Semantic Alignment for Multimodal Large Language Models

Tao Wu*
Zhejiang University
Hangzhou, China
twu22@zju.edu.cn

Mengze Li*
Zhejiang University
Hangzhou, China
mengzeli@zju.edu.cn

Jingyuan Chen†
Zhejiang University
Hangzhou, China
jingyuanchen@zju.edu.cn

Wei Ji
National University of Singapore
Singapore, Singapore
jiwei@nus.edu.sg

Wang Lin
Zhejiang University
Hangzhou, China
22151241@zju.edu.cn

Jinyang Gao
Alibaba Group
Hangzhou, China
jinyang.gjy@alibaba-inc.com

Kun Kuang
Zhejiang University
Hangzhou, China
kunkuang@zju.edu.cn

Zhou Zhao
Zhejiang University
Hangzhou, China
zhaozhou@zju.edu.cn

Fei Wu†
Zhejiang University
Hangzhou, China
wufei@cs.zju.edu.cn

Abstract

Research on **M**ulti-modal **L**arge **L**anguage **M**odel**s** (**MLLMs**) towards the multi-image cross-modal instruction has received increasing attention and made significant progress, particularly in scenarios involving closely resembling images (*e.g.*, change captioning). Existing MLLMs typically follow a two-step process in their pipelines: first, extracting visual tokens independently for each input image, and then aligning these visual tokens from different images with the Large Language Model (LLM) in its textual feature space. However, the independent extraction of visual tokens for each image may result in different semantics being prioritized for different images in the first step, leading to a lack of preservation of linking information among images for subsequent LLM analysis. This issue becomes more serious in scenarios where significant variations exist among the images (*e.g.*, visual storytelling). To address this challenge, we introduce **S**emantic **A**lignment for **M**ultimodal large language models (**SAM**). By involving the bidirectional semantic guidance between different images in the visual-token extraction process, SAM aims to enhance the preservation of linking information for coherent analysis and align the semantics of different images before feeding them into LLM. As the test bed, we propose a large-scale dataset named **MmLINK** consisting of 69K samples. Different from most existing datasets for MLLMs finetuning, our MmLINK dataset comprises multi-modal instructions with significantly diverse images. Extensive experiments on the group captioning task and the storytelling task prove the effectiveness of our SAM model, surpassing the state-of-the-art methods by

a large margin (+37% for group captioning and +22% for storytelling on CIDEr score). Project page: https://mccartney01.github.io/SAM.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Multi-modal Large Language Models; Multi-image Reasoning; Semantic Alignment; Bidirectional Semantic Guidance Mechanism

*Equal Contribution
†Corresponding Author

1 Introduction

Multi-modal Large Language Models (MLLMs) [9, 23, 24, 29, 51, 52] have shown great potential in processing multi-image cross-modal instructions, with GPT-4V(ision) [35] standing out as a leading example. Upon reviewing the preliminary exploration report [49] of GPT-4V, we observe that its successful applications in handling multi-image tasks primarily focus on closely resembling images, as shown in an example from the report (see Figure 1(a)). In such cases, where images share almost identical background contexts, MLLMs can effectively address the reasoning tasks by aligning the similarities and pinpointing the differences between the visual contents. However, the effectiveness of MLLMs, including GPT-4V, diminishes when faced with significantly different images in terms of content, context, or style. This issue is particularly evident with educational images, as illustrated in Figure 1(b). In such cases, MLLMs may struggle to establish meaningful semantic connections between the images and generate accurate responses.

We attribute the deficiency to the absence of semantic alignment in existing MLLMs when processing multiple images. To see this, we revisit the two stages in the processing pipeline of MLLMs for multi-image instructions:
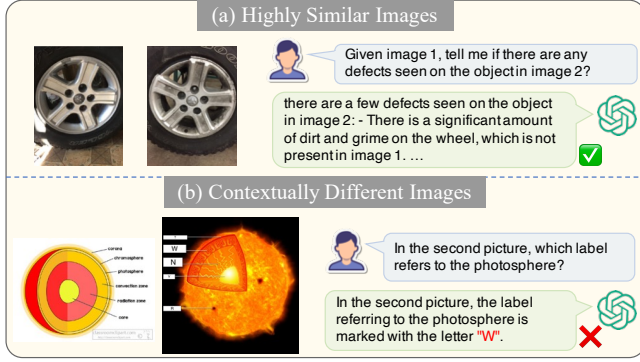
Figure 1: GPT-4V shows great performance in pinpointing differences between highly similar images (a), but struggles to align knowledge concepts across images featuring notably varied contexts (b). Example in (a) is sourced from the preliminary exploration report of GPT-4V [49].

- **Independent Perception.** This step employs a visual tokenizer (*e.g.*, Q-Former [23], linear projection [29] or the Resampler [1]) to map the visual features encoded by a pre-trained vision backbone (*e.g.* EVA-CLIP [10]) to visual tokens within the feature space of the foundation LLM. In this way, the projected visual tokens of an image carry its visual features.

- **Integrated Analysis.** After obtaining the projected visual tokens of each image, the model fuses them with the textual tokens of the instruction using cross-modal attention mechanisms [45]. This step allows the model to combine information from different modalities and images.

In this fashion, MLLMs mainly tackle multi-image reasoning in the integrated analysis step, which leads to semantic misalignment problems that hinder the discovery of inter-image correlations. This is because 1) the visual tokens serving as input to the integrated analysis step may lack crucial 'linking' information (*e.g.*, knowledge concepts in Figure 1(b)) necessary for identifying correlations, especially when the input images exhibit diverse contexts. Specifically, the visual tokens are influenced by the inductive bias from training data like image-caption pairs, which lead them to selectively express salient visual contents [12, 24]. Therefore, affected by complex context, the focus of visual tokens of different images may be different, hindering semantic alignment. 2) In the integrated analysis stage, LLMs perform attentions on visual tokens and textual tokens jointly, which leads to the semantic alignment between visual tokens being overwhelming in a large number of token interactions [7, 39].

