Datasheet for JourneyBench

Zhecan Wang[♠] Junzhang Liu[♠] Chia-Wei Tang[†] Hani Alomari[†]
Anushka Sivakumar[†] Rui Sun[♠] Wenhao Li[♠] Md. Atabuzzaman[†] Hammad Ayyubi[♠]
Haoxuan You[♠] Alvi Ishmam[†] Kai-Wei Chang[♦] Shih-Fu Chang[♠] Chris Thomas[†]
[♠]Columbia University [♦]UCLA [†]Virginia Tech

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. Please see the most updated version here.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset, JourneyBench, was created to rigorously assess the fine-grained multimodal reasoning abilities of state-of-the-art models using challenging, human-annotated, and generated images. The specific tasks in mind include complementary multimodal chain-of-thought (MCOT), multi-image visual question answering (VQA), imaginary image captioning, VQA with hallucination triggers, and fine-grained cross-modal retrieval with sample-specific distractors. The dataset aims to fill the gap in existing benchmarks, which often contain limited visual diversity and scenarios, by requiring models to perform complex reasoning in unusual and fictional visual contexts where typical language biases and shallow visual understanding are insufficient.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by a collaborative team of researchers from Columbia University, UCLA, and Virginia Tech. The authors include Zhecan Wang, Junzhang Liu, Chia-Wei Tang, Hani Alomari, Anushka Sivakumar, Rui Sun, Wenhao Li, Md. Atabuzzaman, Hammad Ayyubi, Haoxuan You, Alvi Ishmam, Kai-Wei Chang, Shih-Fu Chang, and Chris Thomas, representing their respective institutions.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

The support for the creation of the dataset primarily came from institutional funding provided by Columbia University, UCLA, and Virginia Tech.

Any other comments?

NA

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the JourneyBench dataset represent a variety of generated images, along with corresponding annotations for different vision-language understanding tasks. These instances include:

1. Images: Generated visual content that is unusual, fictional, or abstract, designed to challenge models in fine-grained visual understanding.
2. Image-Question Pairs: For the multimodal chain-of-thought (MCOT) and VQA (HaloQuest) tasks, these pairs require models to perform reasoning by integrating information from both visual and textual modalities.
3. Multi-Image Sets: For the multi-image VQA task, sets of images are used to test models' abilities to reason across multiple visual contexts.
4. Image-Text Pairs: Used for the fine-grained cross-modal retrieval task, where models must retrieve the correct text given an image, and vice versa, with the presence of sample-specific distractors.
5. Image-Caption Pairs: For the imaginary image captioning task, where models are required to generate descriptions for unusual or fictional images.

These instances are designed to test various aspects of vision-language understanding, including reasoning, retrieval, captioning, and handling hallucinations.

How many instances are there in total (of each type, if appropriate)?

Overall, JourneyBench has $13,631$ unique image-text samples across five tasks, which consist of $12,405$ unique images and $13,664$ unique text.

JourneyBench includes 2,600 image-question pairs for complementary multimodal chain-of-thought, categorized into 10 fine-grained types based on visual contexts and multimodal co-referencing. All collected images in JourneyBench fall into 11 fine-grained categories based on their level of unusualness or fictionality. For multi-image VQA, there are 316 image-question pairs across three fine-grained categories. The image captioning dataset contains 1,000 images paired with 5,000 captions, with each image having five captions. For visual question answering, JourneyBench comprises 7,748 image questions, categorized into three fine-grained types of hallucination triggers. The fine-grained cross-modal retrieval task contains two subtasks. For image-to-text retrieval, there are 1,000 query images paired with 11,121 texts, averaging five positive texts (ground-truth captions) and six negative texts (sample-specific text distractors) per image. For text-to-image retrieval, there are 1,000 samples, each with five ground-truth captions, resulting in approximately 5,000 query texts against 6,323 images. Each sample has one ground-truth matching image and five negative images (sample-specific image distractors).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The JourneyBench dataset is a curated sample of instances rather than an exhaustive collection of all possible instances. It is designed to represent a diverse range of unusual and fictional visual scenarios to test the fine-grained multimodal reasoning abilities of models. The larger set from which these instances are drawn includes a vast array of generated images available on platforms like Midjourney, a crowd-based image generation service.

