

✨Sparkles: Unlocking Chats Across Multiple Images for Multimodal Instruction-Following Models

Yupan Huang¹² Zaiqiao Meng²³, Fangyu Liu², Yixuan Su², Nigel Collier², Yutong Lu¹

¹Sun Yat-sen University ²University of Cambridge ³University of Glasgow

{huangyp28@mail2, luyutong@mail}.sysu.edu.cn {zm324, fl399, ys484, nhc30}@cam.ac.uk

Abstract

Large language models exhibit enhanced zero-shot performance on various tasks when fine-tuned with instruction-following data. Multimodal instruction-following models extend these capabilities by integrating both text and images. However, existing models such as MiniGPT-4 face challenges in maintaining dialogue coherence in scenarios involving multiple images. A primary reason is the lack of a specialized dataset for this critical application. To bridge these gaps, we present **SparklesChat**, a multimodal instruction-following model for open-ended dialogues across multiple images. To support the training, we introduce **SparklesDialogue**, the first machine-generated dialogue dataset tailored for word-level interleaved multi-image and text interactions. Furthermore, we construct **SparklesEval**, a GPT-assisted benchmark for quantitatively assessing a model's conversational competence across multiple images and dialogue turns. Our experiments validate the effectiveness of SparklesChat in understanding and reasoning across multiple images and dialogue turns. Specifically, SparklesChat outperformed MiniGPT-4 on established vision-language benchmarks, including the BISON binary image selection task and the NLVR2 visual reasoning task. Moreover, SparklesChat scored 8.56 out of 10 on SparklesEval, substantially exceeding MiniGPT-4's score of 3.91 and nearing GPT-4's score of 9.26. Qualitative evaluations further demonstrate SparklesChat's generality in handling real-world applications. All resources are available at <https://github.com/HYPJUDY/Sparkles>.

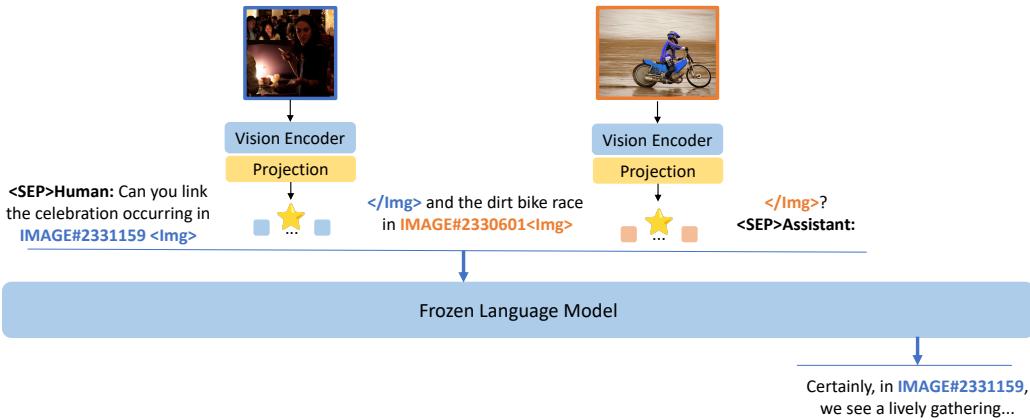


Figure 1: The architecture of SparklesChat. SparklesChat integrates multiple images at the word level within the dialogue, facilitating a fine-grained and human-like multimodal interaction.

When we view IMAGE#2352049 next to IMAGE#2362551, what differences can we identify regarding the nature and level of activity in both scenarios?

User

There are indeed some differences when comparing IMAGE#2352049 and IMAGE#2362551. IMAGE#2352049 showcases a well-organized professional kitchen setup, indicating the level of attention and detail. The image shows the kitchen counter filled with a large assortment of tools, kitchenware, and food items, suggesting that it is a high-activity commercial kitchen, such as a restaurant. We see the chef in this image attentively working with different dishes, indicating the high level of preparation and planning in the kitchen.

In contrast, IMAGE#2362551 depicts an office setup. The image shows a desk, a laptop, a mouse, a cell phone, and several whiteboards. We see the laptop and phone on the desk and whiteboards on the wall and windows, suggesting a working space, not a commercial kitchen. The laptop screen showcases a map and a document, which could be related to project planning, marketing, or research. The overall setup in IMAGE#2362551 is more focused on the digital and office activities.

So, the level of activity, as well as the type and scale of preparation are different in these two images.

SparklesChat

Introducing IMAGE#2386507 into our discussion, could you explain how the change in environment and activity influences the interpretation of this image when considered alongside the previous ones?

User

Certainly. In IMAGE#2386507, we see a man riding a wave. This is an outdoor, water-based scene which implies a different environment from the kitchens and office settings we discussed previously. It indicates a different type of activity, namely surfing, which involves a different level of physical and mental engagement.

The man in IMAGE#2386507 is in the middle of a wave, his body and face full of concentration, as he reaches out for the wave. This depicts a heightened level of awareness and focus, which is different from the tranquil, organized work environment in IMAGE#2352049 and the methodical digital setup in IMAGE#2362551. The image shows a more raw and dynamic setting, emphasizing the power and thrill of surfing, and suggesting a heightened sense of nature and adventure.

This comparison gives us a glimpse into how different settings and activities can affect our interpretation of images, making them unique and diverse. Each image conveys its own energy and style, offering an insight into the unique perspectives they represent.

SparklesChat

VS.

Both images show a **kitchen with a computer** on the counter. The first image shows a man standing at the counter with a **laptop** on it. The second image shows a **man sitting** at the counter with a laptop on it.

MiniGPT-4

The change in environment and activity in the second image when compared to the first image, in which the man is standing at the counter, can be seen by the fact that the man in the second image is **sitting at the counter** with the laptop on it. This change can influence the interpretation of the image because it provides a different perspective and focus on the activity being depicted, in this case, **the man working at the counter**.

MiniGPT-4

GPT-4 Judges SparklesChat

- Evaluating A1
 - (C1) Explanation: The assistant demonstrated excellent understanding of the images' content. It effectively compared the scenes in both the kitchen and office settings and accurately described the contexts of each environment. The assistant also understood the level of activity in both scenarios. There was also a clear logical reasoning behind the differences pointed out between the two images. Rating: [[10]]
 - (C2) Explanation: The assistant maintains a very coherent narrative from the user's question to the images, accurately reflecting the description in the first image, transitioning to the second one, and comparing the differences as requested. Rating: [[10]]
 - (C3) Explanation: The assistant addressed the user's question completely, providing a comprehensive insight into the differences between the nature and level of activity in the compared scenarios. The assistant accurately pulled details from the images to describe how each environment is visually presented and adds aspects of the workplaces indicated by the images. This covers both the relevancy and completeness of the assistant's response. Rating: [[10]]
- Evaluating A2
 - (C1) Explanation: The assistant again showed excellent ability to analyze and interpret the image, accurately identifying all major elements (the man, the wave, the surfers' concentration, etc). The assistant interpreted not only the objects present but also the activity, referring to the heightened awareness levels in such a setting as compared to the previous offices. The change in color and visual representation is also provided depth. Rating: [[10]]
 - (C2) Explanation: The assistant successfully related the new image to the previous ones in a coherent manner. It made a clear distinction between the previous indoor, work-focused images, and the outdoor, action-focused surfing image, providing a clear narrative across multiple turns. Rating: [[10]]
 - (C3) Explanation: The assistant comprehensively answered the user's question regarding how the change in environment and activity influences the interpretation of the new image in comparison to the previous images. The detailed reference to surfing and how different it is from the previous images shows complete understanding of the user's question and the relevance of the answer. Rating: [[10]]

Therefore, the overall rating of A1 is [[10]].

Therefore, the overall rating of A2 is [[10]].

GPT-4 Judges MiniGPT-4

- Evaluating A1
 - (C1) Explanation: The assistant fails to accurately identify and describe the images correctly. The first image is described as a commercial kitchen with chefs at work, while the second is a workspace with a laptop and other items. The assistant, however, sees both as kitchens with men and laptops. This lacks understanding and reasoning of the provided image descriptions. Rating: [[1]]
 - (C2) Explanation: The assistant does not establish any coherence between the two images. The user asked for differences in nature and level of activities in both scenarios. The assistant does not capture the busyness of the kitchen or the quietness of the workspace as described in the images. Rating: [[1]]
 - (C3) Explanation: The assistant's responses are unrelated to the user's inquiries. The user asked for the differences regarding activity in the two settings. The assistant's response does not provide a comparative or detailed answer. Rating: [[1]]
- Evaluating A2
 - (C1) Explanation: There is no reasoning or understanding of the third image from the assistant's response. The third image portrays the thrilling scene of surfing in the ocean but the assistant still talks about a man working on the counter which is not related to this image. Rating: [[1]]
 - (C2) Explanation: The assistant maintains no coherence with the user's question. The user asked for the influence of change in environment and activity on the interpretation of the third image. The assistant does not establish any connections among the three images and fails to refer to the scenes in the third image. Rating: [[1]]
 - (C3) Explanation: There is no relevance of the assistant's responses to the user's inquiries. The assistant does not discuss the contrast between the dynamism of the surfing scene and the static settings of the first two images as the user requested. Rating: [[1]]

Therefore, the overall rating of A1 is [[1]].

Therefore, the overall rating of A2 is [[1]].

Figure 2: Comparison between our SparklesChat (left) and MiniGPT-4 [59] (right) on an example from SparklesEval. We adapt MiniGPT-4 to accept multiple images as input. SparklesChat shows conversational competence in open dialogues across three criteria: (C1) image understanding and reasoning, (C2) maintaining cross-image and cross-turn coherence, and (C3) generating relevant and complete responses. In contrast, MiniGPT-4 faces challenges in these aspects, leading to difficulty following user instructions across various images and dialogue turns.

1 Introduction

Large language models (LLMs) have shown remarkable progress in zero-shot performance across a variety of tasks when fine-tuned using instruction-following data [36, 35, 45, 7, 49, 48, 53]. In the multimodal domain, multimodal instruction-following models such as MiniGPT-4 extend these capabilities by integrating pretrained vision encoders with instruction-following LLMs using projection layers [59]. MiniGPT-4 adapts the projection layer to align vision and language domains by training on concatenated embeddings of images and their descriptions. The training occurs in two stages: first, on a large-scale collection of image-text pairs and then on a smaller dataset of detailed, human-like image descriptions [59]. With this training method, MiniGPT-4 learns alignments between individual images and sentences and performs single-image understanding and reasoning. However, models such as MiniGPT-4 struggle to capture interactions among multiple images and text. This capability is crucial for user-assistant conversations, where users often refer to multiple images with text snippets to convey their instructions in detail. As shown in Figure 2, MiniGPT-4 mixes up the content of multiple images, fails to establish coherence between images, and consequently falls short in following user instructions during open dialogues.

One key limitation hindering progress in this area is the lack of specialized datasets designed for multimodal dialogues that involve multiple images and fine-grained, word-level text interactions. Existing models such as Flamingo can adapt to various image understanding tasks when prompted with a few relevant examples due to their training on image-text interleaved web data [2]. However, these models often fall short in following intricate human instructions because they are trained to predict the next word on a large web dataset rather than perform the task the user wants [36].

To address these gaps, we present **SparklesChat**, a multimodal instruction-following model for open-ended dialogues across multiple images. Unlike previous approaches such as MiniGPT-4 that takes the concatenation of a single image with sentence-level text as input (e.g., "★ Can you describe this image as detailed as possible?" - where '★' denotes a single image), SparklesChat, as shown in Figure 1, integrates multiple images at the word level (e.g., "Can you link the celebration occurring in IMAGE#2331159★ and the dirt bike race in IMAGE#2330601★?"). This innovation enables fine-grained integration of images and text, mimicking natural human communication more closely.

To support the training of SparklesChat, we introduce **SparklesDialogue**, the first machine-generated dialogue dataset designed for word-level interleaved multi-image and text interactions. We use OpenAI's GPT-4 [35] to simulate user-assistant conversations with visual capabilities by leveraging detailed image descriptions. Our dataset achieves greater robustness and diversity by incorporating two subsets, namely SparklesDialogueCC and SparklesDialogueVG, constructed from different image and description sources.

Furthermore, we introduce **SparklesEval**, a GPT-assisted benchmark to quantitatively evaluate a model's conversational competence in multimodal, open-ended dialogues across multiple images and dialogue turns. SparklesEval features a comprehensive and interpretable scoring system based on three distinct criteria: *Image Understanding and Reasoning*, *Cross-Image and Cross-Turn Coherence*, and *Relevance and Completeness of Responses*.

For quantitative evaluation, we validate the effectiveness of SparklesChat through extensive experiments. We conduct zero-shot evaluations on two standard vision-language tasks, including binary image selection on the BISON dataset [15] and visual reasoning on the NLVR2 dataset [43]. On the BISON dataset, SparklesChat achieved an accuracy of 56.7%, surpassing MiniGPT-4's 46.0%. On the NLVR2 dataset, SparklesChat reached an accuracy of 58.0%, outperforming MiniGPT-4's 51.3%. In our SparklesEval benchmark, SparklesChat scores 8.56 out of 10, significantly exceeds MiniGPT-4's score of 3.91, and closely approaches GPT-4's score of 9.26. Qualitative evaluations further demonstrate SparklesChat's applicability in real-world scenarios.

2 Related Works

Our work in constructing a dialogue dataset for instruction tuning is related to developments in multimodal image-text alignment, dialogue datasets, and instruction tuning. We provide a concise summary of these developments herein and direct readers to Appendix A for a detailed discussion.

Multimodal alignment datasets. Various datasets such as Visual Genome [24] and Conceptual Captions [40] align images with corresponding descriptions, forming the foundation for multimodal alignment [20, 37, 18, 38, 29]. Advancements such as the Common Crawl Interleaved data [16] and

Table 1: Prompt and response sequence formats used to train SparklesChat.

Dialogue Turn One
$\mathbf{X}_{\text{prompt}}^{i,1} = \mathbf{X}_{\text{system}} <\text{SEP}> \text{Human: } \mathbf{X}_{\text{q}}^{i,1} <\text{SEP}> \text{Assistant: }$
$\mathbf{X}_{\text{response}}^{i,1} = \mathbf{X}_{\text{a}}^{i,1} <\text{SEP}>$
<hr/>
Dialogue Turn Two
$\mathbf{X}_{\text{prompt}}^{i,2} = \mathbf{X}_{\text{system}} <\text{SEP}> \text{Human: } \mathbf{X}_{\text{q}}^{i,1} <\text{SEP}> \text{Assistant: } \mathbf{X}_{\text{a}}^{i,1} <\text{SEP}> \text{Human: } \mathbf{X}_{\text{q}}^{i,2} <\text{SEP}>$
Assistant:
$\mathbf{X}_{\text{response}}^{i,2} = \mathbf{X}_{\text{a}}^{i,2} <\text{SEP}>$

the Multimodal C4 dataset [60] expand conventional datasets by integrating multiple images and sentences from web corpora. Models including Flamingo [2] and Kosmos-1 [16] trained on them can adapt to various tasks using multiple image-text examples. However, they fall short in following intricate instructions as they are trained to predict the next word on a large web dataset rather than perform the task the user wants [36].

Multimodal dialogue datasets. Datasets such as Visual Dialog [9], created by crowd workers, and LLaVA [32], generated by LLMs, focus on image-driven conversations inquiring about image attributes or factual knowledge. Conversely, datasets such as OpenViDial [33] and PhotoChat [55] integrate images within daily human conversations sparsely. Nonetheless, these datasets are not explicitly designed for instructive, in-depth multi-image analysis dialogues, posing challenges in dealing with real-world analytical scenarios.