Our solution to address the aforementioned semantic misalignment problem is to introduce contextual semantics from contextual images (*i.e.*, images other than the currently perceived image) as guidance in the perception stage. By utilizing contextual images as references, we may accurately align the extracted visual tokens of the currently perceived image with contextual semantics before the integrated analysis step. However, it is crucial to note that not all contextual images in the input multi-modal instruction may exhibit direct correlations with the currently perceived image due

to irrelevant or even noisy information. In such cases, directly incorporating the whole information from all contextual images may pose challenges to the perception process. Therefore, in addition to the semantic misalignment problem mentioned above, how to accurately extract the contextual semantics highly relevant to the currently perceived image is another crucial problem deserving careful consideration.

To tackle the above issues, we propose the **S**emantic **A**lignment method for **M**ultimodal large language models (**SAM**) towards the multi-image cross-modal instruction. The SAM model tackles the misalignment problem between multiple images in the input multi-modal instruction by incorporating the bidirectional semantic guidance mechanism. Specifically, this mechanism involves two interacting steps: (1) SAM extracts visual tokens from the currently perceived image based on the natural language prompt, with the guidance of the contextual semantics from the contextual images (*i.e.*, the other images in the input multi-modal instruction). (2) To obtain representative contextual semantics from the contextual images, SAM incorporates a novel visual tokenizer called W-former, which is specifically designed to extract synchronous semantic information from multiple images. With the visual tokens from the currently perceived image, the W-former first employs an adaptive adjustment module at the patch level of each contextual image in the multi-modal instruction, to enhance the prominence of critical information. Subsequently, the W-former incorporates the aligned patch features and the language prompt to extract the contextual semantics for the visual-token perception from the currently perceived image.

Most existing multi-image datasets applied for fine-tuning the multi-modal large language models may demonstrate substantial similarities between images of the multi-modal instruction [24], lacking associated images with significant difference (*e.g.,* different contexts, variant styles). To this end, we introduce a novel large-scale multi-modal dataset named **MmLINK** for the MLLMs research. It contains 69K vision-language samples, specifically designed to enhance the model's abilities of cross-modal multi-image semantic alignment and correlation mining. Different images within each multi-modal sample of our dataset exhibit both semantic-level correlations and significant visual differences (*e.g.*, the same person in different images with different poses and contexts). Through extensive experiments, we showcase that training on MmLINK leads to significant performance improvements in our SAM model, surpassing the current state-of-the-art models.

Overall, our contributions are summarized as follows:

- To the best of our knowledge, we embark on the early exploration of the semantic correlation between substantially diverse images in the multi-modal instruction, within the MLLMs field. As the research foundation, we construct a large-scale comprehensive dataset called MmLINK comprising 69K samples.
- We propose the multi-modal model named SAM, which employs bidirectional semantic guidance between images of the input multi-modal instruction to align the semantics of different images in the perception stage.
- Extensive experiments prove the effectiveness of our SAM model by significantly outperforming the state-of-the-arts.
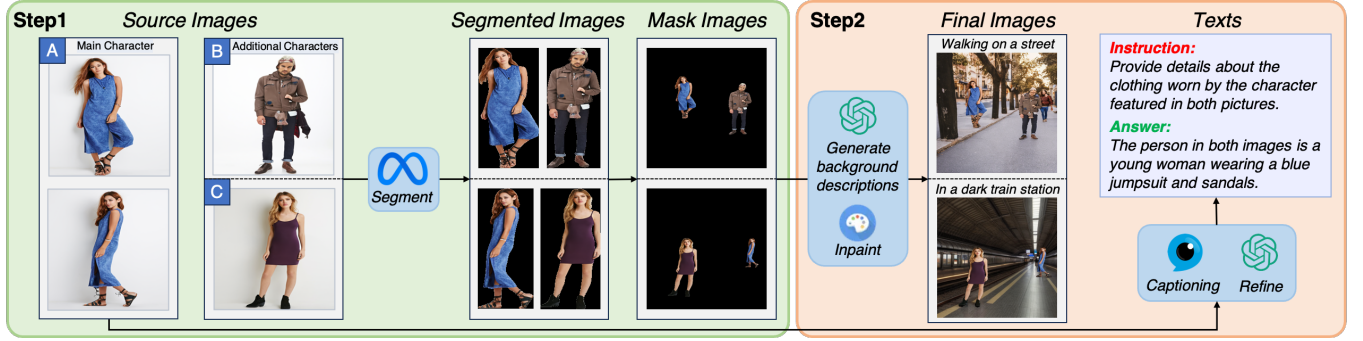
Figure 2: Demonstration of our proposed 2-step sample synthesis pipeline. We begin by selecting images featuring characters in different poses (A), along with 2 another distinct characters (B, C). The selected images are segmented to isolate each character, after which they are merged into mask images. 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.

2 Related Work

2.1 Multi-modal Large Language Models

Multi-modal pretraining [8, 17, 32, 33, 42, 43, 50] aims to train models on varied datasets across multiple modalities to capture cross-modal correlations. Building upon the concept of multi-modal pretraining, Multi-modal Large Language Models represent a fusion of multi-modal pretraining techniques with the advanced capabilities of Large Language Models. Flamingo [1] proposes architectural innovations that enable the integration of powerful pre-trained vision-only and language-only models, managing sequences of interleaved visual and textual data, and seamlessly ingesting images or videos as inputs. BLIP2 [23] employs a lightweight Querying Transformer (Q-Former) that bridges the modality gap through a two-stage pre-training process. Following closely are LLaVA [29], MiniGPT-4 [52] and InstructBLIP [9], which conduct instruction tuning [36] on MLLMs. They utilize different modules to bridge the gap between image modality and text modality, like Q-Former used in InstructBLIP and an linear layer employed in LLaVA. Their training data includes a wide range of single images and textual question-answer pairs, enhancing model's instruction-following ability. Cheetah [24] and MMICL [51] extend their work into the field of multi-image reasoning, featuring a visual tokenizer for visual token extraction and leverage LLMs for integrated analysis. However, almost all these MLLMs tackle multi-image understanding in the integrated analysis step, leading to ignorance of contextual semantics in the perception step and absence of crucial 'linking' information necessary for identifying correlations. To tackle this, we propose a bidirectional semantic guidance mechanism that incorporates contextual semantics during visual token extraction, enhancing multi-image correlation alignment.