Representativeness:

- Selection Criteria: The dataset was curated by retrieving popular and highly rated images from Midjourney, focusing on those that are unusual, fictional, or abstract.

- Diversity and Quality Control: Images were filtered and selected based on specific criteria (unusualness, fictionality, comprehensibility) by human annotators to ensure a diverse and challenging dataset.

- Validation: The representativeness was ensured through multiple rounds of human annotations and verifications. At least four annotators assessed each image to ensure it met the specified criteria, with a high agreement rate of over 72

Not Fully Representative:

- Focus on Diversity and Challenge: The dataset aims to cover a broad and diverse range of challenging scenarios, which may not fully represent the typical distribution of generated images on Midjourney but is intentionally biased towards more complex and less common visual scenes.

- Specific Research Purpose: The sample was curated to address the limitations of existing benchmarks and to provide a more rigorous test of multimodal reasoning capabilities, rather than to represent the full set of generated images comprehensively.

In summary, while the JourneyBench dataset is a curated sample designed for specific research purposes, it is representative in terms of its goal to challenge and evaluate advanced multimodal models across a diverse range of unusual and fictional visual scenarios.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance in the JourneyBench dataset consists of the following data types:

Raw Data:

- **Generated Images**: High-quality, unusual, fictional, or abstract images retrieved from the Midjourney platform. These images are the primary visual content used across different tasks.

Annotations and Features:

- **Image-Question Pairs**: For the complementary multimodal chain-of-thought (MCOT) task, each instance includes an image and a corresponding question designed to require multimodal reasoning.
- **Multi-Image Sets**: For the multi-image VQA task, each instance consists of a set of related images and a question that requires reasoning across these multiple images.
- **Image-Text Pairs**:
  - **Image-to-Text Retrieval**: Each instance includes an image and a set of textual descriptions, with one being the correct caption and others serving as sample-specific distractors.
  - **Text-to-Image Retrieval**: Each instance includes a text query and a set of images, with one being the correct match and others serving as sample-specific distractors.
- **Image-Caption Pairs**: For the imaginary image captioning task, each instance includes an image and one or more corresponding captions that describe the unusual or fictional elements of the image.
- **Visual Question Answering (VQA) with Hallucination Triggers**: Each instance includes an image and a VQA pair, where the question is designed to test the model's ability to handle hallucination-inducing scenarios.

Description of Features:

- **Images**: Visual content in a variety of formats, typically high-resolution, capturing unusual, fictional, or abstract scenarios.
- **Questions and Captions**: Textual content that describes the image, poses a question related to the image, or serves as distractors for retrieval tasks.

- **Distractors**: For retrieval tasks, carefully crafted text or image distractors that are relevant but incorrect, designed to challenge the model's fine-grained differentiation capabilities.

Overall, each instance in the JourneyBench dataset comprises raw visual data (images) and associated textual annotations (questions, captions, and distractors), all designed to test different aspects of multimodal reasoning and understanding.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, there is a label or target associated with each instance in the JourneyBench dataset. These labels or targets are designed to evaluate different aspects of multimodal reasoning and understanding. The descriptions are as follows:

Labels or Targets:

- **Image-Question Pairs (MCOT Task)**:
  - **Target**: The correct answer to the question, which requires integrating information from both the image and the question text.
- **Multi-Image Sets (Multi-Image VQA Task)**:
  - **Target**: The correct answer to the question, which requires reasoning across multiple related images.
- **Image-Text Pairs (Cross-Modal Retrieval Task)**:
  - **Image-to-Text Retrieval**:
    - **Target**: The correct caption corresponding to the given image, with sample-specific distractors included.
  - **Text-to-Image Retrieval**:
    - **Target**: The correct image corresponding to the given text query, with sample-specific distractors included.
- **Image-Caption Pairs (Imaginary Image Captioning Task)**:
  - **Target**: One or more correct captions that accurately describe the unusual or fictional elements of the given image.
- **Visual Question Answering (VQA) with Hallucination Triggers**:
  - **Target**: The correct answer to the question, designed to test the model's ability to handle hallucination-inducing scenarios.