Multimodal instruction tuning. Multimodal instruction tuning developed with datasets like Multi-Instruct [50] offering benchmarks for diverse multimodal tasks and models like MiniGPT-4 [59] being fine-tuned on detailed image descriptions to align better with user intentions. Some techniques leverage LLMs to interpret image annotations and generate instruction-following datasets [32, 56, 31, 54]. Our dataset, SparklesDialogue, and model, SparklesChat, build upon these developments and explore complex interactions between multiple images and word-level text.

3 SparklesChat

We present a multimodal instruction-following model SparklesChat to foster interactions between users and AI assistants across multiple images and illustrate the framework in Figure 1.

3.1 Model

The foundation for SparklesChat is the MiniGPT-4 architecture, which connects a pretrained vision encoder and a pretrained LLM with a projection layer [59]. The language decoder, Vicuna [7], is based on the LLaMA framework [45], which can handle diverse language tasks. For image processing, we use the visual encoder from BLIP-2, combining a pretrained EVA-ViT in Vision Transformer (ViT) backbone with a pretrained Q-Former [29, 11, 10]. In the MiniGPT-4, the input to the language model is a single image representation followed by a sentence embedding of the image description. In SparklesChat, image representations of different images are embedded between text according to their positions in dialogues. Only the projection layer is trainable in the model while other vision and language components are frozen. More implementation details can be found in Appendix B.

3.2 Instruction-tuning

We represent an i -th T -turn dialogue as $\mathbf{X}^i = (\mathbf{X}_{\text{q}}^{i,1}, \mathbf{X}_{\text{a}}^{i,1}, \dots, \mathbf{X}_{\text{q}}^{i,T}, \mathbf{X}_{\text{a}}^{i,T})$, where each pair of $(\mathbf{X}_{\text{q}}^{i,t}, \mathbf{X}_{\text{a}}^{i,t})$ includes a question from the user and an answer from the assistant in turn- t . For each \mathbf{X}^i , we construct T training samples by organizing each pair of questions and answers as a sequence. Given a predefined system prompt $\mathbf{X}_{\text{system}}$, the prompt $\mathbf{X}_{\text{prompt}}^{i,t}$ and response $\mathbf{X}_{\text{response}}^{i,t}$ at the t -th turn are defined as the following:

$$\mathbf{X}_{\text{prompt}}^{i,t} = \begin{cases} \mathbf{X}_{\text{system}} <\text{SEP}> \text{Human : } \mathbf{X}_{\text{q}}^{i,1} <\text{SEP}> \text{Assistant : ,} & \text{if } t = 1, \\ \mathbf{X}_{\text{prompt}}^{i,t-1} <\text{SEP}> \text{Human : } \mathbf{X}_{\text{q}}^{i,t} <\text{SEP}> \text{Assistant : ,} & \text{if } t > 1. \end{cases} \quad (1)$$

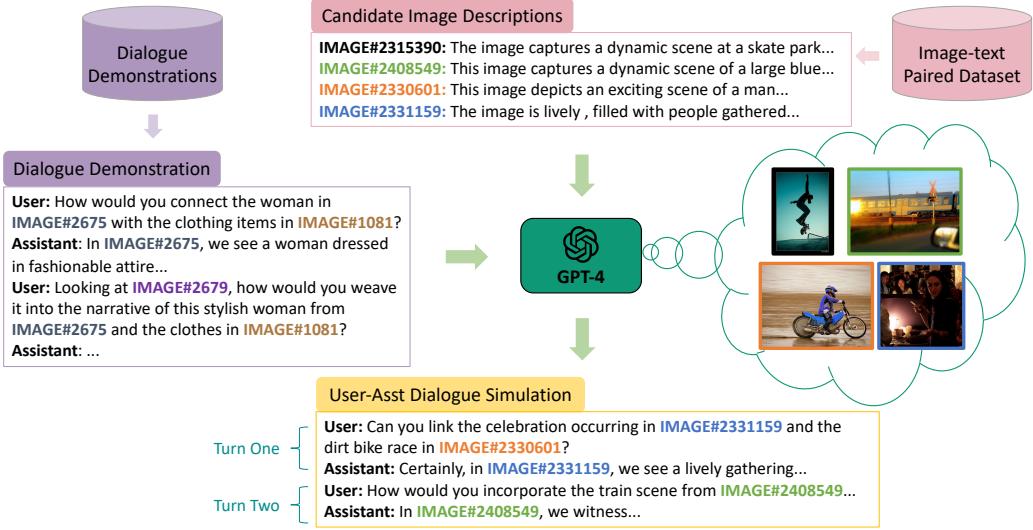


Figure 3: The GPT-assisted data construction process. GPT-4 simulates dialogues between a user and an assistant using multiple images. Dialogue Demonstrations act as learning examples for generating well-formatted dialogues, and Candidate Image Descriptions provide a pool of images for discussion. No visual images are sent to GPT-4 during this process.

$$\mathbf{X}_{\text{response}}^{i,t} = \mathbf{X}_{\text{a}}^{i,t} \langle \text{SEP} \rangle. \quad (2)$$

Table 1 illustrates the unified format for two-turn dialogue training sequences. We train the LLM on the prediction tokens using the auto-regressive training objective. Specifically, for a sequence of length L , we compute the probability of generating target responses $\mathbf{X}_{\text{response}}$ by:

$$p(\mathbf{X}_{\text{response}} | \mathbf{X}_{\text{prompt}}) = \prod_{l=1}^L p_{\theta}(x_l | \mathbf{X}_{\text{prompt}, < l}, \mathbf{X}_{\text{response}, < l}), \quad (3)$$

where θ is the trainable parameters, $\mathbf{X}_{\text{prompt}, < l}$ and $\mathbf{X}_{\text{response}, < l}$ are prompt and response tokens in all turns before the current prediction token x_l , respectively.

4 SparklesDialogue and SparklesEval

We introduce SparklesDialogue and SparklesEval for training and evaluation to enhance the conversational abilities of multimodal models across multiple images and dialogue turns.

4.1 GPT-assisted data construction

We aim to construct a multimodal dialogue dataset that offers fine-grained interactions between multiple images and words, mimicking user-assistant conversations. These dialogues should cover real-world concepts, objects, and entities, spanning scenarios that involve generating text materials, seeking advice, guidance, assistance, and much more. To achieve this, we use GPT-4 as the primary tool in our dialogue data generation, given its capabilities to follow complex instructions and extensive world knowledge. The data collection process is visualized in Figure 3. We instruct GPT-4 to simulate realistic and diverse dialogues between a user and an assistant with advanced image understanding and reasoning capabilities following a structured two-turn pattern. In the first turn, the user initiates a reasonable and creative message regarding some images. In response, the assistant generates detailed answers that include comprehensive reasoning regarding the visual content. In the second turn, the user introduces a new image for further discussion, referencing both the new and previous images. Again, we carefully design prompts for the assistant to respond with highly helpful and exceptionally detailed answers that provide comprehensive reasoning to better align with human preference.

Dialogue Demonstration and Candidate Image Descriptions are crucial components in this process. Dialogue Demonstrations serve as in-context learning examples, steering GPT-4

Table 2: Statistics of SparklesDialogue and SparklesEval.

Dataset Name	Image Source	Caption Source	#Dialogue	#Image Turn one	#Image Turn two	#Unique/Total Image
SparklesDialogue	CC [40]	MiniGPT-4 [59]	1,653	1	1	2,067/3,306
			1,799	2	1	2,642/5,397
			1,069	3	1	3,373/12,979
SparklesEval	VG [24]	SVIT [56]	1,000	2	1	3,000/3,000
			1,000	3	1	4,000/4,000
SparklesEval	VG [24]	SVIT [56]	50	2	1	150/150
			50	2	2	200/200
			50	3	1	550/550

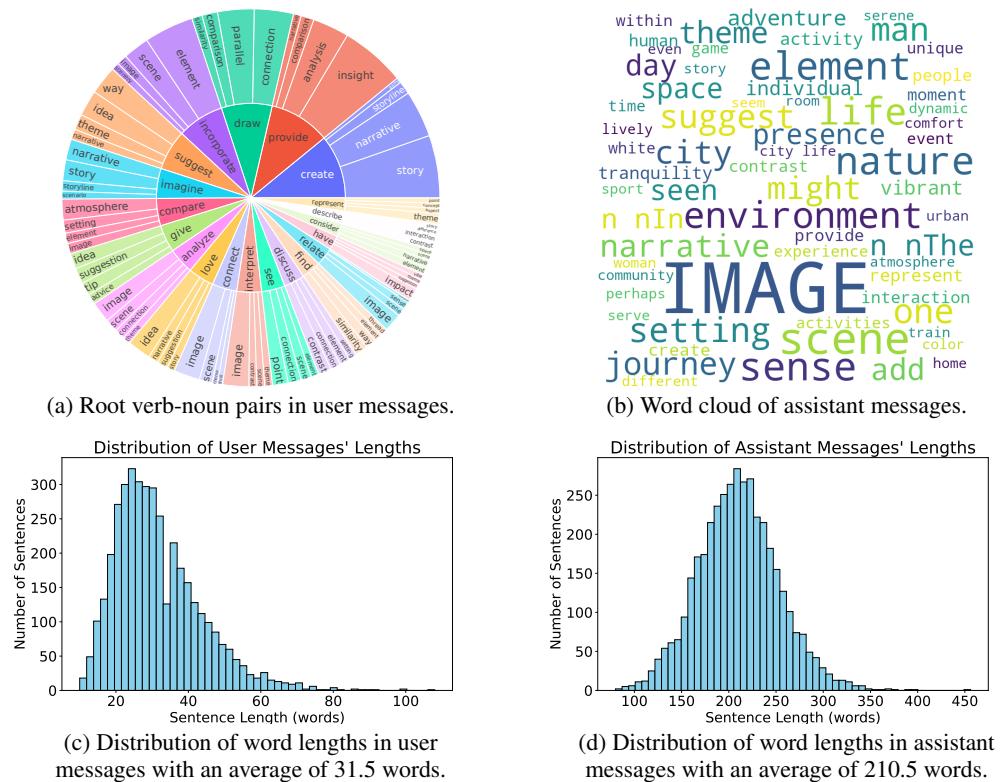


Figure 4: Characteristics of SparklesDialogueVG.

towards generating well-formatted and diverse responses. We initiated the creation of hundreds of demonstration dialogues with GPT-4’s assistance, using similar prompts and checking their quality. A small subset of them is randomly chosen each time. Candidate Image Descriptions serves as a candidate pool for relevant image selection. From the pool of image-text paired dataset, we randomly select a small subset as candidates each time. We include image IDs in dialogues to avoid reference ambiguity. Given that the publicly accessible GPT-4 API only accepts text input, we represent images with detailed descriptions. These descriptions, sourced from various image annotations such as image captions, bounding boxes, and region descriptions, comprehensively portray image content [59, 56, 32]. We parse and retain well-structured responses that conform to our desired format. More details, such as prompt templates and visualized examples, are in Appendix I.

4.2 Statistics and characteristics

We use the data construction process introduced in subsection 4.1 to collect a robust and diverse dataset named SparklesDialogue, which consists of two subsets: SparklesDialogueCC and Sparkles-

DialogueVG. The respective detailed descriptions, provided in MiniGPT-4 [59] and SVIT [56], correspond to image sources from Conceptual Captions (CC) [40] and Visual Genome (VG) [24]. SparklesDialogueVG is of high quality as the VG image descriptions generated by GPT-4 benefit from human-annotated captions, objects, and regions [56]. On the other hand, SparklesDialogueCC enriches SparklesDialogue by drawing from a more extensive set of images – 3.3 million in CC compared to 0.1 million in VG. However, the CC image descriptions are generated by a multimodal model with image features but not human annotations and are more prone to object hallucination issues [59]. Our ablation study elaborated in section 5.2 demonstrates that combining these two subsets improves SparklesChat’s capacity for understanding and reasoning across images and text. SparklesEval emphasizes more on accuracy and is thus constructed using the same source as SparklesDialogueVG. To encourage diversity in evaluation, SparklesEval was curated by analyzing the verb-noun distribution in user questions and selecting those that appear only once.

Table 2 provides the data statistics for SparklesDialogue and SparklesEval. SparklesDialogueCC comprises 4.5K dialogues, each consisting of at least two images spanning two conversational turns. On the other hand, SparklesDialogueVG includes 2K dialogues, each with at least three distinct images across two turns. SparklesEval includes 150 dialogues, with one-third containing two images in both the first and second conversational turns.

Figure 4 shows the characteristics of our dataset using SparklesDialogueVG as a representative subset. We explore key elements such as the root verb-noun pairs in user messages, a word cloud of assistant messages, and the length distributions. The questions from users are diverse, ranging from generating text materials to seeking advice or discussing the relationships between images, such as comparison and connection. The dialogues span various real-world topics, including the environment, nature, life, cities, etc. The high average word count in assistant messages suggests that the responses in SparklesDialogue are thorough and detailed. For details on the characteristics visualization of SparklesDialogueCC and extracting root verb-noun pairs and their visualization based on image count in each turn, please refer to Appendix G.

4.3 GPT-assisted evaluation: SparklesEval

While previous research, such as visual storytelling, has leaned toward human evaluations as superior to quantitative measures, these evaluations are often subjective, costly, and time-consuming [17]. Inspired by the consistency of recent LLMs with human assessment in evaluating output quality [57], we developed SparklesEval, a GPT-assisted benchmark to quantitatively assess a model’s conversational competence across multiple images and dialogue turns.

Each dialogue under evaluation comprises two turns of questions from SparklesEval and model-generated answers, considering both the current question and any preceding dialogue history. To evaluate a dialogue, we provide the judge model (i.e., GPT-4) with the complete dialogue but only require the judge model to assess the two answers. Descriptions corresponding to images mentioned in the dialogue are also provided. Each assessment is based on three distinct criteria across two turns, providing reasons and ratings on a scale of 1 to 10. These criteria scores include C1 for Image Understanding and Reasoning, C2 for Cross-Image and Cross-Turn Coherence, and C3 for Relevance and Completeness of Responses, with details in Appendix C. In particular, we further introduce averaged scores **A1** and **A2** for each turn, and an average **Score** computed by averaging **A1** and **A2**. The evaluation prompt and score computation process are elaborated in Appendix C.

Our evaluation approach differs from prior GPT-assisted evaluations in two aspects. First, it employs a combined score for a more comprehensive and interpretable assessment instead of a singular one [32]. Second, it is less biased and more efficient by assessing a single dialogue per prompt rather than contrasting multiple dialogues within one prompt [57]. Our approach eliminates position bias - the potential favor to certain positions when multiple dialogues are assessed within a prompt [57]. It enhances efficiency by avoiding the recalculation of combined scores for multiple dialogues.

5 Experiments

Beyond assessing conversational competence using SparklesEval, we evaluate models’ capabilities in zero-shot understanding and reasoning across multiple images through two vision-language tasks: binary image selection on BISON [15] and visual reasoning with natural language on NLVR2 [43].

Table 3: Model comparison on BISON, NLVR2 and SparklesEval. We investigate training models on different data sources, including detailed descriptions, complex reasoning, and dialogue data.

Model	Data	BISON	NLVR2	Score	SparklesEval					
					Turn one			Turn two		
		A1	C1	C2	C3	A1	C1	C2	C3	
GPT-4	-	-	-	9.26	9.26	9.23	9.18	9.38	9.26	9.25
MiniGPT-4	description [59]	46.0%	51.3%	3.91	3.55	3.67	3.53	3.44	4.28	4.38
MiniGPT-4*	description [59]	51.3%	46.7%	3.50	3.14	3.27	3.16	3.00	3.86	3.97
SparklesChat	description [32]	52.0%	48.0%	3.06	2.64	2.79	2.67	2.46	3.48	3.76
SparklesChat	reasoning [32]	52.7%	54.0%	6.71	6.55	6.63	6.42	6.59	6.87	6.89
SparklesChat	SparklesDialogue	56.7%	58.0%	8.56	8.76	8.81	8.67	8.81	8.35	8.37
										8.28
										8.41

The evaluation metric is accuracy for BISON and NLVR2; SparklesEval is rated 1-10. All models are adapted to accept multiple images for fair comparison. MiniGPT-4* is our reimplementation of MiniGPT-4 with the same setup as SparklesChat. Description and reasoning datasets from LLaVA [32] are adapted using formats similar to SparklesDialogue, with overlapping samples removed between train and evaluation sets.