2.2 Multi-image Captioning

Multi-image captioning [6, 14, 34, 44] is an advanced domain within the field of multi-modal research that focuses on generating descriptive text for a set of images, rather than just a single image. This area extends the capabilities of traditional image captioning by considering multiple images simultaneously, aiming to produce coherent and contextually relevant captions that encapsulate the

collective content and narrative of the images. Significant advancements have been made in multi-image captioning tasks, such as group captioning [5, 25, 27] and visual storytelling [4, 15, 28, 47, 48]. The challenge of multi-image captioning lies in its requirement for a deeper understanding of the relationships between images, the story they tell together, and how elements from each image relate to one another. Existing models have introduced methods to explore inter-image correlations. For instance, Li et al. [25] fuse visual features from multiple images using self-attention [45] mechanism. Liu and Keller [28] focus on the importance of characters in visual storytelling and propose two task: important character detection and character grounding in visual stories, and develop unsupervised models for these tasks using distributional similarity and pre-trained vision-language models. However, these proposed fusion methods usually require training the entire model, which will lead to significant training costs for large language models. Therefore, we propose a more precise structural design that achieves semantic alignment by introducing fewer trainable parameters.

3 Training Data Curation

Most training datasets for Multi-modal Large Language Models (MLLMs) primarily focus on single-image multi-modal instructions. However, the development of multi-modal datasets that incorporate correlations across multiple images for training is still in its early stages of exploration. The current methods for constructing multi-image multi-modal datasets mainly focus on creating highly similar images for correlation establishment. For instance, some researchers [18, 24] modify minor details of the original image while leaving the majority unchanged, to generate other images in the multi-image multi-modal instruction. However, these construction methods overlook the crucial 'linking' information across diverse contexts (*e.g., the same person in different images with different actions and poses*), which are often more significant for training the semantic alignment ability of MLLMs.

Toward this issue, we introduce a novel 2-step sample synthesis pipeline as shown in Figure 2 to construct our multi-modal dataset, **MmLINK**. In the first step, we curate different images featuring the same object but with varying poses, lighting conditions, or view angles (*e.g.,* the same person with different actions) to create

the image groups, where each group corresponds to one object. We then segment the shared object from the images within each image group to obtain our segmented objects and combine these segmented objects to generate mask images. In the second step, we inpaint the mask images based on descriptions generated by LLM to create a large number of semantically correlated images with different contexts. This approach ensures that the dataset includes diverse correlations across multiple images, enhancing the training process for MLLMs. In order to create the textual components of a training sample, we automatically generate a language query and its corresponding answer for two images depicting the same object but in different contexts.

2-step Sample Synthesis Pipeline. Each sample in our Mm-LINK 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 the images with variance objects (*e.g.,* characters, furniture items, icons, and book covers) from different image sources. Notably, for different types of objects, slightly various pipelines are employed to create the corresponding samples. We take the 'character' as an example to illustrate the sample construction pipeline, as depicted in Figure 2. In addition, we provide information on the construction pipeline for samples involving other objects in supplementary Section 1.1.

In the first step, we construct image groups for different characters from the DeepFashion dataset [30], in which the same character exhibit identical clothing but different poses and view angles. We select two images of the same character with varied poses as the main character (*e.g.,* A in Figure 2), along with two another distinct characters as additional characters (*e.g.,* B and C in Figure 2). Next, we segment each character from the source image using Segment Anything [19]. We combine the main character with each of the two additional characters and place them onto a 512×512 mask image with diverse positions and sizes. In this way, we obtain two mask images featuring the same character, *i.e.,* the main character.

In the second step, we inpaint the two mask images with diverse background descriptions generated by ChatGPT[1]. This is achieved by Stable-Diffusion-Inpainting [41], which can generate background pixels smoothly integrated with the characters 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 character across images despite the varied contexts and presence of additional characters. Finally, we employ InstructBLIP [9] to generate descriptions of the main character and use ChatGPT to obtain refined descriptions, which serve as the textual components in the final multi-modal samples.

Quality Control. To guarantee the quality of our dataset, we filter out the noise samples generated in the image-segmentation and image-inpainting steps in our pipeline. Please refer to supplementary Section 1.3 for details.

We compare MmLINK with previous multi-image datasets in Table 1. MmLINK stands out by including highly relevant yet contextually distinct images. It is created from single-image datasets

---

Table 1: Comparisons with previous multi-image datasets.

Name	Sample Number	Automated Generation	Images in a Group with Distinct Contexts
Spot-the-Diff [16]	13K	✗	✗
CLEVR-Change [38]	79K	✔	✗
ImageCode[20]	21K	✗	✗
VPG-C[24]	64K	✔	✗
MmLINK	69K	✔	✔

using a highly automated pipeline, enabling the low-cost generation of a larger number of samples.

4 Method

In this section, we will introduce our proposed **S**emantic **A**lignment method for **M**ultimodal large language models (**SAM**) in detail.

4.1 Overview

As illustrated in Figure 3, the pipeline of SAM can be formulated in three steps. Firstly, given N input images, the vision encoder transforms them into patch-level features. For the i-th image, we represent P corresponding patch-level features with $\mathcal{I}_i = \{\mathcal{I}_{i,j}\}_{j=1}^P$, where P is the patch number and j denotes the patch index. Then, we utilize the *Bidirectional Semantic Guidance* mechanism in the image perception stage to generate the visual tokens from the patch-level features with enhanced semantic alignment. Finally, in the integrated analysis stage, the large language model processes the visual tokens together with the input textual query and generates final prediction. In the following sections, we use subscript i to indicate symbols relevant to the currently perceived image.

Bidirectional Semantic Guidance. Towards the semantic misalignment problem in existing MLLMs when processing contextually different images, we introduce the bidirectional semantic guidance mechanism in the image perception stage. It comprises two interactive processes, including *Part A: Assisted Visual Token Extraction* and *Part B: Contextual Semantic Generation*. Initially, in Part A, the Q-former layers are applied to process the currently perceived image based on the natural language query and extract the initial visual tokens $\mathbf{h}_i$. Next, in Part B, we propose a novel visual tokenizer called W-former. Guided by initial visual tokens $\mathbf{h}_i$, W-former extracts synchronous contextual semantics $\mathbf{c}_i$ from contextual images (*i.e.,* images other than the currently perceived image). Finally, the contextual semantics $\mathbf{c}_i$ are passed back to the Q-former layers in Part A to guide the updating of visual tokens by aligning them with the contextual information based on the natural language query.

