

Enhanced Visual Instruction Tuning with Synthesized Image-Dialogue Data

Yanda Li^{1,2*}, Chi Zhang^{2,6* †}, Gang Yu^{2‡}, Wanqi Yang¹, Zhibin Wang², Bin Fu²,
Guosheng Lin³, Chunhua Shen⁴, Ling Chen¹, Yunchao Wei^{5,7‡}

¹ University of Technology Sydney ² Tencent ³ Nanyang Technological University
⁴ Zhejiang University ⁵ Beijing Jiaotong University ⁶ Westlake University ⁷ Peng Cheng Laboratory
¹liyanda95@gmail.com, ¹wanqi.yang-1@student.uts.edu.au ¹ling.chen@uts.edu.au
²{johnczhang, skicyu, billzbwang, brianfu}@tencent.com
³gslin@ntu.edu.sg ⁴chunhua@me.com ⁵wycho1987@gmail.com ⁶chizhang@westlake.edu.cn

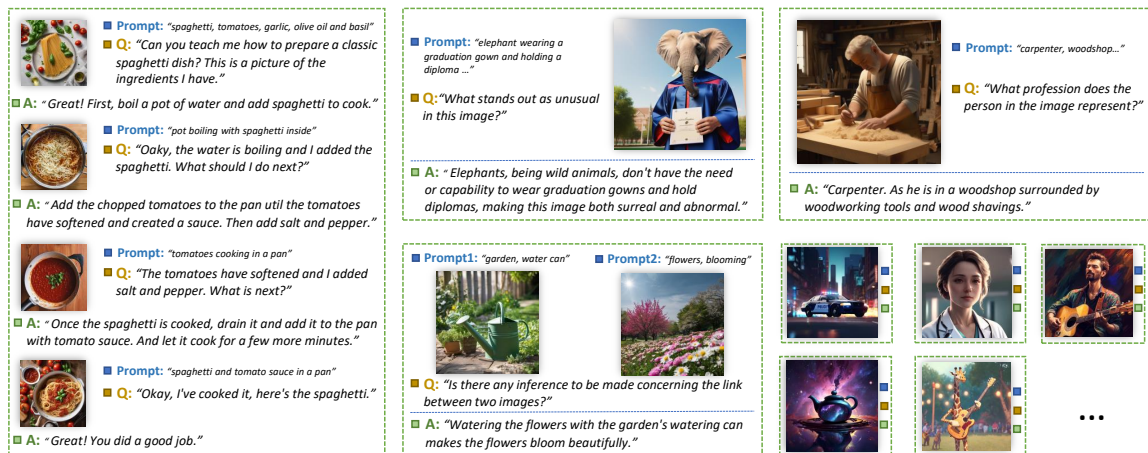


Figure 1: **Examples of synthesized visual instruction data.** We use ChatGPT and text-to-image generation models to synthesize various forms of visual instruction tuning data, such as multi-round dialogue data, multi-image reasoning data, and anomaly detection data. These data are used to train the multimodal large language models.

Abstract

The remarkable multimodal capabilities demonstrated by OpenAI’s GPT-4 have sparked significant interest in the development of multimodal Large Language Models (LLMs). A primary research objective of such models is to align visual and textual modalities effectively while comprehending human instructions. Current methodologies often rely on annotations derived from benchmark datasets to construct image-dialogue datasets for training purposes, akin to instruction tuning in LLMs. However, these datasets often exhibit domain bias, potentially constraining the generative capabilities of the models. In an effort to mitigate these limitations, we propose a novel methodology for data collection, which synchronously synthesizes images and dialogues for visual instruction tuning. This approach leverages the combined capabilities of generative text-to-image models and large language models, facilitating the

creation of a dataset that is both diverse and scalable, and more importantly, customized to enhance the models’ performance across a broad spectrum of tasks. Our research includes comprehensive experiments conducted on various datasets. The results emphasize substantial enhancements in more than ten commonly assessed capabilities. Additionally, our model achieves state-of-the-art results across multiple widely recognized multimodal benchmarks.

1 Introduction

The launch of OpenAI’s ChatGPT (OpenAI, 2023) has marked a significant milestone in artificial intelligence (AI), showcasing the advanced capabilities of Large Language Models (LLMs). These models, exemplified by GPT-4 (OpenAI, 2023), demonstrate exceptional versatility by handling not just images but also excelling in tasks once difficult to accomplish. This includes understanding humor within images and acting as intelligent agents (Zhang et al., 2023a), aspects that highlight its revolutionary potential.

*Equal contributions

†Project Leader

‡Corresponding Author

Project page: <https://github.com/icoz69/StableLLaVA>

However, despite these notable achievements, a crucial aspect remains undisclosed: the specific mechanics underlying GPT-4, particularly concerning the seamless integration of multimodal information into LLMs. This knowledge gap has prompted a concerted research effort to address this puzzle. Among the promising approaches, an emerging method receiving considerable attention involves the utilization of adapter-based techniques (Zhang et al., 2023b; Gao et al., 2023; Luo et al., 2023), which allow the training of a visual-to-text adapter that converts features from pre-trained visual models into LLM tokens, showing promise in achieving results comparable to GPT-4. The effectiveness of adapter-based methods stems from their ability to leverage the extensive pre-existing knowledge in large visual models and LLMs. By focusing on training a lightweight adapter, these methods avoid the computational expense of training comprehensive models from scratch, thereby offering a more efficient pathway to enhancing LLMs’ multimodal integration capabilities.

A prerequisite for implementing these frameworks is the availability of paired vision-text image data. Such datasets are essential for aligning visual and textual information, facilitating the LLMs’ understanding of complex human instructions. Analogous to instruction tuning in LLMs (Wei et al., 2021), this process is commonly referred to as visual instruction tuning. Existing methods (Dai et al., 2023; Liu et al., 2023c; Peng et al., 2023; Liu et al., 2023a; Chung et al., 2022) typically construct visual instruction tuning datasets by leveraging existing datasets, extracting information such as image captions, spatial locations, and categories to form dialogues. This approach maximizes resource utilization, creating a comprehensive and efficient training dataset for multimodal LLMs.

