TopoPerception: A Shortcut-Free Evaluation of Global Visual Perception in Large Vision-Language Models

Wenhao Zhou Hao Zheng Rong Zhao*
Center for Brain-Inspired Computing Research (CBICR), Tsinghua University, Beijing, China
Department of Precision Instruments, Tsinghua University, Beijing, China
IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China
zwh20@mails.tsinghua.edu.cn, hao_z@mit.edu, r_zhao@mail.tsinghua.edu.cn

Abstract

Large Vision-Language Models (LVLMs) typically align visual features from an encoder with a pre-trained Large Language Model (LLM). However, this makes the visual perception module a bottleneck, which constrains the overall capabilities of LVLMs. Conventional evaluation benchmarks, while rich in visual semantics, often contain unavoidable local shortcuts that can lead to an overestimation of models' perceptual abilities. Here, we introduce TopoPerception, a benchmark that leverages topological properties to rigorously evaluate the global visual perception capabilities of LVLMs across various granularities. Since topology depends on the global structure of an image and is invariant to local features, TopoPerception enables a shortcut-free assessment of global perception, fundamentally distinguishing it from semantically rich tasks. We evaluate state-of-the-art models on TopoPerception and find that even at the coarsest perceptual granularity, all models perform no better than random chance, indicating a profound inability to perceive global visual features. Notably, a consistent trend emerge within model families: more powerful models with stronger reasoning capabilities exhibit lower accuracy. This suggests that merely scaling up models is insufficient to address this deficit and may even exacerbate it. Progress may require new training paradigms or architectures. TopoPerception not only exposes a critical bottleneck in current LVLMs but also offers a lens and direction for improving their global visual perception. The data and code are publicly available at: https://github.com/Wenhao-Zhou/TopoPerception.

1. Introduction

The field of artificial intelligence (AI) has undergone a revolutionary transformation with the advent of Large Vision-Language Models (LVLMs). Recent state-of-the-art (SOTA) models combine advanced visual understanding with the powerful reasoning capabilities of Large Language Models (LLMs), demonstrating remarkable performance on a wide range of multimodal tasks, including Visual Question Answering (VQA), captioning, and complex visual reasoning [2–4, 20–22, 43–46]. The dominant architecture in these models involves a powerful visual encoder that processes visual inputs into a set of feature vectors or tokens, which are then fed through a small projection or query network into an LLM to generate responses based on the visual content [36, 56].

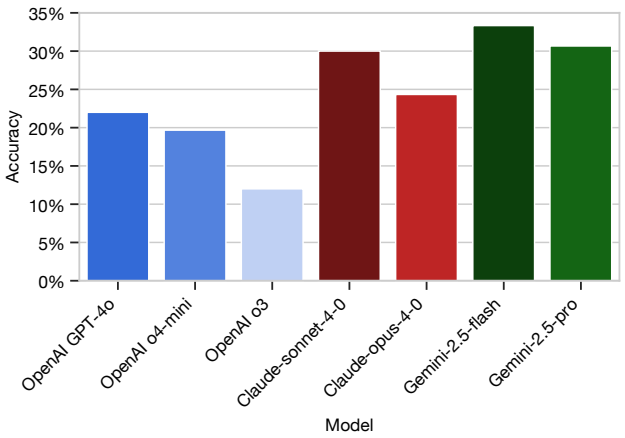

Figure 1. Accuracy of various LVLMs on the easiest level of TopoPerception. Models from the same family are represented by the same color scheme.

Despite its success, it introduces a critical and often overlooked perceptual bottleneck. The conversion of a high-dimensional, continuous visual data into a discrete sequence of tokens is inherently a lossy compression process. Semantically, the training objectives for many visual encoders, such as contrastive learning in models like CLIP [51], aim

*Corresponding author.

