

Supplementary Materials: Semantic Alignment for Multimodal Large Language Models

Anonymous Authors

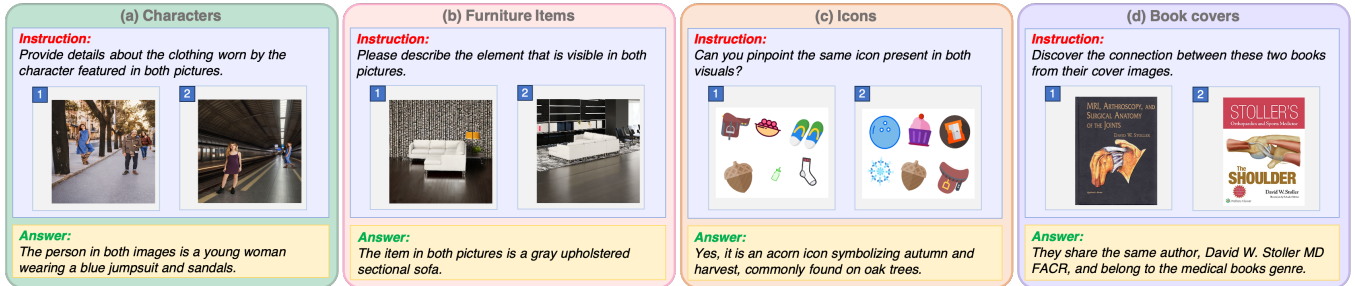


Figure 1: Demonstration of MmLINK’s samples containing each kind of objects. Each sample in our MmLINK dataset consists of an image pair, a corresponding language instruction, and a textual answer for the instruction. The images in the same pair depict the same object with different poses, lighting conditions or view angles. To ensure the diversity of the image objects in our MmLINK dataset, we collect images with variance objects (e.g., characters, furniture items, icons, and book covers) from different image sources.

1 DATASET CONSTRUCTION DETAILS

As illustrated in Section 3 in the paper, each sample in our MmLINK dataset consists of an image pair, a corresponding language instruction, and a textual answer for the instruction. The two images in the same pair depict the same object with different poses, lighting conditions or view angles. To ensure the diversity of the image objects in our MmLINK dataset, we collect images with variance objects (e.g., characters, furniture items, icons, and book covers) from different image sources, as shown in Figure 1. Number of samples containing each kind of objects in our MmLINK dataset is shown in Table 1.

1.1 Construction details for different objects

For different types of objects appeared in our MmLINK dataset, slightly various pipelines are employed to create the corresponding samples. In this section, we provide information on the construction pipeline for samples involving all kind of objects.

1.1.1 Characters. The construction detail of character data is illustrated in Section 3 of the paper.

1.1.2 Furniture Items. Compared to characters, the construction detail of samples involving furniture items has only minor differences as illustrated in Figure 2. Specifically, in the first step, we construct image groups for different furniture items from the Amazon Berkeley Objects (ABO) Dataset [5], which contains a comprehensive 360-degree perspective of furniture items. We select two different perspectives of the same furniture items as the source images. Different from character data, we do not include two other different furniture images here, considering that multiple items may pose challenges for inpainting. When the source images are prepared, the remaining steps are identical to those featuring characters: firstly, we segment each item from the source image using Segment Anything [12], and place them onto a 512×512 mask

Table 1: The distribution of samples containing each type of object in MmLINK.

Object	Sample Number
Characters	3543
Furniture Items	16079
Icons	27741
Book Covers	21012
Total	68375

image separately with diverse positions and sizes. In this way, we obtain two mask images featuring the same furniture item, which exhibit different view angles and sizes between the images.

In the second step, we inpaint the two mask images with diverse background descriptions generated by ChatGPT. The inpainting process is also achieved by Stable-Diffusion-Inpainting [26], which can generate background pixels smoothly integrated with the items according to the background descriptions. The two inpainted images serve as the visual components of each training sample in MmLINK, challenging models to identify the same furniture item across images despite varied contexts. Finally, we employ Instruct-BLIP [6] to generate descriptions of the shared furniture items and use ChatGPT to obtain refined descriptions, which serve as the textual components in the final multi-modal samples.

1.1.3 Icons. The Icon645 dataset [19] offers 645K colored icon pictures across 377 categories, such as cake and dog. We construct 2 groups of icon images, where each group contains 4 to 6 icon categories. The categories are randomly selected, and we ensure that only one category will appear in both groups. In this way, the two groups of icon images have only one icon images in common, serving as the linking information between them. Then we place