Despite the efficiency and simplicity of this approach to dataset construction, certain limitations still persist. Existing large-scale vision-text datasets, such as LAION (Schuhmann et al., 2022) and CC (Changpinoy et al., 2021), often contain noise. Consequently, training only a subset may inadequately align visual-text features for immediate user requirements. Moreover, benchmark datasets (Changpinoy et al., 2021; Schuhmann et al., 2022; Lin et al., 2014) often exhibit a domain bias, primarily in terms of image styles. For instance, prevalent datasets such as COCO (Lin et al., 2014) predominantly feature images from everyday life, while stylized images like cartoons

are rarely represented. Additionally, these vision annotations may also constrain the types of dialogues generated from them. For example, almost none of the current datasets contain data to directly enhance the model’s ability to comprehend jokes presented in the images, an impressive feature of GPT-4. Moreover, as multi-image dialogues become increasingly integral to the practical application of multimodal LLMs, the current lack of comprehensive multi-image datasets further underscores the need for data enrichment in this domain.

In addressing these limitations, we propose a novel data collection approach to enhance visual instruction tuning. Building upon recent successes in generative AI, we leverage generative models to produce image-dialogue pair data for visual instruction tuning. More concretely, we employ ChatGPT to create data that includes image-generating prompts and content-based dialogues. We then utilize the text-to-image diffusion model, StableDiffusion (Rombach et al., 2022), to generate images based on these prompts. Finally, the synthesized images and generated dialogues are employed to train multimodal LLMs. Simultaneously generating both image content and dialogues enables the production of diverse training data, affording greater control over its nature and quality. This flexibility allows us to construct multi-turn dialogues and datasets involving multi-image reasoning, which are challenging to obtain from other benchmarks.

Additionally, our methodology can potentially integrate more advanced image generative models, such as Dall-E 3 (Betker et al., 2023), to provide higher-level control over image contents like specifying complex spacial relations. This advanced control could generate more complex instructions to enhance image understanding capabilities. Examples from our synthesized visual instruction tuning datasets are shown in Figure 1. *Building upon the flexible pipeline outlined above, users can tailor the generation of data to enhance specific capabilities based on their task requirements. Furthermore, our method of generating both images and dialogues eliminates constraints on data volume, thereby facilitating the production for limitless scaling of the datasets.*

To demonstrate the effectiveness of our proposed pipeline, we conducted extensive experiments. Our main contributions are as threefold:

- We develop a novel pipeline for generating vi-

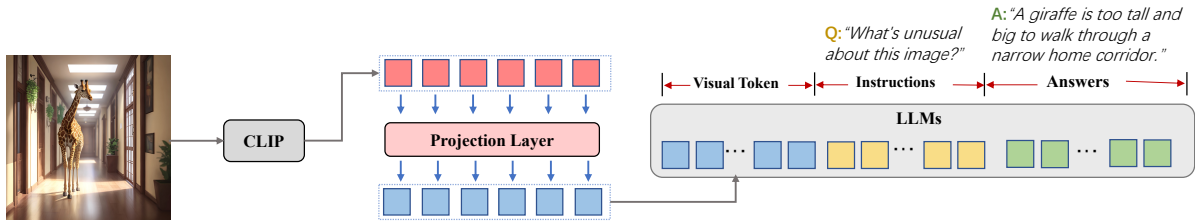


Figure 2: **Architecture of LLaVA.** We use the open-source LLaVA model as a testbed for our proposed data generation pipeline. The model is trained to predict the next tokens in the answers given the visual tokens and instruction tokens in an auto-regressive manner.

sual instruction tuning datasets by leveraging text-to-image diffusion models.

- To showcase its flexibility, we have built a dataset with various form of capabilities including multi-image data, and our results have shown improvements across all abilities.
- Extensive experimental analysis on multiple benchmarks shows the effectiveness of the proposed method, outperforming baseline and existing SOTA approaches.

2 Related Work

Recent research (Zhu et al., 2023; Liu et al., 2023c) efforts in multimodal Large Language Models (LLMs) have yielded promising strategies to efficiently align the embeddings of other modalities with language tokens. This has made it possible to effectively utilize pre-trained encoders from other modalities and LLMs, which effectively reduces the computational burden and training time. While there are alternative research approaches that include training-free methods leverage expert models (Wu et al., 2023; Shen et al., 2023; Yang et al., 2023a), these are not the focus of our work here.

Adapter-based LLMs represent a significant research direction, introducing methods to connect modalities through learnable interfaces with minimal training efforts. These approaches (Zhang et al., 2023b; Gao et al., 2023; Luo et al., 2023; Liu et al., 2023c; Zhu et al., 2023; Yang et al., 2023b) allow for the use of pre-trained modal encoders, reducing the need for training from scratch. Variations include direct training of projection layers for embedding alignment and the use of learnable queries for extracting modality-specific information, as seen in models like LLaVA (Liu et al., 2023c) and Flamingo (Alayrac et al., 2022). Innovations such as the LLaMA-Adapter (Zhang et al.,

2023b) and LaVIN (Luo et al., 2023) have introduced lightweight and mixed-modality adapters, respectively, enhancing the field’s diversity.

Visual instruction tuning datasets are crucial for training multimodal LLMs, focusing on aligning modalities and enabling instruction following. Most existing methods (Liu et al., 2023c; Rotstein et al., 2023; Li et al., 2023a; Yin et al., 2023; Li et al., 2023c; Liu et al., 2023a; Zhang et al., 2023c) rely on benchmark datasets for constructing visual instruction tuning datasets, which may be limited by the fixed categories in annotations. Our method leverages well-trained image generation models to produce controllable image data, enhancing multimodal LLM capabilities and allowing for the integration of advanced generative models for specific guidance forms, offering a more flexible and diverse approach to dataset construction. Recently, the concept of employing synthesized multimodal datasets has been extended to the training of models for chart and video understanding (Han et al., 2023; Song et al., 2024), showcasing its broad applicability and potential in advancing model training across various domains.

