Comparison Visual Instruction Tuning
Supplementary Materials

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

1. Introduction

As additional results, we include more **evaluations on the open-ended CaD-QA** with different LLMs, in-context examples of scoring cases and human study in Sec. 2.1. Then we report results of CaD-VI on **two more general vision-language benchmarks** (Sec. 2.2). Further include the evaluation of a **video LMM** in Sec. 2.3. We report the **error bars** (Sec.2.4), analyze the **Phase-2 data collection on Out-Of-Distribution data** (Sec. 2.5). Finally, we show **qualitative results** of the collected CaD summaries (Sec. 2.6), and compare LMM predictions on our CaD-QA benchmark (Sec. 2.7), and LMM predictions on the BISON dataset (Sec. 2.8).

For further insights into our approach CaD-VI, we report more **statistics on our generated data** (Sec. 3.1), and **statistics on the external evaluation datasets** (Sec. 3.2). We provide more **implementation details** (Sec. 4) including the specifics of baseline methods, data generation, training and evaluation details.

At last, we provide the **list of assets** (Sec. 5) used in this project.

2. Additional Results

2.1. Additional Evaluations of Open-Ended CaD QA

Different LLMs. In order to mitigate the bias from the same LLM used for evaluation and show the impact of different LLMs on the LLM-assisted evaluation, we further employ LLaMA 3.1 70B and GPT4o mini for the evaluation of CaD QA and report the resutls in Tab. 1. In case of LLaMA 3.1 70B and GPT4o mini, CaD-VI still outperforms all the other competitors. However, there is a drop in the margin of its outperformance in comparison to the case of Mixtral model assisted evaluation.

Model	Mixtral 8×7B	LLaMA 3.1 70B	GPT4o mini
SparklesChat	3.01	2.91	2.62
Otter	2.20	1.70	1.66
MMICL	2.01	1.97	2.00
EMU2-Chat	1.20	1.26	1.34
InternLM-XComposer2-VL	2.90	2.79	2.61
LLaVA 1.6 7B	3.10	2.80	2.54
LLaVA 1.6 13B	3.19	3.00	2.67
LLaVA 1.5 7B	2.54	1.98	1.86
LLaVA 1.5 13B	2.65	2.11	1.98
CaD-VI 7B	<u>3.29</u>	<u>3.02</u>	<u>2.72</u>
CaD-VI 13B	**3.34**	**3.10**	**2.78**

Table 1. Impact of different LLMs on the LLM-assisted evaluation of the open-ended CaD QA benchmark.

Model In-context	Mixtral 8×7B No	Mixtral 8×7B Yes	LLaMA 3.1 70B No	LLaMA 3.1 70B Yes
SparklesChat	3.01	2.08	2.91	3.14
Otter	2.20	1.17	1.70	2.02
MMICL	2.01	1.72	1.97	2.40
EMU2-Chat	1.20	1.01	1.26	1.42
InternLM-XComposer2-VL	2.90	2.52	2.79	3.15
LLaVA 1.6 7B	3.10	2.06	2.80	2.97
LLaVA 1.6 13B	3.19	2.16	3.00	3.13
LLaVA 1.5 7B	2.54	1.56	1.98	2.18
LLaVA 1.5 13B	2.65	1.77	2.11	2.33
CaD-VI 7B	<u>3.29</u>	<u>2.54</u>	<u>3.02</u>	<u>3.20</u>
CaD-VI 13B	**3.34**	**2.68**	**3.10**	**3.31**

Table 2. Impact of in-context examples of scoring cases on the LLM-assisted evaluation of the open-ended CaD QA benchmark.

Scoring standard descriptions. We further explore the impact of scoring standard descriptions in the evaluation of open-ended CaD QA. We provide in-context examples for cases of different scores. In Tab. 2, we report the evauation results with and without in-context examples of scoring cases. In all cases. CaD-VI still outperforms the other competitors. Evaluation with in-context examples of ratings leads to drop of ratings on Mixtral 8×7B but slight increase of rating on LLaMA 3.1 70B. This could due to the better in-context learning capability of LLaMA 3.1.

Model	CaD-VI 13B	LLaVA 1.6 13B	LLaVA 1.5 13B	InternLM-XComposer2-VL	SparklesChat
Rating	3.61	3.42	2.84	3.05	3.30

Table 3. Human evaluation on 150 randomly sampled questions from the open-ended CaD QA benchmark.

Human study. Furthermore, we randomly sampled 150 open-ended questions from the evaluation benchmark and asked three volunteers to manually rate the predictions of the compared LMMs in the range between 0 and 5. To reduce the rating efforts, we include the 13B version of CaD-VI and LLaVA models in this task.

As shown in Tab. 3, the results indicate the human preference of answers from CaD-VI, which is aligned with the choice of LLMs. In the analysis of feedback from the human study, we also have some interesting conclusions: (1) The verbose descriptions with hallucinations from the talkative SparklesChat are better rated by humans than LLMs (2) InternLM-XComposer2-VL could generate correct and concise descriptions of visual contents but is not good at the task of comparison (3) LLaVA 1.6 could see more visual details than LLaVA 1.5 due to the AnyRes (any-resolution) pipeline which benefits the comparison reasoning. In this case, using an architecture with more visual tokens to focus on local regions of images would allow comparison of more visual details via the comparison visual instruction tuning.

2.2. Additional Evaluations on General Vision-Language Benchmarks

Model	MMBench	MME Perception	MME Cognition
LLaVA 1.5 7B	65.80%	1498.09	274.64
CaD-VI 7B	65.38%	1493.21	328.57
LLaVA 1.5 13B	69.07%	1541.69	300.36
CaD-VI 13B	68.27%	1530.61	306.07

Table 4. Evaluation of CaD-VI on general vision-language benchmarks MMBench and MME.

In the main paper, we report performance of CaD-VI on the general vision-language benchmark SEED-Bench image (Tab. 2 in the main paper) and SEED-Bench video(Tab. 3 in the main paper), which verifies that introducing multi-image CaD data into tuning does not lead to catastrophic forgetting of general single-image input LMM capabilities.

Additionally, we compare the performance of CaD-VI to the original LLaVA models on MME [8] and MM-Bench [28] in Tab. 4. We see that after introducing CaD data into tuning, there is only a slight performance drop of CaD-VI in comparison to the original LLaVA on MMBench and MME Perception. On MME Cognition tasks, CaD-VI even has some performance improvements.