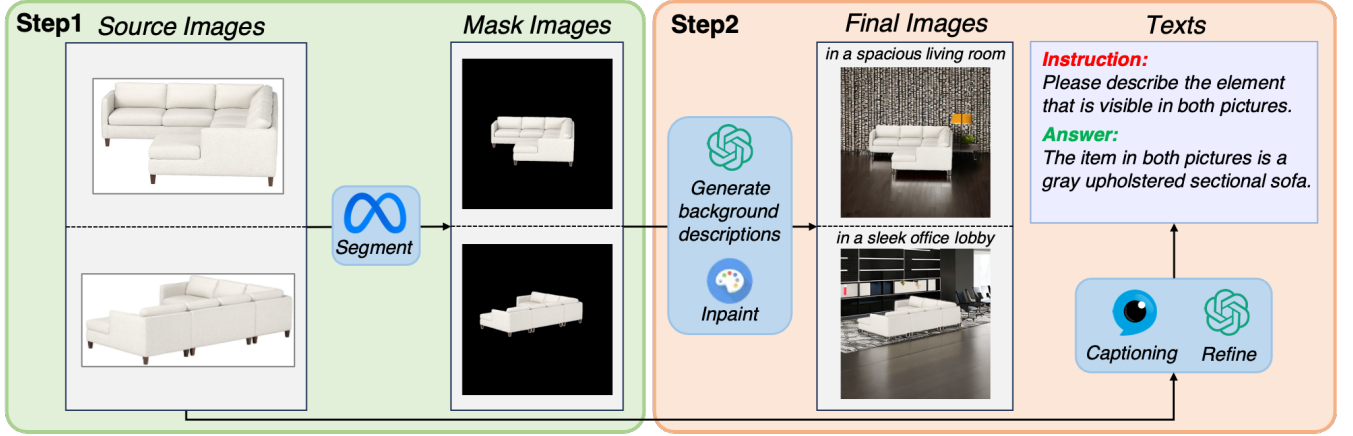


Figure 2: Demonstration of our proposed 2-step sample synthesis pipeline when creating samples containing furniture items as the linking information. We begin by selecting images featuring furniture items in different view angles. The selected images are segmented to isolate each item, after which they are merged into mask images. Inpainting technology is then utilized to fill in the background areas of these mask images to obtain the final images, using descriptions generated by ChatGPT. Text annotations are generated by InstructBLIP and further refined with ChatGPT.

each group of icon image onto a white background image, where each icon image has a random size and rotation angle. The resulting two images serve as the visual components of each training sample, challenging models to identify the same icon across images despite varied contexts and presence of other icon images. The construction detail of textual components are the same as the one of furniture items, that is, we employ InstructBLIP [6] to generate descriptions of the shared icon image and use ChatGPT to obtain refined descriptions as the final answer.

1.1.4 Book covers. We construct samples involving book covers from OCR-VQA [20] dataset. The origin dataset contains about 207K book cover images, with attributes including author, publish year and genre. We randomly select two images, which share at least one identical attribute, as the visual components. Then we extract the identical attributes of the two images from the original dataset, and utilize ChatGPT to organize them into the final answer. These training sample involving book covers aim to enhance model's semantic alignment between objects recognized by its OCR ability.

1.2 Prompts for dataset construction

In our 2-step sample synthesis pipeline, we utilize ChatGPT for diverse background descriptions generation and final answer refinement. We list the detailed prompts in Figure 4. Note that the italicized text enclosed in braces in Figure 4 is the content that needs to be filled in for each sample.

2 FRAMEWORK OF ADAPTIVE ADJUSTMENT

As introduced in Section 4.3 in the paper, we illustrate the detailed framework of the adaptive adjustment module in our proposed bidirectional semantic guidance mechanism in Figure 3. Please recall that P is the patch number in the visual features extracted by vision encoder, i denotes the currently perceived image index and $m \neq i$ denotes the contextual image (i.e., images other than the currently perceived image) index. The patch-level features of

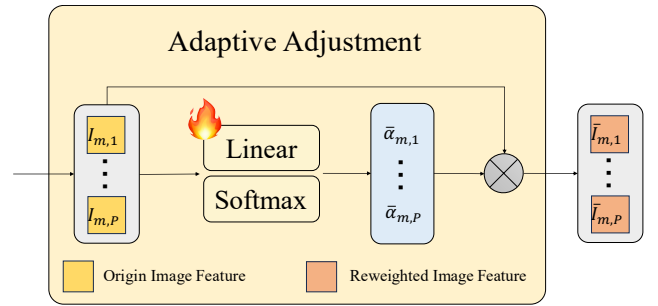


Figure 3: Detailed framework of the adaptive adjustment module in our proposed bidirectional semantic guidance mechanism. For each image feature extracted by the vision encoder, the adaptive adjustment module uses a linear layer and the softmax function to generate normalized patch-level weights, and then reweights each patch feature using the normalized weights.

the m -th ($m \neq i$) image are denoted as $\{I_{m,j}\}_{j=1}^P$. To eliminate the prominence of irrelevant patches, we assign weights to each image patch with the adaptive adjustment module. Firstly, it uses the linear layer $\text{Linear}(\cdot)$ and the softmax function $\text{softmax}(\cdot)$ to generate normalized patch-level weights $\tilde{\alpha}_{m,j}$:

$$\tilde{\alpha}_{m,j} = \text{softmax}(\text{Linear}(I_{m,j})) \quad (1)$$

Then, the adaptive adjustment module re-weights each contextual patch features and sums up the re-weighted patch features to obtain the merged contextual features \hat{I} :

$$\hat{I} = \{\hat{I}_j\}_{j=1}^P, \quad \hat{I}_j = \sum_{m \neq i} (\tilde{\alpha}_{m,j} * I_{m,j}) \quad (2)$$

This process reduces the influence of irrelevant patches and amplifies important ones, improving patch-level alignment between the merged contextual features \hat{I} and the currently perceived ones.

1. Background description generation of character images

Given diverse descriptions of the possible environments where two standing people might be. For example, "in a dark train station". Output 100 such descriptions directly and make sure your descriptions are diverse.

2. Background description generation of furniture item images

First, I will provide you with a statement about an object, which may be quite lengthy, containing many attributes and details and may not be described in English. I need you to provide me with some descriptions of where this object can be placed. Firstly, these descriptions must have a subject, and the subject is the object itself. So, you first need to identify what the object is from the given statement and use up to three English words as the subject. Then, you need to describe its surroundings, which must be brief, like 'in a living room' or 'on a slightly messy floor'. Here is an example:

the statement is: Amazon Brand – Rivet Goodwin Modern Sofa, 88.6"W, Light Grey

Sample descriptions are:

1. The sofa is in a living room
2. The sofa is on a spacious patio
3. The sofa is in a modern office
4. The sofa is in a rustic cabin.

Now the statement is: *{original furniture item statement}*. Only output 4 descriptions without any other words, and the scenarios in these descriptions should be diverse.