Overall, each instance in the JourneyBench dataset is associated with a specific label or target that provides the ground truth for evaluating the model's performance in various multimodal tasks.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, the JourneyBench dataset is designed to be comprehensive, with each instance containing all the necessary information required for its respective task.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No, relationships between individual instances in the JourneyBench dataset are not made explicit. Each instance is designed to be self-contained for its specific task, focusing on evaluating the model's performance on individual visual and textual inputs. Each instance is annotated and curated to ensure it provides the necessary context and information for the task it is associated with, without relying on relationships with other instances.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

JourneyBench dataset serves as a standalone testing set. There are no specific splits for training and validation within JourneyBench itself. Instead, the dataset is used exclusively for evaluating models trained on other datasets.

Testing Split:

- **Purpose**: Used exclusively for testing and evaluating the performance of models.
- **Composition**: The entire JourneyBench dataset, including all instances from tasks such as MCOT, multi-image VQA, image captioning, VQA with hallucination triggers, and fine-grained cross-modal retrieval.
- **Rationale**: To provide a challenging and comprehensive evaluation benchmark that tests the fine-grained multimodal reasoning capabilities of pre-trained models on unusual and fictional visual content.

Models are expected to be trained on other datasets and then evaluated on JourneyBench to assess their ability to handle the complexities and unique challenges presented by the Midjourney-generated images.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The JourneyBench dataset has undergone extensive quality checks and multiple rounds of human annotations to minimize errors, sources of noise, and redundancies. However, as with any large dataset, some potential issues may still exist:

Potential Issues:

- **Annotation Errors**: Despite thorough verification, there may be occasional errors in annotations, such as incorrect captions or answers.
- **Noise in Generated Images**: Some generated images might contain artifacts or visual inconsistencies due to the limitations of the image generation process.
- **Subjectivity in Annotations**: The interpretation of unusual or fictional elements in images can be subjective, potentially leading to variations in annotations.

- **Redundancies**: Although efforts were made to ensure diversity, there might be some instances where similar images or questions appear more than once.

These issues are inherent to the complexity and scale of the dataset. Continuous efforts are made to identify and correct such issues to maintain the dataset's high quality and utility for evaluation purposes.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The JourneyBench dataset is self-contained and does not rely on external resources such as websites, tweets, or other datasets. All necessary data, including generated images and their corresponding annotations, are included within the dataset itself.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No, the JourneyBench dataset does not contain any data that might be considered confidential. The dataset is composed entirely of publicly posted and shared generated images under the community guidelines of the Midjourney platform and their corresponding annotations (e.g., questions, captions, and distractors) created for the purpose of evaluating multimodal models.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No, the JourneyBench dataset does not contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety. The dataset consists of publicly posted and shared generated images and annotations created specifically for evaluating multimodal models. The images were filtered first by the community guidelines of the Midjourney platform and then underwent multiple rounds of annotation during our process to ensure that all content was appropriate and non-distressing. This multi-layered filtering and review process ensures that the dataset is suitable for a broad audience and free from harmful content.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No, the JourneyBench dataset does not relate to people. It consists entirely of generated images and corresponding annotations created for the purpose of evaluating multimodal models. There is no personal, sensitive, or identifiable information about individuals included in the dataset. Any similarity to actual persons, living or dead, or to actual events is purely coincidental. No images within JourneyBench are created to look like any particular individual.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No, the JourneyBench dataset does not identify any subpopulations such as by age, gender, or other demographic characteristics. The dataset is focused on generated images and corresponding annotations for evaluating multimodal models, and it does not include any personal or demographic information about individuals.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No, it is not possible to identify individuals, either directly or indirectly, from the JourneyBench dataset. The dataset consists entirely of generated images and annotations created for evaluating multimodal models. There is no personal, sensitive, or identifiable information about individuals included in the dataset, and the content is designed to be fictional and abstract without any real-world personal references.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
No, the JourneyBench dataset does not contain data that might be considered sensitive in any way. It does not include information that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions, union memberships, locations, financial or health data, biometric or genetic data, forms of government identification, or criminal history. The dataset consists entirely of generated images and corresponding annotations created for the purpose of evaluating multimodal models, ensuring that all content is non-sensitive and appropriate for research and benchmarking purposes.