4.2 Part A: Assisted Visual Token Extraction

Most existing Multi-modal Large Language Models (MLLMs) often generate visual tokens for each input image in the multi-image cross-modal instruction independently. This may lead to the absence of semantic alignment between images with diverse contexts, thereby hindering the ability to identify inter-image correlations. To address this issue, we introduce contextual semantics from contextual images of the multi-modal instruction and the natural language query to guide the extraction of visual tokens for the currently perceived image. These contextual semantics are provided from Part B, which also receives initial visual tokens of currently perceived image generated in Part A as guidance, thereby forming an interactive mechanism between the two processes.

---

[1]https://openai.com/blog/chatgpt. The prompts of ChatGPT used for dataset construction are provided in supplementary Section 1.2.
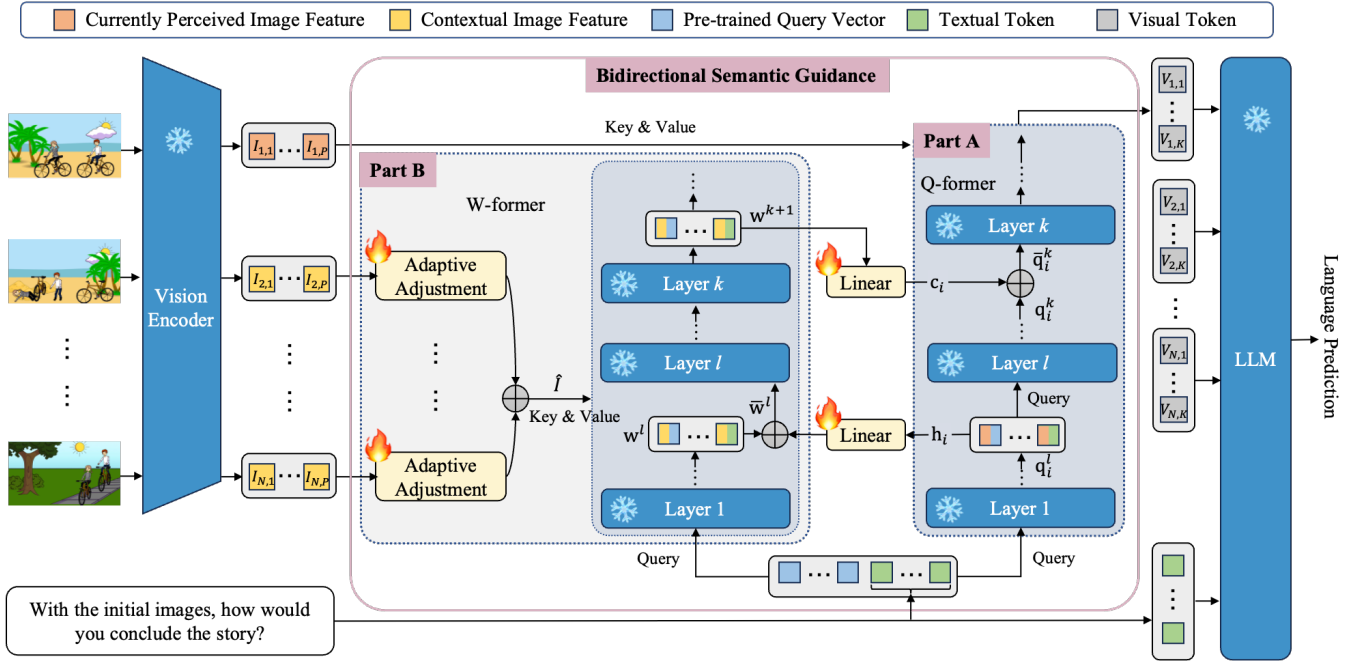
Figure 3: Overview of SAM. The core mechanism of our SAM model is the Bidirectional Semantic Guidance mechanism with two interactive processes: Assisted Visual Token Extraction (Part A) and Contextual Semantic Generation (Part B). In Part A, the Q-former module leverages the contextual semantics c_i, which are generated from contextual images (*i.e.*, images other than the currently perceived image) in the multi-modal instruction in Part B, to guide the extraction of visual tokens from the currently perceived image features. In Part B, the W-former module is utilized to select the contextual semantics from the visual context of contextual images. This selection process is facilitated by the attention mechanism in the adaptive adjustment, along with assistance from the initial visual tokens h_i, which are extracted from the currently perceived image in Part A.

Specifically, Part A is implemented with the transformer-based visual tokenizer, Q-former [9]. Q-former consists of several cross-attention layers, which concatenates pre-trained query vectors and textual tokens as the Query and the current image features $\mathcal{I}_i$ as the Key and Value, allowing them to interact with each other to obtain final visual tokens. We define the initial visual tokens $\mathbf{h}_i$ as the inputs at the l-th intermediate Q-former layer $\mathbf{q}_i^l$: $\mathbf{h}_i := \mathbf{q}_i^l$. The initial visual tokens will serve as guidance to help Part B generate contextual semantics strongly related to the currently perceived image, by effectively reducing redundant details of contextual images in the multi-modal instruction.

After Part B generating contextual semantics $\mathbf{c}_i$, we select a subsequent layer k ($k \geq l$) and add $\mathbf{c}_i$ to the input queries of the k-th Q-former layer $\mathbf{q}_i^k$ as a guidance: $\bar{\mathbf{q}}_i^k = \mathbf{q}_i^k + \mathbf{c}_i$, where $\bar{\mathbf{q}}_i^k$ denotes the updated input queries. The subsequent Q-former layers will refer to the contextual semantics to generate the final visual tokens for the currently perceived image with enhanced alignment.

4.3 Part B: Contextual Semantic Generation

The contextual semantics $\mathbf{c}_i$ from contextual images of the multi-modal instruction are needed to guide the extraction of visual tokens for the currently perceived image in Part A. However, it is important to note that not all contextual images in the input may have a direct correlation with the currently perceived image and

the natural language query. Furthermore, even highly related images may still contain irrelevant or noisy information. To tackle these issues, we involve the W-former module to Part B. W-former contains Q-former layers which share the same parameters with the one in Part A, along with an novel adaptive adjustment design.