3 Preliminary

To assess the effectiveness of our data generation strategy, we chose the open-sourced LLaVA (Liu et al., 2023c,b) as our multimodal LLM model. It should be noted that our pipeline is model-agnostic, making it applicable for various models. This section serves as a foundation, briefly summarizing the LLaVA model’s design and training methods to prepare for a thorough exploration of our pipeline. The reader may refer to the original publication (Liu et al., 2023c) for detail.

Architecture. The LLaVA model integrates Vicuna-13B (Chiang et al., 2023) as the language model with a pre-trained CLIP visual encoder ViT-L/14 (Radford et al., 2021) for extracting visual

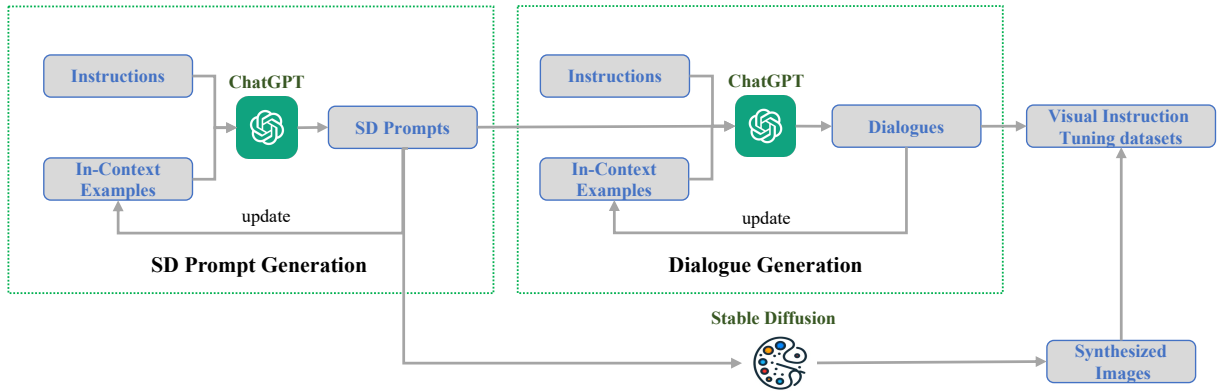


Figure 3: **Our proposed pipeline for generating visual instruction tuning datasets.** We instruct ChatGPT to generate both StableDiffusion prompts and the associated dialogues. For specific generation templates, please refer to the supplementary materials.

features, transforming these features into language embedding tokens through a linear layer. This linear layer was updated in LLaVA-1.5 (Liu et al., 2023b) with a two-layer MLP, replacing Vicuna-13B with Vicuna-13B-v1.5 and increasing input image size to 336x336. A detailed illustration of this model structure can be found in Figure 2.

Training and datasets. LLaVA’s training focuses on visual instruction tuning with data triplets: images, questions, and answers, aiming for predictive accuracy in an autoregressive manner. The training comprises two stages: the first emphasizes modality alignment using 595K image-text pairs, optimizing the linear layer with static visual encoder and LLM weights. The second stage, using 158K multimodal dialogue data from COCO, extends optimization to the LLM’s weights for comprehensive modality integration. LLaVA-1.5 further enriches the dataset by incorporating additional data like Region-level VQA (Kazemzadeh et al., 2014; Krishna et al., 2017; Mao et al., 2016) and GQA (Hudson and Manning, 2019), expanding the second-stage dataset to 665K examples.

4 Methods

This section outlines our dual-generation approach for creating visual instruction tuning datasets, which synthesizes images and their corresponding dialogues, as illustrated in Figure 3. We detail each component below.

4.1 Image Generation

We employ StableDiffusion (Rombach et al., 2022) to generate images based on prompts that include weighted keywords reflecting aspects like subject, scene, style, and visual elements such as image quality and lighting. Keywords at the prompt’s

start are prioritized, with the possibility of adding emphasis using brackets. To encourage diversity and stability during image generation, we add capability-specific instructions and cautions during prompting ChatGPT. For instance, in the task of generating images for joke understanding, we direct ChatGPT to create prompts that would result in the generation of abnormal images, like a “*giraffe walking through a narrow corridor*”, which are unlikely to be found in reality. When generating multi-image data, pairs of prompts can be generated concurrently based on predefined specific criteria. For maximum effect, we ensure that the most crucial keywords are placed at the beginning of the generated prompts, which are double-bracketed for additional emphasis. Furthermore, we instruct ChatGPT to avoid generating prompts that are non-visual, such as the act of growing. The instruction template for prompt generation is provided in the left part of Figure 4. The generated prompts are then used with StableDiffusion to produce visually realistic images, which are subsequently encoded by LLaVA’s vision encoder into visual tokens for LLMs.

4.2 Dialogue Generation

Following the generation of images, we utilize ChatGPT to generate dialogues based on the same prompts used for image synthesis, aligning with LLaVA’s training objectives: the first stage focuses on aligning visual-text data, and the second on processing diverse instructions. Dialogues for the initial stage describe the images, where ChatGPT generates answers to predefined questions about the images’ content. Taking the example of the “*giraffe walking through a narrow corridor*”, a representative dialogue might be: “*Question: What*