Any other comments?
NA

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance in the JourneyBench dataset was acquired as follows:

Data Acquisition:

- **Generated Images**: The images were retrieved from the Midjourney platform, where they were publicly posted and shared under the community rules.
- **Annotations (Questions, Captions, Distractors)**: These were created by human annotators specifically for the purpose of evaluating multimodal models.

Validation and Verification:

- **Images**: The images were initially filtered by the Midjourney platform and further filtered by human annotators to ensure they met the criteria of being unusual, fictional, or abstract, and were appropriate for the dataset.
- **Annotations**: Multiple rounds of human annotation were conducted to ensure the quality and accuracy of the questions, captions, and distractors. Annotations were cross-verified by additional annotators to minimize errors and biases.

Overall, both the generated images and the corresponding annotations underwent rigorous validation and verification processes to ensure the dataset's quality and suitability for research and benchmarking purposes.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

The data for the JourneyBench dataset was collected over a period from January 2023 to March 2023. This timeframe matches the creation timeframe of the data associated with the instances, as the generated images were retrieved and annotated during this period.

Timeframe:

- **Data Collection Period**: January 2023 to March 2023.
- **Data Creation Period**: The generated images and corresponding annotations were created and verified during the same timeframe.
- **First Publication Date**: The dataset will be first published in June 2024.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data for the JourneyBench dataset was collected using the following mechanisms and procedures:

- **Generated Images**:
  - **Procedure**: Images were retrieved from the Midjourney platform using web scraping tools and metadata analysis to identify popular and highly rated images.
  - **Validation**: Images were manually filtered and verified by human annotators to ensure they met the criteria of being unusual, fictional, or abstract.
- **Annotations (Questions, Captions, Distractors)**:
  - **Procedure**: Annotations were created through manual human curation via a human-machine-in-the-loop mechanism by a team of annotators.
  - **Validation**: Multiple rounds of annotation and cross-verification by additional annotators were conducted to ensure accuracy and quality. Detailed instruction manuals and examples were provided to annotators to maintain consistency.

Overall, the data collection mechanisms and procedures involved a combination of automated tools (for image retrieval) and extensive human curation (for filtering images and creating annotations). These mechanisms were validated through rigorous manual verification and quality control processes.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[2] for approaches in this area.)

The resource cost of collecting the data for the JourneyBench dataset included computational resources, financial costs, and energy consumption. Here is an estimate of these costs:

Computational Resources:

- **Web Scraping and Data Retrieval**: Standard computational infrastructure was used for web scraping and metadata analysis.
- **Data Storage**: Approximately 100 GB of storage was required to store the generated images and annotations.
- **LLM and VLM Processing**: Significant computational resources were used for running LLMs and VLMs to generate and verify annotations. This included access to cloud-based GPU instances for model inference.

Financial Costs:

- **Human Annotation**: The total cost of human annotation was approximately $20,000, including payment to annotators on platforms like Amazon Mechanical Turk (MTurk).
- **Computational Costs**:
  - **Web Scraping, Data Processing, and Storage**: Approximately $1,000.

  - **LLM and VLM Inference**: The cost for running large models, such as GPT-4V, for human-machine-in-the-loop processes was approximately \$5,000, considering cloud-based GPU rental and associated infrastructure costs.

Energy Consumption and Carbon Footprint:

- **Energy Consumption**:
  - **Web Scraping, Data Processing, and Storage**: Approximately 500 kWh.
  - **LLM and VLM Inference**: The estimated additional energy consumption for running LLMs and VLMs is around 1,500 kWh.
- **Carbon Footprint**:
  - **Total Energy Consumption**: 2,000 kWh (500 kWh for initial processes + 1,500 kWh for LLM and VLM inference).
  - **Carbon Footprint**: Based on an average carbon intensity of 0.233 kg CO_2 per kWh, the total estimated carbon footprint is approximately 466 kg CO_2.