Adaptive Weights. Specifically, we denote the contextual image set (*i.e.*, images other than the currently perceived image) as $\{\mathcal{I}_m\}_{m \neq i}$. The patch-level features of the m-th ($m \neq i$) image are denoted as $\{\mathcal{I}_{m,j}\}_{j=1}^P$. Then, we assign weights to each image patch with our proposed *Adaptive Adjustment* module, eliminating the prominence of irrelevant patches. Firstly, the adaptive adjustment module uses the linear layer $\mathbf{Linear}(\cdot)$ and the softmax function $\mathbf{softmax}(\cdot)$ to generate normalized patch-level weights: $\bar{\alpha}_{m,j} = \mathbf{softmax}(\mathbf{Linear}(\mathcal{I}_{m,j}))$. 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 $\hat{\mathcal{I}}$:

$$\hat{\mathcal{I}} = \{\hat{\mathcal{I}}_j\}_{j=1}^P, \quad \hat{\mathcal{I}}_j = \sum_{m \neq i} (\bar{\alpha}_{m,j} * \mathcal{I}_{m,j}) \tag{1}$$

This process reduces the influence of irrelevant patches and amplifies important ones, improving patch-level alignment and further alignment during the bidirectional semantic guidance. The detailed framework of the adaptive adjustment module is illustrated in supplementary Section 2.

Table 2: Performance on group captioning and storytelling tasks. The best is bolded and the second best is <u>underlined</u>.

(a) Group Captioning

Model	Conceptual			Animal			Vehicle			Average		
	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4
MiniGPT-4	9.65	4.66	0.25	9.6	2.11	0.36	12.43	2.13	0.41	10.56	2.97	0.34
LLaVA	13.87	13.76	1.21	17.45	8.4	<u>2.26</u>	16.82	9.55	2.11	16.05	10.57	1.86
BLIP2	15.03	19.31	**2.78**	14.88	5.66	0.84	14.33	5.77	0.97	14.75	10.25	1.53
InstructBLIP	12.61	13.69	1.11	14.46	2.81	2.15	15.55	2.45	1.79	14.21	6.32	1.68
Otter	14.08	<u>20.15</u>	1.31	14.93	5.41	0.68	15.21	4.72	0.72	14.74	10.09	0.9
Cheetah	12.13	13.35	0.67	13.22	2.73	0.24	12.45	2.21	0.24	12.6	6.1	0.38
MMICL	<u>18.6</u>	4.06	1.44	17.99	<u>12.34</u>	1.97	<u>19.57</u>	11.84	<u>2.16</u>	<u>18.72</u>	9.41	<u>1.86</u>
GPT-4V	12.47	19.94	0.92	<u>19.11</u>	10.64	1.87	18.96	<u>11.99</u>	2.05	16.85	<u>14.19</u>	1.61
Gemini Pro	12.97	20	<u>1.6</u>	18.84	6.53	0.83	17.73	4.36	0.91	16.51	10.3	1.11
SAM	**20.93**	**20.19**	**2.78**	**19.4**	**19.92**	**3.27**	**20.36**	**18.35**	**3.62**	**20.23**	**19.49**	**3.22**

(b) Storytelling

Model	AESOP			VIST			DM800K			Average		
	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4
MiniGPT-4	15.42	2.29	0.53	10.36	1.4	0.17	12.01	0.89	0.49	12.6	1.53	0.4
LLaVA	14.52	2.92	0.98	8.66	0.22	0.3	11.35	0.35	0.97	11.51	1.16	0.75
BLIP2	<u>21.49</u>	19.88	2.57	<u>18.31</u>	31.7	2.38	12.35	8.47	1.43	<u>17.38</u>	20.02	2.13
InstructBLIP	19.12	18.28	2.72	16.96	28.79	2.38	8.21	7.75	1.22	14.76	18.27	2.11
Otter	11.4	5.41	0.74	7.32	2.35	0	9.5	3.65	0.69	9.41	3.8	0.48
Cheetah	20.03	<u>21.5</u>	2.87	17.89	<u>31.76</u>	<u>2.63</u>	11.67	9.26	1.01	16.53	<u>20.84</u>	2.17
MMICL	15.2	13.04	1.39	12.13	2.87	0.68	11.79	7.4	0.82	13.04	7.77	0.96
GPT-4V	17.45	18.48	1.41	11.76	23.75	0.86	10.37	<u>11.03</u>	0.59	13.19	17.75	0.95
Gemini Pro	20.52	<u>21.5</u>	<u>3.37</u>	13.84	22.1	1.74	<u>12.96</u>	9.9	<u>1.56</u>	15.77	17.83	<u>2.22</u>
SAM	**23.45**	**25.92**	**4.13**	**20.85**	**39.32**	**3.65**	**14.33**	**11.16**	**1.97**	**19.54**	**25.47**	**3.25**

Contextual Semantics. After obtaining merged contextual features $\hat{\mathcal{I}}$, the W-former module generates the contextual semantics c_i. To enhance the specificity of c_i with respect to the currently perceived image, we introduce its initial visual tokens h_i, defined as the input queries of the l-th Q-former layer in Part A, as a guiding factor. This is achieved by incorporating the initial visual tokens with the input queries of the corresponding l-th W-former layer via a linear layer $\text{Linear}(\cdot)$: $\bar{w}^l = w^l + \text{Linear}(h_i)$, where w^l denotes the original input queries of the l-th W-former layer and $\bar{w}^l$ denotes the updated ones. Guided by the initial visual tokens of currently perceived image, the remaining W-former layers could efficiently extract relevant details from merged contextual features $\hat{\mathcal{I}}$. At the corresponding k-th ($k \geq l$) layer, we derive contextual semantics from the output tokens w^{k+1} using a linear layer $\text{Linear}(\cdot)$: $c_i = \text{Linear}(w^{k+1})$. The resulting c_i aggregates pertinent information to the currently perceived image from the entire context, facilitating the generation of visual tokens aligned with contextual semantics.

4.4 Efficient Training

We perform both interactions at the final layer of the Q-former and W-former, which means $l = k = 12$ (defined in Section 4.2). We compare the different choices of l and k in Section 5.3. For efficient training, we freeze the parameters of the vision encoder and the LLM. In addition, the Q-former layers are also frozen. This means that only three linear layers and the LLM projection layer require fine-tuning, constituting just **4.3M (0.05%)** trainable parameters.