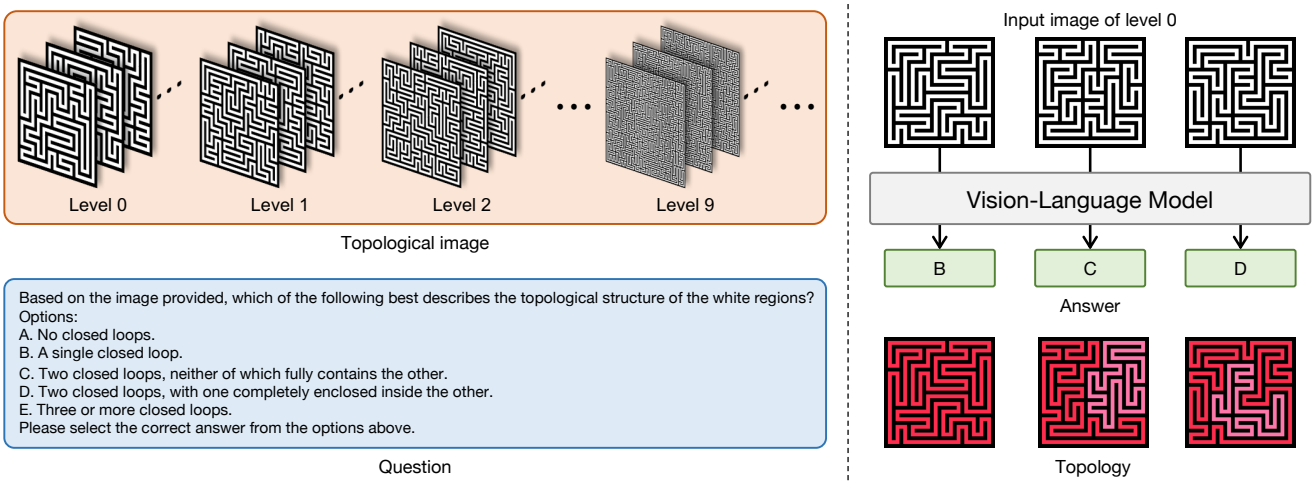

Figure 2. An illustration of the TopoPerception benchmark. The text question and options are fixed (left). The input image belongs to one of three categories, corresponding to options B, C, and D, respectively (right). Options A and E serve as distractors. The input images can be extended to arbitrary difficulty levels.

to create representations that align with natural language descriptions. However, this semantic compression process can inadvertently discard crucial global visual information that lacks a straightforward linguistic correlate. Structurally, the design of visual encoders imposes strong inductive biases, such as fixed input resolutions and aspect ratios, which often necessitate scaling or patching the image [15]. Such operations can corrupt global structures and distort spatial relationships, compromising the integrity of the visual representation before it even reaches the language model. Furthermore, long sequences of visual tokens impose a significant computational burden, motivating research into token reduction techniques [52] that may further risk discarding subtle yet important visual details.

Most current evaluation methods for LVLMs, while comprehensive, are predominantly focused on downstream, semantically rich tasks. Benchmarks for VQA, compositional reasoning, and open-world dialogue assess a combination of perception, reasoning, and language generation capabilities [34]. However, models' success or failure on these tasks does not isolate the fidelity of its initial visual perception. This capability confounding can mask fundamental weaknesses. For instance, models might answer questions by leveraging parameterized knowledge stored within the LLM rather than relying on the visual input. This stored knowledge may be correct [11, 24, 28, 30] or incorrect [25, 35]. Even the design of the question itself may be flawed, allowing a correct answer to be derived from the input text alone, without any need for the visual input [11, 39].

Moreover, these semantically rich tasks introduce unavoidable local shortcuts, where models might correctly answer questions by identifying a specific object or scene element without understanding the image's global structure [6, 17–19, 26]. This reliance on local shortcuts can lead to an overestimation of the model's true visual perception, as it may be exploiting local cues instead of processing the image globally. Therefore, there is an urgent need for an evaluation paradigm that can disentangle visual perception from textual reasoning and directly measure global visual perception in a shortcut-free manner.

Topology offers a promising avenue for addressing this challenge. As a branch of mathematics, topology studies the properties of space that are preserved under continuous deformations, such as stretching, twisting, crumpling, and bending [42]. When considering the two-dimensional manifold projected onto the human retina, these invariant properties can be categorized into three main topological attributes: connectivity, the number of holes, and interior/exterior relationships. Crucially, these attributes are independent of local features. Therefore, topological features capture the global structure of an image, making them an ideal proxy for evaluating models' global visual perception without the influence of local shortcuts. Furthermore, topology is closely related to processes in the human visual system. For humans, global features enable rapid object detection without requiring an in-depth analysis of local features, a critical skill for survival in natural environments. Given the pronounced sensitivity of the human visual system to global features, it has been hypothesized that global features are processed first and subsequently guide the interpretation of local features [9, 10].

In this paper, we introduce TopoPerception, a diagnostic benchmark that employs the topological features of images to isolate and quantify global visual perception without shortcuts. To isolate the visual input and prevent the model from relying on textual cues to answer, TopoPerception uses

a fixed text-based question and a fixed set of options that do not change with the visual input. Furthermore, to ensure that no local visual shortcuts are introduced, TopoPerception utilizes synthetically generated images that possess pure topological properties without confounding semantic features. As previously stated, because topological properties are independent of local features, TopoPerception can precisely control the granularity of the global features being evaluated without being susceptible to local shortcuts. By testing whether LVLMs can differentiate between images based on their topological class across various levels of perceptual granularity, TopoPerception probes whether the models' visual encoder and subsequent alignment process preserve the global features of an image or merely convert them into a language-compatible format at the expense of global fidelity.

The main contributions of this study are threefold:

- We present TopoPerception, a shortcut-free benchmark for evaluating the global visual perception of LVLMs. It defines tasks based on topological properties across a range of perceptual granularities, providing a scalable difficulty hierarchy for stress-testing LVLMs.
- We conduct a comprehensive evaluation of SOTA LVLMs, revealing systematic and severe deficiencies in their global visual perception. Even at the simplest, coarsest evaluation granularity in TopoPerception, all models performed at a level statistically consistent with random guessing, answering almost entirely based on their intrinsic biases (Fig. 1). This highlights a substantial loss of global visual features in current LVLMs.
- We discover an unexpected trend: within the same model family or architecture, larger models with stronger reasoning capabilities tend to exhibit lower accuracy on TopoPerception (Fig. 1). In other words, prompting models to generate more detailed, step-by-step answers is positively correlated with a decrease in accuracy when perceiving global visual features. This suggests that scaling up reasoning may actually interfere with or override the models' already fragile visual signal, a finding that raises concerns about the interplay between chain-of-thought reasoning and visual grounding.