3. Refinement of character image captions

Assuming there is a task that requires you to find the person appearing in both images and describe his or her attire. The correct answer is available, but it is quite lengthy and derived from the analysis of a single image. Your first step is to extract relevant information and compress it. Remember that information only about attire is need, other information like background or actions should be discarded; secondly, you need to organize the response format to make it appear as an answer to the task that find the person appearing in both images and describe his or her attire. Your language organization should be diverse, do NOT use the same pattern such as 'The person appearing in both images is ...' all the time. Only output your answer without any other words. Now the answer is *{captions of character images}*

4. Refinement of furniture item image captions

There is a game where people describe identical items in two pictures. I have one answer, but it may be incomplete or contain many other details. You need to rewrite this answer, retaining only the relevant descriptions related to the items, discarding other details (like its surroundings), to make the response more formal and concise, attempting to describe it in one sentence. But don't forget that you are looking for common items in two pictures, so your answer should reflect this. Now the description is *{captions of furniture item images}*

5. Refinement of icon image captions

I will provide you with a description of an image, and you need to extract the key information from it and then paraphrase it. Your paraphrased description should be a little concise. Directly start with your descriptions, like 'an airplane of ...' or 'a red apple'. Don't contain prefixes, such as 'The image depicts' and so on. Description: *{captions of icon images}*

6. Refinement of identical attributes of book cover images

Imagine there is a task, which requires you to find the common parts between the covers of two books. There is a correct answer, maybe contains their author, genre and publish year. You need to rewrite this answer, ensuring that the meaning remains unchanged and information undiscarded while expressing it in a more varied manner. Answer: *{identical attributes of two book covers}* Directly output your answer without any other words.

Figure 4: Prompt templates of ChatGPT for dataset construction. ChatGPT is utilized in our 2-step sample synthesis pipeline for diverse background description generation (1, 2) and response refinement (3, 4, 5, 6). The italicized text enclosed in braces is the content that needs to be filled in for each sample. Specifically, *{original furniture item statement}* in 2 is the item descriptions from the original ABO dataset [5].

3 IMPLEMENTATION DETAILS

3.1 Test dataset details

We conduct zero-shot evaluation on Group Captioning and Storytelling tasks to evaluate SAM’s generalization ability. Since the chosen datasets are not naturally suited for testing multi-modal large language models, we preprocess the raw data and filter out some dirty data, forming the final test sets, each comprising 500 samples. In the group captioning task, the model is tasked with generating descriptions that capture commonalities within a given set of images. Specifically, for this task, we select three datasets: Conceptual [16], Animal and Vehicle [9]. Note that each sample in the original Conceptual dataset contains 20 images, requiring models to describe a group of target images (5 images) in the context of another group of related reference images (15 images). We only keep the 5 target images during our test. The storytelling task provides N related pictures and a story description for each of the first $N - 1$ pictures. The model is required to generate a story that corresponds to the content of the last picture and maintains coherence with the proceeding stories. For this task, we select three public datasets: AESOP [25], VIST [10] and DM800K [3].

In addition, we conduct experiments on the change captioning task, demonstrating that SAM still retains the ability of discovering correlations between highly similar images. We select 4 change captioning datasets used by Li et al. [15]: IEdit [2], Spot-the-Diff [11], Birds-to-Words [8] and CLEVR-Change [23]. Each dataset contains 500 samples, challenging models to describe subtle differences between two highly similar images.

3.2 Metric details

To comprehensively evaluate the model, we use various evaluation metrics.

3.2.1 ROUGE-L. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [17] is a metric commonly used to evaluate the quality of text summarization or machine translation outputs by comparing them to a set of reference summaries. ROUGE-L calculates the longest common subsequence (LCS) between the generated summary and the reference summaries. It measures the recall of the generated summary by computing the ratio of the LCS length to the length of the reference summary. This recall score is then used to evaluate the quality of the summary. In essence, ROUGE-L emphasizes the importance of capturing the longest common subsequence of words between the generated and reference summaries, providing a measure of how well the generated summary covers the content of the reference summaries.

3.2.2 CIDEr. CIDEr (Consensus-based Image Description Evaluation) [28] is a metric commonly used to evaluate the quality of image descriptions generated by machine learning models. It assesses the consensus between generated descriptions and human reference descriptions. The calculation process involves computing cosine similarity between n -grams (sequences of n words) in the generated descriptions and reference descriptions. These similarities are then aggregated and transformed into a final score using various techniques like TF-IDF weighting and inverse document frequency smoothing. CIDEr emphasizes consensus by considering

multiple reference descriptions and has been shown to correlate well with human judgments of description quality.

3.2.3 BLEU-4. BLEU (Bilingual Evaluation Understudy) [22] is a metric commonly used to evaluate the quality of machine-translated text by comparing it to one or more reference translations. BLEU is based on the idea that “good” translations are those that are similar to reference translations and use similar n -grams (contiguous sequences of n items, typically words). BLEU calculates precision by counting the number of n -grams in the candidate translation that match any n -gram in the reference translations. BLEU scores range from 0 to 1, where 1 indicates a perfect match between the candidate and reference translations. In this paper we use BLEU-4, which specifically refers to the use of 4-grams in the calculation and provides a more comprehensive evaluation compared to lower n -gram values.

3.3 Baseline details

For a comprehensive comparison, we consider open-source state-of-the-art MLLMs as well as industrial multi-modal chatbots.

3.3.1 MiniGPT-4. MiniGPT-4 [30] enhances vision-language understanding by integrating a pre-trained vision encoder with an advanced LLM through a single projection layer. It is pre-trained on 5M image-text pairs and then fine-tuned with a 3.5K visual instruction dataset for improved natural language generation. The tested version is “MiniGPT-4 (LLaMA-2 Chat 7B)”.