2.3. Evaluation of VideoLLaMA2

Model	BISON	SVO	NLVR2	EQBEN	COLA	CaD-QA
VideoLLaMA2	58.00%	61.00%	64.00%	11.67%	16.67%	2.22
CaD-VI 7B	95.33%	92.73%	66.67%	39.17%	40.95%	3.29
CaD-VI 13B	96.67%	93.00%	69.33%	42.50%	43.33%	3.34

Table 5. Evaluation of VideoLLaMA2[3] on the benchmark datasets.

In the main manuscript, we include five models that train on samples with multiple input images, *i.e*. SparklesChat, Otter, MMICL, EMU2-Chat, InternLM-XComposer2-VL. We additionally report the performance of a recent video LMM VideoLLaMA2 [3] on the benchmark datasets. As shown in Tab. 5, our CaD-VI could outperform VideoLLaMA2 on all the benchmarks. The reason that the video LMM does not perform well on benchmarks of CaD capabilities could be that it is trained to understand a video as a spatio-temporal entity instead of multiple individual images.

2.4. Error Bars

Training Data	BISON	SVO	EQBEN	COLA	CaD-QA
LLaVA mix + CaD-LLaVA^{V1}	91.78% ± 1.02%	92.33% ± 0.57%	33.06% ± 0.96%	34.64% ± 2.09%	3.270 ± 0.002%

Table 6. Average performance of the Phase-1 model CaD-LLaVA^{V1} on multiple runs of training.

We run the training of the Phase-1 model CaD-LLaVA^{V1} multiple times and report the average performance with standard deviation in Table 6. In most evaluation cases, the standard deviation is within around 1%.

2.5. Ablation on Phase-2 Data Collection - OOD CaD refinement

In Section 6 (main paper), we perform ablation the Phase-2 data collection. Here we further explore applying our phase-2 data collection on out-of-distribution (OOD) data of Spot-the-diff (SpotDiff) dataset. The dataset contains distant-view frame pairs with very subtle changes from video-surveillance footage, which are OOD from most LMM training data.

	Training Data	BISON	SVO	Difference Spotting	CaD-QA
A:	LLaVA mix (L)	**54.00%**	46.80%	49.50%	<u>2.54</u>
B:	L + SpotDiff orig. annot.	<u>51.33%</u>	<u>52.27%</u>	<u>60.48%</u>	2.51
C:	L + SpotDiff our annot. (refined from orig. annot.)	**54.00%**	**54.87%**	**66.67%**	**2.86**

Table 7. Ablation of phase-2 data collection from 15K pairs of video frames in Spot-the-diff (SpotDiff). We use CaD-LLaVA^{V1} to generate CaD on SpotDiff by refining from the original human-annotated difference descriptions.

In Table 7, we train with SpotDiff original human-annotated difference description (row B) and with our CaD-LLaVA^{V1} generated CaD summaries which is refined from the original annotation (row C). We also evaluate on the Difference-Spotting partition on SEED-Bench 2 [19] which contains multi-choice questions based on frame pairs from SpotDiff. In data collection and training for this experiment, we only used the 15K training image pairs from SpotDiff which are not included in the Difference-Spotting SEED partition. The results in Table 7 verify that our phase-2 data collection using CaD-LLaVA^{V1} is also effective on OOD data.

2.6. Qualitative Results of CaD Summaries

In Fig. 2 (main paper), we illustrate the pipeline of our two-phase CaD-VI together with two examples of Phase-1 LLM-collected CaD summary and Phase-2 LMM-collected CaD summary. Here in Fig. 1, we provide two additional examples. Note that in Fig. 1(a), we only pass the captions with the instruction prompt (in Fig. 11) into the LLM. In Fig. 1(b), we pass the original annotation and both images with the instruction prompt (in Fig. 12) into the Phase-1 model. In the main paper (Table 5), we demonstrate the generated CaD summary without using the original annotation also leads to effective results.

2.7. Qualitative Results on CaD-QA

In Fig. 2, Fig. 3 and Fig. 4, we show examples of Q&A pairs in our CaD-QA, together with the predicted answers from CaD-LLaVA^{V2} model and the vanilla LLaVA 1.5 model. We also report the LLM ratings for the predicted answers. The vanilla LLaVA model has incorrect answers by either mistakenly combining the contents in two images (Fig. 2(b), *the man is standing in front of the toilet while holding an umbrella*), omitting one of the images (Fig. 3(a), Fig. 4(a)), attending to the incorrect image (Fig. 3(c)) or hallucinating non-existent contents (Fig. 4(b)). The failure demonstrates the lacking of capability of properly comparing two images. At the same time, our CaD-Inst^{V2} manages to correctly differentiate between the two images, attend to the corresponding content asked in the question and draw a summary of comparison.

2.8. Qualitative Results on BISON

In Fig. 5, we illustrate some examples of the binary image selection task on BISON. We instruct the LMMs to give both the selection answer and also the reasoning for the selection. Here we compare the vanilla LLaVA 1.5 and our CaD-LLaVA^{V2}. The LLaVA model, even if it captures the relevant content in some cases, has confusion differentiating the two images (Fig. 5(a)(b)). For our CaD-LLaVA^{V2}, the key reasoning that leads to the correct answer is always covered in the structured difference summary.

Image 1 Image 2

(a)

Caption1: In this image there is a dining table on which there is a flower vase. There is a Champagne bottle. There is laptop. There is a headphone. There at the left side of the image there is a chair on which there is a bag. At the background of the image there is a wall and a window.

Caption2: In the image in the center we can see four persons were standing and they were smiling,which we can see on there faces. In front of them,there is a table. On table,we can see different types of books. In the center we can two persons were standing and the left lady is holding paper bag. In the background there is a brick wall,banner,belts,hanger,photo frames,toys,sign board,clothes,table and few other objects.

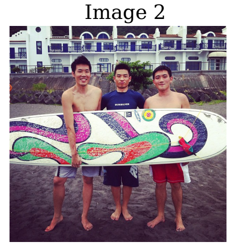
LLM :

Commonalities:
Both images include indoor scenes with man-made objects. There is a table in both images, and the tables are surrounded by other objects.

Differences:
1. Object types: Image 1 contains a dining table, flower vase, Champagne bottle, laptop, headphone, chair, bag, wall, and window. Image 2 contains four people, a table, books, paper bag, brick wall, banner, belts, hangers, photo frames, toys, sign board, clothes, and other objects.
2. Counting the objects: There are 1 dining table, 1 flower vase, 1 Champagne bottle, 1 laptop, 1 headphone, 1 chair, 1 bag, 1 wall, and 1 window in Image 1. In Image 2, there are 4 people, 1 table, many books, 1 paper bag, 1 brick wall, 1 banner, several belts, hangers, photo frames, toys, 1 sign board, and multiple clothes.
3. Object actions: There are no specific actions involving objects in Image 1. In Image 2, four people are standing and smiling.
4. Object locations: In Image 1, the objects are placed on the dining table and the floor. In Image 2, the people are standing in front of the table, and the books are on the table.
5. Relative positions between objects: In Image 1, the chair is located to the left of the table. In Image 2, the paper bag is being held by one of the women, and the other objects are placed on the table or hung on the wall.