2. Related work

2.1. Evaluation benchmarks for LVLMs

The rapid development of LVLMs has been accompanied by the establishment of a diverse ecosystem of evaluation benchmarks, each designed to probe different dimensions of multimodal intelligence [16, 39, 57, 58]. While this rich ecosystem has been instrumental in driving progress, these benchmarks share a common orientation: they predominantly evaluate the semantic interpretation of visual content. Success in these tasks is contingent on the correctness of a textual response to a query about the meaning, composition, or context of an image. However, the complexity of semantic content introduces unavoidable local shortcuts. This leaves a critical gap in our understanding of a more fundamental, upstream capability: the fidelity of global visual perception itself. TopoPerception aims to fill this gap by directly measuring the degree to which global visual features are preserved, independent of their semantic content.

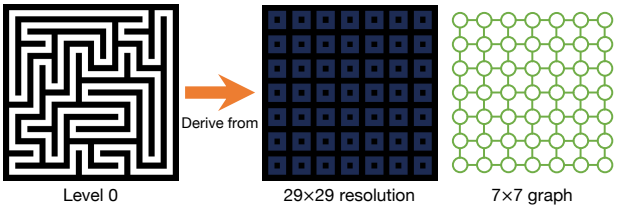

Figure 3. The visual topological images are generated by constructing a uniform spanning tree on a connected graph. Each node in the graph corresponds to a 3×3 pixel block in the image. A connected graph with $n \times n$ nodes generates an image with a resolution of $(4n + 1) \times (4n + 1)$.

2.2. Benchmarking global perception in LVLMs

Current benchmarks for evaluating the global perception capabilities of LVLMs can generally be classified into two main categories:

- **Shape-based**: These methods indirectly express global attributes through the shape of objects [17, 26]. However, the shape itself may contain local shortcuts, such as distinguishing between circles and squares, which can be determined based on the local curvature of the shape rather than its overall form.
- **Maze-based**: Maze tasks typically specify a start and end point and assess the connectivity between them [41]. When the points are not connected, the model must explore all possible paths to determine the result. In contrast, when the points are connected, the model only needs to identify one valid solution, which significantly reduces the demand for global perception. This introduces local shortcuts in the evaluation of global perception abilities and makes it difficult to control the difficulty of the dataset.

In contrast, TopoPerception utilizes topological properties, which effectively eliminate the possibility of local shortcuts, offering a more robust evaluation of global perception. This innovation allows us to directly assess the model's ability to preserve and process the global structure of an image, bypassing the confounding influence of local cues.

2.3. The visual pipeline in LVLMs: sources of information loss

The standard architecture of an LVLM consists of three main components: a visual encoder, a language model,

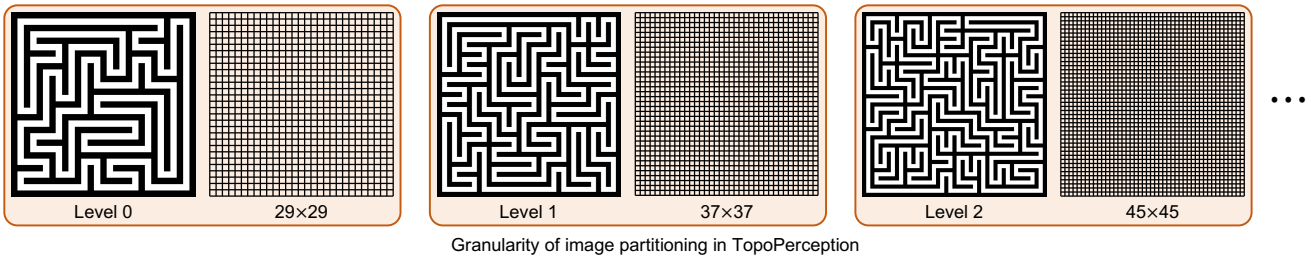

Figure 4. Illustration of the granularity of image partitions at different levels in TopoPerception. The difficulty level for an image with a partitioning granularity of $(4n+1) \times (4n+1)$ is defined as $(n-7)/2$.

and a projection module to connect the two [1, 5, 7, 12–14, 23, 27, 33, 37, 38, 40, 47–50, 54]. The visual encoder, responsible for converting visual inputs into a sequence of feature vectors or tokens, is a critical source where significant information loss can occur. Pre-trained visual encoders often introduce strong inductive biases, such as being trained on fixed-size square images. To handle arbitrarily shaped inputs, models must resort to operations like resizing, padding, or patching the image [15], which can severely distort the image's global layout and corrupt the structural relationships between different parts of a scene, leading to a compromised representation before it is processed by the language model. Furthermore, as the image is transformed into a dense sequence of tokens, the model may risk discarding important visual details in an effort to reduce the computational load [52], compounding the challenge of accurately preserving the global visual structure.

