prompt). Finally, we randomly select one Q&A per conversation.

Benchmark Statistics. There are 7520 QA pairs with an average answer length of 26 words. Among these, we also include 2916 questions asking about the content of only one of the two images. It requires the precise attention of the LMM on the corresponding image to correctly answer these questions. Our CaD-QA covers diverse question types as illustrated in Fig. 3(b).

LLM-assisted Evaluation. Motivated by LLMs' ability

Dataset Random chance	# Instruct. Data	BISON 50%	SVO 50%	NLVR2 50%	EQBEN 25%	COLA 25%
SparklesChat	6.5K	56.70%	43.93%	58.00%	19.17%	20.00%
Otter	2.8M	40.67%	47.33%	52.00%	8.33%	8.10%
MMICL	5.8M	80.00%	88.13%	56.67%	20.83%	25.71%
EMU2-Chat	1.3M	46.00%	47.93%	60.00%	7.50%	13.33%
InternLM-XComposer2-VL	>600K	80.67%	82.07%	66.67%	25.00%	32.38%
LLaVA 1.6 7B	<1M	66.00%	70.40%	58.67%	20.83%	11.90%
LLaVA 1.6 13B	<1M	81.33%	82.13%	60.00%	17.50%	24.76%
LLaVA 1.5 7B	665K	54.00%	46.80%	61.33%	17.50%	7.62%
LLaVA 1.5 13B	665K	59.33%	56.27%	66.00%	16.67%	12.38%
CaD-VI 7B	1M	95.33%	92.73%	66.67%	39.17%	40.95%
CaD-VI 13B	1M	**96.67%**	**93.00%**	**69.33%**	**42.50%**	**43.33%**

Table 1. Performance on closed-ended VQA tasks with image pairs in accuracy. Here the method CaD-VI denotes our Phase-2 model CaD-LLaVA^{V2}.

to judge response quality consistently with human assessment [49], we employ the Mixtral 8×7B LLM to compare the generated responses to the collected open-ended QA responses. We feed the question, correct answer, and the predicted answer into the LLM and instruct it to provide a rating between 0 and 5 for the predicted answer quality. We provide the prompt in the supplementary. In order to mitigate the bias from the the same LLM used for evaluation, we include additional evaluations with different LLMs, in-context examples of scoring cases and human study in the supplementary.

5. Experiments

Evaluation Datasets We evaluate on several VQA benchmarks of closed-ended and open-ended questions. For **closed-ended VQA on image pairs**, we include BISON [15] and SVO Probes [13] both consisting of samples with an image pair and a text query that needs to be matched with one of the images in the pair (chance is 50%). EQBEN [41] and COLA [36] contain samples composed of a pair of two images together with the two textual descriptions. The goal is to correctly match images with corresponding texts (chance is 25%). Furthermore, we evaluate on NLVR2 [37] which comprises samples of a pair of two images and a reasoning sentence. The task is to assess the correctness of the reasoning and has a random chance of 50%. We also evaluate SEED-Bench Video [24] with two frames sampled from the video to explore the generalization value of our CaD tuning for video understanding. SEED-Bench Video contains three partitions from SEED-Bench and has multi-choice questions on action recognition/prediction or procedure understanding with four answer options per question. For **open-ended tasks**, use the LLM-as-a-judge metric (Sec. 4). We evaluate open-ended QAs on our CaD-QA. Furthermore, we also directly evaluate the quality of LMM predicted CaD summaries for 210 image pairs in COLA with shorter summaries generated from brief captions, and for the 7.5K lengthy summaries from CaD-QA generated from detailed VG captions. More details and statistics of the datasets are provided in the supplementary.

Dataset	CaD-QA	VG comm.	VG diff.	COLA comm.	COLA diff.
SparklesChat	3.01	2.41	3.12	1.52	1.22
Otter	2.20	1.88	1.97	1.37	0.81
MMICL	2.01	1.79	1.94	1.73	0.59
EMU2-Chat	1.20	1.04	1.08	1.22	0.41
InternLM-XComposer2-VL	2.90	2.08	2.69	1.72	**1.36**
LLaVA 1.6 7B	3.10	2.23	2.73	1.71	1.22
LLaVA 1.6 13B	3.19	2.19	2.69	1.93	1.01
LLaVA 1.5 7B	2.54	1.79	1.75	1.44	1.02
LLaVA 1.5 13B	2.65	2.16	2.41	1.57	1.10
CaD-VI 7B	3.29	2.32	**3.85**	**2.14**	1.25
CaD-VI 13B	**3.34**	**2.58**	3.68	2.13	1.31

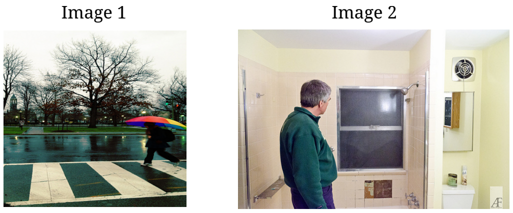
Table 2. Performance on CaD-QA and tasks of CaD summary prediction evaluated using LLM-as-a-judge ratings (range 0 to 5). Here the method CaD-VI denotes our Phase-2 model CaD-LLaVA^{V2}.