Image 1 Image 2

(b)



Original annotation: "*The surfboard in the first image is a very colorful one, while the surfboard in the second image is a decorated one. In the first image, there is only one person holding the surfboard, while in the second image, there are three people holding the surfboard.*"

CaD-LLaVA :

Commonalities:
Both images contain a surfboard as the main object. The surfboard is being held by people in both images. The images also have a natural background, with the first image showing water, a mountain, and the sky, and the second image showing a building, trees, and a fence.

Differences:
1. The number of people holding the surfboard is different. In the first image, there is only one person, while in the second image, there are three people.
2. The background in the first image is more natural, with water, a mountain, and the sky, while the background in the second image is more urban, with a building, trees, and a fence.
3. The first image has a more open and spacious setting, while the second image has a more crowded and enclosed setting.
4. The first image has a more dynamic and active feel, with the person holding the surfboard, while the second image has a more static and posed feel, with the three people standing

Figure 1. Examples of (a) Phase-1 LLM-collected CaD summary and (b) Phase-2 LMM-collected CaD summary

Image 1 Image 2

(a)

Q: What type of outdoor recreational activities are depicted in the two images?

GT: In Image 1, a group of young people are engaged in a game of frisbee in a park, while in Image 2, an ongoing baseball game is shown, with players in uniforms and specialized equipment.

LLaVA: The two images depict a baseball game in progress, with a batter swinging a baseball bat and a catcher wearing a baseball glove. The scene takes place on a baseball field, where multiple people are participating in the game. The presence of a baseball glove and the batter's swinging motion indicate that the game is in action, and the players are actively engaged in the sport. ongoing baseball game is shown, with players in uniforms and specialized equipment.

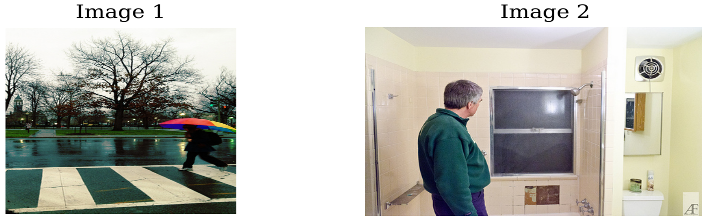
LLM rating: 3

CaD-LLaVA: The two images depict different outdoor recreational activities. Image 1 shows a group of people playing with a frisbee on a grassy field, while Image 2 features a baseball game with a batter, a catcher, and an umpire. The activities in the images are distinct, with one involving a casual, informal game of frisbee, and the other being a more structured and organized baseball game.

LLM rating: 5

Image 1 Image 2

(b)



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.

LLM rating: 2

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.

LLM rating: 3

Image 1 Image 2

(c)

Q: Are there any humans visible in Image 2?

GT: No, there are no visible humans in Image 2, but the presence of a train driver is implied.

LLaVA: Yes, there is a man visible in Image 2, standing next to the train.

LLM rating: 0

CaD-LLaVA: No, there are no humans visible in Image 2.

LLM rating: 4

Figure 2. Examples of Q&A pairs in CaD-QA together with LMM predicted answers and the corresponding LLM evaluation rating for the prediction (Red and green texts denote incorrect and correct description).

(a)

Image 1

Image 2

Q: What is the central figure in each image?

GT: In Image 1, the central figures are several horses and their riders, engaged in a polo game. In Image 2, the central figure is a man holding a mountain bike.

LLaVA : The central figure in each image is a man riding a bicycle.

LLM rating: 2

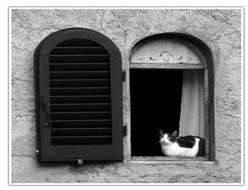
CaD-LLaVA : In Image 1, the central figure is a man riding a horse. In Image 2, the central figure is a man riding a bicycle.

LLM rating: 4

(b)

Image 1

Image 2



Q: What kind of mood do the two images convey?

GT: Image 1 conveys a sense of dynamism and energy in a modern urban setting. Image 2 evokes a feeling of tranquility and curiosity in a traditional, mysterious setting.

LLaVA : The two images convey a calm and peaceful mood. In the first image, a cat is sitting in a window, looking out, which suggests a sense of relaxation and contentment. The second image shows a bus parked on the side of the road, with people waiting at a bus stop. The presence of the bus and the people waiting for it indicate a sense of order and routine, which can also contribute to a calm atmosphere. The combination of these two images creates a serene and tranquil scene, reflecting the everyday life of people and animals in a quiet, urban environment.

LLM rating: 2

CaD-LLaVA : The two images convey different moods. Image 1 has a lively and active mood, with a bus, people, a pole, a dustbin, trees, buildings, and a cloudy sky. The presence of a bus and people suggests movement and activity. Image 2, on the other hand, has a calm and serene mood, with a cat sitting on a window sill and a wall in the background. The cat's presence and the stillness of the scene create a sense of tranquility.

LLM rating: 4

(c)

Image 1

Image 2

Q: What is the atmosphere like in Image 1?

GT: The atmosphere in Image 1 is lively and conveys a sense of joy and adventure for the passengers aboard the train.

LLaVA : The atmosphere in Image 1 is lively and social, with a group of people walking down a dirt road or path, enjoying each other's company.

LLM rating: 3

CaD-LLaVA : The atmosphere in Image 1 is lively and engaging, as there are people sitting on a train and taking pictures of the scenery. The presence of a train and the outdoor setting suggest a sense of adventure and excitement, as the passengers are enjoying their journey and capturing memories of the beautiful surroundings.

LLM rating: 5

Figure 3. Examples of Q&A pairs in CaD-QA together with LMM predicted answers and the corresponding LLM evaluation rating for the prediction (Red and green texts denote incorrect and correct description).

Image 1 Image 2

(a)

Q: Are the settings in the two images similar or different?

GT: The settings in the two images are different. Image 1 is set in a train station during winter, while Image 2 is at a beach during a sunny day.

LLaVA: The settings in the two images are similar, as both depict men sitting on a bench or a wooden chair near a beach.

LLM rating: 0

CaD-LLaVA: The settings in the two images are different. Image 1 features a train station with a train on the tracks, people walking on the platform, and buildings in the background. Image 2, on the other hand, shows two men sitting on a bench under a colorful umbrella, with a body of water and boats in the background.