Subsequently, the long sequence of visual tokens is passed through the projection module and into the LLM. A significant challenge at this stage is the cross-modal alignment problem, where the learned visual representations may not effectively correspond to textual concepts in the LLM's embedding space. This alignment discrepancy can complicate information fusion and may undermine the effectiveness of techniques aimed at reducing the computational cost of long visual sequences, such as token reduction [55]. Our research provides a new diagnostic tool to precisely quantify a key consequence of these architectural choices: the loss of globally perceived visual features.

3. Sources of shortcuts in benchmarks

We classify the potential shortcuts in LVLM evaluation benchmarks into two hierarchical levels:

- **Statistical level**: Due to improper data distribution design, the data may contain unintended statistical regularities, i.e., spurious correlations. This is commonly observed in training-dependent scenarios [19, 60]. For example, in vision-language benchmarks, visual elements frequently co-occur with corresponding textual cues, allowing the task to be solved correctly by relying solely on the language modality without genuinely understanding the image content [24]. In this case, the language modality itself becomes a shortcut. Another example is when certain elements in an image frequently appear together; the model can infer the presence of one element from the other. Here, the co-occurring elements are the shortcut [35].
- **Semantic level**: This level of shortcuts does not depend on statistical co-occurrence but can be inferred from the semantics of the inputs alone. For instance, in a VQA task, the text of the question might contain sufficient information to be answered without reference to the image content [11, 39]. Here, the textual content itself acts as the shortcut. As another example, in an LVLM recognition task, we could ask in two ways:
  1. "What is the category of this image?"
  2. "Based on its texture, what is the category of this image?"

  Compared to the second question, the first allows the model to rely on shape instead of texture, making shape a potential shortcut. The second question explicitly eliminates this shortcut at the prompt level [17]. It is clear that semantic-level shortcuts can be identified from a single sample without resorting to statistical patterns. However, they also manifest as statistical regularities, meaning the semantic level is contained within the statistical level. Therefore, semantic shortcuts exist in both training-dependent and training-free scenarios [60]. Since LVLM benchmarks are typically zero-shot, they are primarily susceptible to semantic-level shortcuts.

4. TopoPerception benchmark

4.1. Benchmark design and task format

As discussed, LVLM benchmarks often face semantic-level shortcuts. To accurately evaluate the global visual perception of LVLMs without being affected by local shortcuts at a semantic level, TopoPerception constructs its tasks using topological properties of images, which are by definition independent of local features. Furthermore, we have designed the benchmark to prevent models from using pa-

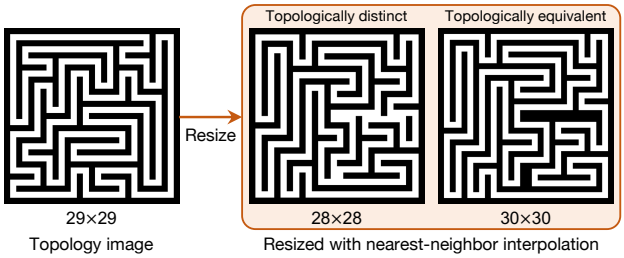

Figure 5. An example of how the topological properties of an image in TopoPerception change across adjacent perceptual granularities.

rameterized knowledge stored from prior training data by addressing both the visual and textual modalities:

- **Visual modality**: To prevent models from memorizing relevant visual features of natural images, we use a synthetic dataset as the foundation for our visual data.
- **Textual modality**: To isolate the visual input and ensure that the model cannot answer based solely on the text, we employ a fixed text question and a fixed set of answer options that remain constant across all visual inputs.

Figure 2 (left) illustrates the task format of TopoPerception, which takes the form of a multiple-choice VQA task. Each query presented to the model consists of one image and a fixed text question with five options. The question asks about the topological properties of the image and provides five lettered options (A, B, C, D, E) as possible answers. The visual modality uses a synthetic dataset of topological images [59], which not only prevents models from relying on memorized natural image features but also offers highly scalable difficulty levels. According to Kirchhoff's theorem [31], the topological image sample space exhibits asymptotic exponential growth $\sim \exp(hn^2)$ with respect to the number of nodes n, where $h \approx 1.16624$ is the lattice tree entropy constant [53]. This ensures that even if TopoPerception is later included in training data, the model cannot solve its extended difficulty levels through memorization. The images have three categories of topological properties, as shown in Fig. 2 (right). The textual modality is a fixed multiple-choice question asking about the image's topological properties, ensuring that the only variable is the image itself. Therefore, the model must rely on the image to select the correct option.

Options B, C, and D in the question correspond to the three topological categories of the images. Options A and E serve as distractors to ensure the question is closed-form, meaning all possible correct answers are among the options. This helps diagnose whether the model is guessing randomly or exhibits a bias towards certain options. If a model's accuracy is close to 20%, it suggests random guessing among all five options (A, B, C, D, E). If the accuracy is close to 33.3%, it suggests a preference for certain options within the valid set (B, C, D).