Implementation Details We leverage the Mixtral 8×7B Instruct v0.1 and set the maximum token size to 750 data collection and 20 for open-ended task evaluation. For visual instruction tuning, we use the official implementation of LLaVA and tune the LLaVA 1.5 7B model with LoRA. We set the batch size to 128 and LoRA learning rate for LLM and the projector is set to 1×10^{-4} and 2×10^{-5} correspondingly. All experiments are run on 4×A100 80G GPUs. More details are in supplementary.

Comparison to State-of-the-Art LMMs

We first compare our final model CaD-LLaVA^{V2} (denoted by CaD-VI in Table) to state-of-the-art LMMs on closed-ended VQA in Table 1. SparklesChat [16], Otter [26], MMICL [48], EMU2-Chat [38], InternLM-Xcomposer2-VL [44] all include samples with multi-image inputs in the visual instruction tuning while LLaVA 1.5 [30] and LLaVA 1.6 [32] are tuned with only single image instructions. The evaluated benchmarks are challenging due to the visually similar image pairs with subtle compositional differences where the LMMs could easily make an incorrect decision leading to performance below random chance. Our CaD-VI 7B model already outperforms all the other baselines on the five benchmarks and our 13B finetuned model further boosts the performance. We also include the comparison to a recent video LMM in the supplementary.

Table 2 demonstrates the comparison to the baseline LMMs on open-ended tasks of CaD-QA and of CaD summary prediction on image pairs. Our CaD-VI models outperform the baselines on four of the five open-ended tasks, with the exception of COLA difference summary where our 13B model achieves a rating (1.31) close to the best performing InternLM-XComposer2 model (1.36). We include additional evaluations with different LLMs, in-context examples of scoring cases and human study in the supple-



Q: What objects are present in Image 2 that are not in Image 1?

GT: In Image 2, there is a well-equipped shower, a bathtub, a commode, a mirror, a wooden shelf, and various items such as a green candle, air freshener spray, and silver faucet. These objects are not present in Image 1.

LLaVA : In Image 2, there is a toilet visible in the bathroom, which is not present in Image 1. The man is standing in front of the toilet while holding an umbrella.

CaD-LLaVA : In Image 2, there are several objects that are not present in Image 1, including a man, a mirror, a sink, a toilet, a shower, a window, a door, and a ceiling. The man is standing in the bathroom, and the mirror, sink, toilet, and shower are all located in the same room. The window and door are also present in the bathroom, and the ceiling is visible at the top of the image.

Figure 4. Example of Q&A pairs in CaD-QA together with LMM predicted answers (Red and green texts denote incorrect and correct description).

# input frames	1	2
SparklesChat	21.81%	19.09% (▼-2.72%)
Otter	18.19%	23.00% (▲**+4.81%**)
EMU2-Chat	**43.43%**	41.09% (▼-2.34%)
InternLM-XComposer2-VL 7B	41.07%	40.16% (▼-0.91%)
LLaVA 1.6 7B	41.95%	42.03% (▲+0.08%)
LLaVA 1.6 13B	41.85%	41.35% (▼-0.50%)
LLaVA 1.5 7B	37.43%	36.68% (▼-0.75%)
LLaVA 1.5 13B	40.12%	38.78% (▼-1.34%)
CaD-VI 7B	38.40%	40.44% (▲+2.04%)
CaD-VI 13B	40.16%	**43.09%** (▲+2.93%)

Table 3. Performance of multi-choice questions on SEED-Bench video partitions by feeding one or two frames into the LMMs.

mentary, which shows that the Mixtral-assisted evaluation is valid as it maintains the same ranking as when using strongest LLMs as judge.

In Fig. 4, we show examples of Q&A pairs in CaD-QA together with predicted answers from CaD-LLaVA^{V2} model and the vanilla LLaVA 1.5 model. The vanilla LLaVA model has an incorrect answer by mistakenly combining the contents in two images (*the man is standing in front of the toilet while holding an umbrella*), demonstrating lacking of capability of properly comparing two images. Our CaD-LLaVA^{V2} manages to correctly differentiate between the two images, attend to the corresponding content queried and draw a summary of comparison. More qualitative results on CaD-QA and BISON can be found in the supplementary.