LLM rating: 5

Image 1 Image 2

(b)

Q: Are there any plants or natural elements in the images?

GT: No, neither Image 1 nor Image 2 contains any plants or natural elements.

LLaVA: Yes, there is a plant in the image, which is placed on the table.

LLM rating: 0

CaD-LLaVA: No, there are no plants or natural elements in the images. Both images are indoor scenes with man-made objects.

LLM rating: 5

Image 1 Image 2

(c)

Q: What objects are present in each image?

GT: Image 1 contains a wooden and metallic bench, a cellphone, and some graffiti, while Image 2 features a blue metal plow, horses, and rocky soil.

LLaVA: In the first image, there is a man and a horse, while in the second image, there are two horses and a man.

LLM rating: 1

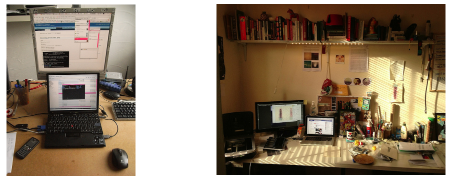
CaD-LLaVA: In Image 1, there is a woman sitting on a bench, and in Image 2, there are two horses, a cart, and three people.

LLM rating: 2

Figure 4. Examples of Q&A pairs in CaD-QA together with LMM predicted answers and the corresponding LLM evaluation rating for the prediction (Red and green texts denote incorrect and correct description).

Image 1 Image 2

(a)



Text query: A large desk has a bookshelf above it.

LLaVA : Image 1.
The image with the bookshelf above the desk is more suitable for the caption "A large desk has a bookshelf above it." because it clearly shows a bookshelf above the desk, which is not visible in the other image. In the first image, the bookshelf is located above the desk, while in the second image, the bookshelf is not visible. The first image also has a laptop and a cell phone on the desk, which are not present in the second image

CaD-LLaVA : Image 2.
Reasons:
* Image 2 contains a bookshelf with books and other items, while Image 1 does not have a bookshelf.
* Image 2 has a desk with a laptop, monitor, and other items, while Image 1 has a desk with a laptop, monitor, and other items.
* Image 2 has a wall with posters and a board, while Image 1 does not have a wall with posters or a board.
* Image 2 has a more cluttered and detailed scene compared to Image 1.

Image 1 Image 2

(b)

Text query: A group of men preparing to fly a kite being filmed by another man.

LLaVA : Image 1.
The image with the man holding a camera captures the essence of the scene better. In this image, there is a man holding a camera, and another man is holding a kite. The presence of the camera indicates that the scene is being filmed, which is the main focus of the image. In contrast, the image with the man holding a camera and two other men standing in the grass does not clearly convey the act of filming. The presence of the kite in the image also adds to the visual interest and context of the scene.

CaD-LLaVA : Image 2.
Reasons:
1. The number of people: Image 1 has two people, while Image 2 has four people.
2. Actions: In Image 1, one person is flying a kite, and the other is standing. In Image 2, all three people are standing, and one person is holding a camera.

Image 1 Image 2

(c)

Text query: A man in a white sweater sitting at a table eating a whole pizza pie.

LLaVA : Image 1.
The image with the man in a white sweater sitting at a table eating a whole pizza pie is more suitable for the caption because it clearly shows the man eating a pizza and wearing a white sweater. The other image, B, shows a man cutting a pizza with a knife and fork, but it does not provide enough context to determine if he is wearing a white sweater or not. The image A provides a more accurate representation of the man eating a pizza and wearing a white sweater, which is the main focus of the caption.

CaD-LLaVA : Image 2.
Reasons:
1. The man in Image 2 is sitting at a table and eating a whole pizza pie, while the man in Image 1 is sitting at a table and holding a glass of beer.
2. The man in Image 2 is wearing a white sweater, while the man in Image 1 is not wearing a sweater.
3. The man in Image 2 is cutting the pizza with a knife, while the man in Image 1 is not cutting anything.

Figure 5. Examples of predictions of the binary image selection task on BISON (red and green texts denote incorrect and correct predictions). We instruct the LMMs to, besides the selection, also give a reasoning for the answer.

3. Dataset Statistics

3.1. Generated Data Statistics

CaD-Inst^{V1} and CaD-Inst^{V2}. In CaD-Inst^{V1}, we collected structured summaries of CaD for 278K image pairs, with an average length of 157 words (40 for commonalities and 117 for differences). In CaD-Inst^{V2}, we collected summaries of CaD for 71K images pairs used in Scene-Difference [21], with an average length of 156 words (28 for commonalities and 128 for differences). We demonstrate the distribution of CaD summary length (number of words) in CaD-Inst^{V1} (Fig. 6(a)) and in CaD-Inst^{V2} (Fig. 6(b)).

In Fig. 7, we also illustrate the cloud of words covered in the CaD summaries in CaD-Inst^{V1} (Fig. 7(a)) and in CaD-Inst^{V2} (Fig. 7(b)).

In the main paper, we mentioned that the collected summaries are structured according to approximate 6 axes of characteristics: *object types, attributes, counting, actions, locations* and *relative positions*. Note that the characteristics appear unevenly on a case-to-case basis based on the LLM decision on individual samples. In Fig. 3(a)(main paper), we illustrate the distribution of these sample-specific characteristics in a Sunburst chart. Here in Fig. 8, we also illustrate the distribution of these characteristics (*e.g*. object types, action of people, surrounding environments, *etc*.) in CaD summaries in the Phase-1 data collection CaD-Inst^{V1}. The structured differences are summarized in terms of these characteristics (see Fig. 1(a) for an example of structured difference summary in terms of several characteristics). The visual instruction tuning guides the model to compare images in terms of these detailed characteristics.

In the main paper, we introduced that we collect 278K image pairs with different levels of similarity between their captions. We measure the similarity between two captions by counting the number of overlapping nouns in the corresponding captions. Here we show the distribution of the number of overlapping nouns in Fig. 9(a). We see that we cover image pairs with different levels of caption-caption similarity. Furthermore, we use the CLIP ViT-B/32 model [30] to compute the similarity scores between the two images in each pair and report the distribution in Fig. 9(b). We verify that image pairs of diverse similarity levels are covered in our Phase-1 data collection CaD-Inst^{V1}.

CaD-QA. Our CaD-QA benchmark contains 7.5K open-ended questions with answers. Here we show the distribution of questions types (first 5 words) and answer types (first 3 words) in Sunburst charts in Fig. 10. There are diverse question categories covered such as *Yes/No* questions, *What* questions on scene characteristics such as objects, attributes and setting, and also requests to describe specific characteristics in details.