3.3.2 LLaVA. LLaVA [18] is an end-to-end trained model that integrates a vision encoder (*i.e.*, ViT-L/14 from CLIP [24]) with a LLM (*i.e.*, LLaMA [27]) using a simple fully-connected layer, enabling it to comprehend and respond to multi-modal instructions. It leverages machine-generated instruction-following data, created using GPT-4, to improve zero-shot capabilities across new tasks. Using this 158K instruction data, LLaVA fine-tunes the FC layer and the LLM. The tested version is “llava-v1.5-7b-lora”.

3.3.3 BLIP2. BLIP2 [14] uses a trainable module called Q-Former in a two-stage pre-training process: first for vision-language representation learning from a frozen image encoder (*i.e.*, ViT-g/14 from EVA-CLIP [7]), and second for vision-to-language generative learning from a frozen LLM (*i.e.*, FlanT5 [4]). BLIP-2 is trained on a dataset comprising 129M images including various sources. The tested version is “blip2-pretrain-flanT5xl”.

3.3.4 InstructBLIP. InstructBLIP [6] introduces an approach to vision-language instruction tuning using the BLIP-2 model, transforming 26 datasets across 11 task categories into instruction format. It features an instruction-aware Q-former for adaptive visual feature extraction, trained on a mix of held-in datasets for instruction tuning and evaluated on held-out datasets for zero-shot generalization. The tested version is “blip2-vicuna-instruct-7b”.

3.3.5 Otter. Otter [13] introduces a multi-modal model that leverages instruction tuning to enhance instruction-following abilities in multi-modal contexts. It employs the MIMIC-IT dataset, comprising image-instruction-answer triplets with contextual examples for training. Otter integrates a LLaMA-7B language encoder and a CLIP

Table 2: Detailed performance on change captioning tasks. R-L indicates ROUGE-L, C is CIDEr and B-4 is BLEU-4. The best is bolded and the second best is underlined.

Model	IEdit			Spot-the-Diff			Birds-to-Words			CLEVR-Change			Average		
	R-L	C	B-4	R-L	C	B-4	R-L	C	B-4	R-L	C	B-4	R-L	C	B-4
MiniGPT-4	8.37	1.46	0.24	14.07	0.97	0.62	15.63	0.86	0.37	14.99	<u>0.39</u>	0.52	13.27	0.92	0.44
LLaVA	7.16	0	0.18	13.75	0.13	1.06	13.48	0.49	0.55	13.72	1.19	0.88	12.03	0.45	0.67
BLIP2	12.81	1.69	0	18.54	<u>7.82</u>	1.02	9.46	1.57	0.35	12.94	0	0	13.44	2.77	0.34
InstructBLIP	10.08	<u>4.51</u>	0.32	11.65	1.82	0.88	12.05	0.78	0.59	9.81	0.38	0.21	10.9	1.87	0.5
Otter	4.23	0.87	0.37	15.35	5.73	0.64	14.95	<u>2.85</u>	0.4	13.48	0	0	12	2.36	0.35
MMICL	10.26	3.14	0.38	18.09	4.56	1.02	14.7	2.39	0.58	15.88	0.05	0.07	14.73	2.54	0.51
GPT-4V	8.45	0.95	0.61	19.47	10.02	<u>2.31</u>	18.7	2.35	1.44	21.84	0.2	0.45	<u>17.12</u>	<u>3.38</u>	1.2
Gemini Pro	11.34	1.23	0.47	<u>19.8</u>	7.15	1.83	<u>18.43</u>	2.8	<u>1.17</u>	<u>20.78</u>	0.31	<u>0.71</u>	17.59	2.87	1.05
SAM	<u>12.24</u>	4.95	<u>0.52</u>	22.32	7.05	3.02	14.8	3.3	0.76	14.97	0.05	0	16.08	3.84	<u>1.08</u>

ViT-L/14 vision encoder, involving approximately 1.3 billion trainable parameters. The tested version is “OTTER-Image-MPT7B”.

3.3.6 Cheetah. Cheetah [15] introduces VPG-C module for enhancing MLLMs to better understand and follow demonstrative instructions by focusing on missing visual details within images. VPG-C operates alongside a frozen LLM and a vision encoder. It employs a synthetic discriminative training strategy, which generates 64K training samples, making the training process efficient and less data-dependent. The tested version is “cheetah-vicuna-7b”.

3.3.7 MMICL. MMICL [29] is designed to enhance MLLMs by efficiently handling multi-modal inputs through a novel context scheme and a constructed multi-modal in-context learning (MIC) dataset. The training involves a two-stage process where the model’s components, except for the image encoder, Q-former, and LLM, are adjusted during the multi-modal in-context tuning phase using part of the 5.8M MIC dataset. The tested version is “MMICL-Instructblip-T5-xxl”.

3.3.8 GPT-4V. GPT-4V [21] is an advanced version of OpenAI’s language models, incorporating multi-modal capabilities that allow it to understand and generate not only text but also images and other forms of media. This enhancement enables GPT-4V to perform a broader spectrum of tasks effectively, ranging from providing detailed descriptions to creating intricate visual content. It can interpret queries that involve both textual and visual contexts, making it ideal for applications in creative industries, educational sectors, and technical fields. With a profound grasp of nuanced language and multimedia content, GPT-4V delivers richer, more contextually aware interactions, showcasing significant strides in AI’s ability to mimic human cognitive functions. The tested version is “gpt-4-vision-preview”.

3.3.9 Gemini Pro. Gemini Pro [1] is Google’s latest breakthrough in artificial intelligence, integrating advanced multi-modal capabilities that enable it to process and understand both text and visual inputs. This cutting-edge technology enhances Gemini Pro’s adaptability across various platforms and tasks, from generating artful visual content to providing sophisticated textual analysis. It’s designed to support a wide range of applications, including creative

endeavors, educational tools, and complex problem-solving scenarios. Gemini Pro leverages deep learning algorithms to deliver more intuitive and context-aware responses, making it a powerful tool for industries seeking to harness the full potential of AI in diverse and dynamic environments. The tested version is “gemini-pro-vision”.