Table 4: Ablation studies on BISON, NLVR2, and SparklesEval analyzing the effects of training SparklesChat with variations of SparklesDialogue on dialogue turn ratios and different subsets.

Data	Turns Ratio	BISON	NLVR2	Score	SparklesEval					
					Turn one			Turn two		
		A1	C1	C2	C3	A2	C1	C2	C3	
SparklesDialogue (CC+VG)	1:0	57.3%	55.3%	8.50	8.65	8.70	8.52	8.73	8.35	8.38
	0:1	50.7%	46.7%	8.24	8.24	8.23	8.18	8.32	8.24	8.15
	1:1	59.3%	51.3%	8.73	8.80	8.83	8.66	8.91	8.65	8.62
	1:2	49.3%	51.3%	8.43	8.54	8.57	8.43	8.63	8.31	8.28
	2:1	56.7%	58.0%	8.56	8.76	8.81	8.67	8.81	8.35	8.37
	3:1	50.7%	48.7%	8.45	8.69	8.74	8.52	8.83	8.20	8.18
SparklesDialogueCC	2:1	44.7%	53.3%	8.18	8.26	8.29	8.16	8.33	8.10	8.10
SparklesDialogueVG	2:1	54.7%	52.0%	8.59	8.71	8.76	8.60	8.78	8.47	8.35

The evaluation metric is accuracy for BISON and NLVR2; SparklesEval is rated 1-10.

5.1 Comparison of model performance

Table 3 compares the performance of SparklesChat and MiniGPT-4 on BISON, NLVR2, and SparklesEval evaluation datasets. We can see from the table that when SparklesChat is trained on reasoning data adapted from LLaVa [32], it achieves improved performance over models trained on description data on all metrics. This emphasizes the importance of reasoning ability in evaluation tasks.

SparklesChat, trained on our SparklesDialogue, outperforms other multimodal models in three tasks. SparklesChat achieves accuracies of 56.7% and 58.0% on BISON and NLVR2, respectively, reflecting its efficacy in handling tasks that require fine-grained visual grounding and compositional visual reasoning over two images. Moreover, SparklesChat significantly excels in the SparklesEval benchmark, scoring 8.56 out of 10. Its highest scores in both the first and second turns across all criteria indicate its superior ability in image understanding and reasoning, maintaining cross-image and cross-turn coherence, and generating relevant and complete responses. In comparison, models trained on description and reasoning data approximate scores of 3 and 6.71, respectively. GPT-4 scores the highest at 9.26, largely attributed to its utilization of detailed ground-truth annotations. SparklesChat's score is about 92% of that of GPT-4, highlighting SparklesChat's conversational competence across various images and dialogue turns.

For a comprehensive understanding, refer to Appendix D for evaluation settings including evaluation protocol and prompt design. Side-by-side comparisons of example outputs for SparklesChat and MiniGPT-4 on SparklesEval, BISON, and NLVR2 are illustrated in Figure 2 and Appendix D. Additionally, Appendix F provides a detailed evaluation of GPT-4, MiniGPT-4, and SparklesChat on SparklesEval using three different versions of judge models.

5.2 Ablation studies

We study the effect of training SparklesChat using data variants of SparklesDialogue on dialogue turn ratios and different subsets and show results in Table 4.

Effect of dialogue turns in SparklesDialogue. We first train models with individual dialogue turns. The model trained solely on the first turn (turns ratio ‘1:0’) performs better across all metrics compared to the model solely trained on the second turn (turns ratio ‘0:1’). Additionally, this model outperforms those trained on the baseline datasets, as evidenced in Table 3. This illustrates that our dataset boosts reasoning and conversational abilities even when used in isolation with just the first turn. In contrast, training only with the second dialogue turn reduces scores on BISON and NLVR2 when compared with baselines in Table 3. This could stem from the extended prompts in the second turn, which includes the content of the first turn, making them less aligned with the short prompt format favored by BISON and NLVR2. Then, we train models with SparklesDialogue blending with the same ratios of samples constructed from two dialogue turns, represented by turns ratio ‘1:1’ in Table 3. The results are better than only training with the second turn in all metrics, while worse than only training from the first turn in the task of NLVR2. An increase in the sampling ratios of the second turn data (turns ratio ‘1:2’) results in a performance drop as expected. Therefore, we increase the sampling ratio of the first-turn data until we cannot observe performance boosting. We finally settled on a 2:1 ratio for the first turn to the second turn as our default setting as it achieves balanced good performance across all benchmarks.

Effect of subsets of SparklesDialogue. Our model has been trained on two subsets of SparklesDialogue: SparklesDialogueCC and SparklesDialogueVG. The model trained on SparklesDialogueVG outperforms that trained on SparklesDialogueCC in both the BISON and SparklesEval evaluations, scoring 54.7% and 8.59, respectively, compared to 44.7% and 8.18. This enhanced performance is partly due to the higher quality of SparklesDialogueVG, which benefits from human-annotated data as discussed in subsection 4.2. It is worth noting that SparklesDialogueVG and SparklesEval use the same sources of images and captions, which could partially account for the higher score achieved by SparklesDialogueVG on SparklesEval. Both subsets demonstrate similar efficacy on the NLVR2 test. Combining both subsets yields higher performance on the BISON and NLVR2 tests, scoring 56.7% and 58.0% respectively. This surpasses the scores achieved by using either subset alone. In addition, the model trained on the combined dataset performs comparably to SparklesDialogueVG in the SparklesEval test, scoring 8.56 versus 8.59. This suggests that combining SparklesDialogueVG’s high-quality data and SparklesDialogueCC’s diverse data results in a more robust and versatile dataset for enhancing models’ capabilities in understanding and reasoning across images and text.

5.3 Demonstrations and applications

We conducted qualitative demonstrations to showcase SparklesChat’s broad applications in free-form scenarios by asking questions such as: “Create a story that takes place in ★ for the characters depicted in ★.”, “Imagine a dialogue between Harry Potter and ★ that takes place in the scene of ★”, “Create a song where the scene twists from ★ to ★.”, “Create a title for this song that takes inspiration from ★.”. The visualization of results is shown in Appendix E.

6 Conclusion

In conclusion, this work unlocks multimodal instruction-following models’ capabilities in open-ended dialogues involving multiple images. We introduced SparklesChat, a model designed to handle word-level text interactions in a multimodal context, offering natural conversational flow and direct context awareness. We also presented SparklesDialogue, the first machine-generated dialogue dataset tailored for multi-image and word-level text interactions. Furthermore, we proposed SparklesEval, a specialized benchmark for quantitatively assessing a model’s multimodal conversational competence. Experimental results demonstrated SparklesChat’s superiority over existing models in both standard vision-language tasks and the newly-introduced SparklesEval benchmark. We also conducted qualitative demonstrations to showcase the model’s broad applications in free-form scenarios. We discuss **limitations** of this work to inspire **future research** in this field in Appendix H. The code, model, and data are publicly available to foster development in the research community.

References

- [1] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.
- [3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- [5] Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*, 2023.
- [6] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *arXiv preprint arXiv:2305.10855*, 2023.
- [7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [9] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [11] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.
- [12] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *ACL*, 2023.
- [13] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- [14] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.
- [15] Hexiang Hu, Ishan Misra, and Laurens Van Der Maaten. Evaluating text-to-image matching using binary image selection (bison). In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- [16] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

- [17] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *NAACL*, 2016.
- [18] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *ACM Multimedia*, 2021.
- [19] Yupan Huang, Zhaoyang Zeng, and Yutong Lu. Be specific, be clear: Bridging machine and human captions by scene-guided transformer. In *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, 2021.
- [20] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, 2021.
- [21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.
- [22] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *ACL*, 2018.
- [23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, 2022.
- [24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [25] Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyon Myaeng. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021.
- [26] Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, and Ho-Jin Choi. Dialogcc: Large-scale multi-modal dialogue dataset. *arXiv preprint arXiv:2212.04119*, 2022.
- [27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.
- [28] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.
- [29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [31] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [33] Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*, 2020.

- [34] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017.
- [35] Openai. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- [39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
- [41] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image-chat: Engaging grounded conversations. In *ACL*, 2020.
- [42] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.
- [43] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.
- [44] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.
- [45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [46] Moskvoretskii Viktor and Kuznetsov Denis. Imad: Image-augmented multi-modal dialogue. *arXiv preprint arXiv:2305.10512*, 2023.
- [47] Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*, 2021.
- [48] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [49] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022.
- [50] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022.
- [51] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *NeurIPS*, 2021.

- [52] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [53] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- [54] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023.
- [55] Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.
- [56] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.
- [57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.
- [58] Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. MMChat: Multi-modal chat dataset on social media. In *Proceedings of The 13th Language Resources and Evaluation Conference*, 2022.
- [59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [60] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.

Appendix

A Related works	14
B Implementation details	15
C SparklesEval Details	16
D Zero-shot evaluation on vision-language tasks	16
E Demonstrations and applications	18
F Judging with different versions of GPT models	22
G Characteristics and verb-noun distribution analysis	22
H Limitations and future works	23
I GPT-assisted SparklesDialogue generation	25
I.1 Single dialogue generation for SparklesDialogueVG	25
I.2 Multiple dialogues generation for SparklesDialogueCC	32

A Related works

Our work exploits image-text pairs to construct a dialogue dataset for instruction-tuning. Thus, we review related works on multimodal alignment datasets, multimodal dialogue datasets, and multimodal instruction tuning, primarily on natural images and text domains.

Multimodal alignment datasets. Various datasets, such as MSCOCO [30], Visual Genome [24], Conceptual Captions [40], Conceptual 12M [4], ALIGN [21] and LAION [39], have been constructed to align images with their corresponding descriptions. These datasets have significantly contributed to the development of multimodal models for image-and-text generation [18, 19, 38, 29, 6] and understanding [21, 20, 37, 51]. We use these datasets in our data construction process. Emerging trends include datasets featuring interleaved images and text sequences from web corpora, such as M3W [2], web and Wikipedia articles [1], Common Crawl Interleaved data [16], and the Multimodal C4 dataset [60]. These datasets extend conventional image-text alignment training by incorporating multiple images and sentences. When trained on these enriched datasets, models such as Flamingo [2], OpenFlamingo [3], Kosmos-1 [16], and EMU [44] can adapt to various image understanding tasks using multiple task-relevant image-text examples. However, these models often fall short in following intricate human instructions because they are trained to predict the next word on a large web dataset rather than perform the task the user wants [36].

Multimodal dialogue datasets. Existing multimodal dialogue datasets broadly fall into two categories. The first comprises datasets where conversations are heavily rooted in and driven by images. Traditional datasets of this type are primarily generated by inviting crowd workers to engage in dialogues about a common image. Notable examples include Visual Dialog [9], which emphasizes question-answering tasks within AI-human chat about visual content, and IGC [34], a compilation of dialogues featuring an image, a corresponding textual description, and a conversation centered on the image. Image-Chat presents image-grounded dialogues crafted around given images [41]. Recently, dialogue datasets, such as LLaVA [32], SVIT [56], and LAMM [54], created by LLMs alongside image annotations have surfaced. Each dialogue in these datasets begins with an inquiry about image attributes or factual knowledge, with responses expected to be brief within 50 words, which may not align with real-world scenarios requiring in-depth multi-image analysis. The second category features

datasets derived from **daily human conversations**, with images interspersed within multi-turn conversations sparsely. For example, OpenViDial [33, 47] is sourced from dialogues in movies and TV series, whereas PhotoChat [55] is a human-human dialogue dataset developed through crowdsourcing and features photo-sharing. Other datasets, such as DialogCC [26], MultiModalDialogue [25], and IMAD [46] enhance text-only dialogues by incorporating semantically relevant images. In addition, MMChat [58] and MMDialog [12] encompass image-grounded dialogues derived from social media interactions. However, these datasets, not being designed for user-assistant interactions, struggle with instructive, problem-solving dialogue requirements.

Multimodal instruction tuning. Multimodal instruction tuning has grown substantially with the advent of multimodal instruction datasets. For instance, MultiInstruct [50] offers a benchmark comprising 62 diverse multimodal tasks unified in a seq-to-seq format. InstructBLIP [8] extended the scope by transforming 26 datasets into instruction-tuning form. Otter [28] is trained on MIMIC-IT [27], a multimodal in-context instruction tuning dataset constructed by grouping multiple similar instructions into a contextual example. To better align with user intentions, MiniGPT-4 is fine-tuned on a small dataset of detailed image descriptions [59] and PF-1M [5] rewrites image annotations in a human-like style across 37 vision-language datasets. Furthermore, techniques such as LLaVA [32], SVIT [56], LRV-Instruction [31], and LAMM [54] have emerged. These methods leverage language-only APIs such as OpenAI’s GPT-4 [35] and self-instruction methods [48] to interpret image annotations (e.g., image captions, region descriptions, object bounding boxes, attributes, and relationships), and generate responses in various forms (i.e., short conversations, image captioning, and visual reasoning). Models such as mPLUG-Owl [52], PandaGPT [42], LLaMAAdapter V2 [13], and Multimodal-GPT [14] further extended this area, incorporating both language-only and vision-language instruction data. These developments are a valuable foundation for our work. Our dataset, SparklesDialogue, is inspired by GPT-assisted data construction techniques and explores the interactions between multiple images and word-level textual content. Training our model, SparklesChat, on this dataset unlocks the capability of multimodal models to interpret complex image-text interactions.

B Implementation details

We implemented SparklesChat on the MiniGPT-4 codebase [59]¹. We refer to MiniGPT-4’s efficient fine-tuning process and tune SparklesChat using 1,500 training steps with a batch size of 8, based on MiniGPT-4’s first-stage pretrained model. Our training data of SparklesDialogue is sampled with the same ratio from SparklesDialogueCC and SparklesDialogueVG, and with sampling ratios of 2 and 1 from the first and second turns of dialogues, respectively.

In our evaluation, we configure the parameters as follows: `temperature` is set to 1.0, `top_p` to 0.9, and `max_new_tokens` to 300, with both `repetition_penalty` and `length_penalty` at 1.0. For demonstration cases, the `beam_size` is 2; for all other evaluations, it is 1.

During instruction-tuning, we follow MiniGPT-4 to use `<ImageHere>` to represent images [59]. In practice, all tags of `<ImageHere>` are replaced by the visual features produced by a linear projection layer. Tags of `` and `` are language tokens that serve as signals for the start and end of images. A system message X_{system} is appended to the beginning of each prompt. We also append `Human:` and `Assistant:` before each user and assistant messages to equip the model with conversation capability. System, user, and assistant messages are separated by a separator `<SEP>`. The system message $X_{\text{system}} = \text{Give the following image: } <\text{Img}>\text{ImageContent}</\text{Img}>. \text{ You will be able to see the image once I provide it to you. Please answer my questions.}$ The separator `<SEP> = ###`.