3.2. Statistics of External Evaluation Datasets

We evaluate on several external VQA benchmarks of closed-ended and open-ended questions. Here we give a brief introduction on the contents and statistics.

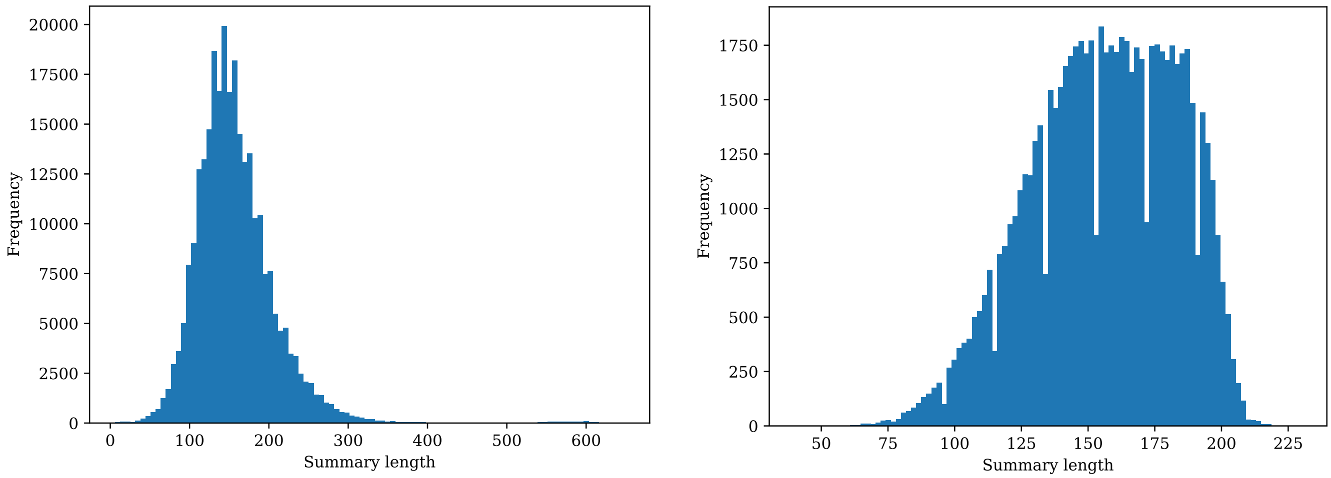
BISON is a dataset for the binary image selection task [12]. There are 150 samples in the evaluation benchmark, each sample consisting of a pair of two visually similar images and a query caption. Only one image correctly matches with the query caption. It measures the ability of the LMMs to relate fine-grained text content in the caption to visual content in the images.

SVO Probes is a benchmark designed to probe for subject, verb and object understanding in vision-language models [11]. In the benchmark, each sample consists of a pair of two images and a query sentence, where only one image correctly matches with the query sentence. The negative image differs from the positive image with regard to either the subject, the verb or the object. There are 36.8K samples in the dataset. For efficient evaluation, we randomly select 1500 samples that can be divided into 3 partitions *subject*, *verb* and *object* where each partition has 500 samples with the image pair contradiction in either subject, verb or object.

EQBEN is a benchmark that focuses on visual minimal change between two images [38]. Each sample in the benchmark consists of a pair of two images with subtle visual changes and two corresponding captions. The dataset is comprised of frames from natural video datasets such as YouCook2 [46], Action Genome [14] and GEBC [39], as well as sythetic image pairs with subtle differences generated by the photo-realistic scene generator Kubric [10] and the diffusion model Stable-Diffusion [32]. We employ an EQBEN subset[1] which is released by the authors in [38] for evaluating the performance of LMMs specifically. The subset consists of 120 samples, comprised of frame pairs from Action Genome and GEBC, image pairs with changes in attributes, count and location generated by Kubric, and image pairs with style change generated by Stable-Diffusion. For each sample, we perform the binary image selection task twice, feeding one of the descriptions for image selection at a time. The sample is considered positively answered only when both selection tasks are correctly solved.

COLA is a benchmark for evaluating the capabilities of vision-language models on representing simple compositions by combing objects with their attributes [31]. Each sample in the benchmark consists of two images with two corresponding captions. The two images have attributes and objects that are swapped in the captions, *e.g. large tree to the right of little short green tree*, and *tall green tree to the*

[1] https : / / entuedu - my . sharepoint . com/ : u : /g / personal / tan317 _ e _ ntu _ edu _ sg / ETkpKSsmun1MpBw7FqfUUS8BwTX2gKkTQkDFsfOGCw - 9yA?e=KGtpg0



(a) CaD-Ins$^{\mathrm{V1}}$ (b) CaD-Ins$^{\mathrm{V2}}$

Figure 6. Distribution of length of CaD summaries (in terms of number of words) in (a) CaD-Inst^{V1} and (b) CaD-Inst^{V2}

(a) CaD-Ins$^{\mathrm{V1}}$ (b) CaD-Ins$^{\mathrm{V2}}$

Figure 7. Word clouds of CaD summaries in (a) CaD-Inst^{V1} and (b) CaD-Inst^{V2}

right of large tall green tree. We employ the partition of *multi-object setting* in the benchmark which consists of 210 image pairs and captions. Similar to evaluation on EQBEN, we perform the binary image selection task twice for each sample.

NLVR2 is a benchmark for evaluation of the visual reasoning with natural language task which aesses the ability of LMMs to predict whether a sentence is true about a pair of images [33]. The task focuses on understanding of compositionalities in terms of relations, comparisons and counting. We use the subset of 150 samples provided in SparklesChat [13] for a fair comparison.

SEED-Bench is an evaluation benchmark on comprehensive vision-language understanding, consisting of 19K multiple choice questions [20]. The are two major categories in the benchmark: *SEED-Image* with 14K samples and *SEED-Video* with 5K samples. SEED-Image consists of 9 dimensions: scene understanding, instance identity, instance attributes, instance location, instance counting, spatial relation, visual reasoning and text understanding. All samples contain only a single input image. SEED-Video consists of 3 dimensions: action recognition, action prediction and procedure understanding. The videos are from Something-Something-v2 [9], EPIC-Kitchen [6] and Breakfast [17].

4. Implementation Details

4.1. Baselines

SparklesChat [13] is finetuned from the first-stage pretrained model of MiniGPT4 [47]. The model is finetuned with their collected multi-image dialogue data. SparklesChat follows the architecture of MiniGPT4 and uses Vicuna 7B [4], EVA-CLIP ViT-G/14 [7] with a Q-Former from BLIP-2 [23]. We use the model weights and instruction templates available at `https://github.com/HYPJUDY/Sparkles`.

