TaskGalaxy Data Card

TaskGalaxy

Dataset: TaskGalaxy link
Data Card Author:
Jiankang Chen,Bin Wen,Tianke Zhang, Changyi Liu, Haojie Ding, Huihui Xiao, Yaya Shi, cheng.feng, Fan Yang, Tingting Gao, Di ZHANG

DATASET SUMMARY

This data card describes the TaskGalaxy (Multimodal VQA 413K) dataset, a large-scale multi-modal instruction fine-tuning dataset that includes 19,227hierarchical task types and 413,648 associated VQA samples.

TaskGalaxy is a generative Visual Question Answering (VQA) dataset based on open-source images, designed for research purposes. It serves as a high-quality and diverse multimodal visual data source, aimed at enhancing the fine-tuning of multimodal models across a wide range of task types.

Dataset Owners

TEAM(S)	CONTACT DETAIL(S)	AUTHOR(S)
TaskGalaxy[Kuaishou Technology] Project	**Dataset Owner(s):** Jiankang Chen,Bin Wen,Tianke Zhang, Changyi Liu, Haojie Ding, Huihui Xiao, Yaya Shi, cheng.feng, Fan Yang, Tingting Gao, Di ZHANG **Affiliation:** Kuaishou Technology	• Jiankang Chen,Bin Wen,Tianke Zhang, Changyi Liu, Haojie Ding, Huihui Xiao, Yaya Shi, cheng.feng, Fan Yang, Tingting Gao, Di ZHANG

Dataset Overview

DATA SUBJECT(S)	DATASET SNAPSHOT	CONTENT DESCRIPTION
Non-Sensitive Data about people Data about natural phenomena Data about places and objects Synthetically generated data Data about animal	TaskGalaxy is a static snapshot. Both images and text are static. Size of Dataset: ~190.73GB Number of Instances: 413648 Number of Fields: 4 Labeled Classes: N/A Number of Labels: Variable[1] Algorithmic Labels: 2[2] Human Labels: Unavailable[3] **Above:** Summary of TaskGalaxy dataset [1]Some fields(such as id has tens of thousands of possible values as unique identification of the sample, task_type has 19227 values in this dataset) [2] `task_type` and `conversations` are machine generated [3] All labels are generated by matching through the TaskGalaxy Pipeline. There are no human-annotated labels in this dataset.	An image, task type related to image conent, and associated question and answer pair generated by GPT-4o. The TaskGalaxy dataset comprises 413Kimages along with associated task types, visual questions, and answer pairs. These annotations are generated using the TaskGalaxy pipeline, which leverages models such as GPT-4, CLIP, GLM-4v-9B, InternVL-Chat-V1.5, and InternVL2-26B. Each data point in the dataset includes an ID, an image (indicating the image path), and conversations related to the task type, encompassing visual questions and their corresponding answers.

Sensitivity of Data

SENSITIVITY TYPE(S)	FIELD(S) WITH SENSITIVE DATA
None	**Intentionally Collected Sensitive Data** No sensitive data was intentionally collected. **Unintentionally Collected Sensitive Data** S/PII, pornographic content, or images depicting violence were not explicitly colleted as a part of the dataset creation process because we collect open-source images which meet unsensitive data and visual question-answer pair is generated by GPT-4o that meets security protocols. Fields that may contain such sensitive data are image_data(pixels of the image) and associate text(visual question-answer pair),

Dataset Version and Maintenance

MAINTENANCE STATUS	VERSION DETAILS	MAINTENANCE PLAN

Actively Maintained No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.	**Current Version:** 1.0 **Last Updated:** 09/2024 **Release Date:** N/A	TaskGalaxy is a multimodal visual Q&A dataset featuring a diverse array of task types. The dataset and its corresponding samples can be continuously expanded using TaskGalaxy's data generation pipeline, allowing for flexible version updates.
	NEXT PLANNED UPDATE(S)	EXPECTED CHANGE(S)
	Version affected: 1.0 **Next data update:** 05/2025 **Next Version:** 1.1 **Next Version update:** 05/2025	**Updates to Dataset:** ● Continuous expansion of potential task types based on existing ones. ● Utilize more open-source image data in the pipeline to continually expand the sample set. ● Re-execute the pipeline steps for task types that failed to match an image, using a generative model to produce the corresponding image.

Example of Data Points

PRIMARY DATA MODALITY	SAMPLING OF DATA POINTS	DATA FIELDS

Multimodal

Below are examples of kind data in the TaskGalaxy dataset.

Task type: analysis~fashion analysis~season identification

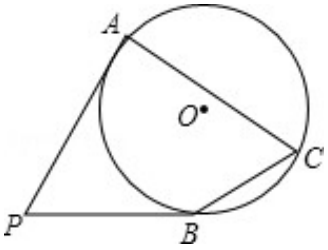
Question: For which season is this jacket most suitable?

Answer: This jacket is most suitable for winter or late fall, given its thick material and protective design.

Task type: suggestions~movie suggestions

Question: What movies would be good to watch if you enjoy 'Holmes & Watson'?

