

## Dataset Documentation and Intended Use

**Motivation** As multi-modal large language models (MLLMs) show promising progress in the visual reasoning domain, to what extent these models have abstract visual reasoning (AVR) abilities is still unknown. Also, the lack of a holistic AVR benchmark limited the current evaluation of these models, which motivates us to propose *MARVEL*, the first comprehensive multidimensional AVR benchmark. The dataset is created intentionally with the task in mind, aiming to evaluate MLLMs’ abstract visual reasoning ability across different patterns, shapes and task configurations. The dataset was created by Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski and Jay Pujara from Information Sciences Institute, University of Southern California, Tencent AI Lab and Vrije Universiteit Amsterdam. This research was sponsored by the Defense Advanced Research Projects Agency via Contract HR00112390061.

**Composition** The instances are abstract visual reasoning puzzle images, together with a set of AVR reasoning questions and perception questions to enable a hierarchical evaluation framework. There are 770 puzzles in total in the dataset. The puzzle is a sample of all possible instances. We root our dataset in human cognitive science to ensure the generality and applicability of our dataset. Each instance consists of a puzzle, an AVR reasoning question, a fine-grained perception question, as well as three coarse-grain questions. Each instance consists of a reasoning label representing the correct answer for AVR questions and a text answer for all perception questions. The whole dataset should be considered as an evaluation benchmark rather than a dataset supporting training, validation and testing. The dataset is entirely self-contained.

*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)?* No.

*Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?* No.

*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.

*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.

*Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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)?* No

**Collection Process** The dataset is collected from public available website using crawlers (mentioned in the main paper). All authors are involved in the data collection process. The instance is collected before Jan 2024.

*Did the individuals in question consent to the collection and use of their data?* No.

*If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?* I N/A.

*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?* N/A.

**Preprocessing/cleaning/labeling** We filter the raw data after collecting data by removing duplicate, low-quality data manually. Three human annotators then choose the puzzle containing proper input shapes and patterns in our predefined settings, which can further be re-organized in different task configurations.

*Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?* Yes.

*Is the software that was used to preprocess/clean/label the data available? If Yes, the python package Pillow is used.*

**Use** *Has the dataset been used for any tasks already?* No.

*What (other) tasks could the dataset be used for?* The task regarding analysing MLLMs’ abstract visual reasoning abilities or visual perception abilities.

*Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?* No.

*Are there tasks for which the dataset should not be used?* No.

### **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?* Yes. The dataset is available on the internet ([https://github.com/1171-jpg/MARVEL\\_AVR](https://github.com/1171-jpg/MARVEL_AVR)).

*How will the dataset be distributed (e.g., tarball on website, API, GitHub)?* Yes. The link is presented in the last questions.

*When will the dataset be distributed?* The dataset was first released in April 2024.

*Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?*

*Have any third parties imposed IP-based or other restrictions on the data associated with the instances?* No.

*Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?* I The dataset is intended for research use only and should not be used to make critical decisions about individuals’ capabilities. Such misuse could cause undue pressure and anxiety for participants and may not accurately reflect their true potential or abilities in real-world scenarios.

### **Maintenance**

*Who will be supporting/hosting/maintaining the dataset?* All the authors: Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski and Jay Pujara

*How can the owner/curator/manager of the dataset be contacted (e.g., email address)?* The contact emails are [yjiang44@usc.edu](mailto:yjiang44@usc.edu), [jzhang37@usc.edu](mailto:jzhang37@usc.edu), [kexuansu@usc.edu](mailto:kexuansu@usc.edu)

*Is there an erratum?* N/A.

*Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?* The update will be posted on the GitHub link ([https://github.com/1171-jpg/MARVEL\\_AVR](https://github.com/1171-jpg/MARVEL_AVR)) and website (<https://marvel770.github.io/>).

*Will older versions of the dataset continue to be supported/hosted/maintained?* Yes, the update will be released using a different version number.

*If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?* Others may do so and should contact the original authors about incorporating fixes/extensions.

**Author Statement** We bear all responsibility in case of violation of rights, and we maintain the dataset for the long term to ensure it is accessible and organized.

**Croissant metadata and Licenses** The data is also released in hugging face (<https://huggingface.co/datasets/kianasun/MARVEL>) under apache-2.0 licenses. The Croissant metadata can be viewed and downloaded via [https://github.com/1171-jpg/MARVEL\\_AVR/blob/main/MARVEL\\_Croissant.json](https://github.com/1171-jpg/MARVEL_AVR/blob/main/MARVEL_Croissant.json)