Otter [22] is finetuned from the OpenFlamingo model [1] with the collected multimodal in-

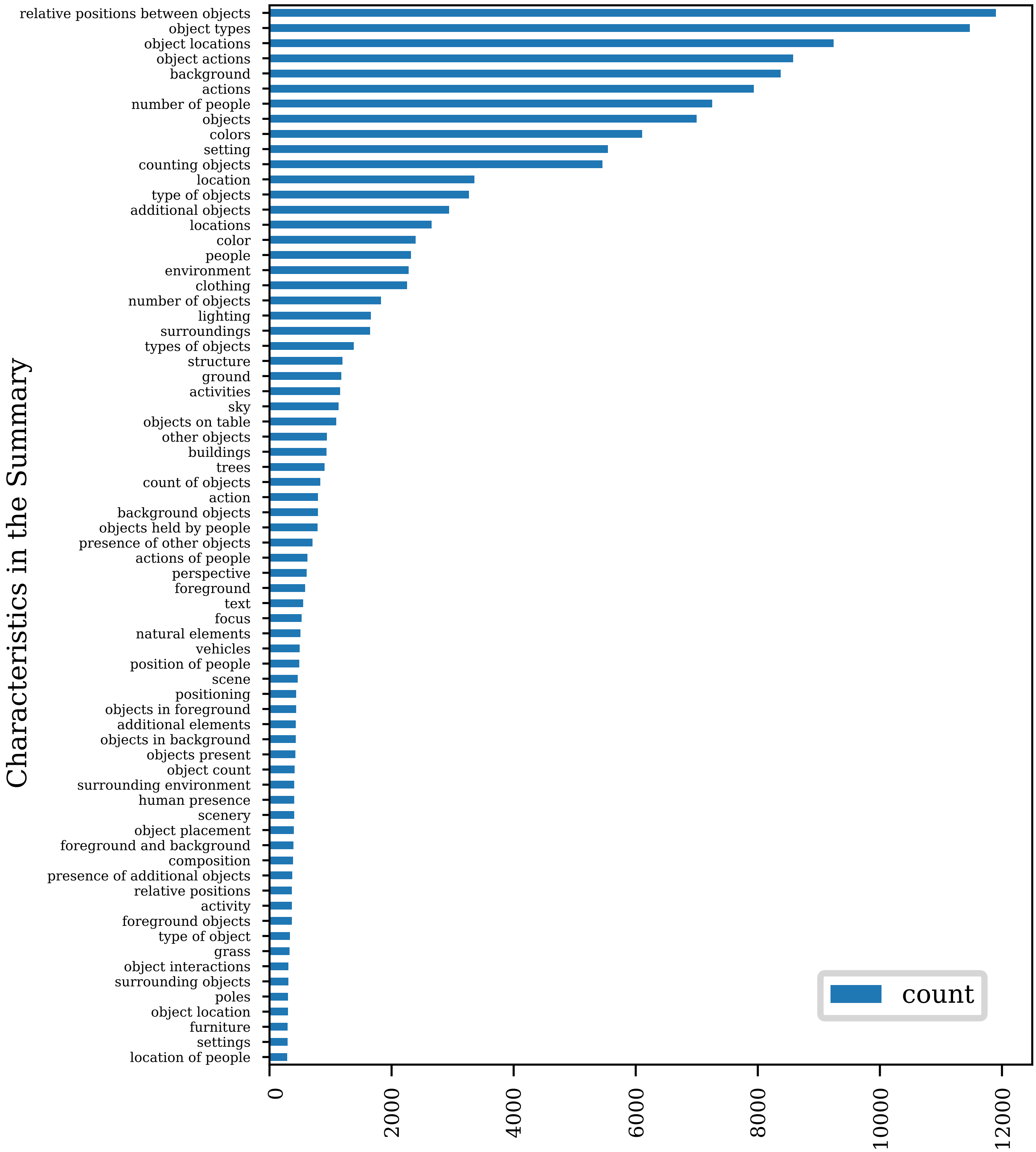
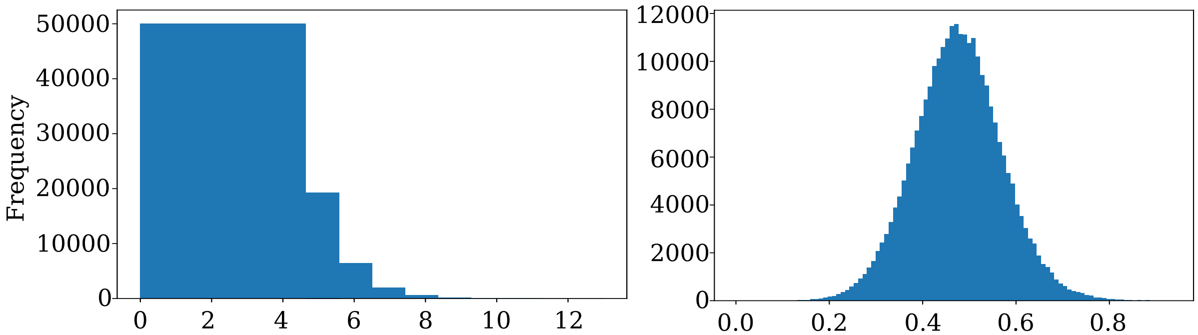


Figure 8. Distribution of sample-specific characteristics (*e.g.* object types, action of people, surrounding environments, *etc.*) in CaD summaries in CaD-Inst^{V1}. The distribution of these sample-specific characteristics is also shown in a Sunburst chart in Fig. 3(a)(main paper).

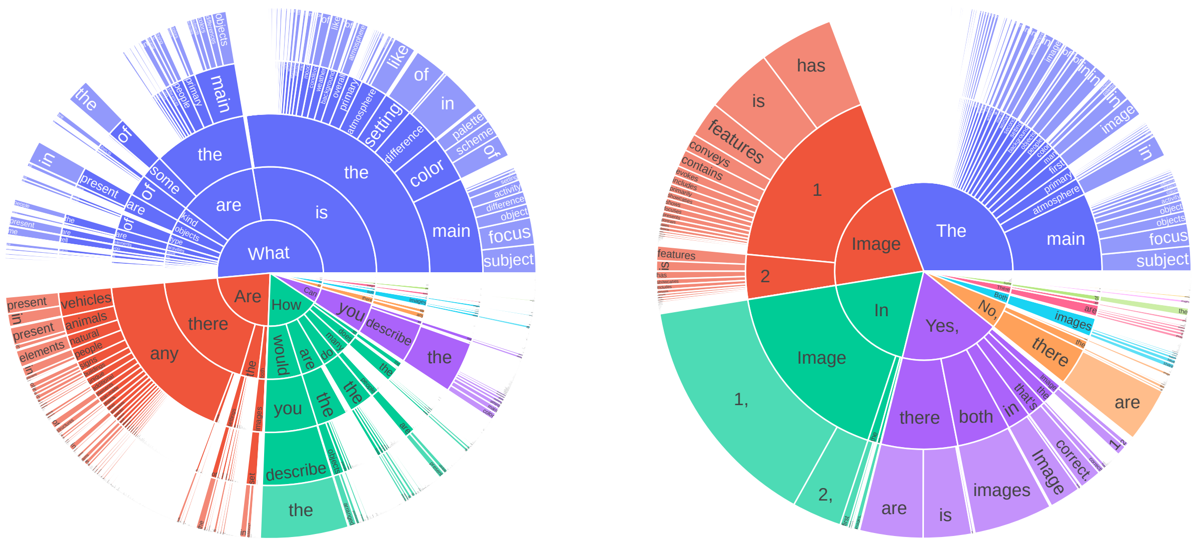
context instruction-response data in MIMIC-IT [21]. We use their most recent open-sourced version Otter-Image-LLaMA7B-LA-InContext available at https://huggingface.co/luodian/OTTER-