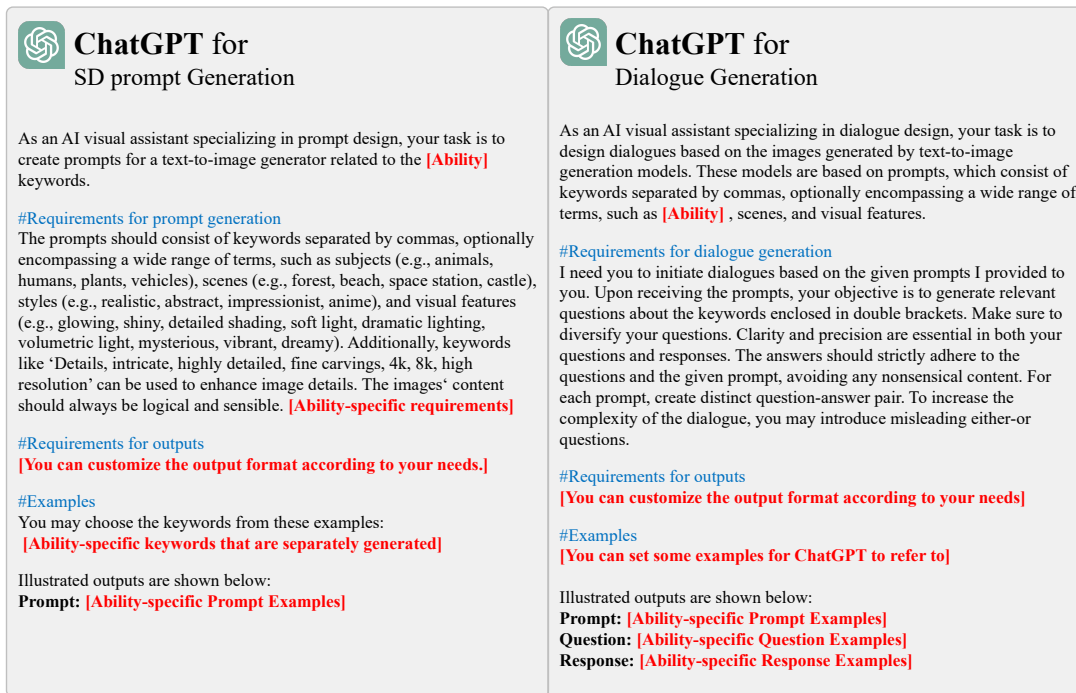


Figure 4: **Templates for guiding ChatGPT to generate StableDiffusion prompts (left) and dialogues (right).** Content in **red** represents ability-specific information. We only provide an example template for constructing dialogues regarding a single image in this figure. For additional forms of data, such as multi-image reasoning and multi-turn dialogues, please refer to our supplementary materials.

is unusual in the image? Response: In reality, a giraffe is too tall and big to walk through a narrow home corridor.” The detailed instruction template for dialogue generation is shown in the right part of Figure 4.

For the second stage, dialogues aim to enhance reasoning across multiple images, addressing similarities, differences, and logical connections, and include multi-turn dialogues that blend image and text. We guide ChatGPT to produce a range of question types, steering clear of inherently ambiguous questions to ensure clarity and accuracy, detailed template can be found in the supplementary materials.

4.3 In-Context Examples

ChatGPT’s in-context learning capability, which allows it to grasp the essence of tasks from a few examples, is leveraged in our methodology. We incorporate in-context examples in the generation of StableDiffusion prompts and dialogues to enhance this learning process. During the data generation process, we observed that ChatGPT sometimes produced a lack of diversity. For example, when generating colors, the outputs frequently revolved around common color categories. To overcome this, we in-

dependently generate ability-related keywords such as color categories with ChatGPT, and utilize them as a reference during the prompting process. This additional step promotes a more diverse range of prompts, thereby enriching our visual instruction tuning dataset.

We further adopt a dynamic strategy to maintain and increase diversity: periodically substituting a portion of the original in-context examples with newly generated data. This continuous update prevents over-repetition and ensures the dataset’s comprehensiveness and representativeness, maintaining a balance that contributes to a richer and more diverse visual instruction tuning dataset.

4.4 Data Filtering Mechanism

To augment the quality and diversity of generated dataset, we introduced a data filtering mechanism. This mechanism encompasses several key aspects:

- **Repetition Rate Filtering** Initially, we filter the generated dataset based on the repetition rate of the extracted prompts. The selected prompts are subsequently utilized to generate corresponding images and dialogues, ensuring novelty and reducing redundancy.
- **Length-based Filtering** We keep prompts

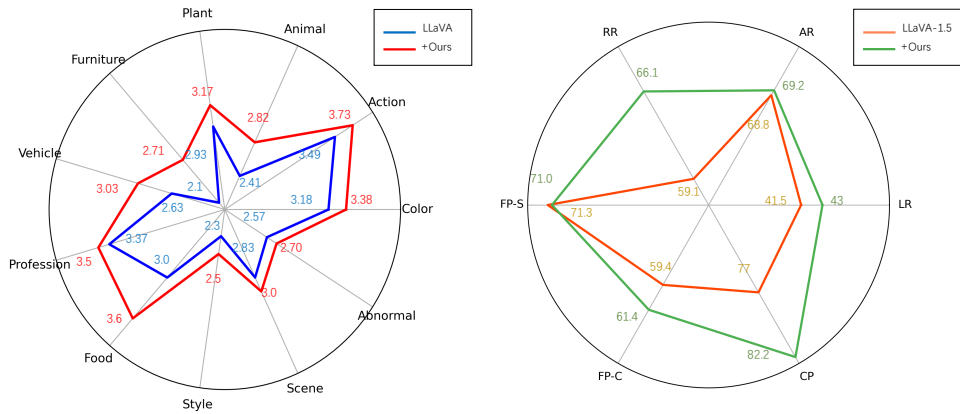


Figure 5: Left: Results on evaluation benchmarks for various abilities (GPT-4 score). Right: Comparison of various subcategories on MMBench (Liu et al., 2023d) with the baseline (Accuracy). Our model outperforms the baselines on both benchmark datasets.

short, with no more than ten keywords, to make sure they’re clear and focused. Long prompts can make it hard for model to create accurate images. We also limit dialogues to 500 characters to keep responses concise.

- **Task-Specific Restrictions** For certain categories, we implemented restrictions based on specific attributes of the capabilities. For example, when generating content related to construction workers, the model tended to focus on buildings. To address this, additional human attributes were incorporated into the prompts to ensure a more accurate prompt.
- **Alignment Check** To ensure a high degree of alignment between generated images and dialogues, we employ the CLIP (Radford et al., 2021) model to compute matching scores for both texts and images. Data entries with scores exceeding a predefined threshold, set at $\gamma = 0.25$, are retained, thereby filtering out less relevant matches and elevating the overall data quality.