4.2. Topological properties and granularity levels

As shown in Fig. 3, the visual topological images used in TopoPerception are generated by constructing a uniform spanning tree on a connected graph, which includes multiple difficulty levels [60]. Each difficulty level corresponds to a different spatial resolution, i.e., a different granularity of partitioning for the image, as illustrated in Fig. 4. At a lower partitioning granularity, the image is divided into coarser regions, placing weaker demands on the model's information retention capacity. At a finer granularity, the demands are stronger. By varying the granularity, TopoPerception allows us to control the difficulty of the task and assess the model's ability to retain global visual information at different levels of abstraction.

As previously mentioned, since topological properties are determined by the global structure of the image and are independent of local features, a model can only preserve the image's topological properties if its perceptual granularity is greater than or equal to the image's own partitioning granularity. To demonstrate this intuitively, we analyze how the topological properties of an image change across two adjacent perceptual granularities, as shown in Fig. 5. Using nearest-neighbor downsampling as an example, when the model's perceptual granularity (i.e., the resolution it effectively processes) is smaller than the image's partitioning granularity, the topological properties may be distorted or lost. Conversely, when the model's perceptual granularity is greater than or equal to the image's partitioning granularity, the topological properties remain stable. This insight allows us to precisely determine the model's true ability to retain global visual information as it processes an image at different levels of resolution during perception.

5. Experiments

5.1. Experimental setup

We selected some of the most powerful recent LVLMs for our evaluation:

- **OpenAI**: GPT-4o [45], o4-mini, o3 [46].
- **Anthropic**: Claude-sonnet-4-0, Claude-opus-4-0 [4].
- **Google**: Gemini-2.5-flash, Gemini-2.5-pro [22].

All models were evaluated using their standard API calls. We chose these models because they represent the current SOTA and have demonstrated outstanding performance on various vision-language tasks.

For each difficulty level and each category of the visual topology dataset, we randomly sampled 100 images to construct the TopoPerception benchmark. This resulted in 300 samples per granularity level, balancing cost with statistical significance.

To observe the processes for the models' answers, we did

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
OpenAI GPT-4o [45]	22.00	33.40	22.00	24.07
OpenAI o4-mini [46]	19.67	28.28	19.67	21.59
OpenAI o3 [46]	12.00	35.36	12.00	16.38
Claude-sonnet-4-0 [4]	30.00	42.51	30.00	**28.07**
Claude-opus-4-0 [4]	24.33	21.50	24.33	22.06
Gemini-2.5-flash [22]	**33.33**	**48.44**	**33.33**	26.34
Gemini-2.5-pro [22]	30.67	33.11	30.67	27.96

Table 1. Results of LVLMs on TopoPerception at the lowest difficulty level (Level 0). The best result in each column is bolded. Precision, Recall, and F1 Score are weighted averages of the per-class metrics, weighted by the number of true samples in each class.

not force them to output only the option letter but allowed them to generate open-ended text responses. We used the default inference settings for each model. It is worth noting that, since all textual questions and answer options were identical across models, we retained the default temperature rather than setting it to 0, in order to preserve the natural stochasticity of the model's generative distribution. This approach enables us to further investigate potential preference patterns when a model shows inclination towards particular options, thereby facilitating an analysis of the strength and stability of its biases across the entire output distribution. By doing so, we can uncover systematic differences at the probabilistic level, rather than being limited to deterministic outputs, thus providing a more comprehensive characterization of the model's behavioral tendencies.

5.2. Results on the TopoPerception

Table 1 summarizes the performance of the evaluated models on the TopoPerception benchmark at Level 0, which corresponds to a resolution of 29×29. This resolution is comparable to the MNIST dataset's 28×28 resolution [32], providing a straightforward reference for difficulty. Figure 1 provides an intuitive comparison of the accuracy of each model, with models from the same family indicated by a consistent color scheme. Remarkably, the performance of most models hovers around the 20% random baseline, indicating that their accuracy is statistically indistinguishable from random guessing. No model's performance exceeds the 33.3% accuracy threshold that would result from a perfect bias towards one of the three correct options. This highlights a significant gap in the current LVLMs' ability to comprehend the global structure of an image. Even the simplest evaluation task, reveals a profound inability to process visual information at the global level.

Notably, Figure 1 reveals that within each model family, variants are close in accuracy. However, a consistent trend is discernible: larger models with more powerful reasoning capabilities tend to have lower accuracy on TopoPerception. This phenomenon holds across different model families. For instance, in the OpenAI family, the accuracy follows GPT-4o $>$ o4-mini $>$ o3. In the Anthropic family, Claude-sonnet-4-0 outperforms Claude-opus-4-0. In the Google family, Gemini-2.5-flash is more accurate than Gemini-2.5-pro. This trend suggests that while increased reasoning ability of models may boost performance on other tasks, they do not automatically improve visual perception capabilities. In fact, it may even exacerbate the issue.

We hypothesize that because topological features are global and abstract, LVLMs that attempt to reason through natural language may be misled by their own priors or by the absence of a precise internal representation of the image. Models with stronger reasoning abilities may be more reliant on language-based reasoning, which could distort their visual understanding. This counter-intuitive finding suggests that for tasks requiring strict perceptual fidelity, prompting step-by-step reasoning is not always beneficial, unlike in arithmetic or commonsense problems. Instead, a more direct vision-to-decision mapping, perhaps akin to a classification head, might be more effective.