(a) Number of overlapping nouns between captions in an image pair

(b) Image-image similarity

Figure 9. Distribution of (a) number of overlapping nouns between captions in an image pair and (b) image-image similarities in the 278K image pairs in CaD-Inst^{V1}



(a) Questions in CaD-QA

(b) Answers in CaD-QA

Figure 10. Distribution of (a) questions (first 5 words) and (b) answers (first 3 words) in the evaluation benchmark CaD-QA.

Image-LLaMA7B-LA-InContext.

MMICL [43] is based on the InstructBLIP model [5]. The model is finetuned their own collected multimodal in-context learning datast consisting of interleaved text-image inputs, inter-related multiple image inputs and multimodal in-context learning inputs. We evaluate with their model of the largest scale MMICL-InstructBLIP-T5-XXL, available at https://huggingface.co/BleachNick/MMICL-Instructblip-T5-xxl.

EMU2-Chat [35] is a generative multimodal model trained on large-scale multimodal sequences. The model consists of pretrained EVA-02-CLIP-E-plus [34] and LLaMA-33B [37]. The model weights and inference code are available at https://huggingface.co/BAAI/Emu2-Chat.

InternLM-XComposer2-VL [41] consists of CLIP ViT-L [30] and InternLM2-7B [36]. The model weights of the InternLM-XComposer2-VL-7B and inference code are available at https://huggingface.co/internlm/internlm-xcomposer2-vl-7b.

LLaVA 1.5 [25] is an improved version from LLaVA [26] with CLIP-ViT-L-336px [30] as the visual backbone and Vicuna 1.5 [44] as the LLM. Our visual instruction tuning is performed using the open-sourced code of LLaVA 1.5. We train on the first-stage pretrained weights of LLaVA 1.5 via LoRA finetuning. We evaluate both LLaVA 1.5 7B lora and LLaVA 1.5 13B lora as baselines. The models are available at https://huggingface.co/liuhaotian/llava-v1.5-7b-lora and https://huggingface.co/liuhaotian/llava-v1.5-

13b-lora.

LLaVA 1.6 [27] is an improved version from LLaVA 1.5 with increased input image resolution and improved mixture of instruction tuning data. The 7B and 13B versions are avaible on Huggingface at https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b and https://huggingface.co/liuhaotian/llava-v1.6-vicuna-13b. However, the training code is not yet available.

4.2. Implementation Details

Data Collection. In Phase-1, we leverage the Mixtral 8x7B Instruct v0.1 model[2] with 8-bit inference for data generation. We set the batch size to 16 and max new token to 750. The prompt for the task of LLM-based CaD summary is given in Fig. 11. The generation with batch 16 fits to an A100 80G GPU.

In Phase-2, we leverage the Phase-1 model CaD-LLaVA^{V1} 13B model to generate CaD summary on additional image pairs. The temporature, max new tokens and number of beams are set to 0, 256 and 1. The prompt for the task of LMM-based CaD summary is given in Fig. 12.

For collecting open-ended QAs in CaD-QA, we first use the LMM to generate the CaD summaries based on the image captions (see Fig. 11). Then we prompt the LLM with both the image captions and the CaD summary, instructing it to generate a multi-turn conversation with several rounds of Q&A. We also provide some in-context samples to demonstrate the desired layout. The prompt for the task of generating Q&A pairs based on both image captions and the CaD summary is illustrated in Fig. 13.

Training. We perform visual instruction tuning following the configuration in LLaVA 1.5. We set the batch size to 128 and train for one epoch. The learning rate for LLM with LoRA and for the projector are set to 1×10^{-4} and 2×10^{-5} correspondingly. The LoRA rank and alpha values are set to 128 and 256. The training experiments are run on $4 \times$A100 80G GPUs.

Inference. For VQA inference, the temperature, max new tokens and number of beams are set to 0, 256 and 1.

LLM-assisted Evaluation We leverage the Mixtral 8×7B model for LLM-assisted evaluation on open-ended questions. We feed the question, correct answer and the predicted answer into the LLM and instruct it to provide a rating between 0 and 5. The prompt for generating the evaluation rating is given in Fig. 14.

[2]Huggingface source: https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

System prompt:
You are an AI visual assistant and you are seeing two images. The two images are provided with two captions, each describing the content of an image. Your task is to summarize the commonalities and differences between the two images. Answer as you are seeing the images. 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.

User prompt:
Please summarize the commonalities and differences between the following two images:
Image 1:<caption1>
Image 2:<caption2>
Commonalities:

Figure 11. Prompt for the task of Phase-1 LLM-based CaD summary.

System prompt:
A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

User prompt:
Image 1: <image>
Image 2: <image>
Here are some context of the difference between the two images:
<description>
Based on the two images and the context, 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.

Figure 12. Prompt for the task of Phase-2 LMM-based CaD summary.

System prompt:
You are an AI visual assistant, and you are seeing two images. The two images are provided with two captions, each describing the content of an image. Additionally, you are provided with a summary of the commonalities and differences between the two images. Design a conversation between you and a person asking about the two images based on the commonalities and differences between the two images. The answers should be in a tone that a visual AI assistant is seeing the two images and answering the question. Ask diverse questions and give corresponding answers. Include questions 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.

User prompt:
Please design a conversation with several rounds of questions and answers, where a person asks about the two images based on the commonalities and differences between the two images, and you answer.