5 Experiments

In this section, we detail the experiments conducted to validate the effectiveness of our novel data collection approach for visual instruction tuning. We describe the training datasets, evaluation strategy, and both quantitative and qualitative outcomes.

5.1 Training Datasets

We generate a diverse and expansive dataset to show its versatility, covering single-image capabilities from basic recognition to complex visual reasoning. This includes understanding physical attributes, life features, and man-made items, among

others, amounting to 38K image-dialogue pairs for initial training. Each ability’s dataset was formulated following a standard template, illustrated in supplementary material. In addition, we also generated a dataset of 3K multi-image instances, encompassing descriptions of image similarity, difference, logical relations, and multi-turn dialogue data for the second stage. These datasets, in combination with the raw LLaVA dataset, provides a comprehensive training set in our experiments.

5.2 Evaluation Metrics

Evaluation datasets. To demonstrate our performance more clearly, we tested on a series of public multi-modal datasets, including VisWiz (Gurari et al., 2018), MM-Vet (Yu et al., 2023), MME (Fu et al., 2023), and MMBench (Liu et al., 2023d).

Subsequently, we established a real-image benchmark to evaluate training effectiveness across a wide range of single-image abilities, including 330 test samples of real images with associated question-answer dialogues, carefully selected and annotated from public repositories. This comprehensive benchmark aims to rigorously test the models’ single-image capabilities.

Simultaneously, we constructed a multi-image test set consisting of 30 dialogues to assess the models’ performance on this specific data type. This dataset evaluates the models across differences, similarities, and reasoning relationships among the images. The test data was sourced from publicly available datasets and manually annotated.

Evaluation strategy In terms of the evaluation process, we employ different testing strategies depending on the benchmarks used.

Our evaluation approach varies with the bench-

Table 1: Quantitative performance (Accuracy) on real-image evaluation benchmark for manual evaluation.

Method	Animal	Action	Color	Abnormal	Scene	Style	Food	Profession	Vehicle	Furniture	Plant
LLaVA	0.63	0.67	0.60	0.40	0.60	0.30	0.70	0.57	0.57	0.30	0.53
Ours	0.70	0.90	0.77	0.50	0.77	0.45	0.83	0.63	0.63	0.50	0.57

mark. For multimodal datasets like VizWiz (Gurari et al., 2018) and MMBench (Liu et al., 2023d), we follow official guidelines, converting test data to a compatible format for our model, and using official scripts or submission portals for assessment, primarily focusing on accuracy.

In evaluating the diverse capabilities we’ve generated, We adopted two evaluation methods, namely manual evaluation and evaluation based on GPT-4 score.

Initially, the participants were instructed to assess the answers produced by our model and those of the baseline for all abilities according to the label (1 for correct, 0 for incorrect), averaging these as the final metric.

Subsequently, inspired by (Chiang et al., 2023; Liu et al., 2023c), we leverage GPT-4 (OpenAI, 2023) to assist in scoring model outputs. We have established six scoring levels, ranging from 0 to 5. Each score level is accompanied by detailed descriptions of the evaluation criteria, and we assist GPT-4 in better assessment by providing a series of scoring examples. In particular, a score of 0 indicates that the predicted answer has no relevance to the reference answer, while a score of 5 signifies that the predicted answer aligns seamlessly with the annotated reference answer without any deviation. Drawing on our manual annotations, we conduct evaluations on the results produced by each model. The average GPT-4 score within each test set serves as the ultimate metric for our benchmark evaluations. The specific instruction template used for result evaluation is illustrated in supplementary material.

5.3 Quantitative Comparison with State-of-the-arts

Public multimodal benchmarks We perform quantitative performance comparisons against various state-of-the-art methods on different benchmarks, as illustrated in Table 2. Utilizing LLaVA-1.5-13B as the baseline, we integrate our synthesized data with its original dataset for training. Training is carried out with identical parameter configurations as LLaVA-1.5. The outcomes demonstrate substantial improvements on many benchmarks, emphasizing the enhanced performance

Table 2: Quantitative comparison with other state-of-the-arts methods on multiple multimodal benchmarks. We achieve state-of-the-art performance on four benchmarks.

Method	VisWiz	MM-Vet	MME	MMB
BLIP2 (Li et al., 2023b)	19.6	22.4	1293.8	-
InstructBLIP (Dai et al., 2023)	33.4	25.6	1212.8	-
IDEFICS-9B (IDEFICS, 2023)	35.5	-	-	48.2
IDEFICS-80B	36.0	-	-	54.5
Qwen-VL (Bai et al., 2023)	35.2	-	-	38.2
Qwen-VL-Chat	38.9	-	1487.5	60.6
LLaVA-1.5 (Liu et al., 2023b)	53.6	35.4	1531.3	67.7
Ours	58.4	36.1	1532.3	69.4

Table 3: Quantitative results (GPT-4 score) on the multi-image benchmark. After the addition of multi-image data, various multi-image capabilities have significantly improved.

Method	Difference	Similarity	Logical relations	Average
LLaVA	2.7	2.2	3.1	2.67
Ours	3.6	2.8	3.7	3.37

achieved by our approach.

Comparison of various abilities. To validate the effectiveness of our generated data, we conducted comprehensive tests on distinct capabilities, employing both manual evaluation and GPT-4 score assessments. Employing LLaVA-13B as our baseline, the quantitative comparison of the baseline results and ours for manual evaluation are shown in Table 1, while the GPT-4 score assessment results are shown in the left part of Figure 5. Notably, our trained model consistently outperforms the LLaVA-13B baseline across all various capabilities on two metrics, which suggests the synthesized datasets’ generalizability and our pipeline’s robustness.

Besides, we conduct a comparison of subcategory performance on MMBench to better validate our superiority, using the LLaVA-1.5-13B as the baseline. The tested subcategories in MMBench encompass six aspects: attribute reasoning (AR), coarse perception (CP), fine-grained perception (cross-instance) (FP-C), fine-grained perception (instance-level) (FP-S), logic reasoning (LR), and relation reasoning (RR). The final results are shown in the right part of Figure 5, indicating better performance of subcategory on MMBench, which also attests to the high quality of our generated data.