Answer: If you enjoy 'Holmes & Watson', you might also enjoy movies like 'Sherlock Holmes' (2009), 'Sherlock Holmes: A Game of Shadows' (2011), 'The Great Mouse Detective', and 'Enola Holmes'.



Task type: logical reasoning~complex reasoning~complex mathematical

Field Name	Field Value	Description
Id	String	Unique id for the data point
Image	String	Path of the image
Task type	String	Types of visual question-answer tasks related to image content.
Conversations	List	Each element of the list is a dictionary with two fields from and value, the first element from is human and value is the question, the second element is from is gpt and value is the answer.

calculations~geometric mathematical operations

Question: If O is the center of the circle and angle AOB is given as x degrees, what is the measure of angle ACB?

Answer: Angle ACB is half of angle AOB because an angle at the center of the circle is twice the angle at the circumference on the same arc. Therefore, ACB = x/2 degrees.

Task type: logical reasoning~complex reasoning~military-related reasoning

Question: What can be inferred about the relationship between the two people based on the individual's attire and their interaction with each other?

Answer: The individual's military attire suggests that they are likely in a military service, and the casual attire and hand-holding suggest a personal, likely romantic relationship.

TYPICAL DATA POINT

Field name	Value
Id	21f1699e-4f7c-4869-90b9-0eec9379d0af
Image	“ALLaVA/a_sharegpt4v_data/sam_images/sa_7686.jpg ”
Task type	“inventory management~product identification”
Conversations	[{“from”: “human”,“value”:”<image>question”}, {“from”: gpt, “value”: “answer”]

ATYPICAL DATA POINT

The dataset does not contain atypical data points as far as we know.

Motivations & Intentions

Motivations

PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
Research	`computer vision`, `multimodal understanding`, `large multimodal model`	- Richer task-diverse visual Q&A data for instruction fine-tuning in multimodal models - A flexible and scalable task-diverse instruction fine-tuning data generation pipeline for the multimodal instruction fine-tuning community（Labor-Free）

Access, Retention, & Wipeout

Access

ACCESS TYPE	DOCUMENTATION LINK(S)	PREREQUISITE(S)
External - Open Access	Dataset link: TaskGalaxy link	• After reviewing, the TaskGalaxy dataset will be publicly released to support future research.

Provenance

Collection

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
API Scraped or Crawled Taken from other existing datasets	**API && Taken from other existing datasets:** **Source:** TaskGalaxy dataset is a multimodal visual Q&A dataset, the whole is composed of GPT-4o and multiple open source models and open source image datasets, where the image dataset is composed of images collected from multiple open source datasets such as ALLaVA, Visual Genome, MathV360K, ShareGPT4v, and the task types are generated by the GPT-4o API, the subsequent task matching, filtering, Q&A pair generation and filtering, etc. are all done by the API and multiple open source multimodal models. **Is this source considered sensitive or high-risk?** [No] **Dates of Collection:** [2024-7 to 2024-9] **Primary modality of collected data:** Multimodal (image and text) **Update Frequency for collected data:** Yearly	The TaskGalaxy dataset leverages only the image component of open-source datasets. The task types, along with the corresponding sample question-answer pairs, are generated using our proprietary pipeline. • **ALLaVA dataset**: A multimodal dataset comprising 664K samples, with average image resolutions of 891 by 770 pixels, sourced from a diverse range of images. • **Visual Genome**: It contains Visual Question Answering data in a multi-choice setting. It consists of 101,174 images from MSCOCO with 1.7 million QA pairs, 17 questions per image on average. • **MathV360K**: It is proposed by Math-LlaVA ,which consists 40K images from 24 datasets and 360K question-answer pairs. • **ShareGPT4v**: A large-scale highly descriptive image-text dataset. A large-scale highly descriptive image-text dataset.
COLLECTION CADENCE	**DATA INTEGRATION**	**DATA PROCESSING**
Static Data was collected once from single or multiple sources.	**All open-source datasets** **Included Fields** Image **Additional Notes:** TaskGalaxy only uses images from open source datasets **Excluded Fields:** Id(python command generation) Task type(GPT-4o API generation) Conversations(GPT-4o, open-source multimodal models generation)	All data is coming from open-source datasets(image part) with other fields(GPT-4o, multiple multimodal models-generated)

Collection Criteria

DATA SELECTION	DATA INCLUSION	DATA EXCLUSION
Records from the open-source dataset are chosen according to the following criteria: ● **No sensitive data and adult content** : images coming from open-source multimodal datasets have guaranteed above requirements. ● **Pluralistic:** To match the diversity of task types, collect multiple open source data covering as wide a range of domains as possible.	Records that are not excluded are in the final dataset.	● Task types that did not match any images are excluded. ● GPT-4o was employed to generate task type-related question-answer pairs for each image. Following this, three open-source models were used to filter out sample examples where the task questions and answers did not match the image.

Extended Use

Use with Other Data

SAFETY LEVEL	KNOWN SAFE DATASET(S) OR DATA TYPE(S)	BEST PRACTICES
Safe to use with other data	TaskGalaxy can be combined with any other command instruction fine-tuning dataset as long as it conforms to the format of the image, question and answer pairs.	TaskGalaxy+LLaVA-665K, +LlaVA-Onevision, +Vision-Flan, ...