The training objective is $\mathcal{L} = -\sum_{i=1}^{t} \log P(T_i|V, T_{1:i-1}; \theta)$, where V stands for visual tokens, $T_{1:i-1}$ stands for the previous textual tokens, T_i stands for the current ones, t stands for textual token number and θ stands for trainable parameters.

5 Experiments

5.1 Experimental Settings

We implement SAM model in LAVIS library [22], building upon InstructBLIP-vicuna7b [9] architecture. We use the AdamW [31] optimizer with $\beta = (0.9, 0.999)$, and a learning rate of 2e-5 along with a weight decay of 0.05. We employ a cosine learning rate decay mechanism, along with a warm-up phase of 240 steps. We conduct training SAM using a batch size of 20 with 4 A100 GPUs.

Datasets. We conduct zero-shot evaluations on Group Captioning and Storytelling tasks to evaluate SAM's generalization ability. Both tasks demand semantic alignment across contextually different images. For the group captioning task, we select three datasets: Conceptual [25], Animal and Vehicle [13]. For the storytelling task, we select AESOP [40], VIST [15] and DM800K [6] as test sets.

Metrics. We measure the performance using a variety of captioning metrics. Following Forbes et al. [11], we report the scores on ROUGE-L [26], CIDEr [46] and BLEU-4 [37].

Baselines. For a comprehensive comparison, we consider open-source state-of-the-art MLLMs including MiniGPT-4 [52], LLaVA [29], BLIP2 [23], InstructBLIP [9], Otter [21], Cheetah [24] and MMICL [51], as well as industrial multi-modal chatbots including GPT-4V [35] and Gemini Pro [2].

Table 3: Ablation study of different modules in SAM over group captioning and storytelling. $+\Delta_{DATA}$ means using our proposed data for training; $+\Delta_{BSG}$ means adding bidirectional semantic guidance without the adaptive adjustment module; $+\Delta_{AA}$ means adding adaptive adjustment module; $+\Delta_{LIN}$ means replacing the entire W-former with a simple linear layer; $+\Delta_{VPG-C}$ means training with the VPG-C architecture [24]; $+\Delta_{MMICL}$ means utilizing the training pipeline proposed in MMICL [51].

	$+\Delta_{DATA}$	$+\Delta_{BSG}$	$+\Delta_{AA}$	$+\Delta_{LIN}$	$+\Delta_{VPG-C}$	$+\Delta_{MMICL}$	Group Captioning			Storytelling		
							ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4
1							14.21	6.32	1.68	14.76	18.27	2.11
2	✓						18.79	11.86	2.45	18.64	22.5	2.78
3	✓	✓					19.49	16.73	3.04	19.1	23.97	3
4	✓			✓			18.24	13.01	2.31	15.66	13.67	1.75
5	✓				✓		19.07	17.03	2.57	18.44	22.94	2.49
6	✓					✓	18.79	14.72	2.67	17.89	21.07	2.41
7 (ours)	✓	✓	✓				20.23	19.49	3.22	19.54	25.47	3.25

Please refer to supplementary Section 3 for detailed experimental settings and Section 4 for more experimental results.

5.2 Performance Comparison

The overall results of SAM and baselines are listed in Table 2. From it we can observe that SAM consistently achieves superiority across all datasets on all evaluation metrics. In particular, SAM outperforms other baselines on CIDEr score by a large margin of **5.3 (37%)** on group captioning and **4.63 (22%)** on storytelling, which demonstrates that SAM generates more accurate and comprehensive answers. Beyond this, we find that the improvement of our method mainly manifests in two aspects:

1) Enhance the ability to follow instructions that require identifying multi-image correlations. We observe that many MLLMs struggle to follow instructions for identifying correlations, such as interpreting storytelling tasks through image captions. This leads to their answers containing only some keywords rather than coherent story descriptions. We speculate that this issue stems from their training data lacking semantic correlations between images. For instance, MMICL relies on in-context learning data that might not be inherently related. In contrast, our training data emphasizes inter-image connections across contextually different images, enabling our model to excel at discerning associations and responding effectively to tasks that require identifying correlations, which further indicates the effectiveness of our training dataset.

2) Enhance semantic alignment between contextually different images. We observe that the state-of-the-arts often misinterpret the relations between images with diverse contexts. For instance, they may mix up character identities across images or invent details not present in the images. We attribute these hallucinatory responses to the absence of semantic alignment in their visual tokens. The existing MLLMs isolate the currently perceived image from its semantic contexts when extract its initial visual tokens, leading to semantic misalignment. Interestingly, similar issues also arise in industrial multi-modal chatbots such as GPT-4V and Gemini Pro, resulting in suboptimal performance. This indicates that these chatbots may not pay sufficient attention to multi-image semantic alignment. However, our approach, with the bidirectional semantic guidance mechanism, can mitigate the issues by accurately aligning the visual tokens of currently perceived image with its contexts using contextual semantics.

5.3 In-Depth Analysis

We further validate 4 vital issues of SAM as follows.

1. Each component in SAM contributes to performance improvement. We conduct experiments to illustrate the effectiveness of each component in Table 3. We start with the baseline model in line 1 of the table that uses only the Q-former module to extract visual tokens for each image. **1)** We fine-tune the baseline model on MmLINK. The experiment results shown in line 2 demonstrate that *our synthetic training data enhances model's semantic alignment ability across contextual different images.* **2)** Next, we adopt the bidirectional semantic guidance mechanism based on averaged contextual image features, which significantly improves the performance as shown in line 3. It demonstrates that *bidirectional semantic guidance mechanism can accurately align visual tokens of currently perceived image with its contexts.* **3)** Finally, replacing average contextual image features with merged contextual features processed through our adaptive adjustment module leads to enhancement in all tasks, as indicated by the results in line 7. This demonstrates that *the adaptive adjustment module effectively filters out irrelevant information from contextual images, thus improving patch-level alignment and the quality of bidirectional semantic guidance.*

2. SAM demonstrates superiority in training strategies. We compare SAM with related finetuning approaches and present the results of different training strategies in Table 3. In line 4, we introduce the contextual semantics extracted by a simple linear layer to guide the visual token extraction process. The results show that this straightforward approach brings either no improvement or even a decline in performance. We speculate that this simple method introduces uncertain contextual semantics with heavy noise, which misguides the visual token extraction process. In line 5, we utilize the VPG-C architecture [24] to enhance semantic alignment by re-extracting the visual contents in the medium layer of the LLM. In line 6, we implement the training pipeline proposed in MMICL [51], which involves joint training of the projection layer and the query and value vectors. To maintain consistent training parameters, we limit the training to the query and value vectors in the first two attention layers of the LLM. Compared to these strategies, SAM (line 7) consistently demonstrates superior performance, highlighting its effectiveness in multi-image semantic alignment.