Comparison on multi-image benchmark In order to validate the effectiveness of multi-image

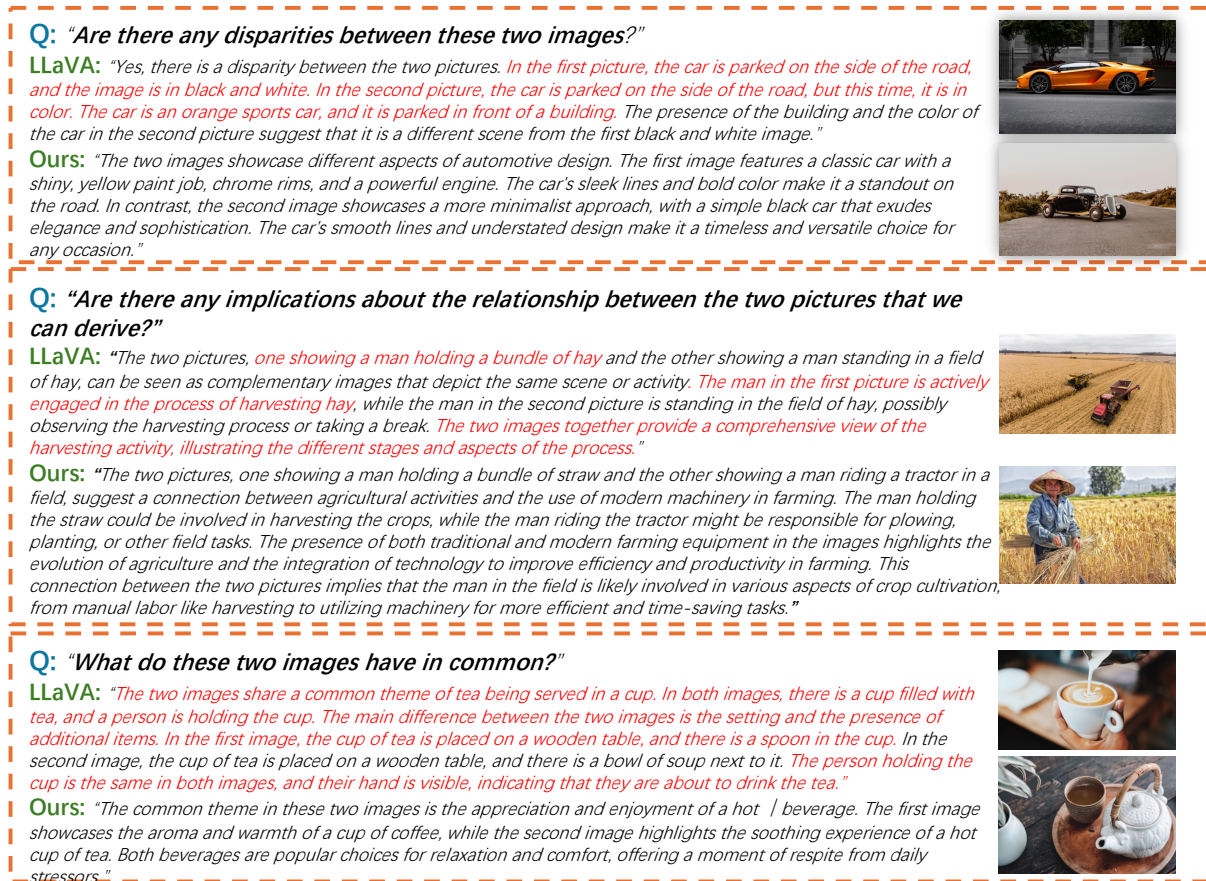


Figure 6: Comparison of the results generated by LLaVA and our trained model. Content in red represents inaccurate information. Our model can better adhere to question instructions, rendering more precise answers.

capabilities, we manually curated a benchmark of real images. The evaluation metric used was the GPT-4 score mentioned. We used LLaVA-13B as the baseline and incorporated multi-image data in the second training phase. Since LLaVA itself lacks the capability for multi-image input, we modified the testing code for LLaVA to enable it to accept multiple sets of images simultaneously. The comparison with LLaVA results is shown in the Table 3, indicating a notable improvement across various multi-image capabilities despite adding less multi-image data in the process.

5.4 Qualitative Results

Supplementing the quantitative analysis, we provide a qualitative comparison between our model's results and LLaVA-13B in Figure. 6 on multi-image data. Our model exhibits a heightened ability to adhere to question instructions, rendering more precise answers.

We compare our approach with the LLaVA-13B baseline, revealing its limitations: it struggles to differentiate between multi-image contents and

provides incomplete answers to questions. Our method, incorporating multi-image data, enhances the model's understanding of multiple images, demonstrating its effectiveness. Additional qualitative results will be included in the supplementary materials.

6 Conclusion and Future Work

In the rapidly evolving realm of Large Language Models, efficiently integrating multimodal information is a key research focus. In this study, we introduced an innovative data collection method to enhance visual instruction tuning for multimodal models. Compared to existing strategies, our approach uniquely combines image and dialogue generation, effectively addressing limitations found in benchmark datasets. By carefully crafting instruction templates, our method ensures high-quality training data covering a broad range of crucial capabilities for multimodal models and users can generate customized data based on their specific requirements.

Our research opens avenues for exploration.

Moving forward, we aim to leverage advanced generative models to enhance model abilities, including spatial comprehension and fine-grained recognition. With promising results from our dual-generation method, forward-thinking data collection techniques are poised to play a significant role in the future of LLM research.

7 Limitations

Due to constraints in text-to-image models like stable diffusion, generating certain data types, such as text-rich images and tables, is not effective in the current pipeline. We anticipate these constraints will be addressed with ongoing advancements in text-to-image generation techniques.