We tailored the OpenAI’s GPT-4 API (gpt-4-0613) parameters to balance diversity and quality for constructing SparklesDialogue and SparklesEval. We set the `temperature` and `top_p` parameters to 1.0, the `max_tokens` parameter to 2,048, and both the `frequency_penalty` and `presence_penalty` parameters to 0.0. In each query to the GPT-4 API, the “system” role was allocated the default instruction `You are a helpful assistant.` As of July 2023, the cost for generating 1,000 tokens was \$0.06 for outputs and \$0.03 for inputs within an 8K context², leading to

¹<https://github.com/Vision-CAIR/MiniGPT-4>

²<https://openai.com/pricing>

a total dataset generation cost of approximately \$500. The cost of evaluating a model on SparklesEval is approximately \$1.4 and \$14 using gpt-3.5-turbo-0613 and gpt-4-0613, respectively.

C SparklesEval Details

The three criteria of GPT-assisted evaluation on SparklesEval are as follows:

Image understanding and reasoning score C1: Assess the assistant’s proficiency in accurately identifying and describing objects, contexts, and relationships within and across the images.

Cross-image and cross-turn coherence score C2: Evaluate the assistant’s ability to maintain consistent understanding across multiple images and dialogue turns.

Relevance and completeness of responses score C3: Determine the extent to which the assistant’s responses are directly related to the user’s inquiries and the images’ content, and whether the responses provide comprehensive and detailed answers.

Following this, we ask the GPT models to assign a combined score for each turn. For each model’s evaluation results, we gather scores for three criteria across two turns. First, we compute the mean scores for all criteria over evaluation samples. Next, we calculate the combined scores **A1** and **A2** by averaging their respective criteria scores, namely $\mathbf{A1} = \text{mean}(\mathbf{C1}, \mathbf{C2}, \mathbf{C3})$ for the first turn and $\mathbf{A2} = \text{mean}(\mathbf{C1}, \mathbf{C2}, \mathbf{C3})$ for the second turn. We refrain from using the **A1** and **A2** scores provided by the judge models, as their calculations may be inaccurate. Ultimately, we derive a final overall **score** by averaging **A1** and **A2**. Through this methodology, our evaluation is more holistic and interpretable.

The prompt template of GPT-assisted evaluation on SparklesEval is presented in Table 5.

D Zero-shot evaluation on vision-language tasks

We chose two vision-language tasks, binary image selection and visual reasoning, to evaluate zero-shot understanding and reasoning capabilities over multiple images. For both tasks, the evaluation metric is accuracy. Side-by-side comparisons of example outputs for SparklesChat and MiniGPT-4 on BISON and NLVR2 can be found in Figure 5.

Binary image selection on BISON. The Binary Image Selection task measures a model’s ability to select the correct image from a pair given a text query that describes one of them [15]. The model’s performance is assessed in terms of binary classification accuracy. For this task, 150 examples were randomly sampled from the COCO-BISON dataset³. The image source of COCO-BISON is COCO images. The image source of SparklesDialogueCC is Conceptual Captions, which should have no overlap with COCO. However, our SparklesDialogueVG originates from the Visual Genome, which includes a subset of COCO images. We carefully eliminate any overlapping images to ensure no overlap between the training and evaluation data.

Visual reasoning with natural language on NLVR2. The evaluation of the Visual Reasoning with Natural Language task assesses the model’s ability to predict whether a sentence is true about a pair of images [43]. This task addresses the challenge of compositional visual reasoning on relations, comparisons, and quantities. The NLVR2 dataset [43] was used for this evaluation, with 150 examples randomly sampled from the public balanced test set⁴. The images in the NLVR2 dataset are sourced from Google Images, distinct from our SparklesDialogueVG’s image source of the Visual Genome [24] and primarily feature images from Flickr.

Evaluation protocol and prompt design. Models are evaluated on these tasks without any additional training. Inspired by [23], we used a simple prompt, “*Let’s think step by step*”, to facilitate step-by-step reasoning before answering each question. We used the phrase “*Therefore, the answer is*” to prompt the answer. Instead of using a two-stage prompting as in [23], we combined the reasoning

³https://github.com/facebookresearch/binary-image-selection/blob/main/annotations/bison_annotations.cocoval2014.json

⁴https://github.com/lil-lab/nlvr/blob/master/nlvr2/data/balanced/balanced_test1.json

Table 5: Prompt format for SparklesEval evaluation.

Users will interact with a conversational assistant. The assistant is designed to understand, analyze, and reason about multiple images across two turns of conversation. The assistant is expected to provide highly helpful and exceptionally detailed answers providing comprehensive reasoning regarding the visual content of the images.

Below are images represented by their image IDs and captions (delimited by triple quotes):

```
```json
{Target Image Descriptions}
```
```

Next is a dialogue between a user and the assistant regarding the images above:

```
```
###User Q1:
{Q1}

###Assistant A1:
{A1}

###User Q2:
{Q2}

###Assistant A2:
{A2}
```

```

Your task as an impartial judge is to evaluate the responses (A1 and A2) provided by the assistant to the user's questions.

Please rate the following three criteria C1, C2, and C3 on a scale of 1-10 for A1 and A2 separately, where a higher score indicates better overall performance:

- (C1) Image Understanding and Reasoning: This measures the assistant's ability to accurately identify and describe objects, context, and relationships within and between the images.
- (C2) Cross-Image and Cross-Turn Coherence: This evaluates the assistant's ability to maintain a consistent understanding across multiple images and dialogue turns.
- (C3) Relevance and Completeness of Responses: This assesses whether the assistant's responses are directly related to the user's inquiries and the images' content, and whether the responses provide thorough, detailed answers.

Begin your evaluation by providing a short explanation for each criterion. Be as objective as possible. After providing your explanation, rate the response on a scale of 1 to 10 by strictly following the format below (note that "5" and "..." are placeholders):

```
```
* Evaluating A1
- (C1) Explanation: "..." Rating: [[5]]
- (C2) Explanation: "..." Rating: [[5]]
- (C3) Explanation: "..." Rating: [[5]]
Therefore, the overall rating of A1 is [[5]]

* Evaluating A2
- (C1) Explanation: "..." Rating: [[5]]
- (C2) Explanation: "..." Rating: [[5]]
- (C3) Explanation: "..." Rating: [[5]]
Therefore, the overall rating of A2 is [[5]]
```

```

Carefully examine a pair of images: the left IMAGE#1  and the right IMAGE#2 .

Determine whether the following statement is true about the pair of images:
'There is a smartphone in the right image.'

Jointly reasoning about the statement grounded in **IMAGE#1** and **IMAGE#2**.
The task requires compositional joint reasoning, including about quantities, comparisons, and relations. Let's think step by step.
Please start your response with "**Let's think step by step.**" and end with "**Therefore, the answer (TRUE or FALSE) is**".

User 

Sure, I'll respond as instructed. Let's think step by step...
1. The statement mentions there's a smartphone in the right image. Based on **IMAGE#2**, we can see a computer and a pair of headphones on the desk, but no obvious smartphone.
2. It could be assumed there might be a smartphone hidden from view or placed out of the frame. However, the image is centered and offers a clear view of everything visible within the frame. Therefore, this assumption appears to be incorrect.
3. The only logical inference we can make from this statement is that the smartphone mentioned is the one in the left image - **IMAGE#1**.
4. Since the left image shows a desk with a computer (white laptop) and a pair of headphones on a desk, we can reason that there's another device present. However, we cannot confirm it is a smartphone. It could be a laptop or a tablet.
Therefore, we can conclude that the statement "**'There is a smartphone in the right image'**" is **FALSE**. It is possible to deduce that the computer in the right image is being used for work or entertainment, and may not necessarily be a smartphone, but it's not certain.

SparklesChat 

VS.

Let's think step by step.
The left image is an Apple MacBook Pro, and the right image is a computer mouse. There are two computers in the images, and one of them is a MacBook Pro. Therefore, the answer is **TRUE**.

MiniGPT-4 

NLVR2 - Statement: There is a smartphone in the right image. **Label:** FALSE

BISON - Statement: a line of people waiting on a catering truck. **Label:** IMAGE#1

Carefully examine the two similar images of **IMAGE#1**  and **IMAGE#2** .

Given the following caption, you must select which of two images best matches the caption. The caption is: '**a line of people waiting on a catering truck**'.
Jointly reasoning about the statement grounded in **IMAGE#1** and **IMAGE#2**.
This task requires fine-grained visual reasoning between the caption and each image. Let's think step by step.
Please start your response with "**Let's think step by step.**" and end with "**Therefore, the answer (IMAGE#1 or IMAGE#2) is**".

User 

Let's think step by step.
We have two images - **IMAGE#1** is a black and white photograph of a group of people standing near a vintage airplane. **IMAGE#2** is an image of a group of people lining up in front of a food truck. **IMAGE#1** contains various visual clues. There is a fence surrounding the crowd, and several people are standing near the plane. The scene is outside, so the sky is visible. We can see people dressed in various attire, including a man in a suit, indicating the event could be formal. Now, considering the caption, **IMAGE#1** doesn't exactly match the description, but it does show people waiting outside. **IMAGE#2**, however, features a large crowd waiting in line for food. Hence, **IMAGE#1** is not entirely accurate, but we can infer that **IMAGE#2** closely matches the caption. Therefore, the answer is **IMAGE#2**.

SparklesChat 

VS.

IMAGE#1 is the better match for the caption because it shows a line of people waiting on a catering truck. The image has a clear view of the catering truck and several people waiting in line for food. In contrast, **IMAGE#2** only shows a group of people standing near a fence, with no clear indication of a catering truck. Therefore, the answer is **IMAGE#1**.

MiniGPT-4 

Figure 5: Comparison between SparklesChat and MiniGPT-4 on examples of NLVR2 and BISON.

extraction and answer extraction stages into a single prompt: "*Please start your response with 'Let's think step by step.' and end with 'Therefore, the answer is'*". The full evaluation prompt templates to evaluate NLVR2 and BISON datasets are presented in Table 6. We regenerated the response if the model failed to follow the instructions to output responses in the specified format. This approach ensures an unambiguous response and allows us to extract a potential answer from the text following the last occurrence of "*Therefore*".

E Demonstrations and applications

We conducted qualitative demonstrations to showcase the model's wide applications in free-form scenarios in Figure 6, Figure 7, and Figure 8.

Create a story that takes place in  for the characters depicted in .



Once upon a time, in the heart of the city, there stood a grand, red brick building - the Hogwarts School of Witchcraft and Wizardry. It was a place where students from all over the world came to learn the magical arts and explore their hidden talents. Harry Potter, the protagonist of the story, had always dreamed of attending Hogwarts. He was an orphan, and his mother's death had left him feeling isolated and alone. But, he was determined to fulfill his mother's dream of becoming a wizard.

The day Harry arrived at Hogwarts, he was struck by the grandeur of the building. The tall, steeped clock tower stood out against the bright blue sky, and the red brick walls seemed to glow in the sunlight. The students were all dressed in their robes, and Harry couldn't help but feel a sense of awe and wonder.

As he made his way through the castle, Harry was greeted by the warm, friendly faces of his new classmates. He was struck by the diversity of the students, each with their own unique talents and personalities. Some were friendly, while others seemed aloof, but Harry was determined to make the most of his time at Hogwarts.

In his first year at Hogwarts, Harry was sorted into Gryffindor House, where he met his best friends Ron Weasley and Hermione Granger. Together, they faced many challenges and adventures, from the dangerous Dementors to the infamous Heir of Salazar Slytherin. Through it all, Harry learned the true meaning of friendship, bravery, and the power of love.

Year after year, Harry continued his magical education at Hogwarts. He learned how to cast spells, brew potions, and explore the mysteries of the wizarding world. He also discovered his own magical talents, such as his ability to speak Parseltongue and his connection to Voldemort's past.

As Harry's time at Hogwarts drew to a close, he realized that he had truly found his home. He had friends, a sense of belonging, and a deep understanding of the magical world. And, most importantly, he had learned the true value of love, friendship, and bravery. In the end, Harry's journey at Hogwarts had transformed him from a shy orphan into a powerful wizard, ready to face the challenges of the future.

SparklesChat

Imagine a dialogue between Harry Potter and  that takes place in the scene of .



Harry Potter, a young wizard, stands in front of the grand dining hall of Hogwarts School of Witchcraft and Wizardry. The room is decorated with wooden tables, chairs, and a large chandelier hanging from the ceiling. The walls are adorned with tapestries and paintings depicting various magical creatures and scenes from Hogwarts' history. Harry's eyes are fixed on the black cat sitting on the table in front of him.

Harry: "Hello there, Mr. Whiskers."

Mr. Whiskers: "Hello, Harry. How are you doing today?"

Harry: "I'm doing well, thank you. I'm just admiring the beauty of this dining hall. It's quite a sight, isn't it?"

Mr. Whiskers: "Indeed, it is. Hogwarts has always been a place of grandeur and magic. I'm glad you appreciate it."

Harry: "I've been looking forward to this day for a long time. It's my first day at Hogwarts."

Mr. Whiskers: "Ah, I remember my first day at Hogwarts. It was quite overwhelming, but also thrilling. I'm sure you'll have a wonderful time here."

Harry: "I hope so. I've heard so much about Hogwarts and the magical world. I can't wait to explore it all."

Mr. Whiskers: "Well, you're in the right place. Hogwarts is full of magic, adventure, and wonder. I'm sure you'll find your place here."

Harry: "Thank you, Mr. Whiskers. I'm looking forward to discovering it all."

SparklesChat

Figure 6: Demonstration of SparklesChat to create a story and a dialogue that connects places and characters.

Create a song where the scene twists from  to 

The song you've imagined is a beautiful blend of imagination and reality. The opening scene depicts a young girl in a red dress standing on the edge of a cliff, looking out at a bustling city below. She's holding onto a long rope, which seems to be tied to a wooden structure on the edge of the cliff. The girl looks peaceful and serene, as if she's contemplating something profound.

As the song progresses, the scene shifts to a vibrant and colorful world of balloons. The girl is now in the middle of a bustling city, surrounded by colorful balloons of all shapes and sizes. The balloons seem to be floating in the air, creating a sense of playfulness and joy.

The lyrics of the song could be something like this:

Verse 1:
 She's standing on the edge of a cliff, looking out at the city below
 She's holding onto a long rope, tied to a wooden structure on the edge of the cliff
 She's lost in thought, as if she's contemplating something profound

Chorus:
 She's floating in a sea of balloons, surrounded by joy and playfulness
 She's laughing and smiling, as if she's found a new sense of freedom
 She's dancing with the wind, as if she's found a new way to fly

Verse 2:
 She's in the middle of the city, surrounded by colorful balloons of all shapes and sizes
 She's bouncing and leaping, as if she's found a new way to live
 She's laughing and smiling, as if she's found a new way to be free

Chorus:
 She's floating in a sea of balloons, surrounded by joy and playfulness
 She's laughing and smiling, as if she's found a new sense of freedom
 She's dancing with the wind, as if she's found a new way to fly

Bridge:
 She's found a new way to soar, a new way to fly
 She's found a new way to live, a new way to be free
 She's found a new way to dream, a new way to be

Chorus:
 She's floating in a sea of balloons, surrounded by joy and playfulness
 She's laughing and smiling, as if she's found a new sense of freedom
 She's dancing with the wind, as if she's found a new way to fly.

 **SparklesChat**

Create a title for this song that takes inspiration from  and tell how they are connected

The song could be titled "Golden Stars" as a nod to the image of the golden stars in the sky. The image of the stars in the sky could be interpreted as a metaphor for the dreams and aspirations of the protagonist in the song. The stars in the sky are bright, shining, and inspiring, much like the protagonist's dreams and aspirations in the song. The song's chorus "She's floating in a sea of balloons, surrounded by joy and playfulness" could be interpreted as a metaphor for the protagonist's journey towards her dreams and aspirations, much like the way the balloons in the image are floating and bouncing in the air. The image of the stars in the sky and the balloons in the song both represent a sense of hope, aspiration, and dreams.