4 MORE EXPERIMENTAL RESULTS

4.1 Detailed performance on the change captioning task

In issue 2 analyzed in Section 5.3 of the paper, we report the average performance on the 4 datasets of the change captioning task, demonstrating that SAM still performs well when reasoning on highly similar images. We provide detailed performance of each model on each dataset in Table 2.

4.2 Detailed performance with different data volumes

In issue 4 analyzed in Section 5.3 of the paper, we report the average performance on the group captioning task and storytelling task with different training data volumes, demonstrating that training SAM is data-efficient. We provide detailed performance of each training data volumes in Table 3.

4.3 Detailed performance of different interaction layers

In issue 5 analyzed in Section 5.3 of the paper, we report the average performance on all 6 test datasets with different bidirectional interaction layer configurations, demonstrating that the final layers of the Q-former and the W-former are the best positions to conduct bidirectional semantic guidance. We provide detailed performance of each interaction layer configuration in Table 4.

4.4 More case studies

In this section, we provide comparisons of our SAM model with the baselines on some group captioning and storytelling samples in Figure 5 - 9.

Table 3: Detailed performance with different data volumes.

(a) Group Captioning												
Data Volumes	Conceptual			Animal			Vehicle			Average		
	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4
0	12.61	13.69	1.11	14.46	2.81	2.15	15.55	2.45	1.79	14.21	6.32	1.68
23K	16.29	12.85	1.68	18.58	8.87	2.48	19.44	8.07	2.74	18.10	9.93	2.30
46K	17.46	13.63	2.10	19.19	14.23	3.24	20.42	12.54	3.28	19.02	13.47	2.87
69K	20.93	20.19	2.78	19.40	19.92	3.27	20.36	18.35	3.62	20.23	19.49	3.22
92K	19.74	16.54	2.46	19.15	16.01	3.19	20.15	15.33	3.17	19.68	15.96	2.94

(b) Storytelling												
Data Volumes	AESOP			VIST			DM800K			Average		
	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4
0	19.12	18.28	2.72	16.96	28.79	2.38	8.21	7.75	1.22	14.76	18.27	2.11
23K	20.36	22.63	3.63	16.68	27.19	2.04	12.07	7.42	1.30	16.37	19.08	2.32
46K	21.76	21.83	3.62	19.69	34.71	3.04	13.76	13.58	2.20	18.40	23.37	2.95
69K	23.45	25.92	4.13	20.85	39.32	3.65	14.33	11.16	1.97	19.54	25.47	3.25
92K	22.97	25.01	3.90	20.70	39.90	3.75	14.08	11.27	1.96	19.25	25.39	3.20

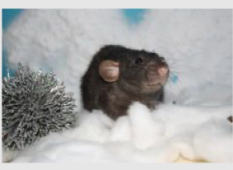
Table 4: Detailed performance with different interaction layers. l is the layer that conveys initial visual tokens, k is the layer that conveys contextual semantics, $k \geq l$ (l and k are defined in Section 4.2 of the paper).

(a) Group Captioning												
l - k	Conceptual			Animal			Vehicle			Average		
	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4
3-3	18.1	12.73	2	19.36	12.36	2.72	20.44	11.98	2.73	19.30	12.36	2.48
3-6	18.41	16.92	2.2	18.85	14.12	3.07	20.55	15.75	3.19	19.27	15.60	2.82
3-9	18.97	16.9	2.3	19.43	16.14	3.2	20.31	16.65	3.4	19.57	16.56	2.97
3-12	19.86	18.16	2.58	19.2	17.3	3.39	19.98	18.87	3.32	19.68	18.11	3.10
6-6	18.51	14.15	2.3	19.26	14.37	2.83	20.62	16.06	3.33	19.46	14.86	2.82
6-9	18.83	15.39	2.21	19.1	16.39	3.04	20.03	17.81	3.4	19.32	16.53	2.88
6-12	19.7	17.17	2.6	18.95	17.45	3.36	20.13	17.55	3.35	19.59	17.39	3.10
9-9	19.35	17.9	2.41	19.23	17.83	3.16	20.07	16.16	3.18	19.55	17.30	2.92
9-12	19.68	16.63	2.6	19.27	16.79	3.22	20.14	16.41	3.22	19.70	16.61	3.01
12-12	20.93	20.19	2.78	19.40	19.92	3.27	20.36	18.35	3.62	20.23	19.49	3.22

(b) Storytelling												
l - k	AESOP			VIST			DM800K			Average		
	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4
3-3	22.14	24.32	3.78	19.16	32.68	2.35	14.35	6.35	1.73	18.55	21.12	2.62
3-6	21.56	22.74	3.32	19.25	34.64	3.05	13.12	11.67	1.44	17.98	23.02	2.60
3-9	22.43	23.89	3.9	19.44	34.65	3.01	14.35	13.37	1.91	18.74	23.97	2.94
3-12	22.67	25.12	3.68	19.85	35.88	3.11	14.25	11.94	2	18.92	24.31	2.93
6-6	22.2	24.25	3.62	19.21	33.48	2.48	14.59	5.92	1.52	18.67	21.22	2.54
6-9	22.64	23.78	3.72	19.64	35.77	3.18	14.13	12.63	1.57	18.80	24.06	2.82
6-12	22.86	25.01	3.75	20	36.27	3.19	14.07	15.16	2.1	18.98	25.48	3.01
9-9	22.57	22.96	3.91	19.44	33.68	2.85	14.32	11.24	1.73	18.78	22.63	2.83
9-12	22.97	24.96	3.85	20.05	37.03	3.24	13.89	13.48	1.89	18.97	25.16	2.99
12-12	23.45	25.92	4.13	20.85	39.32	3.65	14.33	11.16	1.97	19.54	25.47	3.25



Figure 5: Comparisons of SAM model with the baselines on group captioning samples.