8 Ethics Statement

Our method leverages generative models to create synthetic images and dialogues. It is imperative to ensure that the generated content does not perpetuate or amplify biases present in existing datasets or societal prejudices. We have implemented data filtering mechanism to minimize the generation of potentially harmful or biased content. However, continuous vigilance and improvement of these filters are necessary as generative models evolve.

The enhanced capabilities of multimodal LLMs, facilitated by our data generation approach, could potentially be misused for creating deceptive or manipulative content. It is crucial to develop and adhere to guidelines that prevent the misuse of such technology, including transparent disclosure of synthetic content's nature and purpose.

9 Acknowledgements

This project is particularly supported by ARC DP210101347 and DP240101349.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. 2023. [Dall-e 3: Bridging vision and language with few-shot image generation](#).

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). 2, 3, 4, 5, 6.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

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

IDEFICS. 2023. [Introducing idefics: An open reproduction of state-of-the-art visual language model](#).

- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023c. M3 it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023*.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.
- OpenAI. 2023. Chatgpt. <https://openai.com/blog/chatgpt/>. 1, 2.
- OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. 2023. Fusecap: Leveraging large language models to fuse visual data into enriched image captions. *arXiv preprint arXiv:2305.17718*.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugging-gpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.
- Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. 2024. MovieLLM: Enhancing long video understanding with ai-generated movies. *arXiv preprint arXiv:2403.01422*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

- Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023a. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*.
- Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2023b. [Exploring diverse in-context configurations for image captioning](#).
- Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023a. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023c. Llavav: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix for “Enhanced Visual Instruction Tuning with Synthesized Image-Dialogue Data”

A Training Details

During the model training phase, we employed the original LLaVA configuration as the foundation for our training process. In both stages, we utilized 8 NVIDIA V100 GPUs. To conserve GPU memory, we employed deepspeed with zero3 during model training, disabling tf32 and opting for fp16. The remaining parameters, including epochs and learning rates, were set according to the original LLaVA configuration. For specific parameter details, please refer to the the original publication.

B Multi-image Generation Template

We have illustrated the instruction template for generating multi-image data in Figure 7. To enhance the diversity of the data, we first use ChatGPT to generate a series of paired prompts. Subsequently, based on these paired prompts, we generate corresponding captions. As shown in the Figure 7, we can generate different types of data by configuring different examples and replacing similarities with difference or logical relations.

C Interleaved multi-turn dialogues Generation Template

We present a template for generating interleaved multi-turn dialogues as Figure 8. Similar to multi-image generation, we randomly generate a series of non-repeating phases to enhance data diversity. Based on these phases, we generate interleaved multi-turn dialogues. The red sections are replaceable. In our generated dialogues, the focus is primarily on instructional content related to the use of everyday objects and recipes. Users can make specific adjustments according to their needs. Finally, the brackets part represents the prompt used for stable diffusion to generate images. During the model training process, the content within the brackets will be replaced with an **image_placeholder**.

D GPT-4 Scoring Criteria

As shown in the Figure 9, we present our detailed GPT-4 scoring system. We have established a scale of 0-5 with six levels of scores, and for each score, we provide detailed evaluation criteria along with specific examples for assessment. Utilizing the

template in the Figure 9, evaluations are conducted for each model, and the average of the results is taken as the final score.

E More qualitative Results.

We showcase more generated data as shown in the Figure 10 11, which includes various capabilities, as well as multi-image dialogues and interleaved multi-turn dialogues, to validate the effectiveness and diversity of our pipeline.

We present additional outcomes of our methodology on more real images in Figure 12. Spanning diverse capabilities like profession recognition, color categorization, or anomaly detection, our approach consistently exhibits efficacy. In particular, our method demonstrates a robust capability in discerning abnormal images.

Simultaneously, we validated our model on multi-turn interactive dialogues, and the results are shown in the Figure 13 14. It can be observed that our model can provide detailed descriptions of each step, which is attributed to the high-quality data we generated.



ChatGPT for Paired prompt Generation

As an AI visual assistant, your responsibility is to create one pair prompt for a text-to-image generator. These models rely on prompts composed of keywords separated by commas, which can optionally cover a wide variety of terms. The core themes of the pair prompts can be varied, including animals, plants, people, vehicles, buildings, colors, and more. Paired prompts can be similar types of animals, plants, furniture, vehicles, buildings, professions, furniture, and same object in different colors etc.#

#Requirements for outputs

The final output should follow a structured format: first, print the first prompt; second, print the second prompt.

#Examples

[You can set some examples for ChatGPT to refer to.]

Illustrated outputs are shown below:

Prompt 1: **[Prompt Examples]**

Prompt 2: **[Prompt Examples]**



ChatGPT for Multi-image Generation

As an AI visual assistant, your responsibility is to create one caption about **similarity** of the given paired prompt for a text-to-image generator. You are required to describe the **similarity** between the two images based on the paired prompt. The caption should start from the specific content of the image, rather than simply listing a few keywords.

#Requirements for outputs

The final output should follow a structured format: first, print the first prompt; second, print the second prompt; third, print the caption.

#Examples

[You can set some examples for ChatGPT to refer to.]

Illustrated outputs are shown below:

Prompt 1: **[Prompt Examples]**

Prompt 2: **[Prompt Examples]**

Caption: **[Caption Examples]**

Figure 7: Instruction template for multi-image data generation.



ChatGPT for

Interleaved multi-turn dialogues Generation

As an AI visual assistant, your task is to construct a multi-turn conversation in **which one person teaches another person to do something step by step according to the given phrase**. Then the person being taught needs to provide a current picture and then ask the other person. The picture content provided is enclosed in 《》 Get up, inside is the content description of this image. In each step, there should be one picture content. .#

#Requirements for outputs

[You can customize the output format according to your needs.]

#Examples

[You can set some examples for ChatGPT to refer to.]

A: Can you guide me on making a sandwich? Here's a picture of my kitchen counter 《A photo of a kitchen counter with various ingredients such as bread, lettuce, tomatoes, and deli meats. There's also a knife and a cutting board.》 .

B: Great! First, take two slices of bread from the bag on the counter.