 **SparklesChat**

Figure 7: Demonstration of SparklesChat to compose a song containing two scenes and generate a song title inspired by another image.

 User

I'm showing my friends around a building. Its outside looks like 

and the inside looks like 

. Please tell them about the building.

 SparklesChat

The building you've shown to your friends appears to be a modern, multi-level structure with a glass exterior. It has a large glass wall that looks out onto a green space, likely a park or a garden. The interior of the building is equally impressive, with an open floor plan and high ceilings. The building's structure appears to be made of glass and steel, which gives it a sleek and modern look. The interior is well-lit and airy, with large windows that allow natural light to pour in. The building's architecture is unique and eye-catching, making it a prominent landmark in the city.

Here is the exterior 

and interior 

of another building. Could you compare the similarities and differences between the two buildings?

 User

Certainly! Both buildings share a unique architectural style and design. The exterior of the first building, as seen in the first image, features a large glass wall that looks out onto a green space. The building's exterior is modern and sleek, with large windows that allow natural light to pour in. The interior of the first building, as seen in the second image, is open and spacious, with high ceilings and large windows. The building's architecture is modern, with clean lines and a minimalist design. The second building, on the other hand, features a unique and eye-catching exterior, with a large, colorful building that looks like it's been built around a central staircase. The interior of the second building is also unique, with a mix of old and new architectural elements. The building features a mix of wooden floors, modern furniture, and traditional brick walls, creating a unique atmosphere. Overall, both buildings share a modern, sleek style, but the second building has a more eclectic mix of architectural elements.

 SparklesChat

What culture does 

convey?

 User

The building you've shared in the image captures a sense of vibrancy and diversity. The brightly colored tiles, the vibrant rainbow curtain, and the mix of wooden and metal fixtures all suggest a welcoming, energetic atmosphere. The interior of the building is bustling, with a mix of natural light and artificial light sources. The large, open spaces are conducive to gathering and socializing. The colorful tiles and the rainbow curtain could be interpreted as symbols of diversity and inclusivity, reflecting the building's ability to accommodate a diverse range of people and activities. The wooden and metal fixtures, on the other hand, suggest a sense of warmth and comfort, making the building a welcoming and inviting space. Overall, the image suggests a vibrant, energetic, and welcoming culture.

 SparklesChat

Figure 8: Demonstration of SparklesChat to describe and reason about different groups of images.

Table 6: Prompt formats to evaluate NLVR2 and BISON datasets.

| |
|--|
| NLVR2
Carefully examine a pair of images: the left IMAGE#1<ImageHere> and the right IMAGE#2<ImageHere>. Determine whether the following statement is true about the pair of images: ' {Statement} '. Jointly reasoning about the statement grounded in IMAGE#1 and IMAGE#2. The task requires compositional joint reasoning, including quantities, comparisons, and relations. Let's think step by step. Please start your response with "Let's think step by step." and end with "Therefore, the answer (TRUE or FALSE) is". <hr/> |
| BISON
Carefully examine the two similar images of IMAGE#1<ImageHere> and IMAGE#2<ImageHere>. Given the following caption, you must select which of two images best matches the caption. The caption is: ' {Caption} '. This task requires fine-grained visual reasoning between the caption and each image. Let's think step by step. Please start your response with "Let's think step by step." and end with "Therefore, the answer (IMAGE#1 or IMAGE#2) is". <hr/> |

F Judging with different versions of GPT models

As of July 2023, while it is widely recognized that employing gpt-4 as a judge model outperforms alternatives such as gpt-3.5-turbo, the cost of using gpt-4 is significantly higher. Therefore, we also provide scores generated by gpt-3.5-turbo as a reader reference, although we strongly recommend utilizing gpt-4 or more advanced future models as reliable judges. The version of gpt-4-0701 refers to API version 2023-07-01-preview for the GPT-4 model. We adopt the latest version gpt-4-0701 as our default judge model.

We evaluate GPT-4, MiniGPT-4, and SparklesChat using SparklesEval, leveraging three versions of judge models, as presented in Table 7. Both MiniGPT-4 and SparklesChat generate responses based on the question and accompanying visual image. At the same time, GPT-4 is a reference LLM that only uses textual information, including the question, the ground-truth bounding boxes, and captions. From the table, we observe that the more advanced judge models - gpt-4-0613 and gpt-4-0701 - provide higher scores compared to the older gpt-3.5-turbo-0613 when assessing both GPT-4 and OurModel (approximately nine versus eight). However, these advanced judge models yield considerably lower scores for MiniGPT-4 (about three versus five). GPT-4 achieves the highest score of 9.26 out of 10 when evaluated by the default gpt-4-0701 mainly due to its use of detailed ground-truth annotations. Nevertheless, it's worth noting LLM judge models may display a self-enhancement bias, favoring the responses they generate [57]. In contrast, MiniGPT-4 performs behind with a score of just 3.91. SparklesChat achieves a score of 8.56 - about 92% of the GPT-4 score - demonstrating SparklesChat's efficacy in generating responses that are not only relevant and complete but also exhibit cross-image and cross-turn coherence.

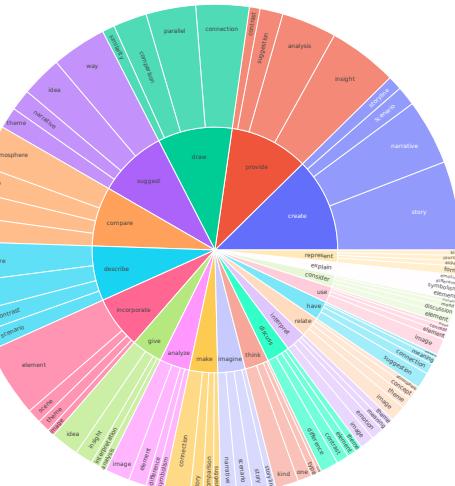
G Characteristics and verb-noun distribution analysis

The characteristics of SparklesDialogueCC are shown in Figure 9. For verb-noun distribution, we follow Self-instruct [48] to extract the verb closest to the root and its first direct noun object and plot the top 20 most common root verbs and their top 4 direct noun objects. We use the Berkeley Neural Parser⁵ [22] to parse user messages. We mainly focus on the last sentence of each message because it usually contains the question. If we can't extract the verb-noun pair from it, we look at the first sentence instead. For SparklesDialogueVG and SparklesDialogueCC, we visualize the

⁵<https://parser.kitaev.io/>

Table 7: Evaluation results on SparklesEval with different judge models.

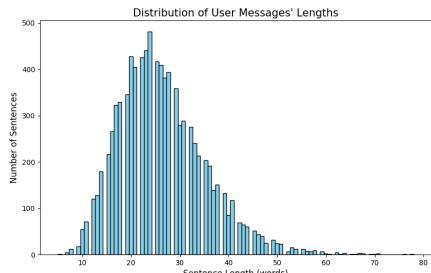
| Model | Judge Model Version | Score | Turn one | | | Turn two | | | | |
|--------------|---------------------|-------|----------|------|------|----------|------|------|------|------|
| | | | A1 | C1 | C2 | C3 | A2 | C1 | C2 | C3 |
| GPT-4 | gpt-3.5-turbo-0613 | 8.48 | 8.61 | 8.69 | 8.18 | 8.95 | 8.35 | 8.27 | 8.08 | 8.68 |
| | gpt-4-0613 | 9.51 | 9.50 | 9.53 | 9.37 | 9.60 | 9.53 | 9.49 | 9.46 | 9.64 |
| | gpt-4-0701 | 9.26 | 9.26 | 9.23 | 9.18 | 9.38 | 9.26 | 9.25 | 9.15 | 9.38 |
| MiniGPT-4 | gpt-3.5-turbo-0613 | 5.51 | 5.46 | 6.11 | 4.78 | 5.48 | 5.55 | 5.92 | 5.23 | 5.51 |
| | gpt-4-0613 | 3.31 | 3.11 | 3.12 | 3.09 | 3.10 | 3.51 | 3.57 | 3.40 | 3.56 |
| | gpt-4-0701 | 3.91 | 3.55 | 3.67 | 3.53 | 3.44 | 4.28 | 4.38 | 4.21 | 4.23 |
| SparklesChat | gpt-3.5-turbo-0613 | 8.37 | 8.51 | 8.59 | 8.04 | 8.89 | 8.24 | 8.16 | 7.92 | 8.64 |
| | gpt-4-0613 | 8.82 | 8.75 | 8.78 | 8.59 | 8.89 | 8.88 | 8.89 | 8.79 | 8.97 |
| | gpt-4-0701 | 8.56 | 8.76 | 8.81 | 8.67 | 8.81 | 8.35 | 8.37 | 8.28 | 8.41 |



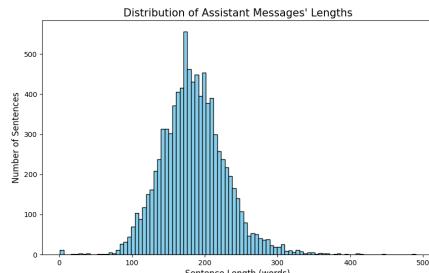
(a) Root verb-noun pairs in user messages.



(b) Word cloud of assistant messages.



(c) Distribution of word lengths in user messages with an average of 26.3 words.



(d) Distribution of word lengths in assistant messages with an average of 184.6 words.

Figure 9: Characteristics of SparklesDialogueCC.

verb-noun distributions regarding different numbers of images in each turn in Figure 10 and Figure 11 respectively.

H Limitations and future works

We discuss some limitations of this work to inspire future research in this field. First, SparklesChat shares common drawbacks with large language models, such as being out-of-date in its knowledge, sometimes providing inaccurate information, and having limited context length and inference speed [35]. Potential solutions may include regular updates to the model’s knowledge base and

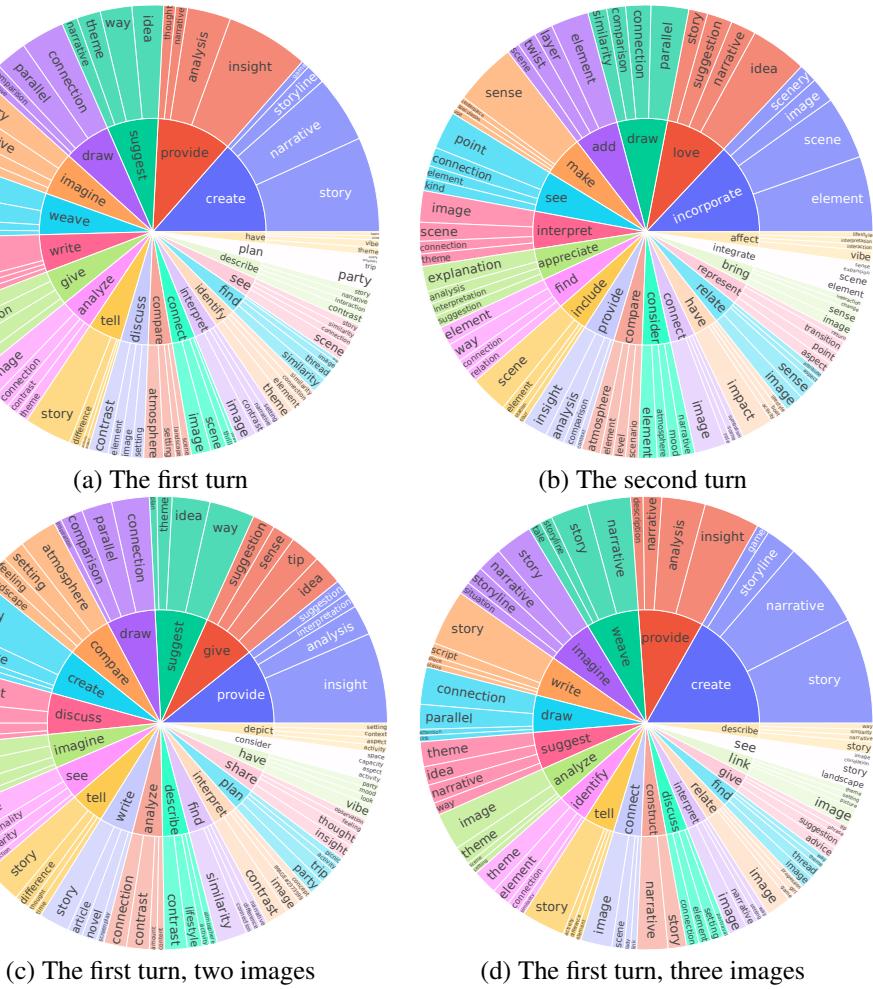


Figure 10: Root verb-noun distributions of SparklesDialogueVG.

fine-tuning with more reliable data sources. Second, SparklesChat inherits weaknesses from vision models, such as inaccurate object recognition, people/places identification, or visual relationships reasoning [29]. This calls for a more powerful visual perception model, and training on more well-aligned image-text datasets. Third, SparklesChat occasionally encounters difficulties maintaining multi-image and multi-turn consistency. Specifically, the model may lose the context of prior images after several dialogue turns or mix up the contents of different images. Potential solutions involve advanced model designs in position encoding and attention mechanisms to enhance the model’s consistency in recalling historical images and dialogues. Fourth, SparklesDialogue primarily concentrates on natural images, which limits its versatility in handling text-rich images such as charts, tables, and receipts, as well as domain-specific images such as medical scans, math illustrations, and satellite photos. Moreover, the dialogues in SparklesDialogue do not cover all possible user scenarios. Therefore, broadening the dataset to cover more diverse image types and user cases is a direction for future work. Fifth, the reliability of SparklesEval is tied to the capabilities of current GPT models. This limitation can be mitigated by incorporating more robust judge models and the assistance of human evaluators. Lastly, further safety considerations are needed to mitigate potential misuse of the model. Future works addressing these issues should make for a more reliable and robust system.

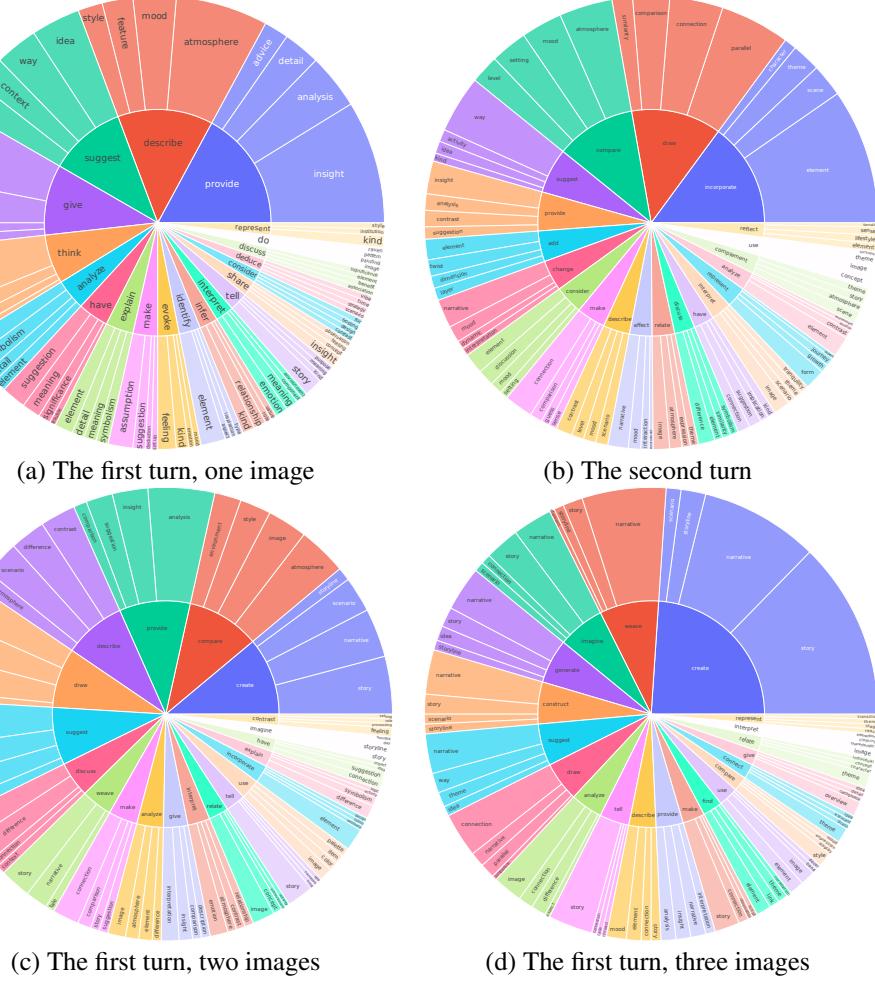


Figure 11: Root verb-noun distributions of SparklesDialogueCC.