Image 1: In this picture we can see two women are taking picture, side there is a glass and bottle are placed on the table, behind we can see some people are sitting on the chairs.
Image 2: In this image, we can see some buildings. There is a bridge on the right side of the image. There is a person and bus in the middle of the image. There is a wall beside the road. There are divider cones at the bottom of the image. At the top of the image, we can see the sky.
Summary: Commonalities: The two images have the presence of people and man-made structures. In both images, the people are engaged in different activities - taking a picture in Image 1 and standing in Image 2. The man-made structures include a table with objects in Image 1 and buildings, a bridge, a wall, and divider cones in Image 2.
Differences:
1. Scene: Image 1 is set in an indoor or semi-enclosed space, such as a café or restaurant, while Image 2 is set outdoors, in a cityscape or urban environment.
2. Objects: Image 1 contains objects like a glass, bottle, and chairs, which are not present in Image 2. On the other hand, Image 2 has elements like a bridge, wall, and divider cones, which are not in Image 1.
3. People: In Image 1, there are two women taking a picture, while in Image 2, there is a single person standing near a bus.
4. Actions: The people in Image 1 are taking a picture, while the person in Image 2 is not engaged in any specific action.
5. Locations: The scene in Image 1 is more confined, with a table and chairs, while Image 2 has a wider view, including a road, buildings, and the sky.
6. Relative positions between objects: In Image 1, objects are placed on a table, while in Image 2, objects are positioned along a road or street.
Question: Are there people in both images?
Answer: Yes, there are people in both images.
Question: what is the difference in the scene of the two images?
Answer: The scene in Image 1 is set in an indoor or semi-enclosed space, such as a café or restaurant, while Image 2 is set outdoors, in a cityscape or urban environment.
Question: What objects are present in Image 1 but not in Image 2?
Answer: Image 1 contains objects like a glass, bottle, and chairs, which are not present in Image 2.
Question: What objects are present in Image 2 but not in Image 1?
Answer: Image 2 has elements like a bridge, wall, and divider cones, which are not in Image 1.
Question: What is the difference between people in the two images?
Answer: In Image 1, In Image 1, there are two women taking a picture, while in Image 2, there is a single person standing near a bus.
Question: What are the people doing in the two images?
Answer: The people in Image 1 are taking a picture, while the person in Image 2 is not engaged in any specific action.
Question: What is the difference in the locations of the two images?
Answer: The scene in Image 1 is more confined, with a table and chairs, while Image 2 has a wider view, including a road, buildings, and the sky.
Question: What is the difference in the relative positions between objects in the two images?
Answer: In Image 1, objects are placed on a table, while in Image 2, objects are positioned along a road or street.

Image 1: <caption1>
Image 2: <caption2>
Summary: <summary>
Question:

Figure 13. Prompt for the task of generating Q&A pairs based on both image captions and the CaD summary.

System prompt:
You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:

##INSTRUCTIONS:
- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

User prompt:
Please evaluate the following question-answer pair:
Question: <question>
Correct Answer: <answer>
Predicted Answer: <prediction>
Evaluate if the predicted answer is correct with yes/no and assign a correctness score between 0 and 5, where 0 indicates incorrect answer, and 5 signifies the highest meaningful match. Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {'pred': 'no', 'score': 0}.

Figure 14. Prompt for the LLM-assisted evaluation.

5. List of Assets

Our image sources and annotations are obtained from public datasets. We release our data in accordance to the source data licenses.

Here is a list of image sources:

- Open Images v6 [18] (https://storage.googleapis.com/openimages/web/download_v6.html): The images are under Creative Commons Attribution (CC BY) 2.0 license.
- COCO 2017 [2, 24] (https://cocodataset.org/#download): The images are under a Creative Commons Attribution 4.0 license.
- Flicker30K [40] (https://shannon.cs.illinois.edu/DenotationGraph/): The images are the property of SmugMug or its third party licensors and are protected by United States and international intellectual property laws. The images are provided for researchers and educators who wish to use the dataset for non-commercial research and/or educational purposes.
- ADE20K [45] (https://groups.csail.mit.edu/vision/datasets/ADE20K/index.html#Download): The images belong to MIT CSAIL and are licensed under a Creative Common BSD-3 License.
- Visual Genome [16] (https://homes.cs.washington.edu/~ranjay/visualgenome/api.html): The images are under a Creative Commons Attribution 4.0 license.

Here is a list of image annotation sources:

- Localized narratives [29] (https://google.github.io/localized-narratives/): The annotations are released under a Creative Common Attribution (CC BY) 4.0 license.
- MIMIC-IT [21] (https://huggingface.co/datasets/pufanyi/MIMICIT): The annotations are released under an MIT license.
- SVIT [42] (https://huggingface.co/datasets/BAAI/SVIT): The annotations are licensed under a Creative Commons Attribution 4.0 license. It should abide by the policy of OpenAI (https://openai.com/policies/terms-of-use). The use of original images and annotations from Visual Genome and MS-COCO should comply with the original licenses.

Here is a list of implementation sources or model weights:

- LLaVA [25, 26] (https://github.com/haotian-liu/LLaVA): The code is released under an Apache-2.0 license. The project utilizes certain datasets and checkpoints that are subject to their respective original licenses, including but not limited to the OpenAI Terms of Use[3] for the dataset and the specific licenses for base language models for checkpoints trained using the dataset (*e.g.* LLaMA community license[4] for LLaMA-2 and Vicuna-v1.5).
- Mixtral 8×7B model [15] (https://huggingface.co/mistralai/Mixtral-8x7B-v0.1): The model is released under an Apache-2.0 license. Usage is subject to the term of use for Mistral products and services[5].

[3] https://openai.com/policies/eu-terms-of-use/

[4] https://ai.meta.com/llama/license/

[5] https://mistral.ai/terms/#terms-of-use

References

[1] 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. 10

[2] 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. 17

[3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 10

[5] 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. 12

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 10

[7] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 10

[8] 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. 2

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 10

[10] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. 9

[11] 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. 9

[12] 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. 9

[13] 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. 10

[14] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 9

[15] 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. 17

[16] 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. 17

[17] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 10

[18] 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. 17

[19] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 3

[20] 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. 10

[21] 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. 9, 11, 17

[22] 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. 10

[23] 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. 10

[24] 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. 17

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 12, 17

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023. 12, 17

[27] 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. 13

[28] 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. 2

[29] 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. 17

[30] 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. 9, 12

[31] 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. 9

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 9

[33] 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. 10

[34] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 12

[35] 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. 12

[36] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023. 12

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 12

[38] 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. 9

[39] Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, pages 709–725. Springer, 2022. 9

[40] 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. 17

[41] 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. 12

[42] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 17

[43] 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. 12

[44] 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. 12

[45] 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. 17

[46] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 9

[47] 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. 10