5.3. Error analysis

To further investigate where the models are failing, we present the confusion matrices for all models on the Level 0 task, as shown in Fig. 6. It is evident that the models exhibit significant biases on TopoPerception, and models from the same family show similar tendencies. For example, both models in the Claude 4 family, Claude-sonnet-4-0 and Claude-opus-4-0, are most inclined to choose option C, followed by options A and D. Similarly, both models in the Gemini 2.5 family, Gemini-2.5-flash and Gemini-2.5-pro, show a strong preference for option C. In contrast, the preferences of the OpenAI family exhibit more varied behavior, likely because they belong to different series. GPT-4o primarily favors options A and B, followed by C; o4-mini's choices are more uniform, resembling random guessing, which is corroborated by its 19.67% accuracy; o3, on the other hand, has a very strong bias towards option A.

To gain deeper insight into the source of these biases, we

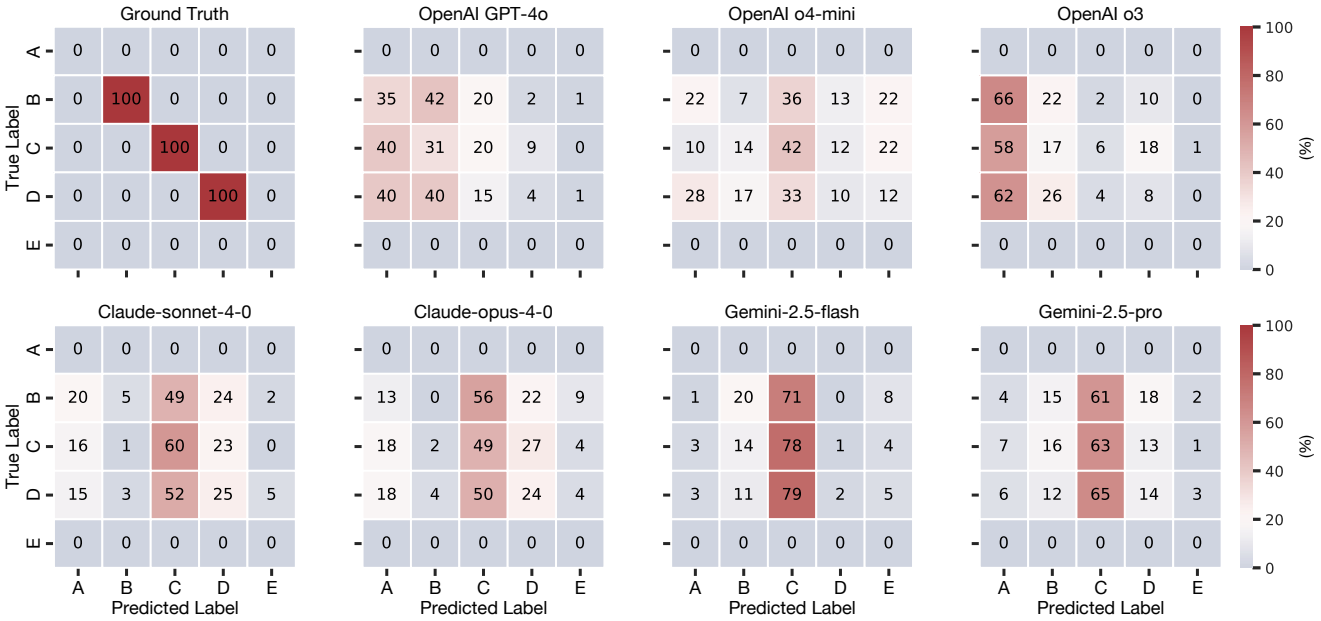

Figure 6. Confusion matrices for all evaluated LVLMs on the easiest level of TopoPerception.

analyze the prediction distributions for each model on each image category for the Level 0 task, as shown in Fig. 7. It can be observed that for each model, the prediction distribution is nearly identical across different categories and consistent with the overall data prediction distribution. This indicates that all models are using their inherent biases to select answers for every question, and their selection strategy does not change based on the input image. Combined with our choice of temperature parameter, we can conclude that this preference is a stable characteristic of the models.

6. Discussion

6.1. Classification of shortcuts in benchmarks

In addition to the previously described classification of shortcuts into statistical and semantic levels, we can also categorize shortcuts in LVLM evaluation benchmarks according to a "Morgan's Canon of data" [60]:

- **Type 1**: Tasks can be solved using only the text modality, without requiring input from the visual modality. In this case, the text modality itself becomes the shortcut.
- **Type 2**: Tasks require visual input, but there is a shortcut within the image itself, such as a local feature shortcut.

Therefore, eliminating Type 2 shortcuts presupposes the elimination of Type 1 shortcuts. Conversely, eliminating Type 1 shortcuts does not affect the existence of Type 2 shortcuts. TopoPerception eliminates Type 1 shortcuts through its fixed question-and-option design and then eliminates Type 2 shortcuts through the introduction of topological properties.