I GPT-assisted SparklesDialogue generation

I.1 Single dialogue generation for SparklesDialogueVG

For SparklesDialogueVG, we generate one two-turn dialogue at a time, with the first turn incorporating two or three images. We derive the demonstration dialogues from SparklesDialogueCC to encourage diversity. However, to minimize redundancy, we retain only those dialogues with unique verb-noun combinations in the user questions. This results in pools of 661 and 441 demonstration dialogues for conversations incorporating two or three images in the first turn, respectively. We pull from an expansive collection of roughly 100,000 image-text pairs for this dataset. We randomly select four candidates each time, and they are not reused by excluding them from future selections.

We first present our designed **prompt** for GPT-assisted Single Dialogue Generation to generate SparklesDialogueVG in Table 8. Then, we show a case of the **Dialogue Demonstration** and **Candidate Image Descriptions** to construct the prompt. Finally, we show the corresponding **generated dialogue** using the example prompt.

Example of dialogue demonstration. We visualize the images corresponding to image IDs in the dialogues in Figure 12 for reference, while these visual images were not sent to GPT-4 for data generation. Note that we abbreviate the message content of the assistant in the second turn as “...” to save space, considering that the previous message contents have provided enough demonstrations.

Table 8: Prompt for GPT-assisted single dialogue generation.

System: You are a helpful assistant.

Users will interact with a conversational assistant that has advanced capabilities of understanding, analyzing, and reasoning about images. This includes discussing a variety of real-world concepts, objects, and entities, generating a range of text materials, seeking advice, guidance, or assistance, and much more.

Below is an illustrative dialogue presented in a JSON format. The dialogue represents a meaningful conversation between a "user" and the "assistant" regarding multiple images. Each "user" message contains an "image_ids" field recording the IDs of newly selected images. The images are referred to in the "content" field as IMAGE#image_id.

```
```json
{Dialogue Demonstration}
```

```

Please note that the user contents in the JSON above may be a counterexample that reveals the content of images and can be answered without looking at the images. Please make sure not to reveal the content of the images or describe the images in the user messages in the conversation that follows.

Please note that the specific "image_ids" and "content" in the JSON above are for illustrative purposes only. The actual candidate images are shown below delimited by triple quotes, each accompanied by an image ID and a caption. Avoid using phrases similar to 'caption' and 'description' in your dialogue as if the user and the assistant have visual capabilities.

```
```json
{Candidate Image Descriptions}
```

```

Each dialogue consists of four messages:

1. A user examines all candidate images, selects **{Number of Images}** highly relevant images, and sends a reasonable and creative message to the assistant.
2. Once the images are provided, the assistant thoroughly perceives and comprehends them, responding with highly helpful and exceptionally detailed answers that provide comprehensive reasoning regarding the visual content of the images.
3. Considering the past dialogue, the user chooses other candidate images for further inquiry. The user should refer to both the newly selected images and those mentioned earlier in the same dialogue.
4. The assistant provides a highly helpful and exceptionally detailed answer providing comprehensive reasoning regarding the visual content of the images.

The following is a dialogue between the user and the assistant, adhering to the given JSON format. Make sure to formulate accurate and diverse "content" that does not follow the illustrative dialogues. And remember to develop the last "content" even though it is shown as "..." in the JSON format provided above.

Dialogue Demonstration

```
[[{ 'role': 'user', 'image_ids': I1,1, 'content': Xq1,1 },
  { 'role': 'assistant', 'content': Xa1,1 },
  { 'role': 'user', 'image_ids': I1,2, 'content': Xq1,2 },
  { 'role': 'assistant', 'content': '...' }]]
```

Candidate Image Descriptions

```
[{ 'image_id': J1, 'caption': C1 },
  .....
  { 'image_id': J4, 'caption': C4 }]
```

Number of Images is "two" or "three".



2675



1081



2679

Figure 12: Reference images corresponding to the image IDs in the demonstration dialogues in section I.I. These images were not sent to GPT-4 for data generation.

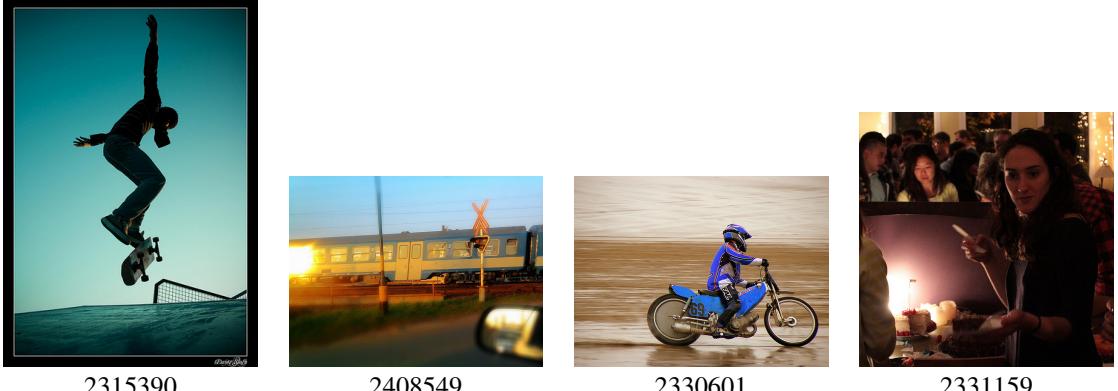
```

1 ````json
2 [
3 [
4   {'role': 'user', 'image_ids': [2675, 1081], 'content': 'How would
      you connect the woman in IMAGE#2675 with the clothing items in
      IMAGE#1081?'},
5   {'role': 'assistant', 'content': 'In IMAGE#2675, we see a woman
      dressed in fashionable attire consisting of a grey turtleneck
      sweater, blue jeans, a black leather belt, a black leather
      jacket with fur detailing and black ankle boots. Her overall
      outfit looks stylish, suggesting she has good fashion sense
      .\\\\\\n\\\\\\nIMAGE#1081 represents a group of clothing items
      hanging on a rack. It shows an outfit comprising a striped
      shirt, a brown skirt, and a pair of black shoes. The clothes
      and shoes have a sophisticated touch to them.\\\\\\n\\\\\\\
      Connecting these images, one could infer that the clothing
      items could belong to the same woman. Maybe she's someone who
      likes keeping her wardrobe organized or could be in the
      fashion industry focusing on outfit planning. Her style in
      IMAGE#2675 is modern and chic, while the ensemble in IMAGE#108
      1 is equally trendy, with classic combinations of stripes and
      brown, which many consider a style staple, and the elegant
      black shoes. Therefore, we could imagine that the outfit from
      IMAGE#1081 may be another stylish option that she might opt
      for. The clothing items could be a part of her personal style
      or they might signal her work attire, given the formal vibe
      the striped shirt and brown skirt ensemble radiate.'},
6   {'role': 'user', 'image_ids': [2679], 'content': 'Looking at IMAGE
      #2679, how would you weave it into the narrative of this
      stylish woman from IMAGE#2675 and the clothes in IMAGE#1081?'}
7   ,
8   {'role': 'assistant', 'content': '...'}
9 ]
10 ````
```

Example of candidate image descriptions. The image IDs in the dialogues refer to the images in Figure 13.

```

1 ````json
2 [
3   {'image_id': 2315390, 'caption': 'The image captures a dynamic
      scene at a skate park, specifically a boy mid-air, performing
      a trick on his skateboard. The boy appears to be a bit on the
      dark side, silhouetted against the lighter sky behind him. He'
```



2315390

2408549

2330601

2331159

Figure 13: Candidate images corresponding to the image IDs in the dialogues generation process in section I.1. These images were not sent to GPT-4 for data generation.

s adorned in casual skate attire: a sweater, blue jeans, and athletic shoes. His arms are raised in the air, which adds to the intensity of the trick he's executing. Interestingly, he is not wearing a helmet.\nThe skateboard itself is flipped, and it appears to be on its side. This unusual positioning gives the impression that the skater is performing an intricate and complex trick. The skateboard has multiple wheels, described as black, and is noticeably detailed in the image.\nThere's another person present in the scene, presumably a photographer or a spectator, positioned toward the right corner of the frame. However, this individual is located quite close to the edge, suggesting that they are not the main focus of the photograph.\nThe skate park consists of a grey concrete ramp that the boy is using for his tricks. It stretches across the majority of the bottom part of the picture, a hard, flat contrast to the dynamic motion taking place above it. There's also a metal gate visible in the scene, possibly part of the boundary or safety measures at the park.\n\nThe backdrop is a vibrant blue sky with clusters of white clouds scattered across it. It seems to be a bright, clear day, perfect for outdoor activities. Lastly, a safety net in the distance lends an additional element of safety to the environment.\nOverall, the photograph encapsulates an exhilarating moment of skill, action, and athleticism at a bustling skate park, set against a serene, blue-skied day.'},

4 {'image_id': 2408549, 'caption': "This image captures a dynamic scene of a large blue train moving rapidly on railroad tracks. The train's hue stands in beautiful contrast with the clear, blue sky overhead. The train is quite long, stretching almost the entire width of the image, and it appears to be well maintained, with grey stripes highlighting its design. The train's lower half is primarily filled with windows and double doors. Three windows are clearly visible, each reflecting the bright sunlight.\nWithin the train, passengers can be seen through the windows. Notably, one person dressed in a white t-shirt is looking out of the window, taking in the scenery or perhaps observing the vehicle whose side mirror is captured in the frame. The double doors, one on the left and the other on the right, stand out on the body of the train. Each door has a number '2' inscribed on it.\nInterestingly, in the right section of the image, the side mirror of a car is in the frame, reflecting a blurry image of another vehicle, further contributing to the sense of movement in this scene.\n\nThe foreground of the image is filled with a wide expanse of green

grass that contrasts nicely with the railroad tracks and a nearby road. To the right of the train, there's a tall pole that rises high into the image, likely used for mounting signs. In this case, the pole hosts a railroad crossing sign with lights and a large X on top. There is also a triangular sign with three lights underneath the X sign, providing important safety information for approaching vehicles.\nBehind the pole, a red metal barrier is barely visible. It appears to be part of the infrastructure that surrounds the tracks. With the beautiful sunny sky overhead, this picture seems to represent a typical day with normal hustle and bustle at this railroad crossing. The sunlight reflecting off the train windows adds a stunning glow to the scene.\nDespite the fast motion of the train, details such as the wheels and even the driver's side view mirror are captured in the image, emphasizing the skill of the photographer in capturing this dynamic and detailed snapshot of a moment in time."},
 5 {'image_id': 2330601, 'caption': 'This image depicts an exciting scene of a man dressed in a blue and black racing suit, riding a dirt bike on a muddy track. The man is prominently positioned in the image, seeming to occupy a considerable portion of it from left to right. His blue helmet, matching his attire, is clearly visible.\nThe motorcycle he's racing is intricately detailed. Its prominent front and back black wheels kick up wet mud as they tear through the track, while the metallic shimmer of the exhaust and the sturdy grey frame suggest its rugged durability. A number, black in color, stands out on the side of the bike, and there's a patch of blue at the bike's back that contributes to the cohesive color scheme.\nThe rider's attire stands out as well. Apart from the matching helmet, he's wearing a blue and orange shirt, black pants, and blue and yellow shoes. A black visor on his helmet and black gloves further accessorize his ensemble. His coat, in shades of blue and grey, fits snugly, outlining his physique.\nThe scene around the bike is as dynamic as the rider. The track underneath is a dark brown, most likely a mix of dirt and water, suggestive of recent rain or the challenging conditions of a dirt bike race. Patches of water and water spots can be seen at various locations, indicating the wetness of the track and the splashing caused by the bike speeding through.\nMoreover, there's an evident sense of motion in the image with water splashing up from the bike and wet sand scattering in its wake. The ground can be seen in patches, displaying its dark brown color. Amidst all this action, the bike stands as a striking subject in the image, catching the eye with its blue frame and detailings, while the rider, dressed in coordinating colors, charges forward.\nAll of these elements combined create an image that is full of life and action, capturing a thrilling moment of a dirt bike race in progress.'},
 6 {'image_id': 2331159, 'caption': 'The image is lively, filled with people gathered possibly for a party or a social event. In the center of the image, a woman dressed elegantly draws attention. She stands prominently, making a distinct statement with her long, dark hair. Her face, sharply defined, features a noticeably distinct nose. She is holding a white plastic spoon in one hand, which also showcases a black wristwatch. As she raises the spoon, it's clear that she is indulging in a delicious treat, a piece of cake resting on a small plate.\nThis cake is a stunning creation, white with red frosting. It's adorned with a delightful mix of red strawberries and an array of white candles. On the cake, there seems to be a flag as well, perhaps signifying a special occasion. An unused serving knife rests in the cake, and it appears that the woman has just served herself a piece.\nAround her, numerous other

```

    faces peer out, all engaged in their individual conversations
. Most of them appear to be men, some notable for their long
hair and glasses. There's an interesting mix of attire in the
scene, from yellow and white-striped shirts to red and black
plaid ones. One Asian man to the right seems to be focused on
the woman with the cake, adding to the collective sense of
attention directed towards her.\nIn the background, several
intriguing details pop out. For instance, the twinkling lights
to the right catch the eye, likely part of the party's
decoration. Additionally, there is a mirror behind the woman,
reflecting the attendees and amplifying the sense of a
bustling crowd. A lamp stands beside a wall, casting a warm
glow, while the silhouettes of patrons in the dark restaurant
form an atmospheric backdrop. A window lets in some additional
light, illuminating a vacant chair.\nThe overall atmosphere
conveys the joyous, friendly nature of the gathering. You can
almost hear the buzz of conversation and feel the warmth of
shared laughter. It's clear that this is an occasion of
happiness and togetherness.'}
7 ]
8 ``-

```

Dialogue example from SparklesDialogueVG. The generated dialogue is visualized in Figure 14. The raw text is shown as follows. The image IDs in the dialogue refer to the images in Figure 13.

```

1 ``- json
2 [
3 [
4   { 'role': 'user', 'image_ids': [2331159, 2330601], 'content': 'Can
      you link the celebration occurring in IMAGE#2331159 and the
      dirt bike race in IMAGE#2330601?' },
5   { 'role': 'assistant', 'content': 'Certainly, in IMAGE#2331159, we
      see a lively gathering with a woman at the centre of the
      attention. She is seen grasping a slice of intricately
      decorated cake from a vibrant party of sorts. Her jovial
      demeanor and elegant apparel suggest she may be the guest of
      honor at this event. The ambiance in the background features
      warm lighting, laughter, conversation, and a sense of
      togetherness and happiness. This scene could possibly be a
      celebration of a personal achievement or milestone.\n\nOn
      the other hand, in IMAGE#2330601, a different form of
      excitement and thrill is captured. This showcases a man in a
      blue and black racing suit, performing at a dirt bike race,
      implying a high level of commitment, skill, and passion. He
      is decisively navigating the muddy track, indicating a race
      in progress.\n\nLinking these images, one possible
      narrative could be that the woman in the first image might be
      celebrating a successful day at the dirt bike race. The man
      in the second image who is skilfully navigating the muddy
      race track could be related to her in some way. Perhaps he is
      her son, husband, or even a team member, and they are
      celebrating his win or a significant accomplishment in his
      racing career at the party. The sense of joy and exhilaration
      is a common thread that connects these two images.' },
6   { 'role': 'user', 'image_ids': [2408549], 'content': 'How would
      you incorporate the train scene from IMAGE#2408549 with the
      earlier party and racing narratives?' },
7   { 'role': 'assistant', 'content': 'In IMAGE#2408549, we witness the
      rush of a large, striking blue train speeding through a
      railroad crossing. A passenger, perhaps contemplating the
      journey, watches the world outside through a window. This
      train scene can signify a journey or transition.\n\nBringing these images together into a coherent story, it
      might be speculated that after the celebration in IMAGE#2331159' }

```

Can you link the celebration occurring in [IMAGE#2331159](#)
 and the dirt bike race in [IMAGE#2330601](#)
 ?