Overall, the resource cost of collecting the JourneyBench dataset, including the use of human-machine-in-the-loop processes with LLMs and VLMs, involved a significant financial investment in human annotation and computational resources, with an estimated total financial cost of \$26,000 and a carbon footprint of approximately 466 kg CO_2. The processes were designed to be efficient while ensuring high-quality data collection and validation.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The JourneyBench dataset is a curated sample from a larger set of generated images available on the Midjourney platform. The sampling strategy used was a combination of deterministic and probabilistic methods:

Sampling Strategy:

- **Deterministic Sampling**:
  - Images were selected based on their popularity and ratings on the Midjourney platform, ensuring a high level of quality and relevance.
  - Specific criteria were applied to filter images, focusing on those that were unusual, fictional, or abstract.
  - Combinations from diverse and fixed sets of topic and attribute words were used to form queries to retrieve images, ensuring a wide range of visual content.
- **Probabilistic Sampling**:
  - Images were randomly sampled from the retrieved set to cover a broad spectrum of visual scenarios and avoid bias towards any specific type of image.

This combined sampling strategy ensured that the JourneyBench dataset included high-quality, diverse, and challenging images that are representative of the unusual and fictional content available on the Midjourney platform.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process for the JourneyBench dataset involved several groups of individuals:

DATA COLLECTION PARTICIPANTS:

- **Crowdworkers**: The majority of the data annotation was conducted by crowdworkers from platforms such as Amazon Mechanical Turk (MTurk).
- **Researchers and Students**: A team of researchers and students from the collaborating institutions oversaw the data collection, annotation, and validation processes.

COMPENSATION:

- **Crowdworkers**: Crowdworkers were compensated at competitive rates for their work. On average, they were paid approximately \$15 per hour, depending on the complexity and time required for each annotation task.
- **Researchers and Students**: Compensation for researchers and students involved in the project was covered through institutional funding and research grants. Specific compensation details vary based on institutional policies and grant provisions.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, the JourneyBench dataset does not require a formal ethical review process by an institutional review board (IRB). The dataset consists entirely of publicly posted and shared generated images and their corresponding annotations created for evaluating multimodal models. There is no inclusion of personal, sensitive, or identifiable information about individuals. All data within the dataset is designed to be publicly accessible and used for research and benchmarking purposes, thereby minimizing ethical concerns related to privacy or harm.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

No, the JourneyBench dataset does not relate to people. It consists entirely of generated images and corresponding annotations created for the purpose of evaluating multimodal models. There is no personal, sensitive, or identifiable information about individuals included in the dataset.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

NA

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

NA

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

NA

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

NA

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

NA

Any other comments?

NA

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, preprocessing, cleaning, and labeling of the data were done for the JourneyBench dataset. The following processes were undertaken:

Preprocessing and Cleaning:

- **Image Filtering**: The generated images retrieved from the Midjourney platform were filtered to ensure they met specific criteria such as being unusual, fictional, or abstract. Images with artifacts or low quality were removed.
- **Data Quality Checks**: Multiple rounds of human annotation were conducted to ensure the accuracy and quality of the images and annotations. This involved verifying the content to ensure it was appropriate and met the desired criteria.

Labeling:

- **Annotations**: Human annotators created various annotations, including questions, captions, and distractors, to accompany the images. These annotations were designed to evaluate the multimodal reasoning capabilities of models.
- **Validation of Annotations**: The annotations were cross-verified by additional annotators to minimize errors and biases. Detailed instruction manuals and examples were provided to annotators to maintain consistency and high quality in the labeling process.

Procedures:

- **Manual Filtering**: Human annotators manually reviewed and filtered the images to ensure they were suitable for inclusion in the dataset.
- **Annotation Creation**: Annotations were manually created by human annotators, followed by validation steps to ensure correctness and relevance.