Furthermore, we explore whether our CaD instruction tuning improves video understanding evaluated using SEED-Bench Video in Table 3. In the evaluation setting of LLaVA, only one frame per SEED-Bench video is passed to the LMM. To explore the impact of our CaD tuning, we compare this to evaluating using two frames as input. As shown in Table 3, although multiple baseline LMMs achieve better performance in single-frame setting, our CaD-VI 13B model performs the best in the two-frame setting with a significant performance improvement of 2.93% on top of the single-frame performance. The only higher improvement is achieved by Otter, which however struggles below the 25% chance level performance. This underlines that our CaD tuning improves the temporal understanding between video frames.

Model	LLaVA 1.5 7B	CaD-VI 7B	LLaVA 1.5 13B	CaD-VI 13B
SEED-Image	67.34%	67.48%	68.83%	69.11%

Table 4. Performance of multi-choice questions on SEED-Bench image partitions for evaluation of general VL capabilities with single-image input.

Additionally, to verify that introducing multi-image CaD data into the tuning does not lead to catastrophic forgetting of general single-image input LMM capabilities, we also evaluate the SEED-Bench Image partitions and report the results in Table 4. Here we directly compare to same architecture baseline of LLaVA 1.5 fine-tuned using its single-image LLaVA mix 665K data. Table 4 demonstrates that our CaD tuning indeed preserves the competence in single-image understanding. Evaluation on more general VL benchmarks like MME [10] and MMBench [33] can be found in the supplementary.

	Training Data	BISON	SVO	EQBEN	COLA	CaD-QA
A:	LLaVA mix	54.00%	46.80%	17.50%	7.62%	2.54
B:	LLaVA mix + ScDiff orig. annot.	92.67%	90.07%	22.50%	33.81%	2.90
C:	LLaVA mix + ScDiff our annot. (from scratch)	88.67%	90.80%	38.33%	36.67%	3.17
D:	LLaVA mix + ScDiff our annot. (refined from orig. annot.)	94.67%	91.80%	32.50%	34.76%	3.17
E:	LLaVA mix + CaD-Inst^{V1}	92.00%	92.27%	34.17%	36.67%	3.27
F:	LLaVA mix + CaD-Inst^{V1} + ScDiff our annot. (refined from orig. annot.)	**95.33%**	**92.73%**	**39.17%**	**40.95%**	**3.29**

Table 5. Ablation of phase-2 data collection from 71K image pairs in Scene-Difference (ScDiff). We use CaD-LLaVA^{V1} to generate CaD on ScDiff either from scratch or by refining from the original annotation of unstructured difference-only summaries. Training settings in E and F lead to our CaD-LLaVA^{V1} and CaD-LLaVA^{V2} models correspondingly.

	Training Data	BISON	SVO	EQBEN	COLA	CaD-QA
A:	LLaVA mix	54.00%	46.80%	17.50%	7.62%	2.54
B:	LLaVA mix + A/G orig. captions only	55.33%	55.67%	3.33%	2.86%	2.78
C:	LLaVA mix + A/G our annot. (from scratch)	**90.00%**	**88.53%**	40.83%	**42.86%**	**3.21**
D:	LLaVA mix + A/G our annot. (given orig. captions)	88.00%	86.87%	**43.33%**	30.48%	3.06

Table 6. Ablation of phase-2 data collection from 66K pairs of video frames in Action Genome and GEBC (A/G). We use CaD-LLaVA^{V1} to generate CaD on A/G either from scratch or with the prior information from the original frame captions.

6. Ablations

Phase-2 Data Collection analysis. Our Phase-2 data collection introduced in Sec. 3.3 can be used to leverage image pairs from various sources for producing effective CaD instructions. We first ablate the data collection from the 71K image pairs in Scene-Difference [25] (ScDiff) which contains annotation of unstructured difference-only summaries. As shown in Table 5, training with original annotation of difference-only summaries (row B) significantly improves on the baseline of training with LLaVA data only (row A). Then we show that using CaD-LLaVA^{V1} to generate CaD instructions on ScDiff remarkably improves further, either if used from scratch (row C) or by refining from the original annotation (row D, also illustrated in Fig. 2 bottom row). Training with our re-annotation from scratch outperforms the original annotation on all datasets except for BISON. Our re-annotation by refining the original annotation leads to a more balanced performance improvement and is used as the phase-2 instruction data CaD-Inst^{V2}. We combine this with our phase-1 data CaD-Inst^{V1} and demonstrate the further performance boost in row F of Table 5.

In order to show the robustness of CaD data collection capability using our CaD-LLaVA^{V1} model, we also explore applying our phase-2 data collection to visually similar frames from user videos in Action Genome and GEBC (A/G). In Table 6, we first train a baseline using original frame captions only and a simple instruction task of image description (row B), which leads to a significant performance drop on EQBEN and COLA, and minimal improvement on other datasets. Then we use our CaD-LLaVA^{V1} to generate CaD instructions on the frame pairs either from scratch (row C) or conditioned on the frame captions (row D). Interestingly, on most datasets CaD instructions generated by our CaD-LLaVA^{V1} from scratch are found to be more effective than ones generated using original captions conditioning, likely due to lack of detail in these captions. This once again demonstrates that our model is effective in generating CaD instructions on unlabeled data. In the supplementary, we further show that our phase-2 data collection is effective on out-of-distribution video-surveillance data of Spot-the-diff (SpotDiff) dataset [20].