3. SAM still performs well when reasoning on highly similar images. We conduct experiments on change captioning task,
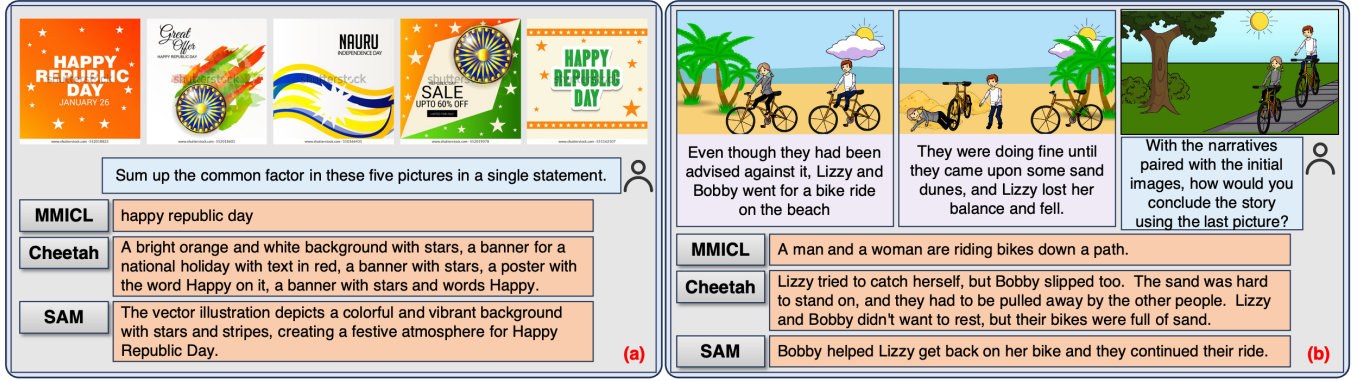
Figure 4: Case examples generated by SAM and other MLLMs. Other MLLMs' answers show either weak instruct-following ability or contain hallucinations, while SAM successfully performs semantic alignment and produces accurate responses.

Table 4: Performance comparison on change captioning tasks. The best is bolded and the second best is underlined.

Model	Average		
	ROUGE-L	CIDEr	BLEU-4
MiniGPT-4	13.27	0.92	0.44
LLaVA	12.03	0.45	0.67
BLIP2	13.44	2.77	0.34
InstructBLIP	10.9	1.87	0.5
Otter	12	2.36	0.35
MMICL	14.73	2.54	0.51
GPT-4V	<u>17.12</u>	<u>3.38</u>	**1.2**
Gemini Pro	**17.59**	2.87	1.05
SAM	16.08	**3.84**	<u>1.08</u>

Figure 5: Average scores on 6 datasets of 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). The original 10 points are marked, and the surface are interpolated from these 10 points.

which requires model to describe subtle differences between highly similar images. We test on 4 datasets: IEdit [3], Spot-the-Diff [16], Birds-to-Words [11] and CLEVR-Change [38]. Dataset details are provided in supplementary Section 3.1. The average scores on the 4 datasets are reported in Table 4. From it, we find that compared to the baseline InstructBLIP, SAM does not show a performance decline. In contrary, SAM demonstrates superior performance comparing with those open-source state-of-the-art MLLMs, and even surpasses industrial multi-modal chatbots on CIDEr score. This indicates that introducing contextual semantics for alignment will not degrade the model's reasoning performance on highly similar images; instead, it improves such capabilities.

4. The final layers of the Q-former and the W-former are the best positions to conduct bidirectional semantic guidance. We investigate the effect of different interaction layers within the Q-former and the W-former, which comprise 12 layers in total. We group every 3 layers into a step, ensuring that the interactive layer k, conveying contextual semantics, does not precede layer l, which conveys initial visual tokens (l and k are defined in Section 4.2). This approach yields 10 distinct configurations of interaction layers. Our findings, presented in Figure 5, highlight that early-stage interactions do not enhance performance, whereas interactions in later layers lead to improvements. We speculate that this is because the initial layers of Q-formers capture less meaningful semantics, lacking the refined semantics needed for effective interaction. In
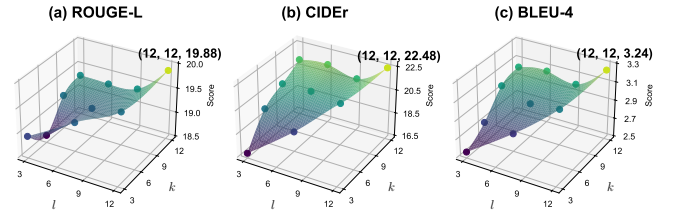
contrast, the later layers provide high-level semantics for both Q-formers, significantly boosting performance.

5.4 Case Study

As illustrated in Figure 4, SAM demonstrates strong abilities to perform group captioning and storytelling tasks. In **(a)**, SAM can identify commonalities between images accurately, while other MLLMs' answers either show weak instruction-following ability or contain redundancy and hallucinations. In **(b)**, while other MLLMs might treat the storytelling task as an image captioning task, SAM successfully discovers the correlation between the characters in the images and matches them with the names of the characters in the text, creating a coherent story. More cases are illustrated in supplementary Section 4.5.

6 Conclusion

In this work, we introduce SAM, an innovative MLLM that utilizes bidirectional semantic guidance to enhance semantic alignment between contextual different images in multi-modal instructions. To effectively train SAM, we propose a 2-step sample synthesis pipeline to construct the MmLINK dataset, ensuring diverse correlations across contextually diverse images. The SAM model significantly outperforms state-of-the-art methods, demonstrating superior semantic alignment and coherence in multi-modal instructions.