6.2. Synthetic data

Some may argue that the use of synthetic data in TopoPerception, which differs significantly from the distribution of natural images. Therefore, one might question why we expect LVLMs to perform well. We emphasize that, on one hand, natural data often contains a complex mixture of properties that make it difficult to assess a specific attribute clearly. In contrast, synthetic data serves as an abstraction of specific objects or properties, designed to decouple and control these factors—something that natural images cannot achieve. Consequently, synthetic and natural data complement each other in scientific research, with synthetic data enabling the evaluation of properties that cannot be directly assessed in natural data. On the other hand, the goal of synthetic data is to minimize interference from prior learning experiences. This concept is similar to how fluid intelligence is measured in human IQ tests, such as Raven's Progressive Matrices [8, 29]. In these tests, researchers design abstract tasks, rather than those based on real-world scenarios, to evaluate cognitive ability. These tasks do not depend on specific knowledge or experience and are applicable across various age groups. In the context of TopoPerception, the introduction of synthetic data represents a pure abstraction of global properties, preventing the model from memorizing certain visual features in natural images that could compromise the reliability of the evaluation.

6.3. Interpretation of experimental results

The consolidated findings of our study paint a concerning picture: despite the large scale of current LVLMs and their

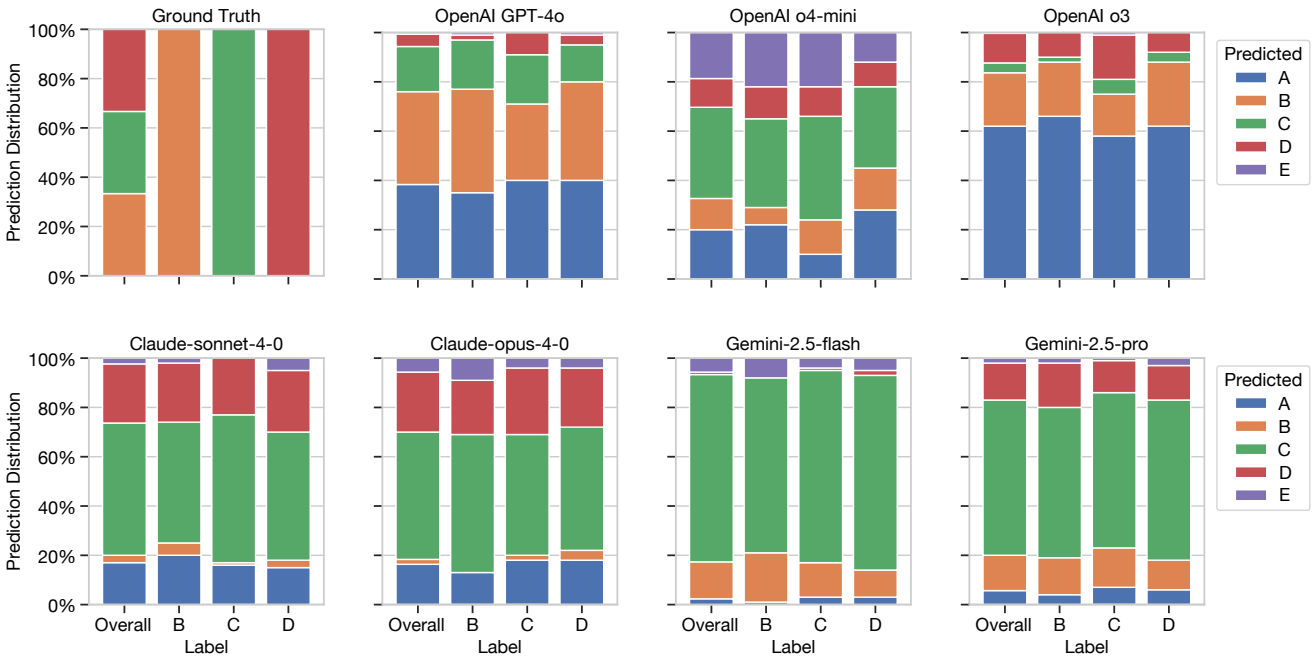

Figure 7. Prediction distributions of LVLMs for data from different ground-truth labels on the easiest level of TopoPerception. "Overall" represents the prediction distribution across all test data.

training on vast image-text corpora, these models fail to preserve or reason about global visual features reliably. A critical aspect of this problem lies in the role of model scale and architecture. We found no clear correlation suggesting that larger models perform better, which indicates that the issue isn't simply one of capacity or knowledge. Rather, it seems to stem from a systemic information bottleneck and a mismatch between training objectives. LVLMs are typically trained or fine-tuned to generate descriptive language or answer general questions, but they are not explicitly designed to retain all visual information. This training bias may lead them to prioritize what humans typically highlight when describing images, resulting in the unintentional filtering out of global structures and treating them as "unimportant details". As a result, these models have learned to "see" more like a describer than a true visual reasoner, focusing on local or salient objects rather than understanding the image in its entirety.