Overall, these preprocessing, cleaning, and labeling steps were essential to maintain the dataset's quality and ensure it was suitable for research and benchmarking purposes.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No, the "raw" data was not saved separately in addition to the preprocessed, cleaned, and labeled data for the JourneyBench dataset. The dataset provided includes only the processed and annotated data that has been prepared for evaluating multimodal models. This approach ensures that the dataset is immediately usable for research and benchmarking purposes without the need for additional preprocessing or cleaning steps.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The critical tools and methods used in the intermediate steps of data collection are saved and made available. You can access the source code for the image scraping tool, human-machine-in-the-loop implementation, and evaluation code through the following repository: `https://github.com/JourneyBench/JourneyBench`.

Any other comments?

NA

USES

Has the dataset been used for any tasks already? If so, please provide a description.

No, the JourneyBench dataset has not yet been used for any tasks. It is a newly created stand-alone testing benchmark specifically designed for evaluating the fine-grained multimodal reasoning capabilities of advanced models.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

We are developing a leaderboard to track and document future works and their model performance utilizing our benchmark. This leaderboard will list all published papers and systems that employ the JourneyBench dataset, along with their respective performance metrics. The leaderboard will be accessible at: `https://github.com/JourneyBench/JourneyBench`.

What (other) tasks could the dataset be used for?

The JourneyBench dataset, while primarily designed for evaluating multimodal models, can also be used for a variety of other tasks, including but not limited to:

- **Transfer Learning**: Leveraging the dataset to pre-train models for better performance on related tasks.
- **Multimodal Representation Learning**: Developing and testing algorithms for learning joint representations of visual and textual data.
- **Cross-modal Retrieval**: Evaluating and improving methods for retrieving relevant images based on text queries and vice versa.
- **Image Captioning**: Enhancing models that generate descriptive captions for images, especially in unusual or fictional contexts.
- **Visual Question Answering (VQA)**: Testing the ability of models to answer questions based on visual content, with a focus on handling complex and abstract scenarios.
- **Hallucination Detection**: Researching techniques to identify and mitigate hallucinations in model outputs when dealing with multimodal inputs.
- **Benchmarking Multimodal Models**: Providing a comprehensive benchmark for evaluating the performance of state-of-the-art multimodal models across various challenging tasks.
- **Educational Tools**: Using the dataset to develop educational tools and resources for teaching multimodal machine learning concepts.
- **Model Robustness Testing**: Assessing the robustness of models to unusual, fictional, and abstract visual scenarios.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

See our Limitations section in our supplementary. In brief, images in our dataset were collected from Midjourney and generated using tools in use at the time. However, the models these tools were trained on may have been trained on data that perpetuate biases and stereotypes. While we performed a filtering on top of this data, it does not mean that it does not persist (e.g. overrepresentation of certain ethnicities). This is due to the training data on which the generative models were trained rather than any issue within our control. Further, because these images were generated from a certain class of generative models, they may have artifacts representative of those particular models. There is no guarantee that future models will have the same artifacts or biases. Thus, future generative images may differ somewhat from the current generative image distribution found in JourneyBench.

Are there tasks for which the dataset should not be used? If so, please provide a description.

JourneyBench is intended as a zero-shot benchmark to test the reasoning capabilities of state-of-the-art generative models. While few-shot or in-context applications may be acceptable, the performance may differ from those reported in our results. Further, JourneyBench is not intended to be used to train models due to its size. Lastly, JourneyBench should not be used for any tasks that are illegal, unethical, perpetrate stereotypes, biases, racial profiling, or monitoring of users without consent. JourneyBench is solely meant as a means for evaluating the performance of state-of-the-art models on tasks within it.

Any other comments?

NA

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the JourneyBench dataset will be distributed to third parties (e.g., Hugging Face) outside of the entities (e.g., companies, institutions, organizations) on behalf of which the dataset was created. The dataset is intended to be publicly accessible for research and benchmarking purposes. It will be made available under a custom license[1], allowing researchers, developers, and other interested parties to use the dataset for evaluating and developing multimodal models. The dataset can be accessed and downloaded from the official repository at `https://github.com/JourneyBench/JourneyBench`.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The JourneyBench dataset will be distributed via GitHub. Interested parties can access and download the dataset from the official repository at `https://github.com/JourneyBench/JourneyBench`. The dataset would also be made available via the Hugging Face platform.

When will the dataset be distributed?