Sum up the common factor in these five pictures in a single statement.



MiniGPT4

A rat is standing in the snow, looking up at the camera. The rat's fur is white and it has a candy cane in its mouth. The rat is on the ground and appears to be enjoying the snow.

LLaVA

The common factor in these five pictures is that they all feature a mouse or a similar small animal.

BLIP2

a rat in the snow is a rat in the snow is a rat in the snow is a rat in the

InstructBLIP

In the image, a small brown mouse is perched on a tree branch in the forest, surrounded by fallen leaves and other debris. The mouse is likely searching for food or exploring its surroundings amidst the natural environment.

Otter

All five images feature a small brown mouse.

Cheetah

All of these animals are white and

MMICL

a rat is standing in the snow with a candy cane in front of it and another rat is standing in the snow with a candy cane next to...[...][...][...][...]

GPT-4V

All five pictures have a rodent in the snow.

Gemini Pro

Rodents in various settings, mostly in the snow.

SAM

In all five images, a small brown rat is standing on a snow-covered surface, surrounded by snowflakes.

Figure 6: Comparisons of SAM model with the baselines on group captioning samples.

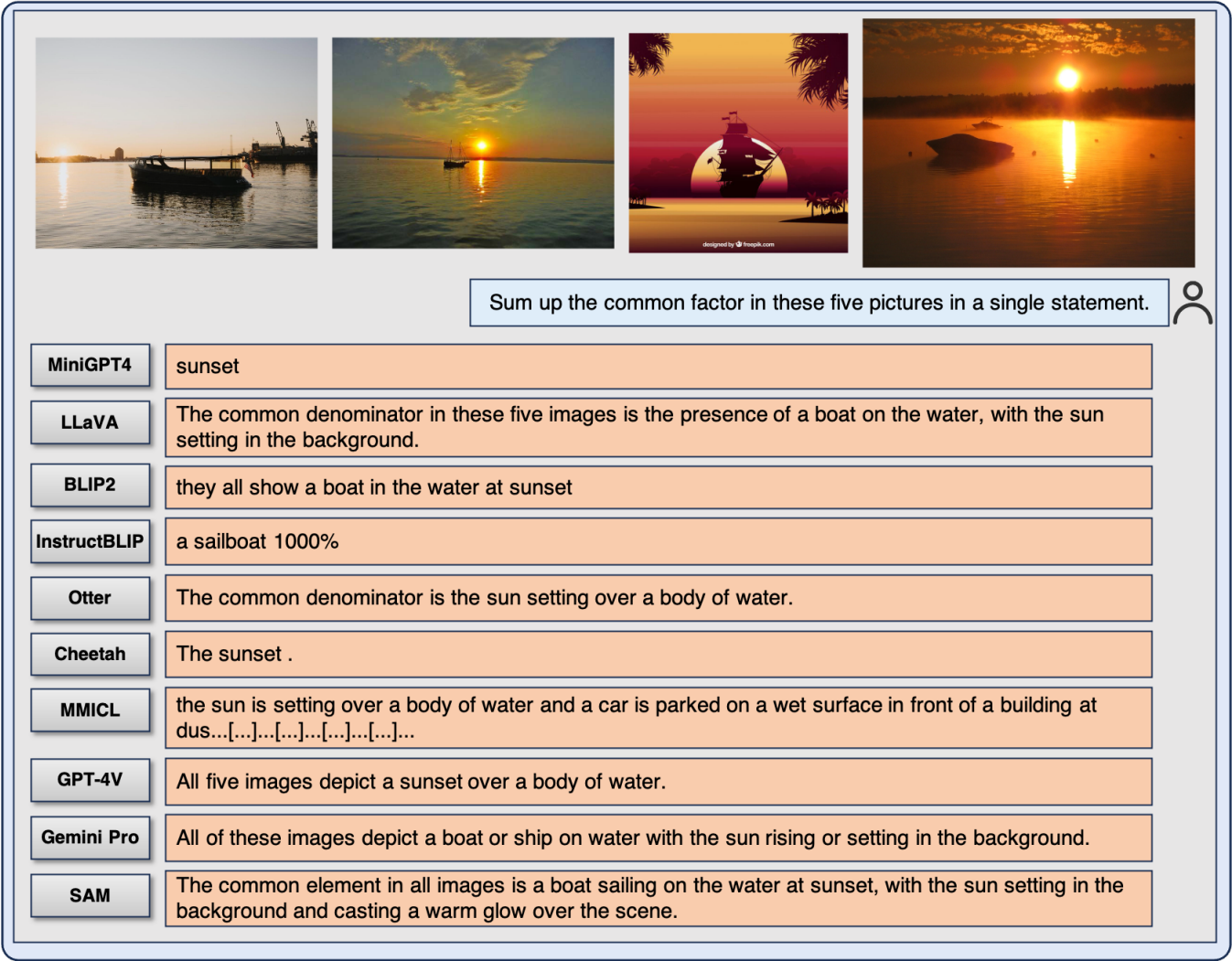


Figure 7: Comparisons of SAM model with the baselines on group captioning samples.