A: I've got the bread slices. What's the next step? 《A photo of the bread slices in hand, ready to be used.》

B: Place the slices on the cutting board. Now, grab some lettuce and tomatoes.

A: Okay, I've placed the bread on the cutting board, and I have lettuce and tomatoes. What's next? 《A photo of the arranged bread slices with lettuce and tomatoes on the cutting board.》

B: Add a layer of deli meat on top of the lettuce and tomatoes.

A: Done! I've added the deli meat. What's the final step? 《A photo of the sandwich with layers of ingredients, ready to be enjoyed.》

B: Place the second slice of bread on top to complete your sandwich. Enjoy your meal!

Illustrated outputs are shown below:

[Interleaved multi-turn dialogues Examples]

Figure 8: Instruction template for interleave multi-turn dialogues generation.



ChatGPT

As an AI visual assistant, your ability is evaluating predicted answer based on the given question and correct answer. Now I have a predicted answer and corresponding question, and I need to compare predicted answer with the correct answer. Here's a scoring standard you can use to assess. A score of 5 indicates that the response perfectly aligns with the answer, while a score of 1 signifies that the response is entirely different from the answer.

****Scoring Key (5-0 Point Scale)****

****Exact Correlation (5 points)****

The predicted answer strictly replicates the reference answer.
It accurately perceives and elaborates on the queried element.

****Almost Exact Correlation (4 points)****

The predicted answer is not verbatim but delivers a notion akin to the reference answer.
Slight variations in phraseology exist, but the core idea remains consistent.
May neglect a minor item, but this does not incite confusion or misinterpretation.

****Moderately Accurate (3 points)****

The predicted answer captures the chief matter but may omit key details.
It could contain data or semantics that may be misleading.
Understands the key intent but lacks exactness.

****Somewhat Accurate (2 points)****

The predicted answer is loosely related to the reference answer with conspicuous discrepancies.
It correctly portrays a segment of the required details but bypasses or distorts major aspects.
Significant corrections are necessary to match this response with the reference answer.

****Barely Related (1 point)****

The predicted answer barely links to the actual object or query.
Mostly incorrect or misleading, exhibiting trivial understanding.
A viewer or user would be predominantly misguided based on this response.

****Completely Inaccurate (0 points)****

The predicted answer shares no relevance to the reference answer.
Utterly misinterprets or overlooks the queried item.
No discernible cognition of the assessment subject is evident.

Figure 9: Score criteria based on GPT-4.



Figure 10: **Qualitative results:** Demonstrations of diverse data we generated.

Q: "Can you teach me how to bake a chocolate cake? This is a picture of my kitchen."

A: "Sure! First, preheat your oven to 350 degrees Fahrenheit."

Q: "Okay, the oven is preheating. What should I do next?"

A: "Next, mix together 2 cups of flour, 2 cups of sugar, 3/4 cup of cocoa powder, 2 teaspoons of baking powder, and 1 1/2 teaspoons of baking soda in a large bowl."

Q: "I've mixed all the dry ingredients together. What's next?"

A: "Now, add 1 cup of milk, 1/2 cup of vegetable oil, 2 eggs, and 2 teaspoons of vanilla extract to the bowl and mix until everything is well combined."

Q: "I've added all the wet ingredients and mixed everything together. What's the next step?"

A: "Grease two 9-inch round cake pans and divide the batter evenly between them."

Q: "I've greased the pans and divided the batter. What should I do next?"

A: "Bake the cakes for 30-35 minutes or until a toothpick inserted into the center comes out clean."

Q: "The cakes are in the oven. What should I do while they bake?"

A: "While the cakes bake, you can make the frosting. Melt 1/2 cup of butter and 2/3 cup of cocoa powder in a saucepan over low heat. Then, stir in 3 cups of powdered sugar and 1/3 cup of milk until the frosting is smooth."

Q: "I've made the frosting. What should I do next?"

A: "Once the cakes are done baking, let them cool completely before frosting them. Then, spread the frosting evenly over one cake layer and place the other cake layer on top. Finally, frost the top and sides of the cake."



Figure 11: **Qualitative results:** Demonstrations of diverse data we generated.



Figure 12: **Qualitative results:** Demonstrations of our method's effectiveness across diverse real-world image scenarios.

Q: *“Can you teach me how to tie a necktie? here's a picture of the tie I want to wear.”*

Ours-A: *“Of course! I'd be happy to help you learn how to tie a necktie. First, Start by holding the tie with the wide end facing upwards and cross the wide end over the narrow end, creating an X shape.”*

Q: *“Okay, I've done it. What's the next step?”*

Ours-A: *“Next, bring the wide end of the tie under the narrow end, and then wrap it around the collar of your shirt.”*

Q: *“Alright, what's next?”*

Ours-A: *“Next, bring the wide end of the tie up over the narrow end, and then tuck it under the narrow end. This will create a loop.”*

Q: *“Got it, I've made the loop, then what should I do, tell me last step?”*

Ours-A: *“Great job! Now, take the wide end of the tie and bring it up through the loop you just d. Then, tuck the wide end under the narrow end of the tie, and finally, adjust the tie to the desired length and position. You should now have a perfectly tied necktie!”*



Figure 13: **Qualitative results:** Demonstrations of our method’s effectiveness across interleaved multi-turn dialogues.

Q: *“Can you teach me how to brew a cup of robust Turkish coffee? This is a picture of my kitchen.”*

Ours-A: *“Sure! First, fill the bottom of the Turkish coffee pot with cold water and place it on the stove.”*

Q: *“Ok, I've done that. What's next?”*

Ours-A: *“Next, add a teaspoon of coffee to the pot and place it on the stove. Turn the heat to medium and let it boil for about 30 seconds. Once it starts to boil, remove the pot from the heat and let it sit for a few seconds.”*

Q: *“Got it, what's next?”*

Ours-A: *“Finally, add a teaspoon of sugar to the pot and stir it in, pour the coffee into a small cup and enjoy! ”*






Figure 14: **Qualitative results:** Demonstrations of our method’s effectiveness across interleaved multi-turn dialogues.