Sampling Methods

Fill out the following block if your dataset employs any sampling methods.

METHOD(S) USED	CHARACTERISTIC(S)	SAMPLING CRITERIA
Multi-stage Sampling Random Sampling Stratified Sampling Unsampled	**Unsampled** Upstream Source[images] open-source datasets Total data sampled ~ 825,161 Sample size ~ 825,161 Using the all images as image source of TaskGalaxy dataset **Stratified Sampling** Upstream Source Dataset version Coming from the unsample described above Total data sampled ~ 825, 161 Sample size ~ 413, 648 After task type and image matching screening, GPT-4 generates task-related question-answer pairs. Three open-source multimodal models are then used to identify and filter out mismatched images and question-answer pairs, ensuring the best matches are retained.	• Random sampling can be done according to all `task types`, and around 19,000 task types can be sampled with different numbers of Q&A samples for different task types on demand. • 613K records were whole used or added to existing multimodal question-answer pairs.

Transformations

TRANSFORMATION(S) APPLIED	FIELD(S) TRANSFORMED	LIBRARY(IES) AND METHOD(S) USED
Cleaning Mismatched Values Cleaning Missing Values	**Cleaning Mismatched Values:** The `task type` field The `conversations` field **Cleaning Missing Values:** `conversations`	**Cleaning Mismatched Values:** 1. `task type`: Firstly, CLIP is used to match image-text pairs, and the task types with high matches are assigned to the images; in order to further filter the task types with better matches, GPT-4o is further filtered with task types with better matches to the image content using the appropriate prompts. 2. `conversations`: Appropriately designed prompts use three open-source multimodal models to score the task type, image content, and question answers, and those with a total score of more than 2 are retained, while the other samples are discarded. **Cleaning Missing Values:** `conversations`: Using hand written rules, discard samples where the answer string is null.

Breakdown of Transformations

CLEANING MISSING VALUE(S)	METHOD(S) USED	COMPARATIVE SUMMARY
Description: The `conversations` filed: Allowing GPT-4o to generate answers based on questions may result in empty answers.	**Platforms, tools, or libraries:** Using hand written rules, discard samples where the answer string is null.	 **Task_type**: image style recognition~decor style recognition **Question**: What decor style is depicted in the image? **Answer**: [] Samples like this are screened out.

CLEANING MISMATCHED VALUE(S)	METHOD(S) USED	COMPARATIVE SUMMARY
Description: The `task type` field and the corresponding image are sometimes mismatched. This may lead to errors later on when combining the image content and task type to generate question answer pairs. The `conversations` field: Depending on the type of task letting GPT-4o generate questions and answers when the questions may not match the images or the answers may not match the questions.	`task type`: Firstly, CLIP is used to match image-text pairs, and the task types with high matches are assigned to the images; in order to further filter the task types with better matches, GPT-4o is further filtered with task types with better matches to the image content using the appropriate prompts. `conversations`: Appropriately designed prompts use three open-source multimodal models to score the task type, image content, and question answers, and those with a total score of more than 2 are retained, while the other samples are discarded.	 `Task type`: **CLIP-Matching**: ▤ product recommendation-sports equipment recommendation ▤ product recommendation-outdoor equipment recommendation ▤ personalization-storytelling-user engagement optimization in storytelling ▤ personal health monitoring-injury recovery monitoring ▤ visual data summarization-user-generated content summarization-blog post summarization ▤ interactive content generation-content personalization engines ▤ image content understanding-content recognition-fitness equipment recognition ▤ content creation-travel blog content creation-offseason travel insights ▤ image content understanding-content recognition-sports and posture-sports posture recognition ▤ remote sensing analysis-coastal change monitoring-sand dune dynamics tracking **GPT-4o filtering**: ▤ Product recommendation-sports equipment recommendation ▤ product recommendation-outdoor equipment recommendation ▤ image content understanding-content recognition-fitness equipment recognition `conversations`: Below are two samples of no-match removed: 1)  "**task_type**": "damage assessment"

"**question**": "What kind of infrastructural damage can be observed in the image due to the flood?"
"**answer**": "The image shows a high level of water submerging the docks and possibly affecting the buildings and facilities near the waterfront, indicating significant infrastructural damage."
{
"intervl2_26B_score": "0",
"glm4v_score": "0",
"internvl_score": "0"
}
Note: The total score is 0 < 2:
The water level did not flood the pier, such that the sample was excluded.

2)

"task_type": "ocr~environmental text ocr~public information boards reading",
"question": "What public information is provided by the board attached to the yellow structure?",
"answer": "The board attached to the yellow structure provides information from NYC regarding collection dates and categories of refuse collection.",
"intervl2_26B_score": "0",
"glm4v_score": "0",
"internvl_score": "1"
Note: The total score is 1 < 2**:**
The yellow board doesn't have any information about garbage collection, but rather the mailboxes, such that the sample with hallucinations was excluded.