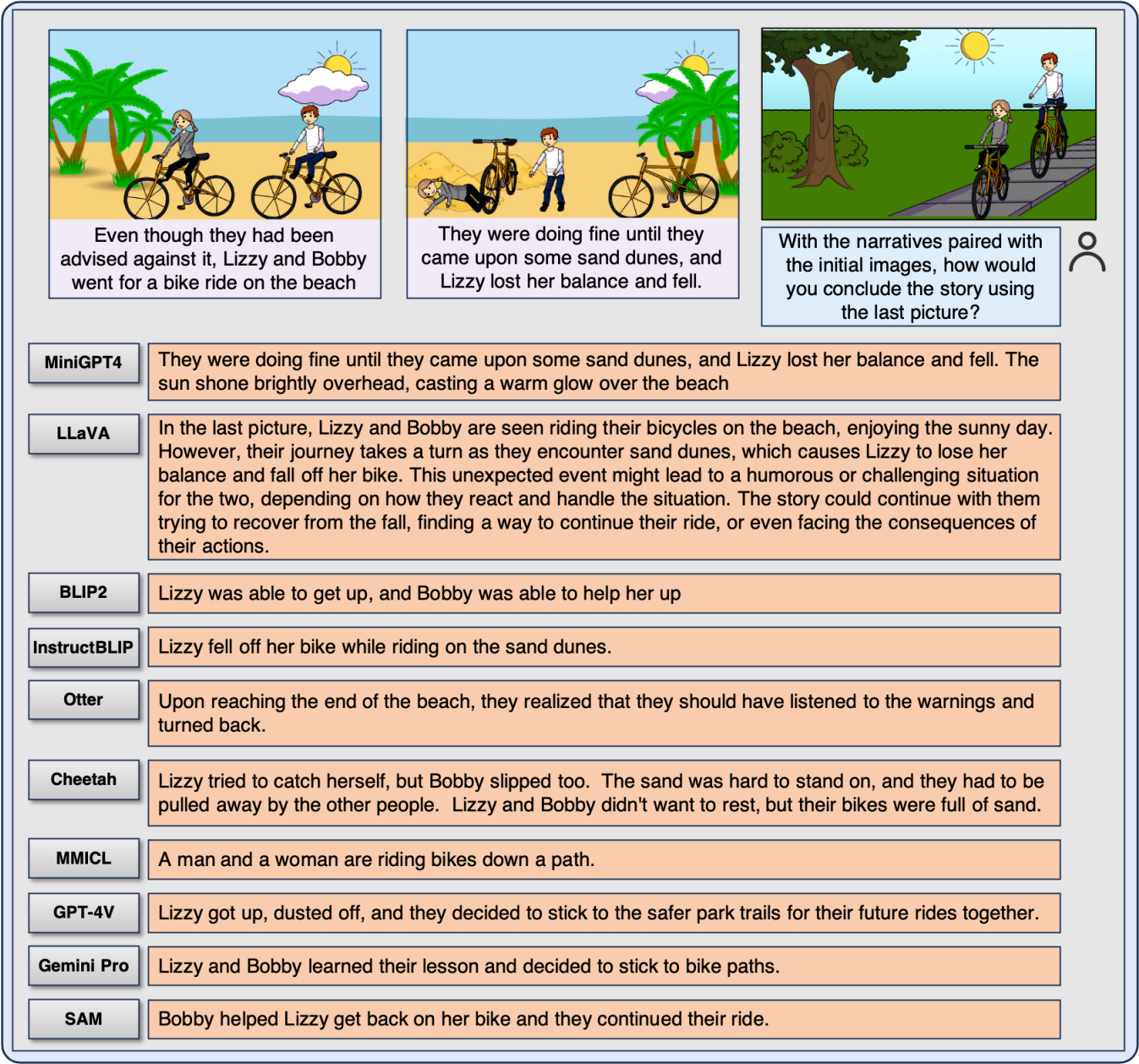
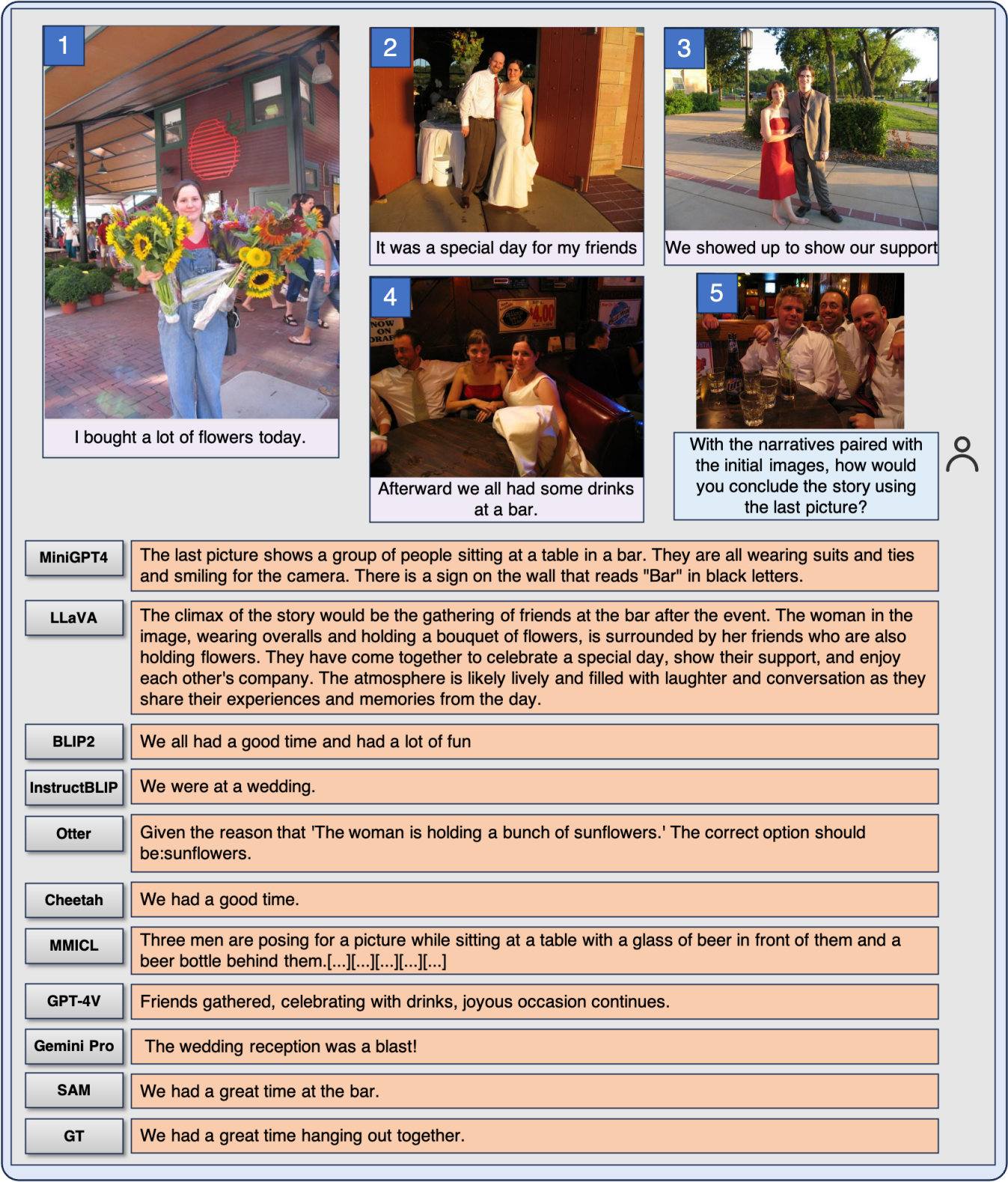