Certainly, in [IMAGE#2331159](#), we see a lively gathering with a woman at the centre of the attention. She is seen grasping a slice of intricately decorated cake from a vibrant party of sorts. Her jovial demeanor and elegant apparel suggest she may be the guest of honor at this event. The ambiance in the background features warm lighting, laughter, conversation, and a sense of togetherness and happiness. This scene could possibly be a celebration of a personal achievement or milestone.

On the other hand, in [IMAGE#2330601](#), a different form of excitement and thrill is captured. This showcases a man in a blue and black racing suit, performing at a dirt bike race, implying a high level of commitment, skill, and passion. He is decisively navigating the muddy track, indicating a race in progress.

Linking these images, one possible narrative could be that the woman in the first image might be celebrating a successful day at the dirt bike race. The man in the second image who is skilfully navigating the muddy race track could be related to her in some way. Perhaps he is her son, husband, or even a team member, and they are celebrating his win or a significant accomplishment in his racing career at the party. The sense of joy and exhilaration is a common thread that connects these two images.

How would you incorporate the train scene from [IMAGE#2408549](#) with the earlier party and racing narratives?

In [IMAGE#2408549](#), we witness the rush of a large, striking blue train speeding through a railroad crossing. A passenger, perhaps contemplating the journey, watches the world outside through a window. This train scene can signify a journey or transition.

Bringing these images together into a coherent story, it might be speculated that after the celebration in [IMAGE#2331159](#) for the successful dirt bike race from [IMAGE#2330601](#), the next chapter of their journey is starting. It could be speculated that the woman or the dirt bike racer or both are now on this train in [IMAGE#2408549](#), heading towards a new location for perhaps the next race or a well-deserved vacation post celebration. This train ride could represent their movement towards the next chapter in their journey, literally as well as metaphorically.

The passenger looking out the window could be contemplating the recent event of the dirt bike race victory and the warm celebration that followed. The train's vigorous motion and the reflected sunlight might be symbolising their sense of achievement and their eager anticipation for upcoming adventures. Thus, these images together weave a tale of triumph, celebration, and anticipation for what lies ahead.

Figure 14: Dialogue example from SparklesDialogueVG. Visual images were not provided to GPT-4 during data generation but will be incorporated during SparklesChat training.

```

59 for the successful dirt bike race from IMAGE#2330601, the
next chapter of their journey is starting. It could be
speculated that the woman or the dirt bike racer or both are
now on this train in IMAGE#2408549, heading towards a new
location for perhaps the next race or a well-deserved
vacation post celebration. This train ride could represent
their movement towards the next chapter in their journey,
literally as well as metaphorically.\n\nThe passenger
looking out the window could be contemplating the recent
event of the dirt bike race victory and the warm celebration
that followed. The train's vigorous motion and the reflected
sunlight might be symbolising their sense of achievement and
their eager anticipation for upcoming adventures. Thus, these
images together weave a tale of triumph, celebration, and
anticipation for what lies ahead.'}]

```

I.2 Multiple dialogues generation for SparklesDialogueCC

For SparklesDialogueCC, we prompt GPT-4 to generate three dialogues in a single response. These dialogues incorporate one, two, and three images in the first turn and a single image in the second. Each prompt includes three demonstration dialogues and nine candidate image descriptions to facilitate this. We curated 150 demonstration dialogues, evenly split with 50 dialogues for each type. The complete image-text dataset comprises about 3,500 pairs.

We first present our designed **prompt** for GPT-assisted Multiple Dialogues Generation to generate SparklesDialogueCC in Table 9. Then, we show a case of the **Dialogue Demonstrations** and **Candidate Image Descriptions** to construct the prompt. Finally, we show the corresponding **generated dialogues** using the example prompt.

Example of dialogue demonstrations. We visualize the images corresponding to image IDs in the dialogues in Figure 15 for reference, while these visual images were not sent to GPT-4 for data generation. Note that we abbreviate the message content of the assistant in the second turn as “...” to save space, considering that the previous message contents have provided enough demonstrations.



3775



4799



301



3122



204



3045

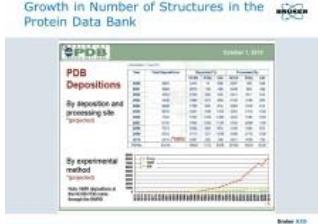


Table 9: Prompt for GPT-assisted multiple dialogues generation.

Users will interact with a conversational assistant that has advanced capabilities of understanding, analyzing, and reasoning about images. This includes discussing a variety of real-world concepts, objects, and entities, generating a range of text materials, seeking advice, guidance, or assistance, and much more.

Below are three illustrative dialogues presented in a JSON format. Each one represents a self-contained conversation between a "user" and the "assistant" regarding multiple images. Each "user" message contains an "image_ids" field recording the IDs of newly selected images. The images are referred to in the "content" field as IMAGE#image_id.

```
```json
{Dialogue Demonstrations}
```

```

Please note that the specific "image_ids" and "content" in the JSON above are for illustrative purposes only. The actual candidate images are shown below delimited by triple quotes, each accompanied by an image ID and a caption. Avoid using phrases similar to 'caption' and 'description' in your dialogue as if the user and the assistant have visual capabilities.

```
```json
{Candidate Image Descriptions}
```

```

Each dialogue consists of four messages:

1. A user examines all candidate images, selects highly relevant ones, and sends a reasonable and creative message to the assistant.
2. Once the images are provided, the assistant thoroughly perceives and comprehends them, responding with highly helpful and exceptionally detailed answers that provide comprehensive reasoning.
3. Considering the past dialogue, the user chooses another candidate image for further inquiry. The user should refer to both the newly selected image and those mentioned earlier in the same dialogue.
4. The assistant provides a highly helpful and exceptionally detailed answer providing comprehensive reasoning regarding the visual content of the images.

The following are three independent dialogues between the user and the assistant, adhering to the given JSON format. In this format, the first message in the three dialogues includes 1, 2, and 3 image IDs respectively.

Make sure to formulate accurate and diverse "content" that does not strictly follow the illustrative dialogues. And remember to develop the last "content" even though it is shown as "..." in the JSON format provided above.

Dialogue Demonstrations

```
[[{"role": "user", "image_ids": "I1,1", "content": "Xq1,1"},
 {"role": "assistant", "content": "Xa1,1"},
 {"role": "user", "image_ids": "I1,2", "content": "Xq1,2"},
 {"role": "assistant", "content": "..."}],
 . . .
 [[{"role": "user", "image_ids": "I3,1", "content": "Xq3,1"},
 {"role": "assistant", "content": "Xa3,1"},
 {"role": "user", "image_ids": "I3,2", "content": "Xq3,2"},
 {"role": "assistant", "content": "..."}]]
```

Candidate Image Descriptions

```
[{"image_id": "J1", "caption": "C1"},
 . . .
 {"image_id": "J9", "caption": "C9}]
```

```

3  [
4    {'role': 'user', 'image_ids': ['3775'], 'content': 'What kind of
      ink is usually used for such tattoos that we see in IMAGE#3775
      , and how long can we expect it to last?'},
5    {'role': 'assistant', 'content': 'Tattoo artists typically use a
      specific type of ink known as tattoo ink for their designs,
      like the one shown in IMAGE#3775. Tattoo ink is usually made
      from metal salts, plastic-based pigments, or heavy metals for
      creating different colors, although black ink, like the one
      used in the "trust yourself" tattoo in the image, is generally
      carbon-based.\n\nHow long the tattoo lasts in its crisp form
      depends on a series of factors such as the depth of the ink
      placement, the type of ink used, how well it\'s cared for
      during the healing process, and how it\'s protected from the
      sun and aging. \n\nGenerally, tattoos are permanent because
      the human body\'s immune system isn\'t capable of dealing with
      the large molecules of tattoo ink. Over several years, a
      tattoo can fade slightly due to the natural regeneration cycle
      of skin cells, but it won\'t disappear completely. The good
      thing about black tattoos, as seen in IMAGE#3775, is that
      black pigment is absorbed well by all skin types and tends to
      last longer without fading. However, the specific longevity of
      a tattoo will differ from person to person. If one wants
      their tattoo\'s health to be maintained, they should keep it
      moisturized, use sunscreen when out in harsh sun, and consider
      touch-ups from a professional tattoo artist when necessary.'}
6    ,
7    {'role': 'user', 'image_ids': ['4799'], 'content': 'Can the ink
      used in the tattoo from IMAGE#3775 potentially cause any skin
      irritations or allergies, similar to how some people have
      reactions to certain foods like in IMAGE#4799?'},
8    {'role': 'assistant', 'content': '...'}
9  ],
10 [
11   {'role': 'user', 'image_ids': ['301', '3122'], 'content': 'What
      kind of environment is depicted in IMAGE#301 and how does it
      contrast with the one in IMAGE#3122?'},
12   {'role': 'assistant', 'content': "IMAGE#301 depicts a tropical,
      beach environment. In the image, there are two kayaks resting
      on a sandy beach, and this detail hints at a location that's
      likely in a warmer climate, suitable for water sports and
      recreational activities. The presence of palm trees further
      supports this, as they are commonly found in regions that are
      hot and sunny most of the year. The blue sky and the small
      building that looks like a beach hut or kiosk suggest a
      relaxed, leisure-oriented environment.\n\nOn the contrary,
      IMAGE#3122 portrays a snowy, mountainous environment. The
      presence of snow suggests a colder climate, likely during
      winter. The pair of black rubber snowshoes with metal spikes
      on the bottom implies activities related to snow, such as
      hiking or mountain climbing. The person wearing a hooded
      jacket and grey pants with their hands in their pockets
      further indicates the cold nature of this environment. The
      trees and mountains visible through the snow complete the
      impression of a wilderness setting, possibly remote and less
      populated.\n\nIn summary, the two images contrast
      significantly - one depicts a warm, tropical beach setting
      that is likely oriented towards leisure and relaxation, while
      the other depicts a colder, mountainous setting associated
      with more strenuous outdoor activities. This contrast goes
      beyond just the climatic conditions; it also extends to the
      kind of activities and lifestyle one might associate with each
      setting."},

```

```

12     {'role': 'user', 'image_ids': ['204'], 'content': 'Does the scene
13         in IMAGE#204 remind you of the environment depicted in IMAGE#3
14         01, or are they entirely different?'},
15     {'role': 'assistant', 'content': '...'}
16 ],
17 [
18     {'role': 'user', 'image_ids': ['3045', '1863', '1336'], 'content':
19         'Generate a promotional text that could fit for a sports
20         event based on IMAGE#3045. Also, please include elements from
21         IMAGE#1863 and IMAGE#1336 to give it some unique angles.'},
22     {'role': 'assistant', 'content': "Join us for an electrifying
23         display of sportsmanship like no other! Just as the fans in
24         IMAGE#3045, cheer on your favorite team in an outdoor stadium
25         filled with a roaring crowd, all decked out in their team's
26         gear, waving pink flags in unity and excitement. Experience an
27         atmosphere buzzing with energy beneath the open sky,
28         surrounded by the vibrant greenery.\n\nDrawing inspiration
29         from IMAGE#1863, watch as the team's performance graph soars
30         high, much like the increasing deposits in a bank account.
31         Witness the thrilling trajectory of action unfold, as each
32         second, each decision, contributes to an upward trend in their
33         game. Be a part of this journey, as they race not just to
34         maintain the pace but to set a new norm, akin to the
35         decreasing withdrawals mark, showcasing improvement and growth
36         over time.\n\nIn the spirit of IMAGE#1336, let your hearts
37         beat as one, with the passion of the red and the harmony of
38         the black and white. Let the dual celebrations of vibrant love
39         for the game and steadfast unity resound in the air. Let the
40         duality of a heartbeat stir a symphony of balance - just like
41         the yin and yang of competitiveness and sportsmanship.\n\nSo,
42         let's delve into this exciting world, where not just the
43         players, but even the spectators embark on a journey of
44         passion, unity, growth, and harmony. Let the games begin!"},
45     {'role': 'user', 'image_ids': ['3302'], 'content': 'Could the
46         atmosphere and energy captured in IMAGE#3045 somehow be
47         invoked in the office space depicted in IMAGE#3302 to make it
48         more lively? Any suggestions?'},
49     {'role': 'assistant', 'content': '...'}
50 ]
51 ]
52 ```

```

Example of candidate image descriptions. An example of Candidate Image Descriptions is shown below, and their corresponding source images are shown in Figure 16 for reference (they are not sent to GPT-4).

```

1 ````json
2 [
3     {'image_id': '2439', 'caption': 'This image shows a kitchen with
4         wooden cabinets, black countertops, and white appliances. The
5         floor is made of tiles and the walls are painted white. There
6         is a large window above the sink that lets in plenty of
7         natural light. The room is spacious and well lit.'},
8     {'image_id': '3065', 'caption': "This is an image of an airplane
9         flying in the sky at sunset. The plane is a large, commercial
10        jet with a white body and red and blue stripes on the tail. It
11        is flying low in the sky, with the sun setting behind it,
12        casting a warm orange glow on the left side of the image and a
13        blue glow on the right. The plane's engines are visible at
14        the bottom of the image, with smoke coming from them. The sky
15        is a deep blue, with clouds in the distance that are tinged
16        with pink from the sunset."},
17     {'image_id': '1093', 'caption': 'The image shows a small room with
18         a wooden shelf on the wall, several rolls of wrapping paper
19         '
20         }
21         ]
22         ```