The JourneyBench dataset will be distributed in June 2024. It will be available for download and use by the research community and other interested parties from the official repository at `https://github.com/JourneyBench/JourneyBench` and the project website at `https://journeybench.github.io/`.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[1] Please refer to the **Sec. Terms of Usage for JourneyBench Dataset** in the supplementary material

Yes, please refer to the **Sec. Terms of Usage for JourneyBench Dataset** in the supplementary material.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

We refer readers to our license in our supplementary material. In brief, users may not use the images provided in JourneyBench to compete with Midjourney or other generative AI platforms, according to their license.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

NA

Any other comments?

NA

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?

The JourneyBench dataset is supported, hosted, and maintained by the research team at Columbia University, UCLA, and Virginia Tech. The dataset and its accompanying resources are available on GitHub at `https://github.com/JourneyBench/JourneyBench` and on the project website at `https://journeybench.github.io/`. The research team will ensure the dataset remains accessible, up-to-date, and properly maintained for use by the research community and other interested parties.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

For any inquiries, please get in touch with us at `journeybench.contact@gmail.com`

Is there an erratum? If so, please provide a link or other access point.

There are two typos in the main paper that should be corrected for the camera-ready submission:

1. "ALBEF-14M" should be corrected to "ALEBF-210M".
2. "CogVLM v2 (Llama3)-17B" should be corrected to "CogVLM v2 (Llama3)-19B".

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The JourneyBench dataset will undergo regular updates and maintenance to ensure its continued relevance and accuracy in evaluating multimodal models. The research team at Columbia University, UCLA, and Virginia Tech will be responsible for these updates, which will include correcting labeling errors, adding new instances, and removing outdated or erroneous data. Updates will be communicated to users through the official GitHub repository at `https://github.com/JourneyBench/JourneyBench`, the project website at `https://journeybench.github.io/`, and a mailing list for subscribed users. The team aims to review and update the dataset at least quarterly or more frequently as needed based on feedback and the identification of new challenges in the field. The maintenance would continue for at least five years after the paper's acceptance. Additionally, a leaderboard will be developed to track and document future works and their model performance using the JourneyBench dataset, fostering a collaborative environment for ongoing research and improvement.

We plan to share the dataset on Hugging Face and host a workshop focusing on a competition via JourneyBench at the upcoming CVPR conference. These initiatives will broaden access to the dataset and encourage active participation and collaboration within the research community.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

JourneyBench does not relate to people. All content within JourneyBench is generated by AI. Any similarity to actual persons, living or dead, or to actual events is purely coincidental. No images within JourneyBench are created to look like any particular individual.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We will keep the older versions of the data in the GitHub repository to enable fair comparison. Should any extensions or changes be made to our dataset, we intend to release versions of JourneyBench with version numbers (v1, v1.1, etc.) to ensure results can be compared consistently.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, there is a mechanism for others to extend, augment, build on, or contribute to the dataset. Potential contributors can submit their contribution requests via pull request or by raising an issue on our GitHub repository.

Contribution Mechanism

1) **GitHub Repository**: Contributors can submit their contributions through pull requests or raise issues on our GitHub repository to propose changes or additions.
2) **Contribution Guidelines**: Detailed guidelines will be provided in the repository to ensure consistency and

quality. These guidelines will include data format, annotation standards, and ethical considerations.

Validation and Verification

All contributions will undergo a stringent validation and verification process to maintain the dataset's quality and integrity:

1) **Initial Review**: Submitted contributions will be initially reviewed by our automated system for format compliance and basic quality checks.
2) **Human Review**: Qualified annotators will manually review the contributions to ensure they meet the dataset's standards.
3) **Expert Verification**: Domain experts will perform a final verification of the data to confirm its accuracy and relevance.

Communication and Distribution

1) **Updates and Versions**: Accepted contributions will be integrated into the main dataset and released as part of updated versions. Contributors will be credited accordingly.
2) **Communication**: Contributors will be notified of the status of their submissions via email and through GitHub notifications. Major updates will be announced on the project website and through our mailing list.
3) **Access**: Updated versions of the dataset will be accessible to all registered users through our official repository and GitHub.

Any other comments?

NA

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.

[2] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.