Acknowledgments

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html

[2] 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

[3] 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

[4] Tom Braude, Idan Schwartz, Alexander G. Schwing, and Ariel Shamir. 2022. Ordered Attention for Coherent Visual Storytelling. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 3310–3318. https://doi.org/10.1145/3503161.3548161

[5] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. 2018. GroupCap: Group-Based Image Captioning With Structured Relevance and Diversity Constraints. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 1345–1353. https://doi.org/10.1109/CVPR.2018.00146

[6] 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

[7] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. *CoRR* abs/2403.06764 (2024). https://doi.org/10.48550/ARXIV.2403.06764 arXiv:2403.06764

[8] Kanzhi Cheng, Wenpo Song, Zheng Ma, Wenhao Zhu, Zixuan Zhu, and Jianbing Zhang. 2023. Beyond Generic: Enhancing Image Captioning with Real-World Knowledge using Vision-Language Pre-Training Model. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 5038–5047. https://doi.org/10.1145/3581783.3611987

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, 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

[10] 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

[11] 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

[12] Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. 2024. Incorporating Visual Experts to Resolve the Information Loss in Multimodal Large Language Models. *CoRR* abs/2401.03105 (2024). https://doi.org/10.48550/ARXIV.2401.03105 arXiv:2401.03105

[13] Yue He, Zheyan 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

[14] Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised Discovery of Multimodal Links in Multi-image, Multi-sentence Documents. 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, 2034–2045. https://doi.org/10.18653/V1/D19-1210

[15] 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

[16] 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

[17] Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Fei Huang, and Ji Zhang. 2023. COPA : Efficient Vision-Language Pre-training through Collaborative Object- and Patch-Text Alignment. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 4480–4491. https://doi.org/10.1145/3581783.3611826

[18] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 6007–6017. https://doi.org/10.1109/CVPR52729.2023.00582

[19] 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

[20] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Maria Ponti, and Siva Reddy. 2022. Image Retrieval from Contextual Descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3426–3440. https://doi.org/10.18653/V1/2022.ACL-LONG.241

[21] 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

[22] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. 2023. LAVIS: A One-stop Library for Language-Vision Intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, Danushka Bollegala, Ruihong Huang, and Alan Ritter (Eds.). Association for Computational Linguistics, 31–41. https://doi.org/10.18653/V1/2023.ACL-DEMO.3

[23] 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

[24] 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]

[25] 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

[26] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[27] Wang Lin, Tao Jin, Ye Wang, Wenwen Pan, Linjun Li, Xize Cheng, and Zhou Zhao. 2023. Exploring Group Video Captioning with Efficient Relational Approximation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 15235–15244. https://doi.org/10.1109/ICCV51070.2023.01402

[28] Danyang Liu and Frank Keller. 2023. Detecting and Grounding Important Characters in Visual Stories. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 13210–13218. https://doi.org/10.1609/AAAI.V37I11.26551

[29] 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

[30] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 1096–1104. https://doi.org/10.1109/CVPR.2016.124

[31] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[32] Zheqi Lv, Shaoxuan He, Tianyu Zhan, Shengyu Zhang, Wenqiao Zhang, Jingyuan Chen, Zhou Zhao, and Fei Wu. 2024. Semantic Codebook Learning for Dynamic Recommendation Models. *CoRR* abs/2408.00123 (2024). arXiv:2408.00123 https://arxiv.org/abs/2408.00123

[33] Zheqi Lv, Wenqiao Zhang, Shengyu Zhang, Kun Kuang, Feng Wang, Yongwei Wang, Zhengyu Chen, Tao Shen, Hongxia Yang, Beng Chin Ooi, and Fei Wu. 2023. DUET: A Tuning-Free Device-Cloud Collaborative Parameters Generation Framework for Efficient Device Model Generalization. In *WWW*. ACM, 3077–3085.

[34] Yangjun Mao, Jun Xiao, Dong Zhang, Meng Cao, Jian Shao, Yueting Zhuang, and Long Chen. 2023. Improving Reference-based Distinctive Image Captioning with Contrastive Rewards. *CoRR* abs/2306.14259 (2023). https://doi.org/10.48550/ARXIV.2306.14259 arXiv:2306.14259

[35] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). https://doi.org/10.48550/ARXIV.2303.08774 arXiv:2303.08774

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[37] 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

[38] 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

[39] Yu Qi, Fan Yang, Yousong Zhu, Yufei Liu, Liwei Wu, Rui Zhao, and Wei Li. 2023. Exploring Stochastic Autoregressive Image Modeling for Visual Representation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 2074–2081. https://doi.org/10.1609/AAAI.V37I2.25300

[40] 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

[41] 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

[42] Zhongfan Sun, Yongli Hu, Qingqing Gao, Huajie Jiang, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2023. Breaking the Barrier Between Pre-training and Fine-tuning: A Hybrid Prompting Model for Knowledge-Based VQA. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 4065–4073. https://doi.org/10.1145/3581783.3612516

[43] Zihao Tang, Zheqi Lv, Shengyu Zhang, Fei Wu, and Kun Kuang. 2024. ModelGPT: Unleashing LLM's Capabilities for Tailored Model Generation. *CoRR* abs/2402.12408 (2024). https://doi.org/10.48550/ARXIV.2402.12408 arXiv:2402.12408

[44] Nicholas Trieu, Sebastian Goodman, Pradyumna Narayana, Kazoo Sone, and Radu Soricut. 2020. Multi-Image Summarization: Textual Summary from a Set of Cohesive Images. *CoRR* abs/2006.08686 (2020). arXiv:2006.08686 https://arxiv.org/abs/2006.08686

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[46] 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

[47] Yuechen Wang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. 2023. Text-Only Training for Visual Storytelling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 3686–3695. https://doi.org/10.1145/3581783.3612179

[48] Dizhan Xue, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2022. MMT: Image-guided Story Ending Generation with Multimodal Memory Transformer. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 750–758. https://doi.org/10.1145/3503161.3548022

[49] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *CoRR* abs/2309.17421 (2023). https://doi.org/10.48550/ARXIV.2309.17421 arXiv:2309.17421

[50] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. RAMM: Retrieval-augmented Biomedical Visual Question Answering with Multi-modal Pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 547–556. https://doi.org/10.1145/3581783.3611830

[51] 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

[52] 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