```

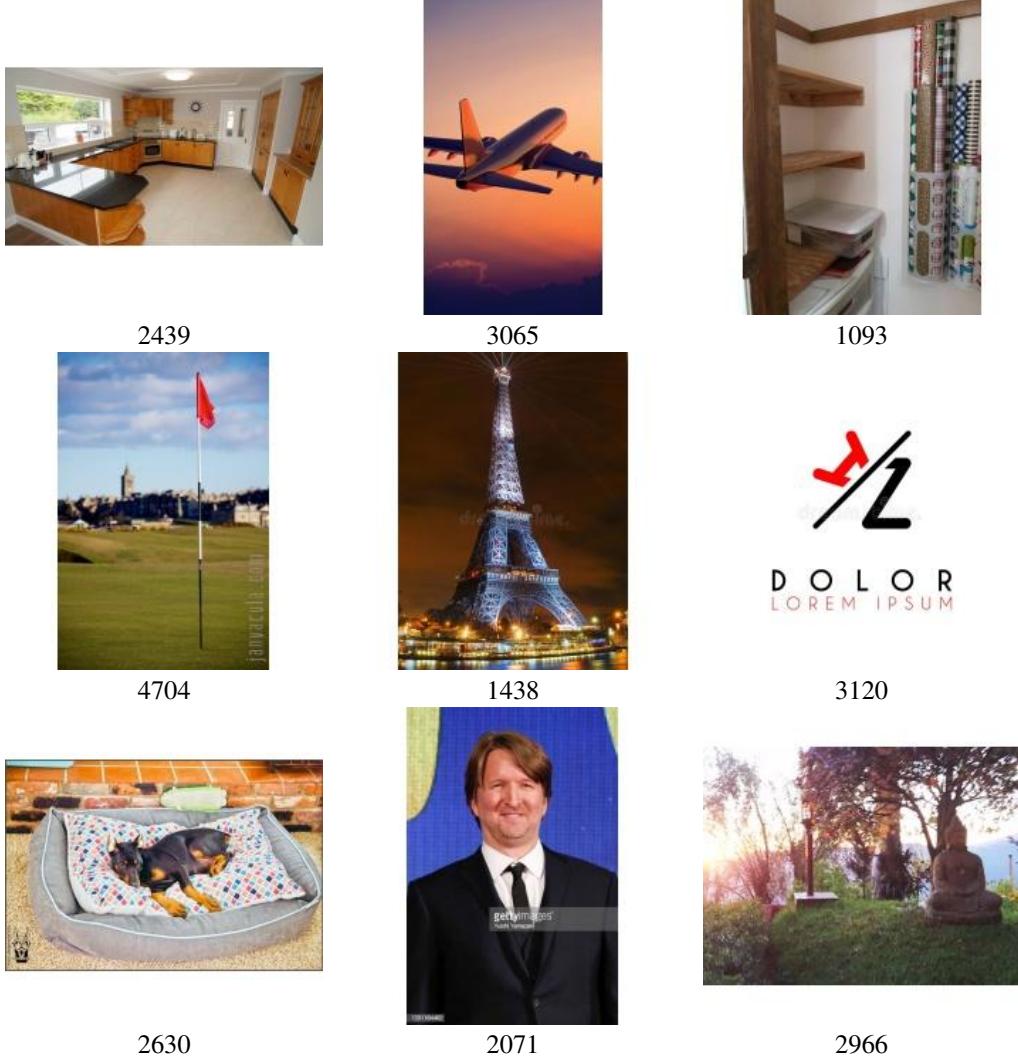


Figure 16: Candidate images corresponding to the image IDs in the dialogues generation process in section I.2. These images were not sent to GPT-4 for data generation.

```

stacked on it, a door on the right side, and a window on the
left side. The walls are painted white and there is a wooden
floor.'},
6   {'image_id': '4704', 'caption': 'The image shows a view of a golf
course with a red flag on the green. In the background, there
is a city skyline with buildings and a church steeple. The
grass on the course is lush and green, and there are trees on
either side of the fairway. The sky is clear and blue, and
there are a few clouds in the distance. The flag on the green
is a small, red flag with a white pole. It is standing upright
in the middle of the green, and it looks like it is blowing
in the wind. The city skyline in the background is quite
impressive, with several tall buildings and a church steeple
visible. The church steeple is quite tall and has a pointed
top.'},
7   {'image_id': '1438', 'caption': 'The Eiffel Tower is a famous
landmark in Paris, France. It is a wrought iron lattice tower
that was built in 1889 to commemorate the centenary of the
French Revolution. The tower stands 324 meters tall and is
located on the Champ de Mars in the heart of Paris. It is one
of the most recognizable landmarks in the world.'}

```

```

of the most visited tourist attractions in the world, with
millions of people visiting it every year. The tower has
become an iconic symbol of Paris and France. The tower is
painted in blue, white, and red, the colors of the French flag
. The lights on the tower create a beautiful and magical
atmosphere, making it a popular spot for romantic walks and
photos. The tower is surrounded by water, with a river running
underneath it.'},
8 {'image_id': '3120', 'caption': 'This image is a logo for a
company or brand. The logo consists of the letters "z" and "l"
in a red and black color scheme. The letters are connected by
a diagonal line through the center of the image, creating a
simple and modern design suitable for various businesses and
industries.'},
9 {'image_id': '2630', 'caption': "The image shows a black and tan
dog lying on a large, gray dog bed. The bed has a geometric
pattern on it and is placed in front of a brick fireplace. The
dog's head is resting on its paws, and it appears to be
sleeping. The image is well lit and the details of the dog and
the bed are clear."},
10 {'image_id': '2071', 'caption': "The image shows a man in a black
suit and tie standing in front of a blue and yellow background
. The man is smiling and has his hands in his pockets.\n\nThe
man is wearing a black suit and tie, with a white shirt
underneath. He has short, dark hair and a well-groomed beard.
He is standing in front of a large blue and yellow background
that looks like a cartoon character's face. The character's
eyes and mouth are wide open, and it looks like the man is
standing in front of the character's face.\n\nThe background
looks like a cartoon character's face, with blue and yellow
colors. The character's eyes and mouth are wide open, and it
looks like the man is standing in front of the character's
face. The character's eyes and mouth are wide open, and it
looks like the man is standing in front of the character's
face."},
11 {'image_id': '2966', 'caption': 'The image shows a garden with a
stone statue of a person sitting under a tree. The statue has
a serene expression on its face, and the sun is setting behind
the trees, casting a warm orange light on the scene. The
foreground grass is green and well-maintained, and there are
no other objects or people in the image.'}
12 ]
13 ...

```

Dialogue examples from SparklesDialogueCC. The generated dialogue is visualized in Figure 17. The raw text is shown as follows. The image IDs in the dialogues refer to the images in Figure 16.

```

1 ````json
2 [
3 [
4   {'role': 'user', 'image_ids': ['3120'], 'content': 'Could you
      tell me what the design of the logo in IMAGE#3120 might
      symbolize or represent?'},
5   {'role': 'assistant', 'content': 'The logo in IMAGE#3120, with the
      combined letters "z" and "l" connected by a diagonal line,
      seems to convey simplicity, modernity, and approachability.
      The use of "z" and "l" could represent the initials of a
      company or individual's name, while the diagonal line may
      symbolize a link or connection, emphasizing unity or
      collaboration. The choice of red and black colors could be
      conveying a sense of power and sleek sophistication,
      respectively. Black often symbolizes solidity and formality,
      while red can represent passion and energy. However, the
      specific meaning could vary as logo interpretations often

```

Generate a promotional text that could fit for a sports event based on [IMAGE#3045](#)



. Also, please include elements from [IMAGE#1863](#) and [IMAGE#1336](#)




to give it some unique angles.

Join us for an electrifying display of sportsmanship like no other! Just as the fans in [IMAGE#3045](#), cheer on your favorite team in an outdoor stadium filled with a roaring crowd, all decked out in their team's gear, waving pink flags in unity and excitement. Experience an atmosphere buzzing with energy beneath the open sky, surrounded by the vibrant greenery.

Drawing inspiration from [IMAGE#1863](#), watch as the team's performance graph soars high, much like the increasing deposits in a bank account. Witness the thrilling trajectory of action unfold, as each second, each decision, contributes to an upward trend in their game. Be a part of this journey, as they race not just to maintain the pace but to set a new norm, akin to the decreasing withdrawals mark, showcasing improvement and growth over time.\n\nIn the spirit of [IMAGE#1336](#) let your hearts beat as one, with the passion of the red and the harmony of the black and white. Let the dual celebrations of vibrant love for the game and steadfast unity resound in the air. Let the duality of a heartbeat stir a symphony of balance - just like the yin and yang of competitiveness and sportsmanship.

So, let's delve into this exciting world, where not just the players, but even the spectators embark on a journey of passion, unity, growth, and harmony. Let the games begin!

Could the atmosphere and energy captured in [IMAGE#3045](#) somehow be invoked in the office space depicted in [IMAGE#3302](#) to make it more lively? Any suggestions?

Absolutely! The outdoor stadium setting of [IMAGE#3045](#) exudes an energy and unity that can certainly be incorporated into [IMAGE#3302](#), not in a literal sense of course, but more so in creating an ambiance and a sense of togetherness and excitement in the workspace.

For starters, we can draw on the crowd energy and unity depicted in [IMAGE#3045](#). Adding some group oriented spaces in the office, like a collaboration zone with comfortable seating, can be a great way of fostering teamwork and unity - much like the fans watching a game together. A big screen showing motivational or industry related dynamic content like live news, updates, or even sports in this area could spark excitement and conversations, stimulating a dynamic work environment.

The greenery in the stadium can be mirrored in the office by including indoor plants. This can bring a touch of nature inside, invoking a sense of calm yet vibrant energy, making the space inviting and friendly.

[IMAGE#3045](#) has fans displaying their team's colors in form of flags and jerseys. Similarly, using brand colors in the office space can evoke a sense of pride and connection to the company's mission. This can even be extended by having artwork or banners that tell the company story or milestones, resembling the signs and banners held up by fans in the image.

From [IMAGE#1863](#), the upward trend of the graph suggests progress and growth. This theme can be maintained in the office by showcasing company progress visuals, awards, or integrated performance dashboards to highlight team victories and achievements, creating a sense of accomplishment among employees.

Taking cues from [IMAGE#1336](#), the office interior can imply balance and harmony. The red heart corresponds to spaces or elements designed for enthusiastic brainstorming and vibrant energy, whereas areas symbolizing the black and white heart can be designed for more focused, calm, and harmonious work.

By blending elements from all three images, the office in [IMAGE#3302](#) can indeed be transformed into a place that is energetic, lively, motivating, while also comforting and balanced - a true field of productivity.

Figure 17: Dialogue example from SparklesDialogueCC. Visual images were not provided to GPT-4 during data generation but will be incorporated during SparklesChat training.

```

    depend on the organization's mission, brand values, and the
    industry it's part of.'},
6   {'role': 'user', 'image_ids': ['1438'], 'content': 'Considering
        the design and message one might infer from the logo in IMAGE
        #3120, would it be a good fit for the place depicted in IMAGE
        #1438? How would they complement each other?'},
7   {'role': 'assistant', 'content': 'The logo from IMAGE#3120 can
        bring a modern, energetic, and connected vibe to the

```

traditional and iconic setting of the Eiffel Tower in IMAGE#1
 438. This can create an interesting juxtaposition between
 historic and contemporary themes.\n\nThe simplicity and
 modernity of the logo design would sharply contrast yet
 harmonize with the complexity and intricacy of the Eiffel
 Tower. This could appeal to audiences who appreciate both the
 historical and futuristic aspects of a brand. Similarly, the
 vibrancy and energy indicated by the red color in the logo
 may align well with the romantic and lively atmosphere
 associated with Paris, the city where Eiffel Tower is
 situated.\n\nIf this was for a brand that intended to evoke a
 sense of modern connection in a place steeped with history,
 or perhaps a company that combines technology with tourism in
 Paris, this logo could indeed complement the place
 effectively.]],
 8 [
 9 { 'role': 'user', 'image_ids': ['2630', '3065'], 'content': 'Can
 you imagine a story that might link the peaceful dog lounging
 in IMAGE#2630 and the airplane soaring in the sunset sky of
 IMAGE#3065?' },
 10 { 'role': 'assistant', 'content': 'Certainly! Here\'s a story
 possibility: The dog in IMAGE#2630, named Boomer, belongs to
 a dedicated pilot who often takes on long haul flights, like
 the airplane depicted in IMAGE#3065. Whenever the pilot sets
 off for a lengthy trip, Boomer retreats to his comfortable
 dog bed in front of the cozy fireplace, waiting for his owner
 \'s return.\n\nThe pilot, while often away, never fails to
 think about Boomer. From high above the earth, amidst
 beautiful sunsets, the canine\'s owner counts the hours until
 they can be reunited. Just the thought of Boomer\'s serene
 expression on the tastefully patterned dog bed provides a
 sense of peace and comfort even thousands of miles away.\n\n
 So despite their geographical separation depicted in these
 images, there\'s an underlying connection of loyalty and
 companionship between Boomer and his pilot owner, which is a
 heartwarming staple in their world filled with frequent
 departures and heartfelt reunions.' },
 11 { 'role': 'user', 'image_ids': ['4704'], 'content': 'How would the
 atmosphere in IMAGE#4704 contribute to the story of our
 pilot-owner and Boomer from IMAGE#2630 and IMAGE#3065?' },
 12 { 'role': 'assistant', 'content': 'The setting in IMAGE#4704 could
 introduce another layer to our narrative. Imagine this: When
 off-duty, our pilot-owner is also an avid golfer who
 cherishes the peacefulness and serenity found on the golf
 course, like the one seen in IMAGE#4704. After returning from
 a long trip, the pilot usually heads to the fairway, and
 Boomer eagerly accompanies him.\n\nThe manicured greens serve
 as a perfect place for Boomer to roam freely, chase golf
 balls, and feel the wind in his fur. The distant city skyline
 seen from the golf course acts as a reminder of the urban
 life our pilot leaves behind when they find solace in the
 peaceful embrace of the game and their cherished canine
 companion.\n\nSubsequently, the lush green golf course fills
 in a missing piece of the puzzle, acting as a mediating space
 that connects the frequent travels of the pilot and Boomer\'
 s waiting. It becomes a place of joy, reconnection, and
 shared experiences between the duo, marking the happy ending
 to their waiting and the start of their time together until
 the next flight.' },
 13 [
 14 { 'role': 'user', 'image_ids': ['2439', '1093', '2071'], 'content':
 'Could you help me design a scenario the place in IMAGE#1093
 is being used as a temporary office for the man in IMAGE#2071,
 who usually works in the kitchen in IMAGE#2439?' },

```

15  {'role': 'assistant', 'content': 'Sure, here\'s a possible scenario
   : In IMAGE#2071, we see a man who typically works from his
   spacious kitchen shown in IMAGE#2439. The kitchen, with its
   wooden cabinets and black countertops, doubles as his office.
   Its large windows provide ample natural light, creating a
   conducive and productive work environment. It is here that he
   plans his strategies, makes important business decisions, and
   leads virtual meetings.\n\nHowever, for a change of scenario -
   perhaps due to a home renovation or family visiting - the man
   decides to shift his workspace to the small room shown in
   IMAGE#1093 for a while. The room is likely a storage space or
   craft room with wrapping paper stacked on a wooden shelf,
   indicating creativity and resourcefulness.\n\nHe sets up his
   temporary office here, positioning his desk near the window to
   replicate the natural light he enjoys in his kitchen-office.
   The surrounding array of vibrant wrapping paper inspires
   creativity and brings a touch of liveliness to his work
   routine. Also, the closed door ensures that he has quiet and
   privacy for his business calls.\n\nSo in this unusual but
   charming workspace, our businessman continues his work,
   exhibiting that all you need to be productive is a bit of
   adaptability, irrespective of where you are.'},
16  {'role': 'user', 'image_ids': ['2966'], 'content': 'Taking into
   account all the previous images, how could we incorporate the
   scene from IMAGE#2966 into the story of the man changing his
   workspace?'},
17  {'role': 'assistant', 'content': 'The serene garden scene in IMAGE#
   2966 could be an excellent addition to the narrative. Let's
   say, towards the back of the house where both the kitchen-
   office and the temporary workspace reside, there's this quiet
   garden depicted in the image, featuring a stone statue sitting
   under a tree.\n\nAt the end of each workday, the man enjoys
   taking a stroll in the garden to unwind. The setting sun casts
   a warm orange light on the scene creating a peaceful
   atmosphere, and he often finds himself sitting next to the
   stone statue, reflecting on his day. The tranquility of the
   garden effectively helps him transition from work mode, and it
   provides him with a sense of balance that complements his
   adaptable work life.\n\nTherefore, despite the change in his
   workspace, the man maintains his evening ritual in the garden,
   capturing the idea that while our surroundings may change,
   there's comfort in keeping certain routines and finding peace
   in nature.'}
18 ]
19 ]
20 ...

```