	Training Data	BISON	SVO	CaD-QA	VG comm.	VG diff.
A:	LLaVA mix	54.00%	46.80%	2.54	1.79	1.75
B:	LLaVA mix + t2i retriev.	58.00%	51.33%	2.47	1.58	1.46
C:	LLaVA mix + comm.	64.67%	79.73%	3.23	**2.67**	2.52
D:	LLaVA mix + diff.	55.33%	72.13%	3.24	1.97	2.89
E:	LLaVA mix + comm. + diff.	72.00%	82.60%	3.24	2.13	3.42
F:	LLaVA mix + comm. + diff. + t2i retriev.	92.00%	92.27%	3.27	2.21	3.69
G:	(F) + CaD-Inst^{V2}	**95.33%**	**92.73%**	**3.29**	2.32	**3.85**

Table 7. Ablation on components in the instruction data. Training settings in F and G lead to our CaD-LLaVA^{V1} and CaD-LLaVA^{V2} models correspondingly. Here *t2i retriev.* refers to the text-to-image retrieval task (see Sec. 3.1). Training settings in F and G lead to our CaD-LLaVA^{V1} and CaD-LLaVA^{V2} models.

Analysis of CaD Instruction Data Components We verify the effectiveness of the components in our instruction data by ablating on the different combinations of our tuning tasks, including: (i) commonality summary (*comm.*); (2) difference summary (*diff.*); and (iii) text-to-image retrieval (*t2i retriev.*) in Table 7. Training solely on the t2i retrieval task (row B) leads to minimum performance improvement on BISON and SVO Probes, and performance degradation on the three benchmarks of the open-ended tasks due to lacking of any CaD learning. Training with the commonality (row C) and difference summary (row D) tasks separately lead to a significant boost on the VG comm (2.67) and VG diff (2.89) tasks correspondingly. Training with combinations of the three tasks (F) boosts the performance in comparison to the case of each single component, except for VG comm where the commonality training (row C) leads to better results on this task. Finally, combining phase-1 and phase-2 data (row G) leads to further performance boosts on most of the benchmarks.

7. Conclusions

We contribute CaD-VI, a two-phase strategy for collecting Commonalities and Differences (CaD) Visual Instruction data, resulting in CaD-Inst with 349K samples that significantly improve CaD and related comparative abilities in LMMs. We also introduce CaD-QA, a 7.6K open-ended QA benchmark for evaluating CaD across image pairs. Our extensive evaluation demonstrates substantial gains using CaD-VI, which complements and enhances existing CaD resources. This work advances the investigation of CaD capabilities in LMMs and paves the way for future CaD VI tuning.

Limitations Our current approach is limited to CaD between two images; extending to three or more remains future work.

References

[1] AI@Meta. Llama 3 model card. 2024. 1

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. 1

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 2

[4] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Percy Liang, and et al. On the opportunities and risks of foundation models, 2022. 1

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2023. 3

[7] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 1

[8] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2024. 1

[9] OpenAI et al. Gpt-4 technical report, 2024. 1

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 7

[11] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 3

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[13] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, 2021. 2, 6

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[15] Hexiang Hu, Ishan Misra, and Laurens Van Der Maaten. Evaluating text-to-image matching using binary image selection (bison). In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 6

[16] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023. 1, 2, 3, 4, 6

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3

[18] Interaction Design Foundation IxDF. What are the gestalt principles?, 2016. 1

[19] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, 2018. 3

[20] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, 2018. 3, 8

[21] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 4

[22] 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. 5

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 3

[24] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 5, 6

[25] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 2, 3, 4, 5, 8

[26] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 1, 2, 3, 6

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[28] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, and Xu Sun. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 3

[29] 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*, pages 740–755. Springer, 2014. 3, 5

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 3, 5, 6

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023. 1, 2, 3, 4, 5

[32] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3, 6

[33] 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*, 2023. 7

[34] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020. 2, 3

[35] 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*, pages 8748–8763. PMLR, 2021. 5

[36] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36, 2023. 2, 6

[37] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019. 2, 6

[38] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023. 1, 2, 6

[39] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. In *CVPR*, 2024. 3

[40] 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*, 2023. 5

[41] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008, 2023. 2, 6

[42] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 1

[43] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3

[44] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 3, 6

[45] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 3

[46] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 3, 4

[47] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 3, 5

[48] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 6

[49] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023. 5, 6

[50] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 3

[51] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[52] 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. 1