7. Conclusion

In this paper, we have introduced TopoPerception, a diagnostic benchmark that rigorously and without shortcuts evaluates the global visual perception of LVLMs across different perceptual granularities. By focusing on the topological properties of images—global features that are independent of local characteristics—we have discovered that current LVLMs suffer from severe deficiencies in global visual perception. Our experiments on SOTA models show that even at the simplest evaluation levels, their performance is close to random chance, with answers based almost entirely on their intrinsic biases. This reveals a fundamental limitation masked by their fluent linguistic output on other problems.

These findings have profound implications for the future development of LVLMs. First, they highlight the urgent need to reconsider the architecture of multimodal models: simply "stitching" a fixed visual encoder to an LLM (with a minimal interface) may be insufficient for tasks requiring deeper image understanding. More expressive or iterative visual encoders, or mechanisms that allow the model to re-examine the image during reasoning, may be necessary. Second, the observation that larger scale and stronger reasoning capabilities can degrade performance suggests that multimodal models require careful calibration between reasoning and perception. When a model "thinks" in language, it may distort what it sees—a mechanism for "fact-checking" its reasoning against the visual input at each step could be beneficial.

In summary, as we strive to build AI that can genuinely "perceive" the world, TopoPerception provides a new lens for analyzing what our models see and what they miss. We hope it will serve as a diagnostic tool to guide the development of the next generation of vision-language models, leading to systems that are not only capable of fluently describing an image but also of grasping its deepest structural truths.

Acknowledgement

This work was supported by the STI 2030–Major Projects 2021ZD0200300.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 4

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://anthropic.com/claude-3-model-card/, 2024. 1

[3] Anthropic. Claude 3.7 sonnet system card. https://www.anthropic.com/claude-3-7-sonnet-system-card/, 2025.

[4] Anthropic. System card: Claude opus 4 & claude sonnet 4. https://anthropic.com/claude-4-model-card/, 2025. 1, 5, 6

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 4

[6] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):1–43, 2018. 2

[7] ByteDance Seed. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 4

[8] Raymond B. Cattell. The measurement of adult intelligence. *Psychological Bulletin*, 40(3):153–193, 1943. 7

[9] Lin Chen. Topological structure in visual perception. *Science*, 218(4573):699–700, 1982. 2

[10] Lin Chen. The topological approach to perceptual organization. *Visual Cognition*, 12(4):553–637, 2005. 2

[11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 2, 4

[12] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 4

[13] DeepSeek-AI. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[14] DeepSeek-AI. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 4

[15] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *Advances in Neural Information Processing Systems*, 37:52545–52567, 2024. 2, 4

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 3

[17] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Can we talk models into seeing the world differently? In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 4

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2, 4

[20] Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[21] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[22] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 5, 6

[23] GLM-V Team. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 4

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017. 2, 4

[25] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024. 2

[26] Arshia Hemmat, Adam Davies, Tom A. Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in plain sight: evaluating abstract shape recognition in vision-language models. *Advances in Neural Information Processing Systems*, 37:88527–88556, 2024. 2, 3

[27] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 4

[28] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and composi-

tional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019. 2

[29] John and Jean Raven. Raven progressive matrices. In *Handbook of nonverbal assessment*, pages 223–237. Springer, 2003. 7

[30] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017. 2

[31] Gustav Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148 (12):497–508, 1847. 5

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002. 6

[33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. 4

[34] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024. 2

[35] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 2, 4

[36] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1587–1606, 2025. 1

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023. 4

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 4

[39] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2024. 2, 3, 4

[40] Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4

[41] Aryo Lotfi, Enrico Fini, Samy Bengio, Moin Nabi, and Emmanuel Abbe. Chain-of-sketch: Enabling global visual reasoning. *arXiv preprint arXiv:2410.08165*, 2024. 3

[42] James R. Munkres. *Topology*. Prentice Hall, Inc., 2 edition, 2000. 2

[43] OpenAI. Gpt-4v(ision) system card. https://openai.com/index/gpt-4v-system-card/, 2023. 1

[44] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[45] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5, 6

[46] OpenAI. Openai o3 and o4-mini system card. https://openai.com/index/o3-o4-mini-system-card/, 2025. 1, 5, 6

[47] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[48] Qwen Team. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[49] Qwen Team. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[50] Qwen Team. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 4

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[52] Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *arXiv preprint arXiv:2507.20198*, 2025. 2, 4

[53] Robert Shrock and Fa Yueh Wu. Spanning trees on graphs and lattices in d dimensions. *Journal of Physics A: Mathematical and General*, 33(21):3881, 2000. 5

[54] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12966–12977, 2025. 4

[55] Rui Xu, Yunke Wang, Yong Luo, and Bo Du. Rethinking visual token reduction in lvlms under cross-modal misalignment. *arXiv preprint arXiv:2506.22283*, 2025. 4

[56] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024. 1

[57] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57730–57754. PMLR, 2024. 3

[58] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024. 3

[59] Wenhao Zhou. Mitigating data bias and ensuring reliable evaluation of ai models with shortcut hull learning. https://doi.org/10.6084/m9.figshare.28794407, 2025. 5

[60] Wenhao Zhou, Faqiang Liu, Hao Zheng, and Rong Zhao. Mitigating data bias and ensuring reliable evaluation of ai models with shortcut hull learning. *Nature Communications*, 16(1), 2025. 4, 5, 7