Figure 8: Comparisons of SAM model with the baselines on storytelling samples.



REFERENCES

- [1] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR* abs/2312.11805 (2023). <https://doi.org/10.48550/ARXIV.2312.11805> arXiv:2312.11805
- [2] Rumeysa Bodur, Erhan Gundogdu, Binod Bhattarai, Tae-Kyun Kim, Michael Donoser, and Loris Bazzani. 2023. iEdit: Localised Text-guided Image Editing with Weak Supervision. *CoRR* abs/2305.05947 (2023). <https://doi.org/10.48550/ARXIV.2305.05947> arXiv:2305.05947
- [3] Jingqiang Chen. 2024. Transform, contrast and tell: Coherent entity-aware multi-image captioning. *Comput. Vis. Image Underst.* 238 (2024), 103878. <https://doi.org/10.1016/J.CVIU.2023.103878>
- [4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *CoRR* abs/2210.11416 (2022). <https://doi.org/10.48550/ARXIV.2210.11416> arXiv:2210.11416
- [5] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. 2022. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 21094–21104. <https://doi.org/10.1109/CVPR52688.2022.02045>
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Hua Tong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *CoRR* abs/2305.06500 (2023). <https://doi.org/10.48550/ARXIV.2305.06500> arXiv:2305.06500
- [7] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 19358–19369. <https://doi.org/10.1109/CVPR52729.2023.01855>
- [8] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge J. Belongie. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 708–717. <https://doi.org/10.18653/V1/D19-1065>
- [9] Yue He, Zheyang Shen, and Peng Cui. 2021. Towards Non-I.I.D. image classification: A dataset and baselines. *Pattern Recognit.* 110 (2021), 107383. <https://doi.org/10.1016/J.PATCOG.2020.107383>
- [10] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual Storytelling. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). The Association for Computational Linguistics, 1233–1239. <https://doi.org/10.18653/V1/N16-1147>
- [11] Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4024–4034. <https://doi.org/10.18653/V1/D18-1436>
- [12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment Anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 3992–4003. <https://doi.org/10.1109/ICCV51070.2023.00371>
- [13] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *CoRR* abs/2305.03726 (2023). <https://doi.org/10.48550/ARXIV.2305.03726> arXiv:2305.03726
- [14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742. <https://proceedings.mlr.press/v202/li23q.html>
- [15] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023. Fine-tuning Multimodal LLMs to Follow Zero-shot Demonstrative Instructions. *arXiv:2308.04152* [cs.CV]
- [16] Zhuowan Li, Quan Tran, Long Mai, Zhe Lin, and Alan L. Yuille. 2020. Context-Aware Group Captioning via Self-Attention and Contrastive Features. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 3437–3447. <https://doi.org/10.1109/CVPR42600.2020.00350>
- [17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=w0H2xGHlkW>
- [19] Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d3d9446802a4259755d38e6d163e820-Abstract-round2.html>
- [20] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual Question Answering by Reading Text in Images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*. IEEE, 947–952. <https://doi.org/10.1109/ICDAR.2019.00156>
- [21] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). <https://doi.org/10.48550/ARXIV.2303.08774> arXiv:2303.08774
- [22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. <https://doi.org/10.3115/1073083.1073135>
- [23] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust Change Captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 4623–4632. <https://doi.org/10.1109/ICCV.2019.00472>
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [25] Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. 2021. AESOP: Abstract Encoding of Stories, Objects, and Pictures. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2032–2043. <https://doi.org/10.1109/ICCV48922.2021.00206>
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. <https://doi.org/10.1109/CVPR52688.2022.01042>
- [27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). <https://doi.org/10.48550/ARXIV.2302.13971> arXiv:2302.13971
- [28] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDER: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 4566–4575. <https://doi.org/10.1109/CVPR.2015.7299087>
- [29] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. *CoRR* abs/2309.07915 (2023). <https://doi.org/10.48550/ARXIV.2309.07915> arXiv:2309.07915

- [30] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *CoRR* abs/2304.10592 (2023). <https://doi.org/10.48550/ARXIV.2304.10592> arXiv:2304.10592

1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450

1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508